UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
INSTITUTO DE ENGENHARIA BIOMÉDIA

GLAUCO CARDOZO

Um Modelo Computacional utilizando Técnicas de Machine Learning e Exames Laboratoriais de Rotina na Triagem e Apoio ao Diagnóstico de Diabetes Mellitus

Florianópolis
2022

Glauco Cardozo

Um Modelo Computacional utilizando Técnicas de Machine Learning e Exames Laboratoriais de Rotina na Triagem e Apoio ao Diagnóstico de Diabetes Mellitus

Tese submetida ao Programa de Pós-graduação em Engenharia Elétrica, área de concentração em Engenharia Biomédica da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Engenharia Elétrica.
Orientador: Prof. Jefferson Luiz Brum Marques, PhD.

Florianópolis

2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Cardozo, Glauco
Um Modelo Computacional utilizando Técnicas de Machine Learning e Exames Laboratoriais de Rotina na Triagem e Apoio ao Diagnóstico de Diabetes Mellitus / Glauco Cardozo ; orientador, Jefferson Luiz Brum Marques, 2022.
183 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Elétrica, Florianópolis, 2022.

Inclui referências.

1. Engenharia Elétrica. 2. Machine Learning. 3. Exames Laboratoriais. 4. Diabetes Mellitus. 5. Diagnóstico. I. Brum Marques, Jefferson Luiz . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

Glauco Cardozo

Um Modelo Computacional utilizando Técnicas de Machine Learning e Exames Laboratoriais de Rotina na Triagem e Apoio ao Diagnóstico de Diabetes Mellitus

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Júlio Cesar Nievola, Dr.
Pontifícia Universidade Católica do Paraná

Profa. Silvia Modesto Nassar, Dra.
Universidade Federal de Santa Catarina

Prof. William Alberto Cruz Castañeda, Dr.
Universidade do Estado de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Engenharia Elétrica.

Documento assinado digitalmente
Telles Brunelli Lazzarin
Data: 29/11/2022 14:58:24-0300
CPF: ***.400.719-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Telles Brunelli Lazzarin, Dr.
Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Jefferson Luiz Brum Marques
Data: 28/11/2022 20:55:57-0300
CPF: ***.624.060-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Jefferson Luiz Brum Marques, PhD.
Orientador

Florianópolis, 2022

Este trabalho é dedicado aos amigos de toda vida.

AGRADECIMENTOS

Agradeço a minha esposa e família pelo apoio em todos os momentos, ao professor Jefferson Luiz Brum Marques, pela confiança, ao Laboratório Médico Santa Luzia, pela parceria e aos professores e colegas do Instituto de Engenharia Biomédica e do Departamento de Informática e Estatística.

Os grandes saltos para frente foram dados pelo homem, nunca experimentalmente, nunca racionalmente, mas por intuição, verdadeiro e grande sistema de pesquisa do futuro (Pietro Ubaldi, 1937).

RESUMO

Estima-se que cerca de 10% da população mundial seja portadora de diabetes mellitus (DM), sendo que em torno de 50% destes indivíduos não foram diagnosticados e não sabem que possuem a doença. Isso ocorre principalmente pelo fato de apenas 40% dos indivíduos com DM tipo2 (DM2) apresentarem sintomas. O DM2 pode permanecer não detectado por vários anos, ocasionando complicações severas e aumentando os custos com o tratamento de saúde. Por esta razão, o diagnóstico precoce é de extrema importância, sendo que quando feito no início, ainda na etapa conhecida como pré-diabetes, o quadro pode ser revertido e estabilizado. O diagnóstico do DM é feito por meio de exames laboratoriais, sendo a glicose no sangue em jejum (FPG) e a hemoglobina glicada (HbA1c) os mais realizados. No entanto eles podem apresentar discrepâncias entre si, uma vez que o FPG exige jejum de oito horas, apresentando variações ao longo do dia. Por apresentar vantagens em relação ao FPG, o exame de HbA1c tem sido mais indicado na busca por um diagnóstico, apesar de, por questões culturais e financeiras, o FPG ainda ser mais utilizado. Nesse contexto, buscou-se utilizar exames laboratoriais de rotina juntamente com técnicas de Machine Learning, tendo como objetivo a predição do HbA1c e a detecção de falsos negativos em exame de FPG. Nesta segunda etapa, foi desenvolvido uma metodologia para, com base na glicose média, calcular um fator de ajuste que melhorasse a concordância entre os exames de FPG e HbA1c. Como ferramenta de predição, foram testados os métodos KNN, SVM, Naïve Bayes, Random Forest e ANN, utilizando exames de 201.338 pacientes. Na predição do HbA1c, o desempenho dos modelos de classificação e regressão foram avaliados utilizando diferentes subconjuntos de dados, como saudáveis, pré-diabéticos, diabéticos, não saudáveis e sem diabetes. Nesta etapa, o melhor desempenho foi obtido com a classificação após a regressão do modelo de rede neural artificial na identificação de indivíduos não saudáveis. O modelo obteve 78,1%, 78,7% e 78,4% para sensibilidade, precisão e F1-Score, respectivamente. Já na detecção de falsos negativos de FPG, o modelo de regressão ANN obteve o melhor resultado no cálculo do fator de ajuste. Com ele foi possível obter um ganho de 16,6% no diagnóstico de diabetes e 35% no de pré-diabetes, sendo estes valores relevantes na identificação de falsos negativos. Desta forma, concluímos que modelos baseados em aprendizado de máquina são capazes de prever valores de hemoglobina glicada a partir de exames laboratoriais de rotina e podem ser usados como uma ferramenta de triagem e apoio no diagnóstico de diabetes. Da mesma forma, por meio do ajuste dos valores do FPG, é possível aumentar a concordância destes com o HbA1c e diminuir a ocorrência de falsos negativos. Assim, durante a rotina de exames laboratoriais, o modelo poderia ser utilizado na triagem de possíveis falsos negativos e consequentemente sugerir a realização do exame de HbA1c para confirmação do diagnóstico de Diabetes Mellitus.

Palavras-chave: Machine Learning; Exames Laboratoriais; Diabetes Mellitus; Diagnóstico.

ABSTRACT

It is estimated that around 10% of the world population has diabetes mellitus (DM), and around 50% of these individuals have not been diagnosed and do not know they have the disease. This is mainly because only 40% of individuals with type 2 DM (DM2) have symptoms. T2DM can remain undetected for several years, causing severe complications and increasing health care costs. For this reason, early diagnosis is extremely important, and when made early, still in the stage known as pre-diabetes, the condition can be reversed and stabilized. The diagnosis of DM is made through laboratory tests, with fasting blood glucose (FPG) and glycated hemoglobin (HbA1c) being the most frequently performed. However, they may have discrepancies between them, since the FPG requires an 8-hour fasting, with variations throughout the day. As it has advantages over the FPG, the HbA1c test has been more indicated in the search for a diagnosis, although, for cultural and financial reasons, the FPG is still more used. In this context, we sought to use routine laboratory tests along with Machine Learning techniques, with the objective of predicting HbA1c and detecting false negatives for the FPG test. In this second stage, a methodology was developed to calculate an adjustment factor based on average glucose that would improve the agreement between the FPG and HbA1c tests. As a prediction tool, the KNN, SVM, Naïve Bayes, Random Forest and ANN methods were tested, using exams from 201338 patients. In predicting HbA1c, the performance of classification and regression models were evaluated on different subsets of data, such as healthy, pre-diabetic, diabetic, unhealthy and non-diabetic. In this step, the best performance was obtained with the classification after the regression of the artificial neural network model in the identification of unhealthy individuals. The model obtained 78.1%, 78.7% and 78.4% for sensitivity, precision and F1-Score, respectively. In the detection of false negatives of FPG, the ANN regression model obtained the best result in the calculation of the adjustment factor. With it, it was possible to obtain a gain of 16.6% in the diagnosis of diabetes and 35% in the diagnosis of pre-diabetes, these values being significant in the identification of false negatives. In this way, we conclude that models based on machine learning are able to predict glycated hemoglobin values from routine laboratory tests and can be used as a screening tool and support in the diagnosis of diabetes. Likewise, by adjusting the FPG values, it is possible to increase their agreement with HbA1c and reduce the occurrence of false negatives. Thus, during routine laboratory tests, the model could be used to screen for possible false negatives and consequently suggest performing the HbA1c test to confirm the diagnoses of Diabetes Mellitus.

Keywords: Machine Learning; Laboratory Tests; Diabetes Mellitus; Diagnosis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC: *Acurracy* - Acurácia

AF: *Adjustment Factor* - Fator de Ajuste

aFPG: *adjusted FPG* - FPG Ajustado

AG: *Average Glucose* - Glicose Média

ANN: *Artificial Neural Network* – Rede Neural Artificial

AUC-PR: Area Under Curve *PR* - Área sob a Curva Precision-Recall

AUC-ROC: *Area Under Curve ROC* - Área sob a Curva ROC

CHCM: Concentração de Hb Corpuscular Média

CLSI: *Clinical and Laboratoty Standards Institute*

Cr: Creatinina

DAG: *Directed Acyclic Graph*

DM: Diabetes Mellitus

DM1: Diabetes Mellitus do tipo 1

DM2: Diabetes Mellitus do tipo 2

eAG: *estimated Mean Glucose* - Glicose Média Estimada

eG: *estimated Glucose* - Glicose Estimada

EHR: *Eletronic Health Record* - Registro Eletrônico de Saúde

FN: Falso Negativo

FP: Falso Positivo

FPG: *Fasting Plasma Glucose* - Glicose Plasmática em Jejum

Hb: Hemoglobina

HbA1c: G*lycated Haemoglobin* - Hemoglobina Glicada

HCM: Hb Corpuscular Média. Quantidade média de Hb

HD: Conjunto de dados com pacientes Saudáveis e Diabéticos

HDL: *High Density Lipoprotein* – Lipoproteínas de alta densidade

HN: Conjunto de dados com pacientes Saudáveis e Não Saudáveis

HP: Conjunto de dados com pacientes Saudáveis e Pré-diabetes

HPD: Conjunto de dados com pacientes Saudáveis, Pré-diabéticos e Diabéticos

Ht: Hematócrito

IA: Inteligência Artificial

KNN: *K Nearest Neighbors* – K Vizinhos mais Próximos

LDL: *Low Density Lipoprotein* - Lipoproteínas de baixa densidade

Lp: *Lipoprotein* - Lipoproteínas

MAE: *Mean Absolute Error* - Erro Médio Absoluto

MeSH: *Thesaurus Medical Subject Headings*

ML: *Machine Learning* – Aprendizado de Máquina

MSE: *Mean Square Error* - Erro Médio Quadrático

NB: *Naïve Bayes*

ND: Conjunto de dados com pacientes Não Diabéticos e Diabéticos

NIH: *National Institutes of Health*

OGTT: *Oral Glucose Tolerance Test* - Teste oral de tolerância à glicose

pAF: *predicted AF* - Fator de Ajuste Preditos

PCA: Principal Component Analysis

PD: Conjunto de dados com pacientes Pré-diabéticos e Diabéticos

PG: *Plasma Glucose* - Glicose Plasmática

PICO: População, Intervenção, Comparação e Resultados

PR: Precisão ou Valor Preditivo Positivo

PRN: Valor Preditivo Negativo

RD: Retinopatia Diabética

RDNP: Retinopatia Diabética Não Proliferativa

RDP: Retinopatia Diabética Proliferativa

RDW**:** *Red cell Distribution Width*

RF: Random Forest – Floresta Aleatória

RMSE: *Raiz Mean Square Error* - Raiz do Erro Médio Quadrático

ROC: *Receiver Operating Characteristic Curve*

SD: *Standard deviation* - Desvio Padrão

SN: Sensibilidade ou Taxa de Verdadeira Positivo

SP: Especificidade ou Taxa de Verdadeiro Negativo

SVM: *Support Vector Machine* – Máquina de Vetor de Suporte

TGC: Triglicerídeo

VCM: Volume Corpuscular Médio das hemácias

VLDL: *Very Low Density Lipoprotein* – Lipoproteínas muita baixa densidade

VN: Verdadeiro Negativo

VP: Verdadeiro Positivo

VPM**:** Volume Plaquetário Médio

SUMÁRIO

1. INTRODUÇÃO

Ao observar os avanços da ciência nos últimos anos, é possível encontrar fortes indícios de que o próximo grande desafio será a análise de grandes quantidades de dados (*big data*). Analisando os estudos de Han e Kamber (HAN; KAMBER; PEI, 2012), é possível observar um rápido crescimento no montante de dados coletados nas mais variadas situações do dia-a-dia. Este crescimento tem gerado uma demanda por novas formas de análise de dados complexos e desestruturados, conhecidos até então como mineração de dados (*data mining*).

Segundo Sharma e Mansotra (SHARMA; MANSOTRA, 2014), a área da saúde tem grande destaque na aplicação de mineração de dados, dando apoio no controle de infecções, analises epidemiológicas, tratamento e diagnósticos de doenças, além de gestão hospitalar, *homecare*, administração da saúde pública e gestão de doenças.

Estas novas possibilidades respondem ao fato de que a área de *Big Data* está crescendo e, com isso, a disponibilidade de dados relevantes para fazer previsões de comportamento, riscos e tendências. Devido a esta nova perspectiva, é possível fazer análises muito mais completas e precisas, tomando decisões futuras com base em previsões muito mais confiáveis do que tínhamos até então (HALL; PHAN; WHITSON, [s.d.]).

O processamento computacional tem sido usado para identificar doenças com base no processamento de dados clínicos (BARON; DIGHE, 2014; CHEN et al., 2017; HOSSAIN et al., 2021; LUO et al., 2016; MA et al., 2018). Extrair conhecimento de dados para apoiar especialistas na tomada de decisões é uma tendência na nova geração de sistemas de saúde inteligentes (FERNÁNDEZ-LLATAS; GARCÍA-GÓMEZ, 2015; PEEK et al., 2015). Métodos computacionais, como mineração de dados e aprendizado de máquina, podem melhorar o diagnóstico junto com os dados do paciente. Nos últimos anos vários estudos vêm utilizando testes de laboratório e técnicas de aprendizado de máquina para buscar novos resultados. No caso do Diabetes Mellitus (DM), por se tratar de condição crônica que acomete uma parcela significativa da população mundial (INTERNATIONAL DIABETES FEDERATION, 2019), a busca pelo diagnóstico tem sido alvo da medicina preditiva. Muitos estudos usaram a inteligência artificial para prever um diagnóstico ou uma propensão futura para desenvolver a doença. Em geral, além dos exames laboratoriais, esses estudos utilizam dados clínicos, histórico do

paciente, exames de imagem e diagnósticos médicos (BERNARDINI et al., 2019; CAMARGO; GROSS, 2004; DE SILVA et al., 2021; DU et al., 2021; LAI et al., 2019; METSKER et al., 2020; RAVAUT et al., 2021; WU et al., 2021; YU et al., 2020a).

Esta análise preditiva está fortemente ligada a evolução de técnicas de inteligência artificial, como é o caso da aprendizagem de máquina ou Machine Learning (ML). Estas técnicas são métodos de análise de dados que automatiza o desenvolvimento de modelos analíticos. Usando algoritmos que aprendem interativamente a partir de dados, permitindo que os sistemas baseados em inteligência computacional encontrem informações incialmente ocultas sem serem explicitamente programados para procurar algo especifico (HALL; PHAN; WHITSON, [s.d.]).

Por causa das novas tecnologias de computação, os conceitos de técnicas aplicados no aprendizado de máquina de hoje tem evoluído. Apesar de muitos algoritmos de machine learning já́ estarem sendo utilizados há bastante tempo, a capacidade de aplicar automaticamente cálculos matemáticos complexos sobre big data é um desenvolvimento recente (HALL; PHAN; WHITSON, [s.d.]).

Na área da saúde, vários estudos vêm apresentando bons resultados com o uso de Machine Learning. Weng (WENG et al., 2017) por exemplo, demonstrou a possibilidade de predição do risco a propensão de patologias cardiovasculares, identificando pacientes que poderiam se beneficiar com um tratamento preventivo. Da mesma forma, Holsbach (HOLSBACH; FOGLIATTO; ANZANELLO, 2014), obteve bons resultados na identificação de câncer de mama.

Atualmente, sistemas de predição (CASTRILLÓN; SARACHE; CASTAÑO, 2017), e apoio a tomada de decisão tem se valido de prontuários on-line e dados clínicos, analisando o histórico de pacientes a fim de propor modelos para identificar situações de alto risco, assim como falsos positivos (LUO et al., 2016). Esta medicina de precisão (in sílico) baseada em registro eletrônico de dados relacionados a saúde (EHR, eletronic health record) tem ganhado força diante da possibilidade de tratamentos mais acessíveis e eficientes, voltados às características particulares de cada indivíduo. Neste sentido, Wong (WONG et al., 2018) propõem o uso de Machine Learning tanto para a estruturação e organização dos dados armazenados quanto para a mineração e auxílio no diagnóstico. Da mesma forma, Roy (ROY et al., 2018) utilizou dados de EHR em um pré-teste para prever resultados de exames laboratoriais.

Também fazendo uso de técnicas de aprendizado de máquina, Oliveira (OLIVERA et al., 2017a) observou a possibilidade na detecção do diabetes do tipo 2, ainda não diagnosticado clinicamente, com base em dados clínicos coletados no dia-a-dia. Neste trabalho os resultados também foram satisfatórios, tendo êxito na utilização de algoritmos baseados em redes neurais artificiais e regressão logística.

De forma bastante semelhante, Zheng (ZHENG et al., 2017) também conseguiu ter bons resultados na detecção do diabetes do tipo 2 fazendo uso de dados EHR e técnicas de Machine Learning.

Essa sequência recente de trabalhos envolvendo aprendizado de máquina e dados de EHR, demostra uma tendência mundial no tratamento destes dados. Aumenta a cada dia o número de sistemas e modelos computacionais auxiliando os profissionais da saúde na tomada de decisão e aprimorando o tratamento de saúde de forma mais precisa, eficiente e acessível.

1.1. MOTIVAÇÃO

Laboratórios de análises clínicas apresentam a maioria dos resultados de exames como valores numéricos individuais. No entanto, os resultados destes exames, vistos isoladamente, geralmente têm um significado limitado na obtenção de um diagnóstico. Em seu estudo com ferritina, Luo (LUO et al., 2016) verificou que exames laboratoriais possuem por muitas vezes informações redundantes. Desta forma, por meio de modelos baseados em *Machine Learning*, conseguiu-se prever os resultados de exames laboratoriais de ferritina a partir do conjunto de resultados de outros exames laboratoriais daquele paciente, como o hemograma, fornecendo informações adicionais permitindo refinar o diagnóstico.

No mesmo estudo Luo verificou que em situações onde a ferritina medida em exames laboratoriais era diferente daquela prevista por meio de modelos computacionais, havia a prévia indicação médica de um diagnóstico de anemia. Isto ilustra que a partir de grandes bases de dados, sistemas inteligentes podem melhorar a interpretação dos resultados de exames laboratoriais, permitindo a extração de informações de diagnóstico aprimoradas além de otimizar a seleção destes exames.

De maneira semelhante, Gunčar (GUNČAR et al., 2018) verificou que modelos de *machine learning* podem ser utilizados para prever doenças hematológicas fazendo

uso apenas de exames de sangue. No estudo, Gunčar afirma que exames laboratoriais possuem mais informações do que aquelas comumente consideradas pelos profissionais da saúde.

Demirci e Rosenbaum (DEMIRCI et al., 2016; ROSENBAUM; BARON, 2018) também utilizaram técnicas de *machine learning* na identificação de possíveis erros no processo clínico de realização de exames laboratoriais. Em ambos os trabalhos, os autores obtiveram resultados satisfatórios, demostrando a capacidade de modelos computacionais baseados em aprendizado de máquina auxiliarem na análise de exames laboratoriais. Baron (BARON et al., 2012), da mesma forma, utilizou um algoritmo para geração de arvore de decisão capaz de identificar exames com possíveis problemas oriundos do processo pré-analítico durante execução de exames laboratoriais.

O uso de testes de laboratório e aprendizado de máquina para buscar novos resultados tem sido amplamente explorado nos últimos anos (LUO et al., 2016; METSKER et al., 2020; RICHARDSON; LIDBURY, 2017; SCHNEIDER et al., 2020; SOUZA et al., 2021; TAMUNE et al., 2020; YANG et al., 2020; YU et al., 2020a). Em particular, chamamos a atenção para o trabalho de Park (PARK et al., 2021), que realizou a predição de diversas doenças por meio de exames laboratoriais, mas não incluindo o DM.

A apresentação destes trabalhos nos faz refletir sobre quanta informação pode estar presente em um conjunto de dados de exames laboratoriais de rotina, assim como a potencialidade de exploração e uso de tais dados.

Nosso interesse nesta linha de estudo é motivado pela possibilidade de que exames laboratoriais podem ser utilizados de forma mais abrangente na busca de informações ocultas, descobrindo patologias até então desconhecidas.

No diagnóstico de Diabetes Mellitus, por exemplo, embora o teste de HbA1c (*glycated haemoglobin*) seja recomendado, o teste de FPG (*fasting plasma glucose*) é o mais utilizado. No entanto, este teste pode apresentar variações e inconsistências (SACKS, 2011; TROISI; COWIE; HARRIS, 2000), gerando resultados falso-negativos. Não são raras as discrepâncias no resultado do diagnóstico de DM realizado com o teste FPG comparado ao teste de HbA1c. Dessa forma, é fundamental prever possíveis diagnósticos de DM e recomendar exames complementares para evitar que um paciente assintomático fique sem tratamento adequado e oportuno. Neste caso, a predição de

HbA1c é uma possibilidade para confirmar o diagnóstico dado pelo teste FPG e, em casos discrepantes, pode propor a realização do teste HbA1c com a mesma amostra de sangue já disponível. Essa abordagem evitaria resultados falso-negativos, economizando tempo e custos com exames e tratamentos adicionais.

1.2. PROPOSTA DA TESE

A possibilidade de usar automaticamente dados de exames laboratoriais para buscar informações de novos pacientes é de grande relevância. Essa metodologia pode impactar diretamente nos processos de análise de resultados de exames laboratoriais, sugerindo exames complementares e mais complexos na triagem de novas patologias e contraprova para casos de falso-negativos. Na maioria dos casos, a amostra de sangue já coletada pode ser utilizada, economizando tempo e diminuindo os custos. Assim, esta metodologia apresenta-se como uma inovação para a realização de exames e diagnósticos em laboratórios médicos com retornos significativos para os indivíduos.

Desta forma, por meio de uma parceria com os Laboratórios Médicos Santa Luzia, propomos uma abordagem baseada em aprendizado de máquina que usa dados laboratoriais existentes para rastrear DM com base nos exames laboratoriais realizados com mais frequência, como é o caso do hemograma completo. Esta abordagem inédita, pode então auxiliar na detecção do DM, direcionando o indivíduo para exames complementares. Assim, este trabalho buscou explorar e avaliar diferentes modelos de aprendizado de máquina e configurações de conjuntos de dados para identificar as melhores formas de apoiar o diagnóstico de DM com base em testes laboratoriais de rotina.

1.3. CONTRIBUIÇÃO

Diferentemente de outros trabalhos, que também utilizam dados clínicos e de diagnóstico no processo de predição. Esta pesquisa adotou apenas o uso de exames laboratoriais na predição de novos exames e apoio ao diagnóstico.

No caso do Diabetes Mellitus, o método pode ser utilizado na predição da hemoglobina glicada e na identificação de falsos negativos em exames de FPG, contribuindo para triagem e apoio ao diagnóstico mais preciso. Desta forma, mesmo

pacientes assintomáticos podem ser identificados e iniciar o tratamento mais precocemente, evitando assim complicações futuras causadas pela doença.

Esta metodologia é inovadora e vantajosa para o processo de diagnóstico de laboratórios médico, uma vez que sistemas inteligentes poderiam analisar automaticamente os exames realizados por um paciente e fazer previsões na busca de patologias ocultas. Em casos positivos, seriam gerados alarmes e sugeridos exames complementares, sendo que, na maioria dos casos, a própria amostra coletada poderia ser utilizada na realização de novos exames. Por fim, o diagnóstico ficaria sob o crivo dos médicos, no entanto, estes teriam informações complementares para apoiá-los na tomada de decisão.

1.3.1 Objetivo Geral

Desenvolver um modelo computacional baseado em técnicas de *machine learning* e exames laboratoriais de rotina capaz de predizer outros exames laboratoriais, possibilitando assim a triagem de pacientes e o apoio ao diagnóstico de DM.

1.3.2 Objetivos Específicos

Analisar e identificar técnicas de *machine learning* capaz de predizer exames utilizados no diagnóstico de diabetes por meio de exames laboratoriais de rotina.

Construir um modelo computacional baseado em técnicas de *machine learning* capaz de identificar falsos negativos em exames laboratoriais utilizados no diagnóstico de diabetes.

2. REVISÃO BIBLIOGRÁFICA

Na busca pelo estado da arte no que se refere ao uso de exames laboratoriais na predição de novas informações, realizou-se uma análise exploratória por meio de uma revisão sistemática.

2.1. MÉTODO DE BUSCA

A metodologia adotada na revisão teve por base a busca de trabalhos que fizeram uso de exames laboratoriais de rotina na predição de novas informações.

2.1.1 Estratégia

As buscas foram realizadas em sete bases de dados eletrônicas em periódicos internacionais nas áreas de Engenharia e Ciências da Saúde: Compendex (Engineering Village), EB-SCO (Medline complete), IEEE, Pubmed (Medline), Science Direct, Scopus e Web da Ciência; na língua inglesa com publicações de 2011 a fevereiro de 2022. Registros adicionais foram identificados ainda durante a fase de triagem desta pesquisa, analisando as referências dos artigos incluídos para elegibilidade.

Os princípios PICO (população, intervenção, comparação e resultados) foram usados para agrupar os termos de busca. Como este estudo abordou exames laboratoriais, três princípios foram considerados, e dois operadores booleanos foram utilizados (OR, AND): população ("Clinical Laboratory Test" OR "Laboratory Diagnosis" OR "Blood Count Complete" OR "Rotina Teste de Diagnóstico") E intervenção ("Aprendizado de Máquina") E resultados ("Tomada de Decisão Clínica" OU "Diagnóstico Assistido por Computador" OU "Valor Preditivo de Testes").

Os termos de busca foram definidos com base na lista de termos utilizados na base de dados MeSH. O *Thesaurus Medical Subject Headings* (MeSH) é um vocabulário controlado e hierarquicamente organizado produzido pela *National Library of Medicine*. Ele é usado para indexação, catalogação e pesquisa de informações biomédicas e relacionadas à saúde (MEDICAL SUBJECT HEADINGS - HOME PAGE, [s.d.]). Foram coletados artigos em bases de dados dos últimos 10 anos até janeiro de 2022; as raízes das palavras foram exploradas e todas as variantes dos termos foram encontradas (singular/plural, pretérito, gerúndio, adjetivo comparativo e superlativo; quando

possível). Foram utilizados os seguintes filtros para a área de atuação: medicina, engenharia (industrial, biomédica, elétrica, manufatura e mecânica), robótica, profissões da saúde e multidisciplinar, conforme disponibilidade na base de dados.

2.1.2 Critérios de Inclusão e Exclusão

Os estudos elegíveis continham os seguintes critérios: (1) Utilização de exames laboratoriais; (2) Usando técnicas de ML; (3) escrito em inglês; (4) artigos em texto completo publicados em revistas especializadas.

Os critérios de exclusão foram estudos: (1) Não utilizar exames laboratoriais. (2) Que não procura prever novos resultados.

Os resultados da busca foram exportados para o software online Mendeley®, onde foram removidos os duplicados/triplicados, a extração dos dados foi obtida em texto completo após análise da possível elegibilidade dos artigos.

2.1.3 Análise do Estudo

Em relação à elegibilidade dos estudos, o processo de revisão ocorreu por meio da análise das palavras-chave dos títulos e leitura dos resumos por dois revisores de forma independente. Em caso de dúvida sobre a elegibilidade, o texto completo foi revisado. Nos casos de desacordo entre os dois revisores, a decisão era tomada por consenso, ou um terceiro investigador fazia uma revisão adicional, e a decisão era tomada por arbitragem

2.1.4 Avaliação da Qualidade Metodológica de Estudos

A qualidade dos estudos elegíveis foi avaliada por meio de ferramentas propostas pelo *National Institutes of Health* (NIH) dos Estados Unidos (NIH, 2014). Este estudo incluiu o instrumento de avaliação Cross-sectional Studies (com 14 critérios). O site do NIH (NIH, 2014) disponibiliza ferramentas de avaliação e diretrizes para avaliar a qualidade de cada tipo de estudo, contendo informações explicativas sobre cada item que deve ser avaliado no artigo.

A qualidade da avaliação foi classificada como boa, regular ou ruim, permitindo a análise geral dos avaliadores considerando todos os itens (NIH, 2014). Cada item do instrumento de avaliação recebeu pontuação "x" quando o estudo foi realizado, negativo

"-" quando não realizado e outras opções (não pode determinar - CD, não aplicável - NA e não relatado - NR).

Segundo Wong (WONG; CHEUNG; HART, 2008), estudos observacionais com classificação ≥ 67% de itens positivos indicaram boa qualidade, com 34-66% de verificações positivas como qualidade regular e ≤ 33% do estudo de baixa qualidade.

2.1.5 Extração de Dados

As seguintes características do estudo foram extraídas e descritas: nome dos autores, ano de publicação, título, descrição, conjunto de dados, características, métodos e principais resultados. Os dados deste estudo foram apresentados de forma descritiva e seguiram um fluxo de ações baseado em evidências para relato em revisões sistemáticas e metanálises conhecido como PRISMA *Statement* (WONG; CHEUNG; HART, 2008) e o NIH *Checklist for Systematic Reviews* (NIH, 2014).

2.2. RESULTADOS DA BUSCA

O resultado da busca incluiu 513 estudos potencialmente elegíveis. Primeiramente, 41 artigos duplicados/triplicados foram excluídos e dos 472 artigos restantes, 43 foram considerados elegíveis com base na revisão de títulos, palavras-chave e resumos. Estudos adicionais (30) foram incluídos após busca das referências e citações elegíveis dos artigos, totalizando 73 textos completos para avaliação. Após a revisão destes, 33 estudos foram inelegíveis, encerrando o processo com 40 estudos para avaliação de qualidade (Figura 1).

A Tabela 2.1 apresenta a avaliação da qualidade metodológica dos estudos. Os artigos estão organizados por ano/autor, por enquadramento das questões e pela média de pontos obtida por esta análise. Na sequência tem-se as questões (traduzidas do inglês) que compõem o questionário de acordo com a numeração apresentada na Tabela 2.1. Avaliação da qualidade (QR) metodológica dos estudos. x (sim), - (não), NR (não relatado), NA (não aplicável), CD (não pode determinar), QR (Classificação de Qualidade) : (≥ 67 % = Bom, 33-66 % = Razoável, ≤ 33 % = Ruim)..

1- O problema ou objetivo da pesquisa foi claramente declarado no artigo?
2- A população do estudo foi claramente especificada e definida?

3- A taxa de participação das pessoas elegíveis era de pelo menos 50%?

4- Todos os indivíduos foram selecionados ou recrutados na mesma população ou em populações semelhantes (incluindo o mesmo período de tempo)? Os critérios de inclusão e exclusão para participar do estudo foram pré-especificados e aplicados uniformemente a todos os participantes?

5- Foi fornecida uma justificativa para o tamanho da amostra, descrição do poder ou estimativas de variação e efeito?

6- Para as análises da pesquisa, as exposições de interesse foram medidas antes do (s) resultado (s) ter (em) sido medido (s)?

7- O prazo de estudo foi suficiente para que se pudesse esperar razoavelmente ver uma associação entre a exposição e o resultado, se existisse?

8- Para exposições que podem variar em quantidade ou nível, o estudo examinou diferentes níveis de exposição em relação ao resultado (por exemplo, categorias de exposição ou exposição medida como variável contínua)?

9- As medidas de exposição (variáveis independentes) foram claramente definidas, validadas, confiáveis e implementadas de forma consistente em todos os participantes do estudo?

10- As exposições foram avaliadas mais de uma vez ao longo do tempo?

11- As medidas de resultado (variáveis dependentes) foram claramente definidas, validadas, confiáveis e implementadas de forma consistente em todos os participantes do estudo?

12- Os avaliadores de resultados não tiveram acesso ao status de exposição dos participantes?

13- A perda de acompanhamento após o início do estudo foi de 20% ou menos?

14- As variáveis de confusão potenciais principais foram medidas e ajustadas estatisticamente para seu impacto na relação entre exposição e resultado?

A Tabela 2.2 apresenta a descrição dos estudos incluídos nesta revisão. Está organizado por ano/autor, título, descrição, conjunto de dados, características, métodos e resultados principais.

Figura 2.1- Diagrama de fluxo PRISMA para triagem e seleção de trabalhos.

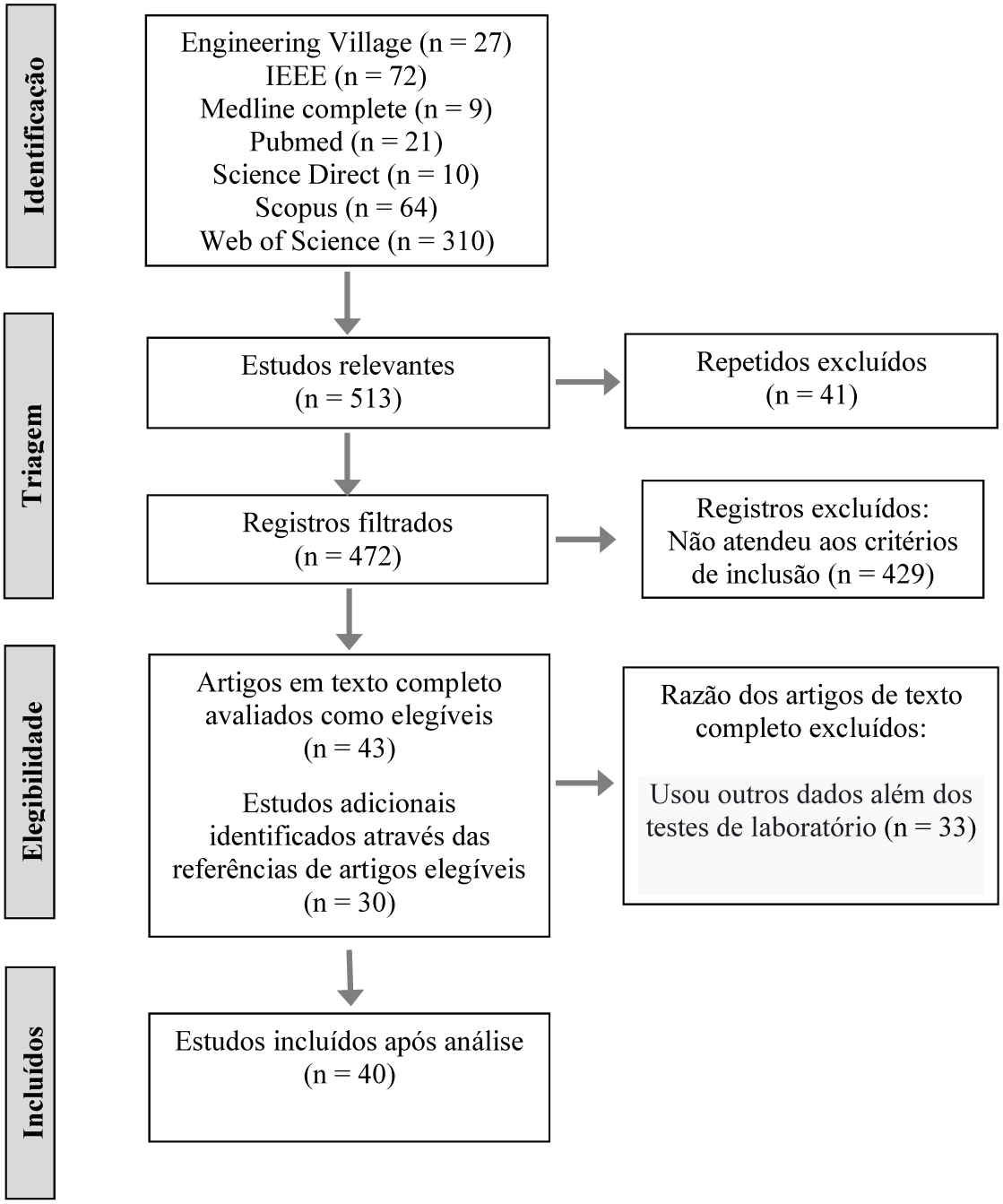


Tabela 2.1. Avaliação da qualidade (QR) metodológica dos estudos. x (sim), - (não), NR (não relatado), NA (não aplicável), CD (não pode determinar), QR (Classificação de Qualidade) : (≥ 67 % = Bom, 33-66 % = Razoável, ≤ 33 % = Ruim).

Ano/Autor	Ferramentas de avaliação de qualidade														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	QR
2013-06 Richardson and Lidbury	X	X	X	X	X	X	X	X	X	NA	X	X	X	X	93%
2013-06 Waljee et al.	X	X	X	X	X	X	X	X	CD	NA	CD	X	X	X	79%
2016-02 Kinar et al.	X	X	X	CD	X	X	X	CD	X	X	X	X	X	X	86%
2016-06 Luo et al.	X	X	X	X	X	X	X	X	X	NA	X	X	X	X	93%
2016-08 Razavian et al.	X	X	X	X	X	X	X	NA	X	X	X	X	X	X	93%
2017-08 Richardson and Lidbury	X	X	X	X	X	X	X	X	X	NR	X	X	X	X	93%
2017-09 Birks et al.	X	X	X	X	X	X	X	X	CD	NA	X	X	X	X	86%
2017-12 Hernandez et al.	X	X	X	X	X	X	CD	CD	X	X	X	X	X	X	86%
2018-05 Roy et al.	X	X	X	X	X	X	X	CD	X	X	X	X	X	X	93%
2019-04 Rawson et al.	X	X	X	X	X	X	X	X	X	NA	X	X	X	X	93%
2019-05 Aikens et al.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	100%
2019-05 Hu et al.	X	X	X	X	X	X	CD	NA	X	CD	X	X	X	X	79%
2019-09 Bernardini et al.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	100%
2019-09 Xu et al.	X	X	X	X	X	X	X	CD	X	X	X	X	X	X	93%
2019-10 Lai et al.	X	X	X	X	X	X	X	X	X	NA	X	X	X	X	93%
2020-02 Tamune et al.	X	X	X	X	X	X	CD	X	X	NA	X	X	X	X	86%
2020-02 Chicco and Jurman	X	X	X	X	X	X	X	X	X	NA	X	X	X	X	93%
2020-04 Yu, Zhang, et al.	X	X	X	X	X	X	X	CD	NR	X	X	X	X	X	86%
2020-06 Banerjee et al.	X	X	X	X	X	X	X	NA	X	NA	X	X	X	X	86%
2020-06 Joshi et al.	X	X	X	X	X	X	X	NA	CD	NA	X	X	X	X	79%
2020-07	X	X	X	X	X	X	X	NA	X	NA	X	X	X	X	86%

Brinati et al.															
2020-10 Metsker et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2020-10 AlJame et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2020-10 Yadaw et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**CD**	**NA**	**X**	**X**	**X**	**X**	**79%**
2020-10 Cabitza et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**NA**	**X**	**X**	**X**	**X**	**86%**
2020-11 Schneider et al.	**X**	**X**	**X**	**X**	**X**	**X**	**CD**	**X**	**CD**	**NA**	**X**	**X**	**X**	**X**	**79%**
2020-11 Yang et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2020-12 Plante et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**CD**	**X**	**NA**	**X**	**X**	**X**	**X**	**86%**
2020-12 Mooney et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**CD**	**X**	**X**	**X**	**X**	**X**	**X**	**93%**
2020-12 Yu, Li, et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2021-03 Kaftan et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2021-04 Park et al.	**X**	**X**	**X**	**X**	**X**	**X**	**CD**	**X**	**CD**	**NA**	**X**	**X**	**X**	**X**	**79%**
2021-05 Souza et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2021-05 Kukar et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2021-06 Gladding et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**CD**	**NA**	**X**	**X**	**X**	**X**	**79%**
2021-08 AlJame et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**NA**	**X**	**X**	**X**	**X**	**86%**
2021-09 Rahman et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**NA**	**X**	**X**	**X**	**X**	**86%**
2021-10 Myari et al	**X**	**X**	**X**	**X**	**X**	**X**	**CD**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**93%**
2021-12 Campagner et al	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**X**	**X**	**X**	**93%**
2022-02 Babaei Rikan et al.	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**NA**	**X**	**NA**	**X**	**X**	**X**	**X**	**86%**

Tabela 2.2 – Descrição resumida dos artigos selecionados na revisão.

Ano Autor	Título Original	Descrição	Conjunto de Dados	Atributos	Métodos	Principais Resultados
2013-06 Richardson and Lidbury (RICHARDSON; LIDBURY, 2013)	**Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data**	Este artigo investiga o efeito do pré-processamento de dados, o uso de conjuntos construídos por ensacamento e uma votação por maioria simples para combinar previsões de classificação de dados laboratoriais de rotina de patologia, particularmente para superar um desequilíbrio significativo de vírus negativos da hepatite B (HBV) e hepatite Casos de vírus C (HCV) versus casos positivos de imunoensaio de HBV ou HCV.	Para atingir esse objetivo, interrogamos um conjunto de dados de 18.625 registros de 1997 a 2007 disponibilizados pela ACT Pathology no The Canberra Hospital, ACT Australia.	Idade, sexo, alanina aminotransferase, gama-glutamil transpeptidase, hemoglobina, hematócrito, hemoglobina corpuscular média, concentração de hemoglobina corpuscular média, volume corpuscular médio, plaquetas, contagem de glóbulos brancos, contagem de glóbulos vermelhos, largura de distribuição de glóbulos vermelhos, neutrófilos, linfócitos, monócitos, eosinófilos e basófilos.	Implementamos a análise usando o algoritmo RPART em R (Decision Tree)	Foi mais fácil prever casos positivos de imunoensaio do que casos negativos de HBV ou HCV.
2013-06 Waljee et al. (WALJEE et al., 2013)	**Comparison of imputation methods for missing laboratory data in medicine**	Compara a precisão de quatro métodos de interpolação para dados laboratoriais faltantes inteiramente aleatórios e compara o efeito dos valores interpolados na precisão de dois modelos clínicos preditivos.	A coorte Cirrose teve 446 pacientes e a coorte Doença Inflamatória Intestinal teve 395 pacientes de uma instituição de nível terciário em Ann Arbor, Michigan.	Parâmetros de hemograma completo (FBC).	MissForest, interpolação média, interpolação do vizinho mais próximo e interpolação multivariada por equações encadeadas (MICE) para interpolar os dados ausentes simulados.	MissForest teve um erro de interpolação menor para variáveis contínuas e categóricas em cada frequência de falta, e teve uma diferença de previsão menor quando os modelos usaram valores laboratoriais interpolados.
2016-02 Kinar et al. (KINAR et al., 2016)	**Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study**	Desenvolver e validar um modelo para identificar indivíduos com risco aumentado de câncer colorretal (CCR).	2 milhões de pacientes dos Serviços de Saúde Maccabi em Israel e da Rede de Melhoria da Saúde do Reino Unido (THIN).	20 parâmetros de hemograma completo mais idade e sexo.	Modelo de aumento de gradiente e classificador de floresta aleatória.	AROC para detectar CRC foi $0,82 \pm 0,01$ para o conjunto de validação israelense.
2016-06 Luo et al. (LUO et al., 2016)	**Using Machine Learning to Predict Laboratory Test Results**	Aprendizado de máquina usado para prever valores de ferritina a partir de	989 pacientes internados no hospital terciário em Boston, Massachusetts,	Idade, sexo, mais 41 exames laboratoriais	Foram utilizadas quatro técnicas de regressão: regressão linear, regressão	O modelo pode prever resultados de ferritina com alta precisão (área sob a

		resultados de testes de laboratório.	coletados ao longo de 3 meses em 2013.		linear Bayesiana, regressão de floresta aleatória (RFR) e regressão de lasso.	curva tão alta quanto 0,97, dados de teste retidos).
2016-08 Razavian, Marcus, and Sontag (RAZAVIAN; MARCUS; SONTAG, 2016)	**Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests**	Uso de medidas longitudinais de testes laboratoriais e avaliação do modelo na identificação do início da doença.	Exames laboratoriais e informações de diagnóstico de 298.000 indivíduos de uma coorte com mais de 4,1 milhões de assinantes de seguros entre 2005 e 2013.	Creatinina, Nitrogênio Ureia, Potássio, Glicose, Alanina Aminotransferase, Aspartato Aminotransferase, Proteína, Albumina, Colesterol, Triglicerídeo, Colesterol.in LDL, Cálcio, Sódio, Cloreto, Dióxido de Carbono, Nitrogênio Ureia/Creatinina, Bilirrubina, Albumina/Globulina.	Treinamos uma rede neural recorrente de Long Short-Term Memory (LSTM) e duas novas redes neurais convolucionais para previsão multitarefa do início da doença.	Descobrimos que as abordagens de aprendizado baseadas em representação superam significativamente essa linha de base, sugerindo um novo caminho para a estratificação de risco do paciente com base apenas em resultados de laboratório.
2017-08 Richardson and Lidbury (RICHARDSON; LIDBURY, 2017)	**Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines**	Explorar o desempenho de três métodos de balanceamento e um método de seleção de recursos para avaliar a capacidade dos SVMs de classificar dados de patologia diagnóstica desequilibrados associados ao diagnóstico laboratorial de infecções por hepatite B (HBV) e hepatite C (HCV).	O conjunto de dados empregado neste estudo originalmente compreendeu 18.625 casos individuais de testes de vírus da hepatite ao longo de uma década de 1997 a 2007.	Idade, sexo, alanina aminotransferase, gama-glutamil transpeptidase, hemoglobina, hematócrito, hemoglobina corpuscular média, concentração de hemoglobina corpuscular média, volume corpuscular médio, plaquetas, contagem de glóbulos brancos, contagem de glóbulos vermelhos, creatinina; potássio, fosfato alcalino, albumina, níveis de bilirrubina total, sódio, uréia sanguínea, largura de distribuição de eritrócitos, neutrófilos, linfócitos, monócitos, eosinófilos e basófilos.	Florestas aleatórias (RFs) para seleção de variável preditora e reformulação de dados para superar um grande desequilíbrio de resultados de testes negativos a positivos em relação aos resultados de imunoensaios de HBV e HCV são examinados.	A geração de conjuntos de dados pela Synthetic Minority Oversampling Technique (SMOTE) resultou em uma previsão significativamente mais precisa do que o downsizing único ou múltiplo (MDS) do conjunto de dados.
2017-09 Birks et al. (BIRKS et al., 2017)	**Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records**	Avalia um algoritmo de risco existente derivado em Israel que identifica indivíduos de acordo com o risco de câncer colorretal usando dados de hemograma completo,	2.550.119 pacientes com mais de 40 anos do Clinical Practice Research Datalink.	Idade, sexo, exame de hemograma.	Aplicação do algoritmo na análise caso-controle de pacientes submetidos a hemograma completo durante o ano de 2012, para estimar valores preditivos.	O algoritmo oferece um meio adicional de identificar o risco de câncer colorretal e pode apoiar outras abordagens para detecção precoce, incluindo triagem e descoberta ativa de casos.

		usando dados do Clinical Practice Research Datalink (CPRD) do Reino Unido.				
2017-12 Hernandez et al. (HERNANDEZ et al., 2017)	**Supervised learning for infection risk inference using pathology data**	Avalia o desempenho de diferentes classificadores binários para detectar qualquer tipo de infecção a partir de um conjunto reduzido de medições clínicas comumente solicitadas.	Dados de patologia e microbiologia para pacientes do Imperial College Healthcare NHS Trust.	Alanina aminotransferase; fosfatase alcalina; Bilirrubina; Creatinina; Proteínas C-reativas Hemograma.	Algoritmos de aprendizado de máquina supervisionados para classificação binária (Gaussian Naïve Bayes, classificador de árvore de decisão, classificador de floresta aleatória e máquina de vetor de suporte).	AUC ROC (0,80-0,83), sensibilidade (0,64-0,75) e especificidade (0,92-0,97).
2018-05 Roy et al. (ROY et al., 2018)	**Predicting Low Information Laboratory Diagnostic Tests**	Descreve a prevalência de exames laboratoriais comuns em ambiente hospitalar e a taxa de resultados "normais" para quantificar as probabilidades pré-teste sob diferentes condições.	Registros médicos eletrônicos (Epic) de 71.000 pacientes internados no Stanford Terciary Academic Hospital entre os anos de 2008-2014.	Testes laboratoriais comuns (por exemplo, hormônio estimulante da tireóide, protocolo de sepse com lactato, ferritina, peptídeo natriurético pró-cérebro N-Terminal).	Fornece um método sistemático orientado por dados para identificar casos em que o valor incremental do teste vale a pena reconsiderar.	Achamos que testes laboratoriais de baixo rendimento são comuns (por exemplo, ~ 90% das hemoculturas são normais).
2019-04 Rawson et al. (RAWSON et al., 2019)	**Supervised machine learning for the prediction of infection on admission to hospital: A prospective observational cohort study**	Um algoritmo SML foi desenvolvido para classificar os casos em infecção versus não infecção usando registros de microbiologia e seis parâmetros sanguíneos disponíveis.	Este estudo foi realizado no Imperial College Healthcare NHS Trust (ICHNT), composto por três hospitais universitários de ensino. O estudo ocorreu entre outubro de 2017 e março de 2018 com 160203 indivíduos.	Proteína C reativa, contagem de leucócitos, bilirrubina, creatinina, ALT e fosfatase alcalina.	O algoritmo classificador binário SVM foi desenvolvido e incorporado ao EPIC IMPOC CDSS para investigação dentro deste estudo após validação e avaliação piloto.	O grupo de infecção teve uma probabilidade de 0,80 (0,09) e o grupo de não infecção de 0,50 (0,29) ($P < 0,01$; IC 95%: 0,20–0,40). ROC AUC foi de 0,84 (IC 95%: 0,76–0,91).
2019-05 Aikens, Balasubramanian, and Chen (AIKENS; BALASUBRAMANIAN; CHEN, 2019)	**A machine learning approach to predicting the stability of inpatient lab test results**	Desenvolvimento de um modelo preditivo que possa identificar exames laboratoriais de baixa informação antes de serem solicitados.	Foram analisados seis anos (2008-2014) de dados de internação do Stanford University Hospital, um hospital universitário terciário.	Troponina, Hormônio Estimulante da Tireóide, Contagem de Plaquetas, Fosfato no soro/plasma, Tempo de Tromboplastina Parcial, NT-PROBNP, Magnésio, Lipase, Lactase, Atividade da Heparina, Ferritina, Creatinina quinase, Proteína C reativa.	Seis modelos diferentes de aprendizado de máquina para classificação: uma árvore de decisão, um classificador de árvore impulsionado (adaboost), uma floresta aleatória, um classificador gaussiano naive Bayes, uma regressão logística regularizada por laço e uma regressão linear seguida de arredondamento para 0 ou 1.	Uma grande proporção de testes repetidos está dentro de ±10% ou ±0,1 SD da medição anterior, indicando que um grande volume de testes repetitivos pode estar contribuindo com poucas informações novas.
2019-05 Hu et al. (HU et al., 2019)	**Using Biochemical Indexes to Prognose Paraquat-Poisoned Patients: An Extreme Learning Machine-Based Approach**	Explorar índices úteis de testes bioquímicos e identificar seu valor preditivo no prognóstico de pacientes envenenados por	Os índices bioquímicos de 101 pacientes intoxicados por paraquat foram hospitalizados na sala de emergência do First	Bilirrubina total, bilirrubina direta, bilirrubina indireta, proteína total, albumina, razão albumina-globulina, alanina aminotransferase,	Um modelo eficaz de máquina de aprendizado extremo (ELM) foi desenvolvido para tarefas de classificação.	Um novo método para prognóstico de envenenamento PQ com precisão de 79,6%.

		QP.	Affiliated Hospital of Wenzhou Medical University de 2013 a 2017.	aspartato aminotransferase, razão de AST para ALT, glicemia, nitrogênio ureico, creatinina.		
2019-09 Bernardini et al. (BERNARDINI et al., 2019)	**TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records**	O estudo visa descobrir fatores clínicos não triviais em dados de EHR para determinar onde a condição de resistência à insulina é codificada.	2.276 registros de 968 pacientes não afetados por DM2. O período de observação longitudinal do paciente foi de 2010 a 2018. (conjunto de dados FIMMG_obs).	2,27 pacientes 968 pacientes6 não registrados por DM2. O período de observação longitudinal do paciente foi de 2010 a 2018. (conjunto de dados FIMMG_obs).	Abordagem de aprendizado de máquina altamente interpretável (ou seja, floresta de regressão de conjunto combinada com estratégias de imputação de dados), denominada TyG-er	Alta concordância (de 0,664 a 0,911 do coeficiente de correlação de Lin ()) da abordagem TyG-er e poder preditivo da abordagem TyG-er (até erro absoluto médio de 5,68% e rc=0,666, p<05).
2019-09 Xu et al. (XU et al., 2019)	**Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests**	Identificar testes laboratoriais de diagnóstico de pacientes internados com resultados previsíveis que provavelmente não fornecerão novas informações.	116.637 pacientes internados no Hospital da Universidade de Stanford de janeiro de 2008 a dezembro de 2017; 60.929 pacientes internados tratados na Universidade de Michigan de janeiro de 2015 a dezembro de 2018; e 13.940 pacientes internados atendidos na Universidade da Califórnia, São Francisco, de janeiro a dezembro de 2018, foram avaliados	Os principais recursos incluíam dados demográficos do paciente, normalidade do teste de interesse mais recente, número de testes recentes de interesse, histórico das categorias do Índice de Comorbidade de Charlson, qual equipe de especialidade estava tratando o paciente, tempo desde a admissão, hora do dia e ano do teste, e estatísticas resumidas de estatísticas vitais recentes e resultados laboratoriais.	Regressão logística regularizada, Naïve Bayes, rede neural perceptrons multicamadas, árvore de decisão, floresta aleatória, AdaBoost e XGBoost.	Os resultados sugerem que os testes de diagnóstico de baixo rendimento são comuns e podem ser sistematicamente identificados por meio de métodos baseados em dados e previsões baseadas no contexto do paciente.
2019-10 Lai et al. (LAI et al., 2019)	**Predictive models for diabetes mellitus using machine learning techniques**	O objetivo deste estudo foi construir um modelo preditivo eficaz com alta sensibilidade e seletividade para melhor identificar pacientes canadenses em risco de ter Diabetes Mellitus com base em dados demográficos do paciente e nos resultados laboratoriais durante suas visitas às instalações médicas.	13.309 pacientes canadenses com idade entre 18 e 90.	Idade, sexo, glicemia de jejum, índice de massa corporal, lipoproteína de alta densidade, triglicerídeos, pressão arterial e lipoproteína de baixa densidade	Modelos preditivos utilizando técnicas de Regressão Logística e Gradient Boosting Machine (GBM).	O AROC para o modelo GBM proposto é de 84,7% com sensibilidade de 71,6% e o AROC para o modelo de Regressão Logística proposto é de 84,0% com sensibilidade de 73,4%.
2020-02 Tamune et al. (TAMUNE et al., 2020)	**Efficient Prediction of Vitamin B Deficiencies via Machine-Learning Using Routine Blood Test Results in Patients with**	Preveja a deficiência de vitamina B usando modelos de aprendizado de máquina a partir das características do paciente e resultados de exames de sangue de rotina	Avaliados 497 pacientes internados no Departamento de Neuropsiquiatria do Tokyo Metropolitan Tama Medical Center, entre	Idade, sexo e mais 29 exames de sangue de rotina.	Modelos de aprendizado de máquina (k-vizinhos mais próximos, regressão logística, máquina de vetor de suporte e floresta aleatória)	O estudo demonstrou que o aprendizado de máquina pode prever com eficiência algumas deficiências de vitaminas em pacientes com

	Intense Psychiatric Episode	que podem ser obtidos em uma hora.	setembro de 2015 e agosto de 2017.			sintomas psiquiátricos ativos.
2020-02 Chicco and Jurman (CHICCO; JURMAN, 2020)	**Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone**	O aprendizado de máquina, em particular, pode prever a sobrevivência dos pacientes a partir de seus dados e pode individualizar os recursos mais importantes entre os incluídos em seus registros médicos.	Registros médicos de 299 pacientes com insuficiência cardíaca coletados no Faisalabad Institute of Cardiology e no Allied Hospital em Faisalabad (Punjab, Paquistão), durante abril-dezembro de 2015.	Idade, anemia, hipertensão arterial, creatinina fosfoquinase, diabetes, fração de ejeção, sexo, plaquetas, creatinina sérica, sódio sérico, tabagismo, período de acompanhamento.	Aplique vários classificadores de aprendizado de máquina para prever a sobrevida do paciente e classificar os recursos correspondentes aos fatores de risco mais importantes.	Nossos resultados desses modelos de dois recursos mostram não apenas que a creatinina sérica e a fração de ejeção são suficientes para prever a sobrevida de pacientes com insuficiência cardíaca a partir de registros médicos, mas também que o uso desses dois recursos sozinhos pode levar a previsões mais precisas do que o uso dos recursos originais do conjunto de dados na sua totalidade.
2020-04 Yu, Zhang, et al. (YU et al., 2020b)	**Predict or draw blood: An integrated method to reduce lab tests**	Proponha um novo método de aprendizado profundo para prever em conjunto futuros eventos de teste de laboratório a serem omitidos.	O conjunto de dados (MIMIC III) continha 598.444 resultados laboratoriais e 5.598.079 registros de sinais vitais de um total de 41.113 pacientes adultos (16 anos ou mais) internados em unidades de terapia intensiva entre 2001 e 2012.	Na (sódio), K (potássio), Cl (cloreto) e HCO3 (bicarbonato sérico), Ca (cálcio total), Mg (magnésio), PO4 (fosfato), BUN (nitrogênio ureico no sangue), Cr (createnina), Hgb (hemoglobina), Plt (contagem de plaquetas), WBC (contagem de leucócitos).	Executamos um novo método de aprendizado profundo sobre quatro combinações de recursos: Lab (dados de teste de laboratório), D (dados demográficos), V (dados vitais que eram média e desvio padrão nas proximidades do teste de laboratório correspondente) e C (codificação para indicar valores ausentes).	Foi capaz de omitir 15% dos testes de laboratório com perda de precisão de previsão <5%.
2020-06 Banerjee et al. (BANERJEE et al., 2020)	**Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population**	O objetivo do estudo foi usar aprendizado de máquina (ML), uma rede neural artificial (ANN) e um teste estatístico simples para identificar pacientes positivos para SARS-CoV-2 a partir de hemogramas completos sem conhecimento dos sintomas ou histórico dos indivíduos.	O conjunto de dados incluído na análise e treinamento contém resultados anonimizados de hemogramas completos de 5.664 pacientes atendidos no Hospital Israelita Albert Einstein (São Paulo, Brasil), de março a abril de 2020, e que tiveram amostras coletadas para realizar o SARS-CoV- 2 teste rt-PCR durante uma visita ao hospital.	Idade, hematócrito, hemoglobina, plaquetas, volume médio de plaquetas, glóbulos vermelhos, linfócitos, concentração média de hemoglobina corpuscular, leucócitos, basófilos, neutrófilos, hemoglobina corpuscular média, eosinófilos, volume corpuscular médio, monócitos e largura de distribuição de glóbulos vermelhos.	Random forest e modelos lineares generalizados regularizados baseados em Lasso e rede neural artificial.	Descobrimos que, com exames de hemograma completo, floresta aleatória, aprendizado superficial e um modelo de RNA flexível predizem pacientes com SARS-CoV-2 com alta precisão entre populações em enfermarias regulares (AUC = 94-95%) e aqueles não internados no hospital ou na comunidade (AUC = 80-86%).

2020-06 **Joshi et al.** (JOSHI et al., 2020)	**A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results**	Preveja a positividade da PCR para SARS-CoV-2 com base nos componentes do hemograma completo e no sexo do paciente.	387 dados de hemograma completo de janeiro de 2020 a março de 2020 solicitados dentro de 24 h de uma PCR SARS-CoV-2 (com base no ensaio da OMS).	Contagem absoluta de neutrófilos, contagem absoluta de linfócitos e hematócrito.	Treinamos um modelo de regressão logística regularizado por L2.	A previsão da positividade do SARS-CoV-2 PCR demonstrou uma estatística C de 78%, uma sensibilidade otimizada de 93%.
2020-07 **Brinati et al.** (BRINATI et al., 2020)	**Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study**	Desenvolva um modelo preditivo, baseado em técnicas de Machine Learning, para prever a positividade ou negatividade do COVID-19.	O conjunto de dados utilizado para este estudo foi disponibilizado pelo IRCCS Ospedale San Raffaele2 e consistiu em 279 casos, extraídos aleatoriamente de pacientes internados naquele hospital no final de fevereiro de 2020 a meados de março de 2020.	Sexo, idade, leucócitos, plaquetas, proteína C reativa, transaminases, transaminases, gama glutamil transferasi, lactato desidrogenase, lactato desidrogenase, neutrófilos, linfócitos, monócitos, eosinófilos, basófilos.	Árvore de Decisão, Árvores Extremamente Randomizadas, K-vizinhos mais próximos, Regressão Logística, Naïve Bayes, Floresta Aleatória, Máquinas de Vetor de Suporte.	Sua precisão varia entre 82% e 86%, e sensibilidade entre 92% e 95%.
2020-10 **Metsker et al.** (METSKER et al., 2020)	**Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study**	Implementação de métodos de aprendizado de máquina para identificação do risco de polineuropatia diabética com base em prontuários eletrônicos estruturados coletados em bancos de dados de sistemas de informação médica.	Registros laboratoriais de 5.425 pacientes entre 2010 e 2017.	Hemoglobina, leucócitos, plaquetas, pH, volume plaquetário médio, creatinina, hemoglobina celular média, neutrófilos, volume corpuscular médio, colesterol, glicose, procalcitonina, largura de distribuição de glóbulos vermelhos, alanina transaminase, bilirrubina, largura de distribuição de plaquetas, lipoproteína de alta densidade, aspartato aminotransferase, leucócitos, troponina, monócitos, bilirrubina, hemácias, triglicerídeos, hematócrito, lipoproteínas de baixa densidade e sangue na urina.	ANN, SVM, árvore de decisão, regressão linear, classificador de regressão logística.	79,82% de precisão, 81,52% de recall, 80,64% de pontuação F1, 82,61% de precisão e 89,88% de AUC usando o classificador de rede neural
2020-10 **AlJame et al.** (ALJAME et al., 2021)	**Ensemble learning model for diagnosing COVID-19 from routine blood tests**	Propomos o ERLX, que é um modelo de aprendizado conjunto para diagnóstico de COVID-19 a partir de exames de sangue de rotina.	Usamos 5.644 amostras de dados com 559 casos confirmados de COVID-19 do conjunto de dados disponíveis publicamente do Hospital Albert Einstein no Brasil.	Hemoglobina, plaquetas, leucócitos, linfócitos, basófilos, eosinófilos, monócitos, neutrófilos, idade, uréia, proteína C reativa, creatinina, potássio, sódio, alanina transaminase, aspartato transaminase, razão normalizada	O modelo proposto utiliza três classificadores: árvores extras, floresta aleatória e regressão logística, combinando suas previsões com um extreme gradient boosting (XGBoost).	O modelo ensemble obteve excelente desempenho com precisão geral de 99,88%, AUC de 99,38%, sensibilidade de 98,72% e especificidade de 99,99%.

				internacional (INR), albumina, D-dímero e tempo de protrombina.		
2020-10 Yadaw et al. (YADAW et al., 2020)	**Clinical Predictive Models for COVID-19: Systematic Study**	O objetivo deste estudo é desenvolver, estudar e avaliar modelos clínicos preditivos que estimam, usando aprendizado de máquina e com base em dados clínicos coletados rotineiramente, quais pacientes provavelmente receberão um teste positivo para SARS-CoV-2 ou exigirão hospitalização ou terapia intensiva	Usamos dados anônimos de uma coorte de 5.644 pacientes atendidos no Hospital Israelita Albert Einstein em São Paulo, Brasil, nos primeiros meses de 2020.	Recebemos 106 medições clínicas, laboratoriais e demográficas de rotina.	Regressão logística (LR), rede neural (NN), floresta aleatória (RF), máquina de vetor de suporte (SVM) e aumento de gradiente (XGB).	Teste positivo para SARS-CoV-2 a priori com sensibilidade de 75% (IC 95% 67%-81%) e especificidade de 49% (IC 95% 46%-51%), pacientes que são SARS-CoV- 2 positivos que requerem hospitalização com área de 0,92 sob a curva característica do operador do receptor (AUC; IC 95% 0,81-0,98) e pacientes positivos para SARS-CoV-2 que requerem cuidados intensivos com AUC 0,98 (IC 95% 0,95-1,00) .
2020-10 Cabitza et al. (CABITZA et al., 2020)	**Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests**	Exames de sangue de rotina podem ser explorados pelo nosso método para diagnosticar COVID-19.	1.925 pacientes internados no pronto-socorro do Hospital San Raffaele (OSR) de fevereiro de 2020 a maio de 2020.	72 características: hemograma completo (CBC), valores bioquímicos, coagulação, hemodiálise e CO-Oximetria, idade, sexo e sintomas específicos na triagem	Random Forest (RF), Naïve Bayes (NB), regressão logística (LR), máquina de vetor de suporte (SVM) e k-vizinhos mais próximos (KNN).	Para o conjunto de dados OSR completo, a área sob a curva característica de operação do receptor (AUC) para os algoritmos variou de 0,83 a 0,90; para o conjunto de dados específico da COVID de 0,83 a 0,87
2020-11 Schneider et al. (SCHNEIDER et al., 2020)	**Validation of an Algorithm to Identify Patients at Risk for Colorectal Cancer Based on Laboratory Test and Demographic Data in Diverse, Community-Based Population**	Validação de uma pontuação preditiva, gerada por um algoritmo de aprendizado de máquina com dados comuns de testes laboratoriais, para identificar pacientes com alto risco de CCR em uma coorte grande, baseada na comunidade e etnicamente diversificada.	A população de coorte elegível do estudo incluiu 2.855.994 membros do Plano de Saúde KPNC entre 1996 e 2015.	Sexo, ano de nascimento e pelo menos 1 teste de hemograma completo, incluindo parâmetros celulares.	Validar a capacidade de um algoritmo que usa informações laboratoriais e demográficas para identificar pacientes com risco aumentado de CCR.	O algoritmo identificou 3% da população que necessita de investigação e identificou 35% dos pacientes que receberam diagnóstico de CCR nos próximos 6 meses.
2020-11 Yang et al. (YANG et al., 2020)	**Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning**	Desenvolvimento de um modelo de aprendizado de máquina integrando idade, sexo, raça e exames de sangue de laboratório de rotina, que estão	5.893 pacientes avaliados no New York Presbyterian Hospital/Weill Cornell Medicine (NYPH/WCM) de março a abril de 2020.5.893 pacientes avaliados no New York Presbyterian	Anion gap, albumina, fosfatase alcalina, bilirrubina indireta, relação creatinina, cálcio, cloreto, globulina, glicose, sódio, proteína total, porcentagem de basófilos, hematócrito,	Foi utilizado um modelo de árvore de decisão de aumento de gradiente (GBDT).	O modelo alcançou uma área sob a curva característica de operação do receptor (AUC) de 0,854 (IC 95%: 0,829-0,878). O modelo também previu a positividade inicial de RT-

		prontamente disponíveis com um TAT curto.	Hospital/Weill Cornell Medicine (NYPH/WCM) de março a abril de 2020.	hemoglobina, leucócitos, contagem de linfócitos, hemoglobina corpuscular média, volume corpuscular médio, monócitos contagem, contagem de neutrófilos, contagem de glóbulos vermelhos, largura de distribuição de glóbulos vermelhos, proteína C reativa, ferritina, ácido lático desidrogenase e magnésio.		PCR para SARS-CoV-2 em 66% dos indivíduos cujo resultado de RT-PCR mudou de negativo para positivo em 2 dias.
2020-12 Plante et al. (PLANTE et al., 2020)	**Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study**	Desenvolvimento de um modelo de aprendizado de máquina para descartar o COVID 19 usando apenas exames de sangue de rotina entre adultos em departamentos de emergência.	O treinamento do modelo usou 2.183 casos confirmados por PCR de 43 hospitais durante a pandemia; controles negativos foram 10.000 pacientes pré-pandêmicos dos mesmos hospitais. A validação externa utilizou 23 hospitais com 1.020 casos confirmados por PCR e 171.734 controles negativos pré-pandêmicos.	Eosinófilos, cálcio total, aspartato aminotransferase, contagem de glóbulos brancos, basófilos, largura de distribuição de glóbulos vermelhos, contagem de glóbulos vermelhos, albumina, bilirrubina total, volume corpuscular médio, hemoglobina corpuscular média, sódio, bicarbonato, nitrogênio ureico no sangue, cloreto.	Modelo de aprendizado de máquina XGBoost.	O modelo encontrou alta discriminação entre subgrupos de idade, raça, sexo e gravidade da doença e teve alto rendimento diagnóstico em pontos de corte de baixa pontuação em uma população de triagem com prevalência de doença <10%. Esse modelo pode identificar rapidamente aqueles com baixo risco de COVID-19 em um método de "exclusão" e pode reduzir a necessidade de testes de PCR nesses pacientes.
2020-12 Mooney et al. (MOONEY et al., 2021)	**Predicting bacteraemia in maternity patients using full blood count parameters: A supervised machine learning algorithm approach**	Uso de ferramentas de aprendizado de máquina para identificar se a bacteremia em mulheres grávidas ou pós-parto pode ser prevista por parâmetros de hemograma completo (hemograma completo) além da contagem de glóbulos brancos.	129 mulheres do Hospital Rotunda em 2019, uma maternidade autônoma de nível terciário na Irlanda.	Contagem de leucócitos (WCC), neutrófilos absolutos, linfócitos, monócitos, eosinófilos, basófilos, razão neutrófilo/linfócito (NLR), plaquetas (PLT), volume plaquetário médio (VPM), razão MPV:PLT e razão monócito/linfócito.	Técnicas de aprendizado de máquina, como particionamento recursivo e árvores de classificação e regressão, foram usadas.	sensibilidade de 27,9% (IC 95% 20,3-36,4), especificidade de 94,1% (93,3-94,8), valor preditivo positivo de 13,9% (10,6-17,9) e valor preditivo negativo (VPN) de 97,4% (97,2-97,7).
2020-12 Yu, Li, et al. (YU et al., 2020a)	**A deep learning solution to recommend laboratory reduction strategies in ICU**	Criação de um modelo de aprendizado de máquina que preveja resultados de testes de laboratório e forneça uma estratégia promissora de redução de testes de laboratório, usando	O conjunto de dados do Medical Information Mart for Intensive Care III com 53.423 internações hospitalares distintas para pacientes adultos internados em unidades de terapia	Sódio, potássio, cloreto, bicarbonato sérico, cálcio total, magnésio, fosfato, nitrogênio ureico no sangue, creatinina, hemoglobina, contagem de plaquetas,	Construiu um modelo de aprendizado profundo com cinco variantes para cada uma das combinações de recursos de entrada: (1) testes de laboratório; (2) exames laboratoriais e	O modelo previu normalidade/anormalidade de testes laboratoriais com uma precisão de 98,27% (AUC, 0,9885; sensibilidade, 97,84%; especificidade, 98,80%;

		correlações espaço-temporais.	intensiva no Beth Israel Deaconess Medical Center.	leucócitos, idade, sexo e raça.	diferenças de tempo entre duas visitas adjacentes; (3) exames laboratoriais e sinais vitais; (4) exames laboratoriais, diferenças de tempo e demografia; (5) exames laboratoriais, sinais vitais, diferenças de tempo e dados demográficos.	VPP, 99,01%; VPN, 97,39%) em 20,26% de testes laboratoriais reduzidos e recomendado 98,10% de transições a serem verificadas.
2021-03 Kaftan et al. (KAFTAN et al., 2021)	**Predictive Value of C-reactive Protein, Lactate Dehydrogenase, Ferritin and D-dimer Levels in Diagnosing COVID-19 Patients: a Retrospective Study**	Avaliar a precisão diagnóstica da PCR, ferritina, LDH e D-dímero na previsão de casos positivos de COVID-19 no Iraque.	O tamanho da amostra foi baseado em sensibilidade e especificidade mínima de 95%, selecionamos aleatoriamente prontuários de 938 indivíduos com suspeita de COVID-entre maio e dezembro de 2020.	Idade, sexo, proteína C reativa, ferritina, LDH e D dime.	Um estudo de coorte observacional retrospectivo baseado nas diretrizes STARD para determinar a precisão diagnóstica do COVID-19	A combinação de biomarcadores laboratoriais de rotina (PCR, LDH e dímero de ferritina ±D) pode ser usada para prever o diagnóstico de COVID-19 com sensibilidade e especificidade aceitas antes de prosseguir para o diagnóstico definitivo por RT-PCR.
2021-04 Park et al. (PARK et al., 2021)	**Development of machine learning model for diagnostic disease prediction based on laboratory tests**	Construção de um novo modelo de conjunto otimizado combinando um modelo DNN (deep neural network) com dois modelos de ML para previsão de doenças usando resultados de testes laboratoriais.	Analisamos conjuntos de dados fornecidos pelo Departamento de Medicina Interna de 5.145 pacientes que visitaram a sala de emergência e aqueles admitidos no Hospital St. Vincent's da Universidade Católica da Coreia em Suwon, Coreia, entre 2010 e 2019.	Utilizou-se um total de 88 atributos, incluindo sexo e idade.	Desenvolvemos um novo modelo de conjunto combinando nosso modelo DL (DNN) com nossos dois modelos de ML (SVM e RF) para melhorar o desempenho da IA.	O modelo de conjunto otimizado alcançou um Escore-F1 de 81% e precisão de previsão de 92% para as cinco doenças mais comuns.
2021-05 Souza et al. (SOUZA et al., 2021)	**Simple hemogram to support the decision-making of COVID-19 diagnosis using clusters analysis with self-organising maps neural network**	Identificar possíveis variáveis em exames de sangue de rotina que podem apoiar a tomada de decisão do médico durante o diagnóstico de COVID-19 na admissão hospitalar.	5644 pacientes alocados no Hospital Albert Einstein em São Paulo, Brasil, na plataforma Kaggle em 2020-03	Durante o processo de treinamento, 14 variáveis presentes no exame de sangue [Hematócrito, Hemoglobina, Plaquetas, Volume Plaquetário Médio, Glóbulos Vermelhos, Linfócitos, MCHC, Leucócitos, Basófilos, MCH, Eosinófilos, MCV, Monócitos e RDW]	Análise de agrupamento não supervisionada com mapas auto-organizáveis de redes neurais (SOM) como estratégia de tomada de decisão.	Foi possível detectar um grupo de unidades do mapa com um poder de discriminação em torno de 83% para pacientes positivos para SARS-CoV-2
2021-05 Kukar et al. (KUKAR et al., 2021)	**COVID-19 diagnosis by routine blood tests using machine learning**	O objetivo do presente estudo é determinar a precisão diagnóstica de um modelo de ML construído	52.306 pacientes admitidos no Departamento de Doenças Infecciosas do University Medical Center	Idade, sexo e mais 35 exames laboratoriais.	Algoritmo Smart Blood Analytics (SBA): um pipeline de aprendizado de máquina baseado em	O modelo exibiu uma alta sensibilidade de 81,9%, uma especificidade de 97,9% e uma AUC de 0,97

		especificamente para o diagnóstico de COVID-19 usando os resultados de exames de sangue de rotina.	Ljubljana (UMCL), Eslovênia, em março/abril de 2020.		CRISP-DM que consiste em cinco estágios de processamento e usa um modelo XGBoost.	
2021-06 Gladding et al. (GLADDING et al., 2021)	**A machine learning PROGRAM to identify COVID-19 and other diseases from hematology data**	Propomos um método para triagem de metadados de hemograma completo para evidências de doenças transmissíveis e não transmissíveis usando aprendizado de máquina (ML).	156.570 dados brutos de hematologia foram coletados entre julho de 2019 e junho de 2020 do Waitakere Hospital e North Shore Hospital.	Um máximo de 247 características FBC de dados CSV foram usados; 134 categóricos, 101 numéricos.	O software Medcalc foi usado para analisar e aplicar modelos de ML, usando árvores de decisão e ensembles, regressão logística e redes neurais profundas (DNN).	Infecção do trato urinário AUROC: 0,68, sensibilidade 52%, especificidade 79%; COVID-19 AUROC: 0,8, sensibilidade 82%, especificidade 75%, IC 95%: 0,79–0,8, p = 0,0006; e área de insuficiência cardíaca sob a curva do operador receptor (AUROC): 0,78, sensibilidade 72%, especificidade 72%, IC 95%: 0,77–0,78; p < 0,0001.
2021-08 AlJame et al. (ALJAME et al., 2021)	**Deep forest model for diagnosing COVID-19 from routine blood tests**	Desenvolvimento de um modelo de previsão de aprendizado de máquina para diagnosticar com precisão o COVID-19 a partir de dados laboratoriais clínicos e/ou de rotina.	5.644 prontuários coletados de março de 2020 a abril de 2020 (Hospital Albert Einstein Israelita localizado em São Paulo, Brasil) e 279 pacientes internados no Hospital San Raffaele, Milão, Itália, do final de fevereiro de 2020 a meados de março de 2020.	Aspartato aminotransferase, contagem de leucócitos, contagem de linfócitos, contagem de neutrófilos, gama glutamil transpeptidase, idade, contagem de basófilos, eosinófilos, alanina aminotransferase, plaquetas, sexo, proteína c-reativa, fosfatase alcalina, lactato desidrogenase e contagem de monócitos.	Modelo de floresta profunda (DF) construído a partir de três classificadores diferentes: árvores extras, XGBoost e LightGBM.	Os resultados experimentais mostram que o modelo DF proposto tem acurácia de 99,5%, sensibilidade de 95,28% e especificidade de 99,96%.
2021-09 Rahman et al. (RAHMAN et al., 2021)	**Mortality Prediction Utilising Blood Biomarkers to Predict the Severity of COVID-19 Using Machine Learning Technique**	Desenvolvimento de modelo de previsão de alto risco de mortalidade para pacientes com COVID-19 e não COVID-19.	654 pacientes com COVID-19 e não COVID-19 foram admitidos no departamento de emergência em Boston (março de 2020 a abril de 2020) e no Hospital Tongji na China (janeiro de 2020 a fevereiro de 2020).	Idade, contagem de linfócitos, D-dímero, PCR e creatinina.	Random Forest, Support Vector Machine (SVM), K-nearest neighbor (KNN), XGBoost, Extra-tree e regressão logística.	Para a coorte de desenvolvimento e coortes de validação interna e externa usando regressão logística, a área sob as curvas (AUCs) foi de 0,987, 0,999 e 0,992, respectivamente.
2021-10 Myari, Papapetrou, and Tsaousi (MYARI; PAPAPETROU; TSAOUSI, 2021)	**Diagnostic value of white blood cell parameters for COVID-19: Is there a role for HFLC and IG?**	Investigar a capacidade da contagem de glóbulos brancos (WBC) e seus subconjuntos, células de linfócitos de alta fluorescência (HFLC), contagem de granulócitos	Estudo retrospectivo de caso-controle realizado com dados coletados de pacientes admitidos no departamento de emergência do Hospital Geral Universitário de Ioannina	Idade, sexo, contagem de basófilos, proteína C reativa, contagem de eosinófilos, células de linfócitos de alta fluorescência, contagem de granulócitos imaturos; razão	Digite a análise de regressão logística binária foi conduzida para determinar a influência dos parâmetros no resultado.	O marcador combinado WBC-HFLC é o melhor marcador diagnóstico para doença leve e grave. CRP e contagem de linfócitos são indicadores precoces de progressão para doença

		imaturos (IG) e proteína C reativa (PCR) para auxiliar no diagnóstico de COVID-19 durante o processo de triagem e como indicadores de progressão da doença para estado grave e crítico.	(Ioannina, Epirus, Grécia) de março de 2020 a março de 2021.	linfócito-monócito; contagem de linfócitos; contagem de monócitos; contagem de neutrófilos; proporção de neutrófilos para linfócitos, proporção de plaquetas para linfócitos, contagem de glóbulos brancos.		grave, enquanto WBC, NEUT, IG e proporção de neutrófilos para linfócitos são os melhores indicadores de doença crítica.
2021-12 Campagner, Carobene, and Cabitza (CAMPAGNER; CAROBENE; CABITZA, 2021)	**External validation of Machine Learning models for COVID-19 detection based on Complete Blood Count**	Avaliar se os modelos de ML para diagnóstico de COVID-19, com base em dados de CBC, podem ser robustos para transportabilidade entre locais e, portanto, podem ser implantados de forma confiável como ferramentas de apoio à decisão médica.	Dados de 1736 pacientes coletados nos departamentos de emergência (ED) do IRCCS Hospital San Raffaele e do IRCCS Istituto Ortopedico Galeazzi de Milão (Itália).	Idade, sexo e 23 exames laboratoriais de rotina.	Floresta aleatória, regressão logística, k-vizinhos mais próximos, máquina de vetor de suporte, Naïve Bayes, conjunto.	Relatamos uma AUC média de 95%. O modelo de melhor desempenho (SVM) relatou uma AUC média de 97,5%.
2022-02 Babaei Rikan et al. (BABAEI RIKAN et al., 2021)	**COVID-19 diagnosis from routine blood tests using artificial intelligence techniques**	Apresentamos o desenvolvimento e comparação de vários modelos para diagnosticar casos positivos de COVID-19 usando três conjuntos de dados de exames laboratoriais de sangue de rotina.	Três conjuntos de dados de estudo de acesso aberto de 2.498 pacientes, contendo dados de exames de sangue de rotina de casos COVID-19 e não COVID-19 foram usados.	Testes laboratoriais de rotina de acordo com cada um dos 3 conjuntos de dados.	Sete métodos de aprendizado de máquina, incluindo Regressão Logística (LR), K Nearest Neighbors (KNN), Árvore de Decisão (DT), Máquina de Vetor de Suporte (SVM), Naïve Bayes (NB), Árvores Extremamente Randomizadas (ET), Floresta Aleatória (RF), e XGBoost, juntamente com quatro métodos de aprendizado profundo, incluindo Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) e Long Short-term Memory (LSTM).	Em média, precisão de 92,11%, especificidade de 84,56% e AUC de 92,20% para o primeiro conjunto de dados, 93,16%, 93,02%, 93,20% para o segundo conjunto de dados e 92,5%, 85%, 92,20% para o terceiro conjunto de dados, respectivamente.

2.3. CONSIDERAÇÕES

O uso de testes laboratoriais e técnicas de aprendizado de máquina aumentou nos últimos anos, principalmente devido à pandemia do COVID-19. Essa metodologia tem o potencial de inovar os processos diagnósticos dos laboratórios médicos e vem despertando o interesse de diversos pesquisadores ao longo do tempo.

Nesta revisão buscou-se trabalhos que utilizaram testes laboratoriais para prever novas informações. No total, encontramos 40 trabalhos referentes à última década, que atendem aos critérios estabelecidos. Destes, 27 estudos foram publicados entre 2020 e 2022, sendo que 19 estão relacionados ao SARS-CoV-2. Todos eles usam testes laboratoriais para prever algumas informações desconhecidas, e a grande maioria dos estudos se concentra na busca pelo diagnóstico.

A pesquisa que inicialmente chamou nossa atenção foi desenvolvida por Luo (LUO et al., 2016) para prever a ferritina para detectar pacientes com anemia. A pesquisa utilizou 41 exames laboratoriais de 989 pacientes internados no Hospital Terciário de Boston, Massachusetts, durante três meses em 2013. O trabalho teve bons resultados, com AUC de 97%. O mais interessante é que mesmo nos casos em que os testes de ferritina foram falsos negativos, o sistema conseguiu detectar anemia. Esse resultado mostra que os exames laboratoriais podem ter mais informações do que aqueles referentes ao exame realizado quando analisados em conjunto.

Rawson (RAWSON et al., 2019) utilizou exames laboratoriais para identificar casos de infecção bacteriana entre 160.203 pacientes hospitalizados ao longo de seis meses. Uma característica interessante nesta pesquisa é que apenas seis testes foram utilizados como parâmetros de entrada (proteína C reativa, contagem de leucócitos, bilirrubina, creatinina, ALT e fosfatase alcalina), obtendo bons resultados, com AUC de 0,84%. A utilização de baixo número de exames é um fator importante na construção do modelo. Essa situação possibilita a utilização de exames já realizados pelos pacientes, tornando o processo de triagem rápido e direto sem coletar mais amostras de sangue desse paciente.

Entre os estudos selecionados, três focaram na predição do câncer colorretal. O câncer colorretal tem alta incidência, sendo responsável por muitas mortes em todo o mundo. A identificação precoce desse tipo de patologia pode ser muito vantajosa para

governos e sistemas de saúde, proporcionando tratamento adequado para evitar o agravamento da doença. Kinar (KINAR et al., 2016) obteve bons resultados na identificação de pacientes com propensão a desenvolver câncer colorretal um ano antes do desenvolvimento da doença. Nesta pesquisa, foram utilizados 20 parâmetros do hemograma completo de aproximadamente 2 milhões de pacientes. Da mesma forma, Birks (BIRKS et al., 2017) utilizou um hemograma completo de 2,5 milhões de pacientes, obtendo uma AUC de 75% para períodos mais longos (3 anos) e 85% para períodos mais curtos (6 meses). Mais recentemente, a Schneider também obteve uma AUC média de 78% em um estudo de cerca de 2,8 milhões de pacientes atendidos durante 1996 e 2015.

Outros três estudos (AIKENS; BALASUBRAMANIAN; CHEN, 2019; ROY et al., 2018; XU et al., 2019) tiveram como objetivo identificar exames que não sofreriam alterações ao longo do tempo, permanecendo classificados como normais, sem necessidade de serem repetidos. Em geral, todos tiveram bons resultados; no entanto, destacamos o trabalho de Xu (XU et al., 2019), que obteve uma AUC superior a 90% em 12 meses de análise.

Uma publicação recente que também chamou nossa atenção foi o trabalho de Park (PARK et al., 2021). Os autores usaram modelos de aprendizado profundo para prever 39 diferenças de doenças na pesquisa, atingindo precisão acima de 90% e uma pontuação F1 de 81% para as cinco mais comuns. Usamos 88 características de 5.145 pacientes que visitaram a sala de emergência.

Todos os estudos apresentados nesta revisão utilizaram exames laboratoriais como dados de entrada, além de alguns dados clínicos como sexo e idade. Alguns estudos utilizaram um número maior de parâmetros, como o trabalho de Schwab (YADAW et al., 2020), que utilizou mais de 100 parâmetros diferentes. Outros utilizaram muito poucos, como é o caso do trabalho de Joshi (JOSHI et al., 2020), que utilizou apenas três parâmetros (contagem absoluta de neutrófilos, contagem absoluta de linfócitos e hematócrito). No entanto, a maioria dos estudos utilizou pouco mais de dez parâmetros, tendo o hemograma completo como fonte primária de dados.

Em relação aos modelos, a grande maioria utilizou métodos de aprendizado de máquina com treinamento supervisionado, quase sempre tendo como alvo o exame responsável pelo diagnóstico. Dentre os modelos mais utilizados, podemos citar: regressão logística, Random Forest, Support Vector Machine e K-nearest neighbor, sendo

treinado como classificador binário. No caso das redes neurais, quase sempre foram utilizadas com técnicas de aprendizado profundo (redes neurais profundas, DNN).

Ao analisar o instrumento de avaliação da qualidade (Tabela 2.1), todos os estudos apresentaram bons resultados, com valor médio de 88%. Como a grande maioria das pesquisas se caracteriza como um estudo de coorte retrospectivo, os dados utilizados foram gerados antes da pesquisa. Assim, as questões 8 e 10 do questionário, referentes aos níveis e quantidades de exposição, foram respondidas principalmente como NA (Não aplicável) ou CD (Não pode ser determinado). Este fato rebaixou um pouco a média no processo de avaliação da maioria dos trabalhos. No entanto, dois estudos (AIKENS; BALASUBRAMANIAN; CHEN, 2019; BERNARDINI et al., 2019) foram avaliados com 100%. Outros 18 trabalhos foram avaliados com 93%, 13 com 86% e 7 com 79%.

A Tabela 2.2 apresenta um resumo das principais características dos estudos encontrados. Além de uma breve descrição da pesquisa, também é possível conhecer de forma simplificada a metodologia e os principais resultados.

Em geral, todos os trabalhos avaliados apresentaram bons resultados, fazendo previsões de acordo com o objetivo da pesquisa. Desta forma, a utilização de testes laboratoriais e técnicas de aprendizado de máquina representam um potencial inovador ao processo de laboratórios médicos, permitindo uma análise mais abrangente dos testes realizados, possibilitando a descoberta precoce de patologias ou erros desconhecidos nos testes realizados. Essa análise automática é muito vantajosa por ser de baixo custo e não interferir nos processos já estabelecidos pelos laboratórios médicos.

3. ASPECTOS CLÍNICOS

Doenças crônicas caracterizam-se por terem uma longa duração e requerer atenção médica contínua e/ou limitando a vida do paciente, sendo muitas vezes assintomáticas no início do seu desenvolvimento. Doenças crônicas, como doenças cardíacas, câncer e diabetes, são as principais causas de morte e incapacidade nos Estados Unidos. Juntas, elas são as principais responsáveis pelos mais de US$ 3,5 trilhões gastos anualmente com assistência médica (ABOUT CHRONIC DISEASES | CDC, [s.d.]; RAGHUPATHI; RAGHUPATHI, 2018), (SBPC/ML | SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA/MEDICINA LABORATORIAL, [s.d.]). No Brasil, as doenças crônicas são responsáveis por um elevado número de mortes prematuras, além de impactar tanto economicamente quanto na qualidade de vida da sociedade. No caso da diabetes, a pesquisa nacional em saúde (ROBERTO NUNES GUEDES SECRETÁRIO ESPECIAL DE FAZENDA WALDERY RODRIGUES JUNIOR et al., [s.d.]) estimou que 7,7% da população com mais de 18 anos é diagnosticado com a doença.

3.1. DIABETES MELLITUS

Sempre que nos alimentamos, os carboidratos que ingerimos são metabolizados em glicose, sendo esta uma das principais fontes de energia utilizadas pelas células. Produzido pelo pâncreas, a insulina é um hormônio que age como uma espécie de chave, permitindo que a glicose presente no sangue seja levada para dentro das células, produzindo energia. O Diabetes Mellitus, ou simplesmente diabetes, é um distúrbio metabólico crônico causado pela deficiência na produção de insulina ou pela dificuldade das células do corpo em fazer uso desta insulina. Com o tempo isto causa o aumento do nível de glicose no sangue, conhecido como hiperglicemia, estando esta relacionada a inúmeras complicações com a saúde do indivíduo (AMERICAN DIABETES ASSOCIATION STANDARDS OF MEDICAL CARE IN DIABETES-2017, [s.d.]; DIRETRIZES DA SOCIEDADE BRASILEIRA DE DIABETES 2017-2018, 2017).

O diabetes é classificado principalmente como do tipo 1, do tipo 2 e gestacional. O diabetes do tipo 1 (DM1) pode ocorrer em qualquer idade da vida, sendo mais frequente em crianças e adolescentes, correspondendo entre 5 a 10% dos casos. O DM1 se caracteriza pela pouca ou nenhuma produção de insulina, decorrente da destruição das

células β pancreáticas, necessitando de injeções diárias de insulina para manter os níveis de glicose sob controle (AMERICAN DIABETES ASSOCIATION STANDARDS OF MEDICAL CARE IN DIABETES-2017, [s.d.]).

Já o diabetes do tipo 2 (DM2) abrange mais de 90% dos casos, ocorrendo geralmente em indivíduos mais velhos (a partir dos 40 anos), embora também possa ocorrer em jovens e crianças (RAO, 2015). O DM2 caracteriza-se principalmente pela resistência dos tecidos a ação da insulina, cujas causas ainda não foram totalmente esclarecidas, mas tendo forte relação com fatores comportamentais do indivíduo, como hábitos alimentares, inatividade física e obesidade (AMERICAN DIABETES ASSOCIATION STANDARDS OF MEDICAL CARE IN DIABETES-2017, [s.d.]; GLOBAL REPORT ON DIABETES WHO Library Cataloguing-in-Publication Data Global report on diabetes, 2016). Por se tratar de uma doença quase sempre assintomática o indivíduo pode passar um longo período de tempo sem ser corretamente diagnosticado. O diagnóstico, por sua vez, é realizado por dosagens laboratoriais de rotina ou do surgimento de complicações crônicas quando a doença já está em estado mais avançado (DIABETES UK, 2018).

Estima-se que aproximadamente 8,8% da população mundial com idade entre 20 e 79 anos, ou seja, em torno de 463 milhões de pessoas estejam com diabetes no mundo. Deste total, cerca de 50% não foram diagnosticados e não sabem que tem a doença, gerando atraso no tratamento e aumentando a morbidade e mortalidade, a redução da qualidade de vida e dos custos em saúde (INTERNATIONAL DIABETES FEDERATION, 2019).

Em todo mundo, os custos com DM representam cerca de 1/8 do total gasto com saúde, estando entre as enfermidades com maior índice de mortalidade, sendo responsável por mais de 10% dos óbitos no mundo (INTERNATIONAL DIABETES FEDERATION, 2019).

Desta forma, um diagnóstico precoce é extremamente importante para se evitar maiores complicações e reduzir os custos com o tratamento. No entanto, não é isto que acontece, já que praticamente metade dos afetados pela doença não tem consciência de que a possuem (INTERNATIONAL DIABETES FEDERATION, 2019).

Isso ocorre muito devido a característica, inicialmente, assintomática da doença, sendo que apenas 4 em cada 10 pessoas com DM tipo 2 apresentam algum sintoma.

Dentre os sintomas apresentados por pessoas com diabetes, podemos citar a necessidade constante de urinar, sede excessiva e perda de peso (DIABETES UK, 2020)

Caso a doença não seja devidamente tratada, o paciente poderá ser acometido com complicações ao longo do tempo. Dentre as principais complicações do DM podemos citar (DIABETES UK, 2020)

- **Retinopatia**: A retinopatia diabética (RD) é uma doença ocular que as pessoas diabéticas podem desenvolver. A RD é uma doença que afeta os pequenos vasos da retina, região do olho responsável pela formação das imagens enviadas ao cérebro. O surgimento da RD está relacionado principalmente ao tempo de duração do diabetes e ao descontrole da glicemia. Existem dois tipos de retinopatia diabética:
    - A Retinopatia Diabética Não Proliferativa (RDNP): é a forma mais frequente e corresponde ao estágio inicial da doença, onde ocorre hemorragia e vazamento de líquidos de pequenos vasos da retina, levando a um edema local.
    - Retinopatia Diabética Proliferativa (RDP): é o estágio mais avançado da doença, onde áreas da retina deixam de receber sangue devido ao dano permanente nos vasos sanguíneos. Neste caso, ocorre a formação de novos vasos (i.e., neovasos) que causam grandes hemorragias e algumas vezes descolamento da retina.
- **Nefropatia**: O funcionamento incorreto dos rins pode prejudicar a regulação de líquidos e sais nos vasos, de modo que à medida que a doença renal progride, os rins se tornam cada vez menos eficientes e a pessoa pode ficar muito doente. Isso acontece como resultado do acúmulo de resíduos no sangue. Com o tempo isso poderá apresentar sintomas como pés e tornozelos inchados, sangue na urina, sensação de cansaço e falta de ar (DIABETES UK, 2020).
- **Neuropatia**: A neuropatia é uma das complicações a longo prazo que afeta o sistema nervoso, podendo afetar ações como ver, ouvir, sentir e se mover. O diabetes pode causar neuropatia como resultado de altos níveis de glicose no sangue, danificando os pequenos vasos sanguíneos que suprem os nervos. Isso

evita que os nutrientes essenciais cheguem aos nervos, danificando as fibras nervosas (TESFAYE; BOULTON, 2009), (TESFAYE; SELVARAJAH, 2012). De modo geral a neuropatia pode ser classificada com sensorial, autonômica e motora.

- Neuropatia Sensorial: Atinge os nervos relacionados ao toque, temperatura, dor e outras sensações da pele, ossos e músculos para o cérebro. Afeta principalmente os nervos dos pés e das pernas, mas as pessoas também podem desenvolver esse tipo de neuropatia nos braços e mãos (DIABETES UK, 2020).
- Neuropatia Autonômica: Afeta os nervos que transportam informações para órgãos e glândulas. Relacionada a funções automáticas no corpo como controle do intestino, batimentos cardíacos e órgãos sexuais (DIABETES UK, 2020).
- Neuropatia Motora: Atinge diretamente os nervos que controlam o movimento podendo levar a problemas como contrações, câimbras, fraqueza e perda de massa muscular (DIABETES UK, 2020).

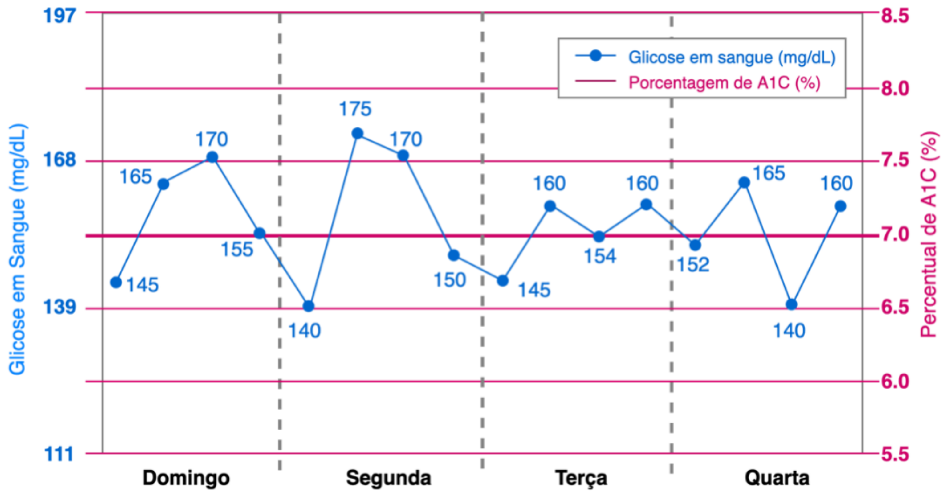
- **Doenças Cardíacas**: O acumulo de açúcar no sangue pode danificar os vasos e comprometer o funcionamento do coração por meio da falta de oxigênio e nutrientes (DIABETES UK, 2020).

3.2. DIAGNÓSTICO

O diagnóstico do DM é realizado por meio da análise de exames laboratoriais como glicose plasmática, glicemia de jejum (*Fasting Plasma Glucose* - FPG) ou glicose plasmática (PG) de 2h, após ingestão de 75g de glicose (*Oral Glucose Tolerance Test*, OGTT), além do HbA1c, também conhecido como hemoglobina glicada. Todos podem ser utilizados no diagnóstico de diabetes, no entanto podem apresentar variações entre eles (ASSOCIATION, 2017; VAN 'T RIET et al., 2010), como mostra a Figura 3.1. Estudos mostram que, em comparação com os pontos de corte do FPG e do HbA1c, o valor de PG de 2 horas diagnostica mais pessoas com diabetes (NATHAN et al., 2009). Entretanto HbA1c possui vantagens como a padronização internacional dos ensaios,

menor variabilidade biológica, não é afetado por estresse agudo, não tem necessidade de jejum, entre outros (ASSOCIATION, 2017; MALKANI; DESILVA, 2012). Desta forma, diante das características existentes entre os diferentes métodos apresentados, o HbA1c tem sido cada vez mais indicado para triagem e diagnóstico da diabetes (SACKS, 2011).

Figura 3.1- Comparação dos exames de glicose em sangue com o HbA1c ao longo de 4 dias. As medidas de glicose em sangue (mg/dL) foram realizadas quatro vezes ao dia (jejum ou antes do café da manhã, antes do almoço, antes do jantar e na hora de dormir).



Fonte: Adaptado de (SACKS, 2011)

Para o teste de glicemia em jejum (FPG), pacientes com níveis de glicose abaixo dos 100 mg/dL são considerados saudáveis. Pacientes com níveis de glicose entre 101 e 125 mg/dL são classificados como pré-diabéticos e pacientes com glicose igual ou acima de 126 mg/dL são considerados diabéticos. No entanto este teste requer que o paciente esteja há pelo menos 8h em jejum (ASSOCIATION, 2017).

Para o teste de glicemia OGTT, de 2 h após a ingestão de 75 g de glicose, os pacientes são considerados saudáveis caso o nível de glicose esteja abaixo de 140 mg/dL, considerados pré-diabéticos caso a glicose enteja entre 140 e 199 mg/dL e diabéticos quando a glicose mensurada for maior ou igual a 200 mg/dL (ASSOCIATION, 2017).

Caso a glicose apresente valores acima de 200 mg/dL, mesmo sem a ingestão de glicose e sem a realização de jejum, mas o paciente também apresente sintomas, então este é considerado com DM (ASSOCIATION, 2017).

No caso da hemoglobina glicada (HbA1c), os pacientes são classificados como saudáveis caso a HbA1c esteja abaixo dos 5,7 % (< 39 mmol/mol), pré-diabéticos caso a

HbA1c esteja entre 5,7 % (>= 39 mmol/mol) e 6,4% (<= 46 mmol/mol) e diabético se for igual ou maior que 6,5 % (>= 48 mmol/mol) (ASSOCIATION, 2017).

Considerando que aproximadamente 60 % dos pacientes não apresentam sintomas na fase inicial da doença (DIABETES UK, 2020), faz-se necessário que alguns destes exames sejam realizados pelo paciente a fim de que ele possa ser diagnosticado. No entanto, geralmente isso não ocorre em situações de rotina com pessoas sem sintomas.

Na Tabela 3.1 é apresentado um resumo com os limites de cada exame diagnóstico para a classificação saudáveis, com pré-diabetes e diabetes.

Tabela 3.1 – Valores limites para a classificação no diagnóstico de diabetes de acordo com cada exames.

Exames	**Saudável**	**Pré-diabetes**	**Diabetes**
FPG (mg/dL)	< 100	100 a 125	> 125
OGGT (mg/dL)	< 140	140 a 199	> 199
HbA1c (%)	< 5,7	5,7 a 6,4	> 6,4

3.3. EXAMES LABORATORIAIS

Os exames complementares são uma parte essencial da prática médica atual. Mesmo representando apenas 2,3 % dos custos com assistência à saúde, os exames laboratoriais são muito importantes na tomada de decisão. Existem hoje mais de 4.000 tipos de exames laboratoriais disponíveis para uso clínico, sendo que cerca de 500 destes são realizados com maior regularidade (WILLIAMSON; SNYDER, 2015). Todo componente de uma amostra que é alvo da análise é chamado de analito.

Os resultados podem ser apresentados de forma qualitativa, como positivo/negativo ou reagente/não-reagente, assim como de forma quantitativa com uso de um intervalo de referência em relação a dosagem de determinado analito (Xavier, 2016). Estes padrões são definidos pelo Instituto de Padronização Clínica e Laboratorial (*Clinical And Laboratoty Standards Institute*, CLSI), mediante a observação e mensuração de analitos em indivíduos selecionados (CLSI STANDARDS & GUIDELINES: SHOP FOR CLSI STANDARDS, [s.d.]).

Para interpretar corretamente um resultado positivo ou negativo, é necessário compreender a acurácia do teste realizado, quantificando as taxas de resultados falso-

positivos e falso-negativo (XAVIER; DORA; BARROS, 2016). A melhor forma de obter isso é por meio de uma matriz de confusão, cujo cruzamento de dados nos fornecerá:

- Taxa de Verdadeiros Positivos ou Sensibilidade: grupo de indivíduos classificados como doentes sendo que realmente eles estão doentes.
- Taxa de Verdadeiros Negativos ou Especificidade: grupo de indivíduos classificados como saudáveis sendo que realmente eles estão saudáveis.
- Taxa de Falso-Positivo: grupo de indivíduos classificados como doentes sendo que eles estão saudáveis.
- Taxa de Falso-Negativo: grupo de indivíduos classificados como saudáveis sendo que eles estão doentes.

A Precisão representa uma concordância entre os resultados, demostrando a capacidade do teste em fornecer resultados próximos entre si. Na prática é a razão de todos os verdadeiros-positivos, por todos os classificados como positivo. Já a Acurácia representa a probabilidade de o teste acertar o diagnóstico (verdadeiro-positivo ou verdadeiro-negativo). O ideal é que tanto a acurácia quanto a precisão sejam altas, sendo completamente inútil o procedimento em que ambas são baixas (XAVIER; DORA; BARROS, 2016).

3.4. EXAMES LABORATORIAIS DE ROTINA

3.4.1 Hemograma Completo

O hemograma completo provê a avaliação dos três componentes principais do sangue periférico (i.e., hemácias, leucócitos e plaquetas), sendo a base de qualquer avaliação hematológica. De forma detalhada, ele relata numericamente o estado de um conjunto de analitos assim como descreve suas características, sendo realizado na maioria das vezes, por máquinas que executam a contagem de forma automática.

O exame é útil na avaliação de anemias, infecções bacterianas e virais, inflamações, leucemias e plaquetopenias (XAVIER; DORA; BARROS, 2016), podendo ocorrer resultados errôneos quando as amostras conterem coágulos, quando o sangue não for adequadamente misturado, ou quando houver eritrócitos aglutinados (WILLIAMSON; SNYDER, 2015).

A seguir uma descrição sucinta dos analitos avaliados no exame de Hemograma.

Hb: Hemoglobina. É o principal componente da hemácia, consistindo em uma metaloproteína conjugada com ferro que realiza o transporte de O_2 e CO_2 pelo sistema vascular.

Ht: Hematócrito. Reflete a concentração de hemácias. É calculado a partir da razão do volume da massa eritrócitos sobre o volume do sangue total.

VCM: Volume Corpuscular Médio das hemácias.

HCM: Hb Corpuscular Média. Quantidade média de Hb pela contagem de hemácias (Hb/hemácias). Consiste na concentração da Hb na hemácia, sendo utilizado principalmente como um instrumento de calibração.

CHCM: Concentração de Hb Corpuscular Média.

RDW: Do inglês *Red cell Distribution Width,* reflete a variação do tamanho das hemácias.

Leucócitos: Contagem global de leucócitos, também conhecidos como glóbulos brancos.

Neutrófilos Segmentados: São os neutrófilos na sua forma madura, formando as células de defesa que compõem o sistema imunológico.

Linfócitos: Tipo de leucócito também relacionado a defesa do organismo.

Monócitos: Tipo de leucócito que se desloca para os tecidos, sendo responsável pela defesa dos mesmos.

Eosinófilos: Tipo de leucócito envolvido com o combate de infecções parasitárias e processos alérgicos.

Basófilos: Liberam mediadores químicos como a histamina e heparina, estando relacionado à processos alérgicos com hipersensibilidade imediata.

Plaquetas: São células responsáveis pelo processo de coagulação. Um indivíduo saudável tem entre 150.000 e 400.000 plaquetas por mm3 de sangue.

VPM: Volume Plaquetário Médio.

Na Tabela 3.2 é apresentado os valores de referência para os principais analitos relacionados às hemácias.

Tabela 3.2 – Valores de referência para medidas relacionadas às hemácias (XAVIER; DORA; BARROS, 2016).

Idade	Hemácias (milhões/mm^3)	Hb (G%)	Ht (%)	VCM (FI)	HCM (PG)	CHCM (%)
Crianças (2-12 anos)	3,8 - 5	11 - 15,5	35 - 44	82 - 95	28 - 32	30 - 34
Mulheres	3,9 - 5,3	12 - 15,5	37 - 46	82 - 96	28 - 32	31 - 35
Homens	4,5 - 5,9	14 - 18	39 - 54	83 - 98	28 - 32	31 - 35

Na Tabela 3.3 são apresentados os valores de referência para os principais analitos da contagem diferencial de leucócitos.

Tabela 3.3 – Valores de referência para contagem diferencial de leucócitos (XAVIER; DORA; BARROS, 2016).

Medidas	Crianças (4-12 anos)	Adultos (homens e mulheres)
Contagem global de leucócitos	6.000-12.000	4.000-10.000
Neutrófilos bastonados	0-700	0-700
Neutrófilos segmentados	2.000-6.000	2.000-7.500
Linfócitos	5.500-8.500	1.500-4.000
Monócitos	700-1.500	200-800
Eosinófilos	300-800	40-400
Basófilos	0-100	0-100

3.4.2 Colesterol

O colesterol é uma substância lipídica presente nas membranas celulares, sendo um precursor dos ácidos biliares e hormônios esteroides. O colesterol circula no sangue em partículas contendo lipídeos e proteínas – as lipoproteínas (Lp). As três principais classes de Lp encontradas no plasma são Lp de baixa densidade (LDL, *low density lipoprotein*), Lp de alta densidade (HDL, *high density lipoprotein*) e Lp de muito baixa densidade (VLDL, *very low density lipoprotein*), que é calculado como sendo a quinta parte do total de triglicerídeos (WILLIAMSON; SNYDER, 2015).

Colesterol LDL: Lipoproteínas de baixa densidade (XAVIER; DORA; BARROS, 2016).

- Ótimo: < 100 mg/dL
- Desejado: 100 a 129 mg/dL
- Limítrofe: 130 a 159 mg/dL
- Alto: 160 a 189 mg/dL
- Muito alto: ≥ 190 mg/dL

Colesterol HDL: Lipoproteínas de alta densidade (XAVIER; DORA; BARROS, 2016).

- Desejável: > 60 mg/dL
- Baixo: < 40 mg/dL

Colesterol VLDL: Lipoproteínas de muito baixa densidade (XAVIER; DORA; BARROS, 2016).

- Normal: de 2 a 30 mg/dL
- Alto: acima de 30 mg/dL

Colesterol Total: O Colesterol Total compreende todas as formas de colesterol encontradas nas lipoproteínas (XAVIER; DORA; BARROS, 2016).

Pediátrico:
- Desejável: < 150 mg/dL
- Limítrofe: 150 a 169 mg/dL
- Alto: > 170 mg/dL

Adulto:
- Desejável: < 200 mg/dL
- Limítrofe: 200 a 239 mg/dL
- Alto: > 240 mg/dL

3.4.3 Creatinina

Cr: A Creatinina é o produto catabólico da fosfato-creatinina, utilizada na contração muscular. Sua dosagem é utilizada para avaliar a função renal e estimar a filtração glomerular, pois sua excreção é realizada pelos rins, principalmente por filtração glomerular (WILLIAMSON; SNYDER, 2015).

- Homens adultos: 0,7 a 1,3 mg/dL
- Mulheres adultas: 0,6 a 1,1 mg/dL
- Pediátrico: 0,3 a 1 mg/dL

3.4.4 Triglicerídeo

TGC: Os TGCs têm como função permitir ao organismo a estocagem de moléculas com longas cadeias de carbono, úteis em processos de formação de energia em estados de jejum prolongado. Essas moléculas altamente energéticas constituem 95% das gorduras estocadas nos tecidos, sendo transportadas no plasma nas lipoproteínas (WILLIAMSON; SNYDER, 2015).

Adultos:
- Normal: até 150 mg/dL
- Limítrofe: 151 a 199 mg/dL
- Alto: 200 a 500 mg/dL
- Muito alto: > 500 mg/dL

Não é raro que pacientes tenham centenas de exames laboratoriais ao longo da vida. No entanto, os médicos podem ignorar resultados importantes ou não observar padrões e tendências presentes no conjunto de dados laboratoriais. Além disso, as próprias limitações humanas ao considerar simultaneamente um grande número de dados, particularmente em associações complexas, podem não conseguir extrair todas as informações úteis dos dados clínicos e laboratoriais existentes, já que por vezes, informações importantes para um diagnóstico, podem estar presentes de forma muito sutil, inter-relacionada e complexa para ser identificada sem um apoio computacional adequado (LOUIS et al., 2014), (DIGHE et al., 2014).

Para interpretar adequadamente estes resultados, os clínicos devem avaliar muitos exames e interpretá-los juntamente com outros dados clínicos, além de considerar o histórico do paciente. Embora esta abordagem manual para interpretação de exames seja o padrão na maioria dos casos, abordagens computacionais para integração e análise de dados laboratoriais podem oferecer um grande potencial na busca por um diagnóstico (LUO et al., 2016).

Com grande evolução nos últimos anos (PEEK et al., 2015), métodos de *Machine Learning* são hoje ferramentas poderosas no apoio ao diagnóstico médico. Estudos recentes [12], [13], [14], têm mostrado que o uso destes métodos são capazes de auxiliar na predição e identificação de doenças com base em exames laboratoriais e dados clínicos de maneira similar a um especialista humano. Da mesma forma, outros estudos (KAVAKIOTIS et al., 2017; ZHENG et al., 2017) têm conseguido auxiliar no diagnóstico de diabetes mellitus, também fazendo uso de técnicas de *machine learning*.

4. MACHINE LEARNING

O aprendizado de máquina (*Machine Learning* - ML) é um campo de estudo em ascensão, cujo principal objetivo é o desenvolvimento de algoritmos que permitem aos computadores aprender (ROYAL SOCIETY, 2017). Watt, em seu livro (WATT; BORHANI; KATSAGGELOS, 2016), define o aprendizado de máquina como um conjunto de métodos que, diante de um conjunto de dados, possam detectar automaticamente padrões e que possa ser utilizado para prever dados futuros ou para executar outros tipos de tomada de decisão.

Aprendizado de máquina é quando um computador, por meio de uma experiência E, melhora sua habilidade em uma tarefa T, de acordo com alguma métrica de performance P (MITCHELL, 1997).

Embora ainda seja uma área com muito mais a ser descoberto do que se sabe atualmente, hoje o aprendizado de máquina pode ser usado para ensinar computadores a executar uma ampla variedade de tarefas úteis. Isso inclui tarefas como a detecção automática de objetos em imagens, reconhecimento de fala, descoberta de conhecimento nas ciências médicas e análises preditivas. Considerado uma subárea da Inteligência Artificial (IA), o aprendizado de máquina utiliza o raciocínio indutivo, metodologia que extrai regras e padrões de grandes conjuntos de dados a fim de alcançar os resultados. Outros métodos da IA também utilizam o raciocínio dedutivo, onde o conhecimento é baseado na lógica de regras pré-definidas. Já a abordagem probabilística do aprendizado de máquina está intimamente relacionada ao campo da estatística, mas difere um pouco em termos de ênfase e terminologia (MURPHY, 2012).

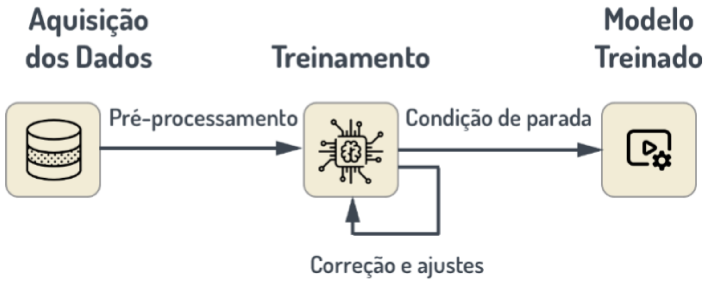
O aprendizado de máquina possibilita tratar problemas formalmente não organizados com o conhecimento humano. Muitos tipos de tarefas podem ser resolvidos com o aprendizado de máquina, como regressão, classificação, reconhecimento de padrões, detecção de anomalias, entre outros (GOODFELLOW; BENGIO; COURVILLE, 2016).

De modo geral, o aprendizado de máquina pode ser dividido em dois tipos principais de abordagem conhecido como aprendizado supervisionado e o não supervisionado. No aprendizado supervisionado ou preditivo, o objetivo é realizar um mapeamento das saídas Y com base nas entradas X. Desta forma, dado um conjunto

rotulado de pares entrada-saída chamado de conjunto de treinamento, é possível construir um modelo que seja capaz de aprender padrões implícitos nestes dados e assim prever ou classificar novos dados (WATT; BORHANI; KATSAGGELOS, 2016).

Um algoritmo de aprendizado supervisionado analisa os dados de treinamento e produz uma função inferida, que pode ser usada para mapear novos exemplos (Figura 4.1). Um cenário ideal permitirá que o algoritmo determine corretamente os rótulos de classe para instâncias desconhecidas (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

Figura 4.1 - Estrutura básica de um modelo de treinamento supervisionado.



4.1. PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento de dados é uma etapa importante no processo de mineração de dados e descoberta de conhecimento, em que as anomalias e inconsistências de dados são detectadas e corrigidas (HAN; KAMBER; PEI, 2012), (PREPROCESSING DATA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]). Durante o pré-processamento também são realizadas a limpeza, integração, redução e transformação dos recursos necessários.

Os recursos são características que definem um determinado conjunto de dados de entrada e definem sua dimensionalidade, permitindo o aprendizado do modelo. De fato, recursos bem definidos são absolutamente cruciais para o desempenho dos modelos de regressão e classificação. No entanto, em termos gerais, a qualidade dos dados utilizados depende do nosso nível de conhecimento sobre o fenômeno estudado. Quanto mais entendermos o processo que gera os dados de entrada, melhor poderemos utiliza-los na modelagem e processo de aprendizagem (WATT; BORHANI; KATSAGGELOS, 2016).

4.1.1 Redução da Dimensionalidade

Muitas vezes temos uma grande variedade de atributos (variáveis de entrada) que podem ser utilizados na construção do modelo estudado. Nestes casos podemos imaginar que quantos mais atributos utilizarmos, melhor será nosso modelo. Mas isso nem sempre é verdade. De fato, após um certo ponto, aumentar a dimensionalidade, do problema adicionando novos atributos prejudicaria o desempenho do modelo. Esse fenômeno conhecido como maldição da dimensionalidade, foi abordado pela primeira vez por Bellman (BELLMAN, 1961), e afirma que o número de amostras necessárias para estimar uma função arbitrária com um determinado nível de precisão cresce exponencialmente em relação ao número de variáveis de entrada. Desta forma, quanto maior o número de atributos utilizados, exponencialmente maior deverá ser o número de amostras necessárias para o treinamento do modelo. Por este motivo, muitas vezes é necessário trabalhar com um número reduzido de atributos, escolhendo apenas os mais significativos e reduzindo a dimensionalidade do problema (ROSS et al., 2009).

De modo geral existem duas formas de se efetuar a redução da dimensionalidade, podendo ser por extração ou por seleção de atributos. Basicamente, os algoritmos de extração de atributos criam novos atributos a partir de transformações ou combinações do conjunto de dados original. Já os métodos de seleção, selecionam os melhores atributos do conjunto de dados original. A escolha entre seleção e extração de atributos depende do tipo de aplicação e do conjunto de dados de treinamento. Geralmente a seleção de atributos reduz o custo de medição de dados, de forma que os dados selecionados mantêm sua interpretação física. Já os atributos transformados por meio da extração podem prover melhor resultado que os melhores atributos selecionados do conjunto de dados original, mas estas novas características podem não possuir um significado físico (BROWNLEE, 2016).

Scikit-learn é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação Python. Nela encontra-se alguns métodos para se trabalhar com a redução de dimensionalidade (SCIKIT-LEARN: MACHINE LEARNING IN PYTHON — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

Uma técnica de seleção de atributos de maior influência, pode ser realizada com qualquer estimador que possua um atributo *coef_* ou *feature_importances,* dentre os métodos disponibilizados biblioteca scikit-learn. Os atributos são considerados sem

importância e consequentemente removidos, se os valores correspondentes estiverem abaixo do limiar fornecido. Métodos baseados em árvore de decisão como *RandomForestClassifier* e *ExtraTreesClassifier*, geralmente são utilizados como estimadores para avaliar a importância dos atributos (FEATURE SELECTION IN PYTHON WITH SCIKIT-LEARN, [s.d.]).

Seleção Univariada

A seleção univariada de atributos funciona selecionando os melhores atributos com base em testes estatísticos univariados, como é o caso dos métodos *SelectKBest* e *SelectPercentile* que basicamente selecionam os atributos de maior pontuação de acordo com as configurações do usuário (FEATURE SELECTION IN PYTHON WITH SCIKIT-LEARN, [s.d.]).

Eliminação Recursiva

A eliminação recursiva de atributos disponíveis por meio do método *REF*, funciona removendo de forma recursiva os atributos e construindo modelos com os recursos que permanecem. Ele usa a precisão do modelo para identificar quais combinações de atributos contribuem mais para construir o melhor resultado (FEATURE SELECTION IN PYTHON WITH SCIKIT-LEARN, [s.d.]).

Análise Fatorial

Outra técnica bastante poderosa é a redução de dimensionalidade com uso de análise fatorial. Neste processo, as variáveis de entrada (atributos) são testadas a fim de se obter o melhor resultado e avaliar o impacto que cada uma tem na variável de saída. Como resultado, as variáveis de entrada são agrupas de acordo com seu impacto sobre o modelo, sendo que um fator de influência é atribuído a cada grupo e a cada atributo dentro dos grupos (FACTOR-ANALYZER · PYPI, [s.d.]; GOODFELLOW; BENGIO; COURVILLE, 2016).

PCA – Principal Component Analysis

O método PCA, provê um algoritmo para análise de componentes principais que utiliza álgebra linear para transformar o conjunto de dados em um formato reduzido. Uma propriedade do PCA é que você pode escolher o número de dimensões ou o componente

principal no resultado transformado (SKLEARN.DECOMPOSITION.PCA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

É importante observar que, se a redução de dimensionalidade for excessiva, o modelo pode perder o poder de discriminação. Por isso se faz necessário analisar a variação do comportamento do classificador com a dimensionalidade, de forma que seja possível estimar a dimensionalidade ideal para determinado classificador e conjunto de dados (SKLEARN.DECOMPOSITION.PCA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

4.1.2 Normalização

Alguns modelos, como por exemplo as redes neurais artificiais, são fortemente sensíveis a variação na escala dos dados de entrada. Assim, sem o ajuste apropriado, dados de entrada com valores maiores, pela própria natureza dos dados, acabariam por ter maior impacto na construção do modelo. Para resolver este tipo de problema, métodos próprios de ajuste dos dados são implementados (PREPROCESSING DATA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]). A padronização de escala (ou normalização do escore Z) é o processo de redimensionar os atributos para que eles tenham as propriedades de uma distribuição gaussiana com média igual a zero e desvio padrão igual a um. Os escores padrão das amostras são calculados conforme Equação 4.1.

$$z = \frac{X-\mu}{\sigma} \qquad \text{Equação 4.1}$$

Na equação, X Representa o conjunto de dados da população, μ a média e σ o desvio padrão da população.

Já a normalização, também chamada de escala Min-Max reduz o intervalo dos dados, de modo que o intervalo é fixo entre 0 e 1 (ou -1 a 1 se houver valores negativos). Este método funciona melhor nos casos em que a padronização pode não funcionar tão bem. Se a distribuição não for gaussiana ou o desvio padrão for muito pequeno, o escalonador *min-max* funcionará melhor. A desvantagem deste método é a sensibilidade que o mesmo tem aos *outliers*, de modo que se houver *outliers* no conjunto de dados, é uma prática ruim (PREPROCESSING DATA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

A biblioteca *scikit-learn* disponibiliza vários métodos para ajuste dos dados de entrada (PREPROCESSING DATA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]):

- *StandardScaler*: remove a média e dimensiona os dados para a variação da unidade. No entanto, os valores extremos influenciam o cálculo da média empírica e do desvio padrão, o que não garante escalas balanceadas.
- *MinMaxScaler*: redimensiona o conjunto de dados de forma que todos os valores estejam no intervalo [0, 1]. No entanto, dependendo dos *outliers*, essa escala comprime a maioria dos valores próximos a zero.
- *RobustTransformer*: neste método, as estatísticas de centralização e redimensionamento são baseadas em percentis e, portanto, não são influenciadas por um número pequeno de discrepantes marginais muito grandes.
- *Normalizer*: por padrão, uma normalização L2 é aplicada a cada observação para que os valores em uma linha tenham uma norma de unidade. Isso significa que se cada elemento fosse elevado ao quadrado e somado, o total seria igual a 1.

4.2. TREINAMENTO E VALIDAÇÃO

O aprendizado supervisionado é um método de aprendizado de máquina onde o processo de aprendizado ocorre pela modelagem de uma função que se ajusta de acordo com os valores de entrada e uma saída esperada.

No cenário mais simples, cada entrada de treinamento X é um vetor de números, representando por exemplo, a altura e o peso de uma pessoa. Estes dados de entrada são chamados de recursos ou atributos. No entanto, X pode ser um objeto estruturado e complexo, como uma imagem, uma frase, uma mensagem de e-mail, uma série temporal, uma forma molecular, um gráfico, etc. Da mesma forma, a variável Y de saída pode ser qualquer coisa, mas a maioria dos métodos assume que Y é uma variável categórica ou nominal de algum conjunto finito, o que indica um problema de classificação, ou um escalar com valor real no caso de um problema de regressão.

Já na abordagem não supervisionada ou descritiva, tem-se apenas os dados de entrada, sendo que o objetivo é encontrar padrões ocultos neste conjunto de dados. Esse tipo de problema é mais indefinido, uma vez que não se sabe o que procurar, assim como

também não se tem métricas de erro com base em um aprendizado (WATT; BORHANI; KATSAGGELOS, 2016).

Os processos de aprendizagem de máquina para construção de modelos computacionais geralmente utilizam três diferentes conjuntos de dados que são divididos em treinamento, validação e teste. O conjunto de treinamento é utilizado para o aprendizado dos modelos, servindo para ajustar os parâmetros durante o processo de aprendizagem (RIPLEY, 2007). Ou seja, é o conjunto de dados de exemplos que ensinam o modelo computacional. Este processo de aprendizagem compara o valor predito durante o treinamento com o valor real. Essa comparação é feita por meio de uma função custo, que tem por finalidade calcular o erro no processo de predição e fazer os ajustes necessários nos parâmetros do modelo. As funções custo podem variar de acordo com o modelo treinado.

O grande desafio no aprendizado de máquina é conseguir um bom desempenho com a utilização de novos dados - não apenas naqueles em que nosso modelo foi treinado. A capacidade de ter um bom desempenho em entradas não observadas anteriormente é chamada de generalização. Neste sentido, durante o processo de treinamento, vários ajustes podem ser realizados a fim de melhorar o desempenho do modelo. A maioria dos algoritmos de aprendizado de máquina possui configurações chamadas hiperparâmetros que devem ser determinados externamente ao próprio algoritmo de aprendizado (GOODFELLOW; BENGIO; COURVILLE, 2016). De forma complementar, o conjunto de dados de validação fornece uma avaliação imparcial dos ajustes do modelo sobre o conjunto de treinamento, auxiliando no ajuste dos hiperparâmetros.

Quando o modelo não consegue alcançar um treinamento satisfatório, diz-se que ocorreu *Underfiting*, ou seja, que o modelo não conseguiu aprender. Já quando o modelo aprende com os dados do conjunto de treinamento, mas não consegue um bom desempenho com novos dados, significa que ocorreu *Overfitting*. Ou seja, o modelo ficou tão especializado nos dados do conjunto de treinamento que acaba por ter um desempenho ruim em outros conjuntos de dados.

O conjunto de dados de validação também é utilizado para regularização em um processo de parada antecipada. Nesta situação, o treinamento será interrompido quando o erro sobre a base de validação aumentar, o que na prática indica um sobreajuste (ou *overfitting*) sobre a base de treinamento.

Outra técnica bastante utilizada, principalmente quando se tem um conjunto de dados reduzido é a chamada validação cruzada. Neste método, os dados são aleatoriamente divididos em vários grupos de treinamento e validação, de forma que ao final do treinamento uma média do desempenho dos treinamentos é computada.

Já o conjunto de teste é utilizado para avaliar o desempenho do modelo treinado, seguindo a mesma distribuição de probabilidade, provendo o cálculo de métricas como precisão, sensibilidade, medida F, entre outras.

4.2.1 Regularização

Durante o treinamento, uma técnica bastante utilizada para evitar o *overfitting* é a regularização. Neste processo uma penalidade é apresentada à medida que o modelo se torna mais especializado melhorando assim a sua generalização (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

Regularização L1

A regularização L1 (Equação 4.2) ou de Lasso adiciona o valor absoluto da magnitude do coeficiente como termo de penalidade à função de perda. Este método reduz o coeficiente do atributo menos importante para zero, removendo completamente esse atributo (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\boldsymbol{\beta_j}| \qquad \text{Equação 4.2}$$

Na equação, o primeiro termo representa a função custo e o segundo ($\lambda \sum_{j=1}^{p} |\boldsymbol{\beta_j}|$) a penalidade L1

Regularização L2

A regularização L2 ou Ridge (Equação 4.3) adiciona magnitude igual ao quadrado do coeficiente como termo de penalidade à função de perda. Ela força os pesos a serem pequenos, mas não os torna zero e faz uma solução não esparsa. L2 não é robusto para discrepantes, pois termos quadrados aumentam as diferenças de erro dos discrepantes e o termo de regularização tenta corrigi-lo penalizando os pesos (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \boldsymbol{\beta_j^2} \qquad \text{Equação 4.3}$$

Semelhante a equação anterior, aqui o primeiro termo representa a função custo e o segundo ($\lambda \sum_{j=1}^{p} \boldsymbol{\beta_j^2}$) a penalidade L2.

4.3. MÉTRICAS DE DESEMPENHO

Avaliar um modelo de aprendizado de máquina é uma das partes mais importantes em qualquer projeto. Seu modelo pode fornecer resultados satisfatórios quando avaliado usando determinada métrica, mas ruins quando outras. Na maioria das vezes, usamos a Acurácia da classificação para medir o desempenho de um modelo, no entanto, nem sempre é o método mais adequado.

Dependendo do tipo de modelo, se é classificação ou de regressão, diferentes métricas de avalição podem ser utilizadas. Para os modelos de classificação podemos ressaltar as seguintes métricas:

4.3.1 Acurácia

A acurácia (ACC) (Equação 4.4) é a métrica de avaliação mais utilizada em problemas de classificação; sendo muitas vezes utilizada de maneira incorreta. Ela é recomendada apenas quando há um número aproximado de observações em cada classe, ou seja, a base esteja balanceada, e que todas as previsões e erros de previsão sejam igualmente importantes, o que geralmente não é o caso (MISHRA, 2018), (IMBALANCED DATA : HOW TO HANDLE IMBALANCED CLASSIFICATION PROBLEMS, [s.d.]).

$$ACC = \frac{Número\ de\ predições\ corretas}{Número\ total\ de\ predições} \qquad \text{Equação 4.4}$$

Por exemplo, considere que existem 98% de amostras da classe A e 2% de amostras da classe B em nosso conjunto de treinamento. Então, nosso modelo pode obter 98% de precisão de treinamento facilmente, prevendo simplesmente todas as amostras de treinamento pertencentes à classe A. O problema maior surge quando o custo da classificação incorreta das amostras da classe de menor número é muito alto. Se lidarmos com uma doença rara, mas fatal, o custo de não diagnosticar a doença de uma pessoa doente é muito maior do que o custo de enviar uma pessoa saudável para mais testes complementares (MISHRA, 2018).

4.3.2 Logarithmic Loss

A perda logarítmica (ou perda de log) é uma métrica de desempenho para avaliar as previsões de probabilidades na associação a uma determinada classe. As previsões corretas ou incorretas são recompensadas ou punidas proporcionalmente à confiança da previsão. Funciona bem para classificação de várias classes. Ao trabalhar com perda logarítmica, o classificador deve atribuir probabilidade a cada classe para todas as amostras. Suponha que haja N amostras pertencentes a M classes, então a perda logarítmica é calculada da seguinte forma (Equação 4.5) (MISHRA, 2018).

$$Logarithmic\ Loss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * log(p_{ij}) \qquad \text{Equação 4.5}$$

Onde y_{ij}, indica se a amostra i pertence à classe j ou não e p_{ij}, indica a probabilidade da amostra i pertencer à classe j.

A perda logarítmica não tem limite superior e existe no intervalo [0, ∞). Perda de log mais próxima de zero indica maior precisão, enquanto que se a perda de log estiver longe de zero, indica menor precisão.

4.3.3 Matriz de Confusão

A matriz de confusão é uma apresentação útil da precisão de um modelo com duas ou mais classes, fornecendo uma matriz como saída e descrevendo o desempenho completo do modelo.

A tabela apresenta valores verdadeiros no eixo y e valores preditos no eixo x. As células da tabela são o número de previsões feitas por um algoritmo de aprendizado de máquina.

Vamos supor que temos um problema de classificação binária representada pelas classes Doente e Saudável (MISHRA, 2018).

Figura 4.2 - Matriz de confusão para classificação binária.

Da matriz de confusão pode-se tirar quatro métricas básicas capaz de indicar o desempenho do nosso modelo:

- Sensibilidade ou Taxa de Verdadeira Positivo (SN)

$$\text{SN} = \frac{VP}{VP+FN} \qquad \text{Equação 4.6}$$

- Especificidade ou Taxa de Verdadeiro Negativo (SP)

$$\text{SP} = \frac{VN}{VN+FP} \qquad \text{Equação 4.7}$$

- Precisão ou Valor Preditivo Positivo (PR)

$$\text{PR} = \frac{VP}{VP+VN} \qquad \text{Equação 4.8}$$

- Valor Preditivo Negativo (PRN)

$$\text{PRN} = \frac{VN}{VN+FN} \qquad \text{Equação 4.9}$$

Onde, VP são os verdadeiros positivos, VN são verdadeiros negativos, FP são os falsos negativos e FN são os falsos negativos.

A sensibilidade é capacidade do modelo detectar a doença. Já a precisão é probabilidade de acertos que o modelo tem sobre os classificados como verdadeiros. Desta forma, se desejamos que nosso modelo detecte determinada doença, o ideal é que nosso o modelo tenha uma sensitividade e precisão alta.

4.3.4 Área sob a Curva ROC

A área sob a curva ROC (ou AUC ROC para abreviar) é uma métrica de desempenho para problemas de classificação binária.

A AUC representa a capacidade de um modelo de discriminar entre classes positivas e negativas. Uma área de 1,0 representa um modelo que fez todas as previsões perfeitamente.

A AUC pode ser dividido em sensibilidade e especificidade. Um problema de classificação binária é realmente uma troca entre sensibilidade e especificidade (SAITO; REHMSMEIER, 2015).

Sensibilidade é a verdadeira taxa positiva também chamada de recall. São as instâncias numéricas da classe positiva (primeira) que realmente previram corretamente.

A especificidade também é chamada de taxa negativa verdadeira. É o número de instâncias da classe negativa (segunda) que foram realmente previstas corretamente (MISHRA, 2018).

4.3.5 Área sob a Curva PR

Devido a característica da base de possuir classes desbalanceadas, é mais recomendado a utilização da área sob a curva do gráfico Precision-Recall. O gráfico Precision-Recall (PR) mostra valores de precisão para os valores de sensibilidade (recall) correspondentes. Semelhante ao gráfico ROC, o gráfico PR fornece uma avaliação de todo o modelo, sendo robusto e eficaz para avaliação de bases desbalanceadas (SAITO; REHMSMEIER, 2015).

4.3.6 Escore-F1

Usado para medir a precisão de um teste, o Escore-F1 é a média harmônica entre precisão e sensibilidade. O intervalo para a pontuação F1 é [0, 1]. Ele mostra o quão preciso é o seu classificador (quantas instâncias ele classifica corretamente) e também o quão robusto é (não perde um número significativo de instâncias).

Quanto maior a pontuação do Escore-F1, melhor é o desempenho do nosso modelo. Matematicamente, pode ser expresso com (Equação 4.10) (MISHRA, 2018):

$$F1 = 2 * \frac{1}{(1/Sensibilidade)+(1/Precisão)} \qquad \text{Equação 4.10}$$

Já para os modelos de regressão tem-se de forma mais abrangente as seguintes métricas:

4.3.7 Erro Médio Absoluto

O erro médio absoluto (MAE) é a média da diferença entre os valores verdadeiros e os valores preditos. Ele nos fornece a medida de quão longe as previsões estavam do resultado real. No entanto, eles não nos dão nenhuma ideia da direção do erro, ou seja, se estamos subestimando os dados ou superestimando os dados. Matematicamente, é representado como (Equação 4.11) (MISHRA, 2018):

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - y'_j\right| \qquad \text{Equação 4.11}$$

Nas equações de erro, y_j, representa os valores originais de saída e y'_j, os valores preditos.

4.3.8 Erro Médio Quadrático

O erro médio quadrático (MSE) é bastante semelhante ao erro médio absoluto, tendo como única diferença o calculo da média do quadrado da diferença entre os valores originais e os valores preditos. A vantagem do MSE é que é mais fácil calcular o gradiente, enquanto o Erro Médio Absoluto requer ferramentas de programação linear complicadas para calcular o gradiente. Como tomamos o quadrado do erro, o efeito de erros maiores se torna mais pronunciado que o erro menor; portanto, o modelo agora pode se concentrar mais nos erros maiores (Equação 4.12) (MISHRA, 2018):

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - y'_j\right|^2 \qquad \text{Equação 4.12}$$

4.3.9 Raiz do Erro Médio Quadrático

Um problema com o MSE é o fato de que a unidade da métrica também é quadrada. Para resolver isso, a Raiz do Erro Médio Quadrático **(**RMSE) é usado para remover a raiz quadrada, mas mantendo a propriedade de penalizar erros maiores (MISHRA, 2018).

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left|y_j - y'_j\right|^2} \qquad \text{Equação 4.13}$$

4.4. MODELOS DE APRENDIZAGEM

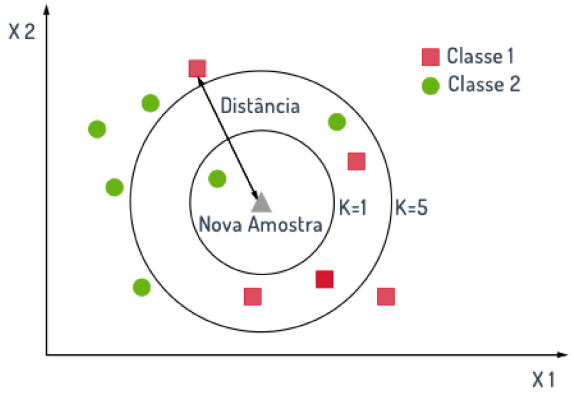
Diferentes tipos de algoritmos e métodos de aprendizado de máquina podem ser utilizados na construção dos modelos. Dependendo do tipo de problema e do padrão dos dados, um ou outro método será melhor empregado na solução do problema.

4.4.1 K Nearest Neighbors

O algoritmo K vizinhos mais próximos (K Nearest Neighbors - KNN) está entre os mais simples de todos os algoritmos de aprendizado de máquina. Neste método, as amostras são classificadas de acordo com a classificação de seus vizinhos (PARSIAN, [s.d.]), (MUCHERINO; PAPAJORGJI; PARDALOS, 2009).

Desta forma, diante de um conjunto de amostras com classificação conhecida, o chamado conjunto de treinamento, cada amostra desconhecida deve ser classificada com base nas k amostras circundantes. Assim, dada uma amostra desconhecida e um conjunto de treinamento, todas as distâncias entre a amostra desconhecida e as amostras circundantes são calculadas. Por fim, a menor distância corresponde à amostra no conjunto de treinamento mais próxima da amostra desconhecida (MUCHERINO; PAPAJORGJI; PARDALOS, 2009), (GOLDSTEIN, 1972).

Figura 4.3 - Exemplo de classificação por KNN.



Para definir quais amostra de K no conjunto de treinamento são mais semelhantes a nova amostra, uma medida de distância é utilizada. Dentre as várias métricas utilizadas para calcular a distância, a mais comum é a distância euclidiana (BROWNLEE, [s.d.]). Além desta, outras funções podem ser utilizadas.

Seja $X = (x_1, x_2, \dots, x_n)$ e $Y = (y_1, y_2, \dots, y_n)$ dois pontos $\in R$.

A distância Euclidiana entre X e Y é dada por:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad \text{Equação 4.14}$$

A distância Manhattan entre X e Y é dada por:

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| \qquad \text{Equação 4.15}$$

A distância Minkowski entre X e Y é dada por:

$$d(x,y) = (|x_1 - y_1|^q + |x_2 - y_2|^q + \cdots + |x_n - y_n|^q)^{1/q} \qquad \text{Equação 4.16}$$

onde $q \in N$.

A distância Minkowski é a generalização das duas distâncias anteriores. Quando q = 1, esta distância representa a distância de Manhattan e quando q = 2, a distância Euclidiana.

Considerando as amostras da Figura 4.3, precisamos prever a classe de saída para a nova amostra.

No primeiro exemplo, usamos K=1, o que significa que a nova amostra pertencerá à classe cujo ponto de dados estará mais próximo. E como aqui estamos falando de apenas um vizinho mais próximo, podemos ver que o vizinho mais próximo da nova amostra é aquele círculo verde. Portanto, é bastante óbvio que a nova amostra será um círculo verde.

Mas no segundo caso em que estamos tomando k=5, precisamos levar as 5 maiores distâncias mais próximas da nova amostra para o vizinho. E depois de calcular a distância euclidiana, é visível que a nova amostra pertencerá a classe de quadrados vermelhos (BROWNLEE, [s.d.]).

A biblioteca *scikit-learn* implementa dois classificadores de vizinhos mais próximos (NEAREST NEIGHBORS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]):

- *KNeighborsClassifier* que implementa o aprendizado baseado nos k vizinhos mais próximos de cada ponto de consulta, sendo que k é um valor inteiro especificado pelo usuário.
- *RadiusNeighborsClassifier* que implementa a aprendizagem com base no número de vizinhos em um raio fixo especificado pelo usuário.

A técnica *KNeighborsClassifier* é a mais usada. Nela a escolha ideal do valor k é altamente dependente dos dados. Em geral, um k maior suprime os efeitos do ruído, mas torna os limites de classificação menos distintos.

Nos casos em que os dados não são amostrados uniformemente, a classificação de vizinhos com base no raio pode ser uma escolha melhor. O usuário especifica um raio fixo de modo que pontos em regiões mais dispersas usam menos vizinhos para a classificação. Para espaços de parâmetros de alta dimensão, esse método se torna menos eficaz devido à chamada "maldição da dimensionalidade" (NEAREST NEIGHBORS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

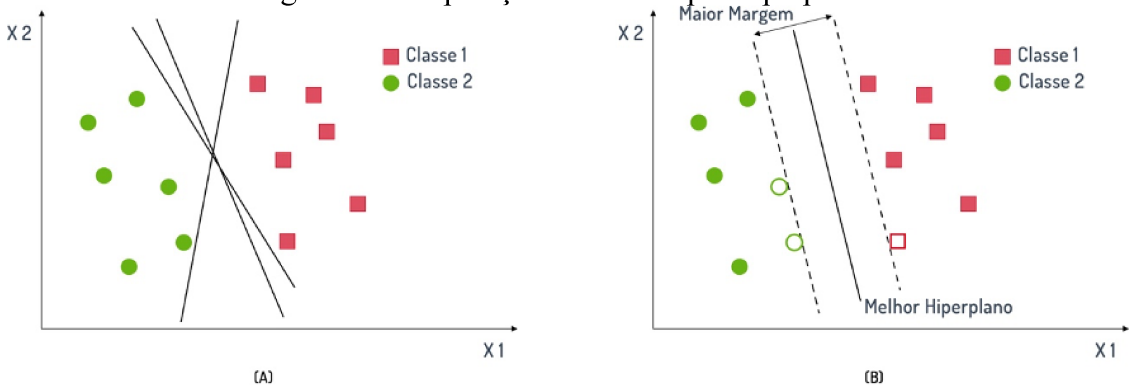
A classificação básica de vizinhos mais próximos usa pesos uniformes, ou seja, o valor atribuído a um ponto de consulta é calculado a partir de uma votação majoritária simples dos vizinhos mais próximos. Sob algumas circunstâncias, é melhor ponderar os vizinhos de modo que os vizinhos mais próximos contribuam mais para o ajuste (NEAREST NEIGHBORS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

O método de vizinhos mais próximos também pode ser utilizado para problemas de regressão, nos casos em que os rótulos de dados são contínuos, em vez de variáveis discretas. A classificação atribuída a um ponto de consulta é calculada com base na média dos rótulos dos vizinhos mais próximos (NEAREST NEIGHBORS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

4.4.2 Support Vector Machine

Máquinas de vetores de suporte (*Support Vector Machine* - SVM) são modelos de aprendizado supervisionado, sendo utilizados em problemas de classificação, regressão e detecção de outliers. SVM utiliza um conjunto de dados de entrada com o objetivo de predizer qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador binário (BOSER; GUYON; VAPNIK, 1992), (CORTES; VAPNIK, 1995), (SAVAN PATEL, 2017).

Figura 4.4 - Separação dos dados por hiperplano.



O objetivo é encontrar um hiperplano em um espaço N-dimencional que faça a separação dos dados de duas classes distintas. Diante das inúmeras possibilidades (Figura 4.4 - A), o objetivo é encontra o plano que tenha a maior margem, ou seja, com a máxima distância entre os pontos mais próximos em relação a cada uma das classes (Figura 4.4 - B) (CORTES; VAPNIK, 1995).

As SVMs são embasadas pela teoria de aprendizado estatístico (TAE), caracterizando-se por boa capacidade de generalização. O modelo é semelhante a uma regressão logística, mas diferente desta, a máquina de vetores de suporte não fornece probabilidades, mas apenas gera uma identidade de classe (LORENA; DE CARVALHO, 2007).

Na regressão logística, pegamos a saída da função linear e comprimimos o valor dentro da faixa de [0,1] usando a função sigmóide. Se o valor compactado for maior que um valor limite (0,5), atribuímos a ele um rótulo 1, caso contrário, atribuímos a ele um rótulo 0. No SVM, obtemos a saída da função linear e, se essa saída for maior que 1, identificamos com uma classe e se a saída for -1, identificamos com outra classe. Como os valores limite são alterados para 1 e -1 no SVM, obtemos esse intervalo de valores de reforço ([-1,1]) que atua como margem (SAVAN PATEL, 2017), (LORENA; DE CARVALHO, 2007).

SVM com Margens Rígidas

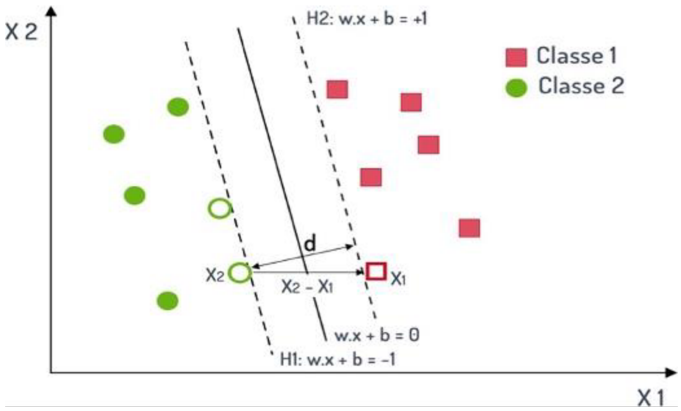
As SVMs lineares com margens rígidas definem fronteiras lineares a partir de dados linearmente separáveis. Seja T um conjunto de treinamento com n dados $x_i \in X$ e seus respectivos rótulos $y_i \in Y$, em que X constitui o espaço dos dados e $Y = \{-1, +1\}$. T é linearmente separável se é possível separar os dados das classes $+1$ e -1 por um hiperplano (SCHÖLKOPF et al., 2002).

A equação de um hiperplano é apresentada pela seguinte equação:

$$f(x) = w.x + b = 0 \qquad \text{Equação 4.17}$$

Onde $w.x$ é o produto escalar entre os vetores w e $x, w \in X$ é o vetor normal ao hiperplano descrito e $\frac{b}{||w||}$ corresponde à distância do hiperplano em relação à origem, com $b \in R$.

Figura 4.5 - Distância entre os hiperplanos de separação.



De acordo com a Figura 4.5, seja x_1 um ponto no hiperplano H1 e x_2 um ponto no hiperplano H2. Projetando $x_1 - x_2$ perpendicular ao hiperplano separador $w \cdot x+b = 0$, é possível obter a distância entre os hiperplanos H1 e H2 (CORTES; VAPNIK, 1995).

Desta forma, a maximização da margem de separação dos dados em relação a $w \cdot x+b = 0$ pode ser obtida pela minimização de $\|w\|$ (BOSER; GUYON; VAPNIK, 1992), o que nos leva ao seguinte problema de otimização (SCHÖLKOPF et al., 2002):

$$\text{Minimizar} \quad \underset{w,b}{\frac{1}{2}} \|w\|^2 \qquad \text{Equação 4.18}$$

Com as restrições: $y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) - 1 \geq 0, \qquad \forall i = 1, \ldots, n$

SVM com Margem Suave

Na prática é raro encontrar aplicações cujos dados sejam linearmente separáveis. Isso pode ocorrer por vários motivos, entre eles a presença de ruídos e outliers nos dados ou então, pela própria natureza de um problema não linear. Nestes casos, uma adaptação dos modelos SVMs lineares de margens rígidas pode ser feita. Basicamente, permite-se que alguns dados possam violar a margem. Isso é feito com a introdução de variáveis de

folga ξi, para todo i = 1, ..., n. Essas variáveis relaxam as restrições impostas ao problema de otimização (SCHÖLKOPF et al., 2002):

$$y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad \forall i = 1, \ldots, n$$

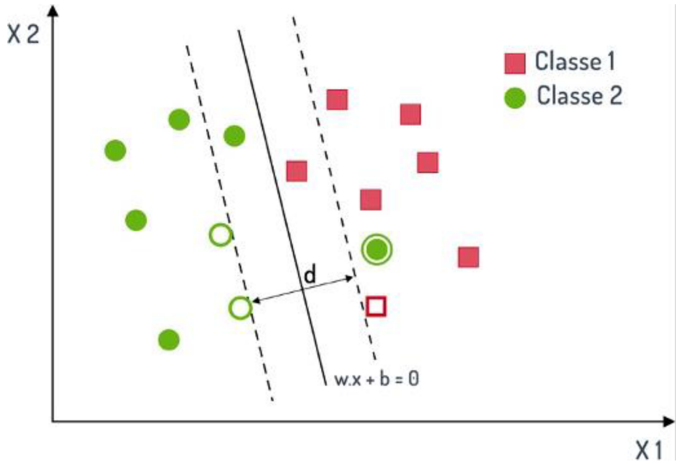
O erro no conjunto de treinamento é indicado por um valor de ξi maior que 1. Logo, a soma dos ξi representa um limite no número de erros de treinamento. Para levar em consideração esse termo, minimizando assim o erro sobre os dados de treinamento, reformulamos a função de otimização (BOSER; GUYON; VAPNIK, 1992), (SAVAN PATEL, 2017):

$$\text{Minimizar}_{w,b,\xi} \frac{1}{2} \|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \qquad \text{Equação 4.19}$$

Onde C é um termo de regularização que atribui um peso para a minimização dos erros sobre o conjunto de treinamento.

Assim como no SVM com margem rígida, os dados que participam da formação do hiperplano separador, são denominados vetores de suporte (SV).

Figura 4.6 - Vetores de suporte na definição dos hiperplanos.

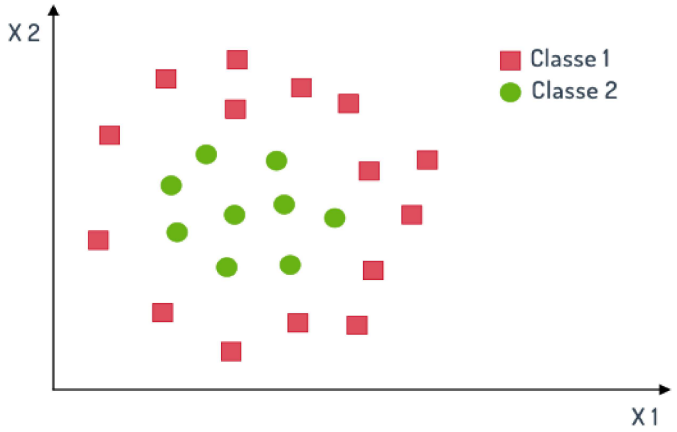


Na Figura 4.6 são ilustrados os possíveis tipos de SVs. Elementos sem preenchimento representam SVs livres. Os elementos dentro da margem são SVs limitados. Elementos com bordas extras correspondem a SVs limitados que são erros de treinamento.

SVM Não Linear

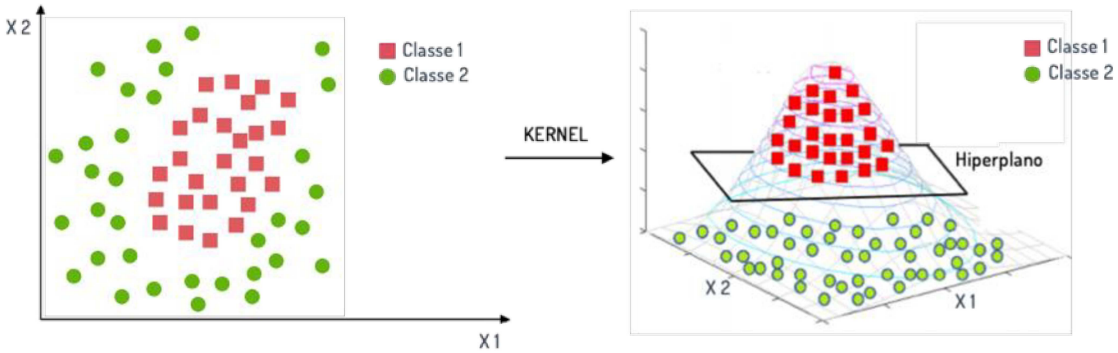
As SVMs lineares são bastante eficientes na classificação de conjuntos de dados que tenham uma distribuição aproximadamente linear, sendo que a SVM de margens suaves aceita a presença de ruídos e outliers. No entanto, existem casos em que não conseguimos dividir os dados de treinamento por um hiperplano (LORENA; DE CARVALHO, 2007). Na Figura 4.7, por exemplo, precisaríamos de uma fronteira curva para uma separação adequada das classes.

Figura 4.7 - Conjunto de dados com uma distribuição não linear



Embasadas pelo teorema de Cover (DUAN; KEERTHI, 2005), as SVMs atuam sobre distribuições não lineares mapeando o conjunto de dados do seu espaço original, conhecido como espaço de entrada, para um espaço de maior dimensão, chamado de espaço de características (*Feature Space*) (LORENA; DE CARVALHO, 2007). Por exemplo, na Figura 4.8 os dados do espaço de entrada em R^2 são mapeados para o espaço de características em R^3, o que torna o conjunto de dados de treinamento linearmente separáveis.

Figura 4.8 - Aumento da dimensão de R2 para R3.



Assim, diante de problemas não lineares, mapeia-se inicialmente os dados para um espaço de maior dimensão e aplica-se a SVM linear sobre o novo espaço. Este mapeamento é realizado por meio de uma função denomidada Kernel (LORENA; DE CARVALHO, 2007). Na prática, os Kernels mais utilizados são os Polinomiais, os Gaussianos ou RBF (Radial-Basis Function) e os Sigmoidais, sendo que dentre estes, o RBF é um dos que mais se destaca.

SVM Multiclasse

A principal abordagem para tratar o problema multiclasse é transformar um problema de várias classes em vários problemas de classificação binária (DUAN; KEERTHI, 2005). Isto pode ser feito com a construção de classificadores binários que distinguem entre um dos rótulos e o restante (um contra todos) ou entre cada par de classes (um contra um). A classificação de novas elementos para o caso um contra todos é feita por uma estratégia onde o vencedor leva tudo. Para a abordagem um contra um, a classificação é feita por uma estratégia de votação, onde a classe que possuir o maior número de vitórias determina a classificação do elemento.

A biblioteca *scikit-learn* implementa modelos SVM, possibilitando o uso de diferentes kernels. *SNC* e *NuSVC* são métodos de classificação não lineares para multiclasses, adotando uma abordagem um contra um. Já o método *LinearSVC*, também para classificação multiclasse, adota uma abordagem um contra todos, sendo utilizado com kernel linear (SUPPORT VECTOR MACHINES — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

SVM também pode ser estendido para resolver problemas de regressão. Esse método é chamado de Regressão de vetores de suporte. O modelo produzido pelo SVC depende apenas de um subconjunto dos dados de treinamento, porque a função de custo para a construção do modelo não se importa com os pontos de treinamento que estão além da margem. Analogamente, o modelo produzido pelo Support Vector Regression SVR depende apenas de um subconjunto dos dados de treinamento, porque a função de custo para a construção do modelo ignora todos os dados de treinamento próximos à previsão do modelo. Assim como no SVC, o SVR implementa três diferentes métodos: *SVR*, *NuSVR* e *LinearSVR* (SUPPORT VECTOR MACHINES — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

As funções kernel linear, polinomial, RBF e sigmoide estão disponíveis para o usuário, no entanto uma função de kernel personalizada também pode ser implementada.

4.4.3 Naïve Bayes

Os métodos Naïve Bayes são um conjunto de algoritmos de aprendizado supervisionado baseados na aplicação do teorema de Bayes com a suposição de independência condicional entre cada par de atributos dado o valor da variável de classe.

Redes bayesianas, são modelos gráficos para raciocínio baseados em incerteza, onde os nós representam as variáveis e os arcos representam suas dependências condicionais através de um grafo acíclico dirigido (*directed acyclic graph* - DAG). Elas foram desenvolvidas por Judea Pearl durante a década de 1980 e se baseiam no do Teorema de Bayes, publicado pelo matemático Thomas Bayes em 1763 (KORB; NICHOLSON, 2010; PEARL, 1985).

Seja o espaço de probabilidades (ε, P) e os eventos A e B $\subseteq$ ε, sendo estes não tenham probabilidade nula, então:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad \text{Equação 4.20}$$

Onde:

$P(A|B) =$ Probabilidade de ocorrer o evento A, dado que ocorreu o evento B.

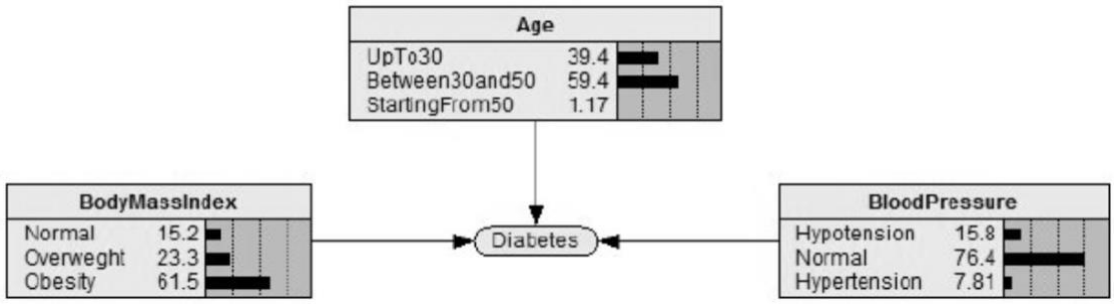
$P(B|A) =$ Probabilidade de ocorrer o evento B, dado que ocorreu o evento A.

$P(A) =$ Probabilidade de ocorrer o evento A.

$P(B) =$ Probabilidade de ocorrer o evento B.

Desta forma, usando o teorema de Bayes, pode-se encontrar a probabilidade de A acontecer, dado que B ocorreu. Aqui, B é a evidência e A é a hipótese.

Figura 4.9 – Exemplo de rede bayesiana para identificação de Diabetes.



Fonte: Editado de (TOWARDS THE APPLIED HYBRID MODEL IN DECISION MAKING: SUPPORT THE EARLY DIAGNOSIS OF TYPE 2 DIABETES, [s.d.]))

Tabela 4.1 - Probabilidade (em %) condicional de um paciente ter Diabetes com base na probabilidade de ocorrência dos critérios (editado de (TOWARDS THE APPLIED HYBRID MODEL IN DECISION MAKING: SUPPORT THE EARLY DIAGNOSIS OF TYPE 2 DIABETES, [s.d.])).

Massa	Idade	Pressão Sanguínea	Prob.(%)
Normal	Até 30	Hipotensão	0.03
Normal	Até 30	Normal	0.18
Normal	Até 30	Hipertensão	0.03
Normal	Entre 30 e	Hipotensão	0.11
Normal	Entre 30 e	Normal	0.7
Normal	Entre 30 e	Hipertensão	0.1
Normal	Acima de	Hipotensão	0
Normal	Acima de	Normal	0.02
Normal	Acima de	Hipertensão	0
Sobrepeso	Até 30	Hipotensão	0.12
Sobrepeso	Até 30	Normal	0.8
Sobrepeso	Até 30	Hipertensão	0.11
Sobrepeso	Entre 30 e	Hipotensão	0.48
Sobrepeso	Entre 30 e	Normal	3.12
Sobrepeso	Entre 30 e	Hipertensão	0.44
Sobrepeso	Acima de	Hipotensão	0.02
Sobrepeso	Acima de	Normal	0.11
Sobrepeso	Acima de	Hipertensão	0.01
Obeso	Até 30	Hipotensão	0.67
Obeso	Até 30	Normal	4.36
Obeso	Até 30	Hipertensão	0.61
Obeso	Entre 30 e	Hipotensão	2.64
Obeso	Entre 30 e	Normal	17.09
Obeso	Entre 30 e	Hipertensão	2.39
Obeso	Acima de	Hipotensão	0.09
Obeso	Acima de	Normal	0.58
Obeso	Acima de	Hipertensão	0.08

De acordo com a Figura 4.9, a probabilidade de cada critério é calculada com base na ocorrência dos mesmos. Desta forma, a probabilidade condicional de um indivíduo ter ou não Diabetes é calculada com base na probabilidade de ocorrência dos critérios relacionados. Na Tabela 4.1 é apresentado os valores da probabilidade condicional de se ter Diabetes para cada combinação de critérios (TOWARDS THE APPLIED HYBRID MODEL IN DECISION MAKING: SUPPORT THE EARLY DIAGNOSIS OF TYPE 2 DIABETES, [s.d.]).

Em aprendizado de máquina, tem-se uma família de classificadores probabilísticos e supervisionado baseados no teorema de Bayes conhecidos como Naïve Bayes. Estes métodos são simples e rápidos, possuindo um bom desempenho como classificador, com a vantagem de precisar apenas de um número reduzido de dados de entrada em comparação a outros modelos. No entanto, ele desconsidera completamente a correlação entre as variáveis de entrada, de modo que um requisito é que os eventos sejam condicionalmente independentes.

A biblioteca *scikit-learn* implementa vários modelos de classificação naïves bayes com *GaussianNB*, *MultinomialNB*, *BernoulliNB*, entre outros. Quando os preditores assumem um valor contínuo e não são discretos, assumimos que esses valores são amostrados a partir de uma distribuição gaussiana. O método *GaussianNB* implementa o algoritmo gaussiano Naïve Bayes para classificação de acordo com a seguinte equação (NAIVE BAYES — SCIKIT-LEARN 0.21.3 DOCUMENTATION, [s.d.]):

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \qquad \text{Equação 4.21}$$

Onde $P(x_i|y)$. é a probabilidade dos valores de x_i em relação a classe y. Já μ_y representa a média dos valores de pertencente a classe e σ_y^2 representa a variância dos valores de x_i em relação a classe y.

O método *MultinomialNB* implementa o algoritmo Naïve Bayes para dados distribuídos multinominalmente. Ele é usado principalmente para problemas de classificação de documentos, ou seja, se um documento pertence à categoria de esportes, política, tecnologia etc. Neste caso os preditores usados pelo classificador são a frequência das palavras presentes no documento. Já o método *BernoulliNB* implementa o

algoritmo de Naïve Bayes com dados em uma distribuição multivariada de Bernoulli. Este método é semelhante ao Multinominal, com a diferença que os preditores são variáveis booleanas (NAIVE BAYES — SCIKIT-LEARN 0.21.3 DOCUMENTATION, [s.d.]).

4.4.4 Random Forest

Proposto por Ho (HO, 1995) em 1995, florestas aleatórias (*Random Forest*) ou florestas de decisão aleatória são modelos de aprendizado de máquina supervisionado, utilizado para classificação e regressão de dados, entre outras tarefas. Basicamente o algoritmo cria várias árvores de decisão e as combina a fim de obter o melhor resultado.

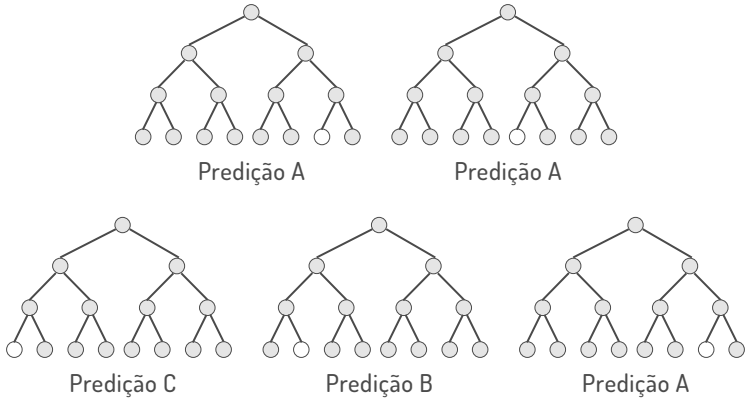
Uma árvore de decisão é um modelo computacional que utiliza a estratégia de dividir para conquistar: onde um problema complexo é decomposto em problemas menores.

Nessa estrutura (Figura 4.10), os nós internos correspondem as variáveis de entrada e as folhas representam a variável de destino. O objetivo é criar um modelo que preveja o valor de uma variável de destino com base em várias variáveis de entrada, sendo que vários caminhos podem ser tomados (BREIMAN, 2001), (YIU, 2019).

Figura 4.10 - Estrutura de árvore de decisão.

Um modelo de floresta aleatória é formado por um grande número de árvores de decisão que funcionam como um conjunto (Figura 4.11). Cada árvore na floresta aleatória encontra um resultado, sendo que o resultado com maior número de ocorrências corresponde ao valor final (BREIMAN, 2001), (YIU, 2019).

Figura 4.11 - Representação de uma floresta aleatória com várias árvores de decisão.



O modelo usa empacotamento e aleatoriedade dos dados de entrada para criar cada árvore de forma independente, tendo como objetivo construir uma floresta não correlacionada de árvores onde o resultado final será melhor do que qualquer árvore individual. O segredo está na baixa correlação entre as árvores de decisão. A razão para esse grande efeito é que o modelo busca se proteger de seus erros individuais. Desta forma, enquanto algumas árvores podem estar erradas, outras estarão corretas, levando o grupo para o melhor resultado (BREIMAN, 2001), (YIU, 2019).

Floresta Aleatória é um excelente algoritmo para ser utilizado nos primeiros estágios de criação do processo de desenvolvimento de um modelo, para se ter uma ideia de performance. Devido à sua simplicidade, é difícil construir um Floresta Aleatória "ruim". Este algoritmo é também uma boa opção se você precisa desenvolver um modelo em curto espaço de tempo. Além disso, ele provê um bom indicador de importância para as características (CHAUHAN, [s.d.]).

A biblioteca *scikit-learn* possui dois algoritmos baseados em árvores de decisão: o *RandomForest* e o *Extra-Trees*. O *RandomForest* possui os métodos *RandomForestClassifier* e *RandomForestRegressor*, podendo ser utilizado tanto para classificação quanto para regressão, respectivamente. Em ambos os métodos, cada árvore do conjunto é criada a partir da base de dados de treinamento. De forma complementar, ao dividir cada nó durante a construção de uma árvore, a melhor divisão é encontrada tanto individualmente para cada dado de entrada como para subconjuntos formados por eles. O objetivo dessas duas fontes de aleatoriedade é diminuir a variação do estimador florestal. Fato é, que geralmente as árvores de decisão individuais apresentam alta variação e tendem a se superestimar. A aleatoriedade atribuída às florestas produz árvores de decisão com erros de previsão um pouco dissociados. Assim, ao fazer uma média

dessas previsões, alguns erros podem ser cancelados, fazendo com que florestas aleatórias alcancem uma variação reduzida ao combinar diversas árvores, o que na prática, resulta modelos mais eficientes (ENSEMBLE METHODS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

Os principais parâmetros a serem ajustados na utilização destes métodos são *n_estimators* e *max_features*. O primeiro é o número de árvores na floresta, de forma que quanto maior, melhor o resultado, mas também maior o custo computacional assim com o tempo necessário para calcular. Além disso, observe que os resultados deixarão de melhorar significativamente quando alcançado um número crítico de árvores. Já o parâmetro *max_features* é o tamanho dos subconjuntos aleatórios de recursos a serem considerados ao dividir um nó. Quanto menor, maior a redução da variância. Por padrão (*max_features=None*), são considerados todos os recursos, em vez de um subconjunto aleatório (ENSEMBLE METHODS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

O método também possibilita a execução paralela tanto para criação das árvores quanto para as previsões do modelo. Isto é feito pelo parâmetro *n_jobs*, que recebe número de núcleos de máquina que atuarão de forma paralela. Caso atribua n*_jobs* = -1, todos os núcleos disponíveis na máquina serão utilizados. Isto é bastante útil na construção de um grande número de árvores assim com na utilização de um grande conjunto de dados de entrada (ENSEMBLE METHODS — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

4.4.5 Artificial Neural Network

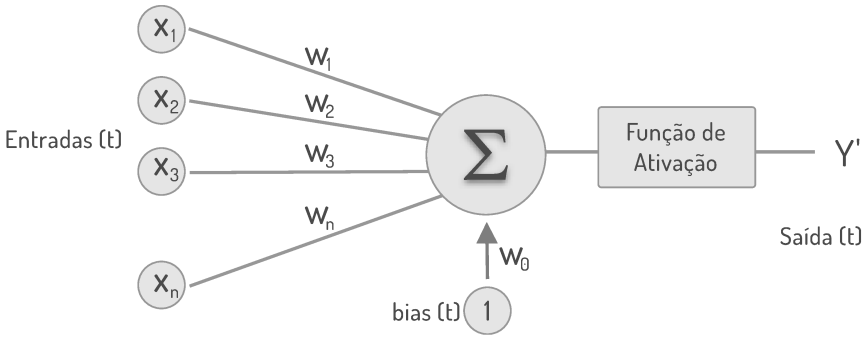
Redes neurais artificiais (Artificial Neural Network - ANN) são modelos computacionais inspirados na estrutura cerebral de animais. Estes modelos são formados por grupos de neurônios, que por sua vez, são modelos matemáticos baseados no funcionamento de uma célula nervosa. Proposto pelo neurofisiologista americano Warren S. McCulloch e pelo lógico americano H. Pitts Jr em 1943, o Threshold Logic Unit foi o primeiro modelo matemático de um neurônio artificial (MCCULLOCH; PITTS, 1990). Basicamente este modelo possui várias entradas que são somadas e comparadas a um valor de limiar pré-definido, fazendo com que o neurônio seja ativado ou não, dependendo do valor somado.

$$c = \sum_{j=1}^{m} x_j . w_j$$ Equação 4.22

Na equação y representa a saída para um conjunto de entradas x_j, com pesos w_j.

Já em 1957, o psicólogo americano Frank Rosenblatt, inspirado pela teoria hebbiana que defende a adaptação plástica do neurônio durante um processo de aprendizagem, propôs um novo modelo de neurônio chamado de Perceptron (ROSENBLATT, 1958), sendo este aperfeiçoado por Marvin Minsky e Seymour Papert em 1969 (MINSKY; PAPERT; BOTTOU, 2017).

Figura 4.12 - Modelo de neurônio Perceptron.



Com algumas melhorias em relação ao neurônio de Pitts e fazendo uso de um método de aprendizado supervisionado e diferentes funções de ativação, o Perceptron é capaz aprender e classificar padrões linearmente separáveis. Conforme é ilustrado na Figura 4.12, além das entradas de dados, o Perceptron possui uma entrada extra chamada de *bias*. Cada entrada também recebe um peso sináptico w, permitindo que algumas entradas tenham maior influência do que outras (HAYKIN, 2007).

No aprendizado supervisionado a rede faz uso de respostas corretas a fim de calcular o erro diante das respostas encontradas pelo modelo. Por meio de um método de retro propagação, os pesos dos neurônios são ajustados com o objetivo de encontrar a saída esperada de acordo com cada entrada de dados (HAYKIN, 2007).

A escolha da função de ativação pode variar dependendo do tipo de dados de entrada e de acordo com a distribuição da saída. Para problemas de regressão, geralmente é utilizada a função *identidade*, já para problemas de classificação binária o mais

apropriado seria a função *sigmoide logística*, de modo que para problemas de várias classes, a função de ativação *softmax* seria mais apropriada (BISHOP, 2016).

Abaixo são apresentadas algumas funções de ativação utilizadas na confecção de redes neurais artificiais (FACURE, 2017).

- **Funções Linear**: É a função de ativação mais básica porque não altera a saída de um neurônio. Geralmente é utilizada nas camadas de saída em redes neurais de regressão.

$$f(x) = x \qquad \text{Equação 4.23}$$

- **Função Sigmoid**: A função de ativação sigmoid é comumente utilizada por redes neurais com propagação positiva (Feedforward) que precisam ter como saída apenas números positivos, em redes neurais multicamadas e em outras redes com sinais contínuos.

$$f(x) = \frac{1}{1+e^{-x}} \qquad \text{Equação 4.24}$$

- **Função Tangente Hiperbólica**: A função de ativação tangente hiperbólica possui uso muito comum em redes neurais cujas saídas devem ser entre -1 e 1.

$$f(x) = \tanh(x) \qquad \text{Equação 4.25}$$

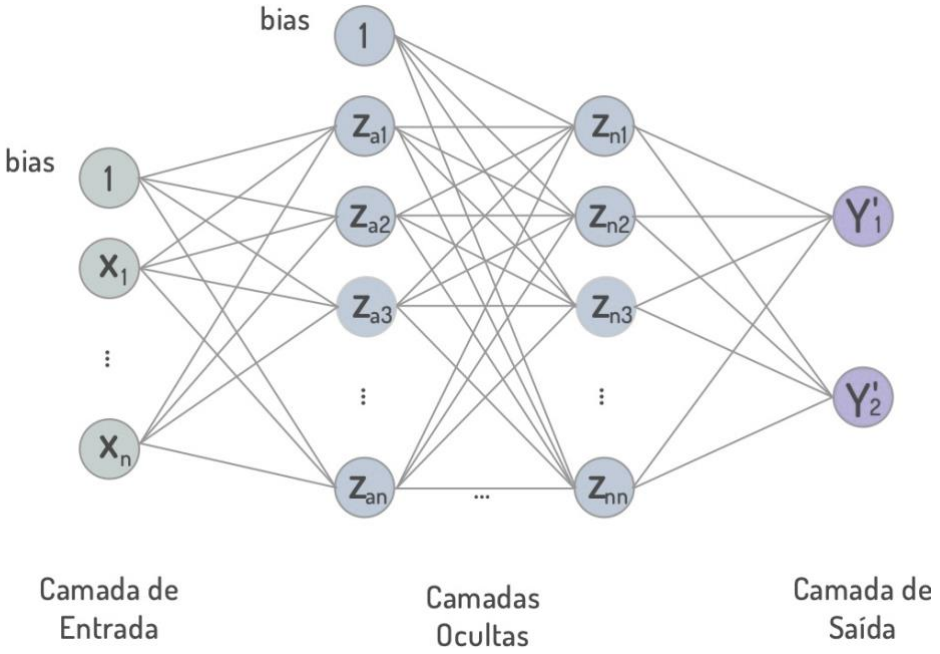
- **Função Softmax**: A função de ativação softmax é usada em redes neurais de classificação. Ela força a saída de uma rede neural a representar a probabilidade de os dados serem de uma das classes definidas. Sem ela as saídas dos neurônios são simplesmente valores numéricos onde o maior indica a classe vencedora.

$$f(x) = \frac{e^{z_i}}{\sum_{j=1}^{J} e^{z_j}} \qquad \text{Equação 4.26}$$

Adicionando mais camadas de neurônios, é possível resolver problemas linearmente não-separáveis. Este tipo de rede é conhecido como Perceptron

Multicamadas (*Multi-layer Perceptron* – MLP). Com uma topologia funcionando com fluxo progressivo (*feedforward*), os dados da entrada são projetados para as camadas posteriores até a saída (HAYKIN, 2007).

Figura 4.13 - Rede neural com múltiplas camadas ocultas e duas saídas.



Essa estrutura de rede possui uma camada de entrada e uma camada de saída, podendo possuir várias camadas ocultas com um número diferenciado de neurônios em cada uma delas. Desta forma, a complexidade de uma rede MLP se dá justamente pela quantidade de camadas ocultas e pelo número de neurônios que estas camadas possuam (HAYKIN, 2007).

O aprendizado é basicamente realizado em duas etapas. Na primeira os resultados encontrados são comparados com os esperados e um erro é encontrado. Para isso pode-se utilizar diferentes funções de custo. Esse erro é utilizado para calcular ajustes nos pesos, que na sequencia são corrigidos da saída em direção a entrada, o que nos arremete ao termo retropropagação (BISHOP, 2016).

A biblioteca *scikit-learn* possui um algoritmo MLP com métodos tanto para classificação quanto para regressão. A classe *MLPClassifier* implementa um algoritmo de perceptron de múltiplas camadas com treinamento por retropropagação. Já a classe *MLPRegressor* implementa um MLP, também com treinamento por retropropagação, mas sem função de ativação na camada de saída, o que também pode ser comparado ao uso da função identidade como função de ativação. Portanto, a classe usa o erro quadrado

como função de perda e a saída é um conjunto de valores contínuos. *MLPRegressor* também suporta regressão de várias saídas, desta forma uma amostra pode ter mais de um destino (NEURAL NETWORK MODELS (SUPERVISED) — SCIKIT-LEARN 0.21.3 DOCUMENTATION, [s.d.])

5. CONJUNTO DE DADOS

O estudo faz uso de uma base de dados de exames laboratoriais realizados por um laboratório de análises clínicas da região de Florianópolis, SC, Brasil, durante todo o período de 2015 a 2019. Os dados são oriundos de diferentes unidades e estavam semiestruturados, necessitando de um pré-processamento.

Considerando que se queria utilizar exames de rotina para predizer outros exames, realizou-se uma análise da frequência como primeiro critério para escolher os parâmetros de entrada. Como nosso alvo é a HbA1c, foram descartados todos os exames ou analitos com frequência de realização menor que a hemoglobina glicada, assim como aqueles com valores não quantitativos, uma vez que alguns métodos utilizados não suportam este tipo de dado. Isto resultou em uma lista com os 41 exames (analitos) mais frequentes.

5.1. ANÁLISE DESCRITIVA

De acordo com essa primeira seleção, a base possui 3.028.074 pacientes com idades entre 18 e 100 anos. A média de idade dos pacientes é de 51,8 anos com desvio padrão (SD) de 18,4, sendo 60,9% (1844360) de mulheres e 39,1% (1183714) de homens.

A idade média entre as mulheres é de 57 anos (SD 18,7) e entre os homens é de 54,3 anos (SD 17,7). Do total de pacientes, 61,9% (1874598) declararam fazer uso de algum tipo de medicamento. Sendo que deste montante, 64,9% (1201434) eram mulheres e 45,1% (673164) eram homens.

Analisando apenas os pacientes que realizaram exames de hemoglobina glicada, essa primeira seleção possui 489406 indivíduos. Tomando o exame de hemoglobina glicada (HbA1c) como padrão ouro e seguindo a classificação de diagnóstico adotada pela comunidade médica (AMERICAN DIABETES ASSOCIATION STANDARDS OF MEDICAL CARE IN DIABETES-2017, [s.d.]) (Tabela 3.1), a base possui 50,6% (247812) de pacientes classificados como saudáveis, 27,6% (135078) de pacientes classificados com pré-diabetes e 21,8% (106516) classificados com diabetes.

Entre os que não faziam uso de algum tipo de medicamento, 59,1% (68869) são classificados como saudáveis, 20,1% (23452) são classificados com pré-diabetes e 20.8%

(24256) são classificados com diabetes. Esta análise é importante, pois o uso contínuo de medicamento pode induzir um resultado que não represente a real saúde o paciente.

Na Tabela 5.1 é apresentado a lista contendo os 41 exames com ocorrência de realização maiores que o exame de hemoglobina glicada. Observa-se que os exames possuem quantidades de registros diferentes uns dos outros. Isso ocorre por que os pacientes não realizam sempre os mesmos exames, assim os pacientes possuem diferentes quantidades de exames realizados.

Tabela 5.1 – Estatística descritiva dos 41 exames com maior frequência na base ordenados de forma decrescente em relação ao total de registros.

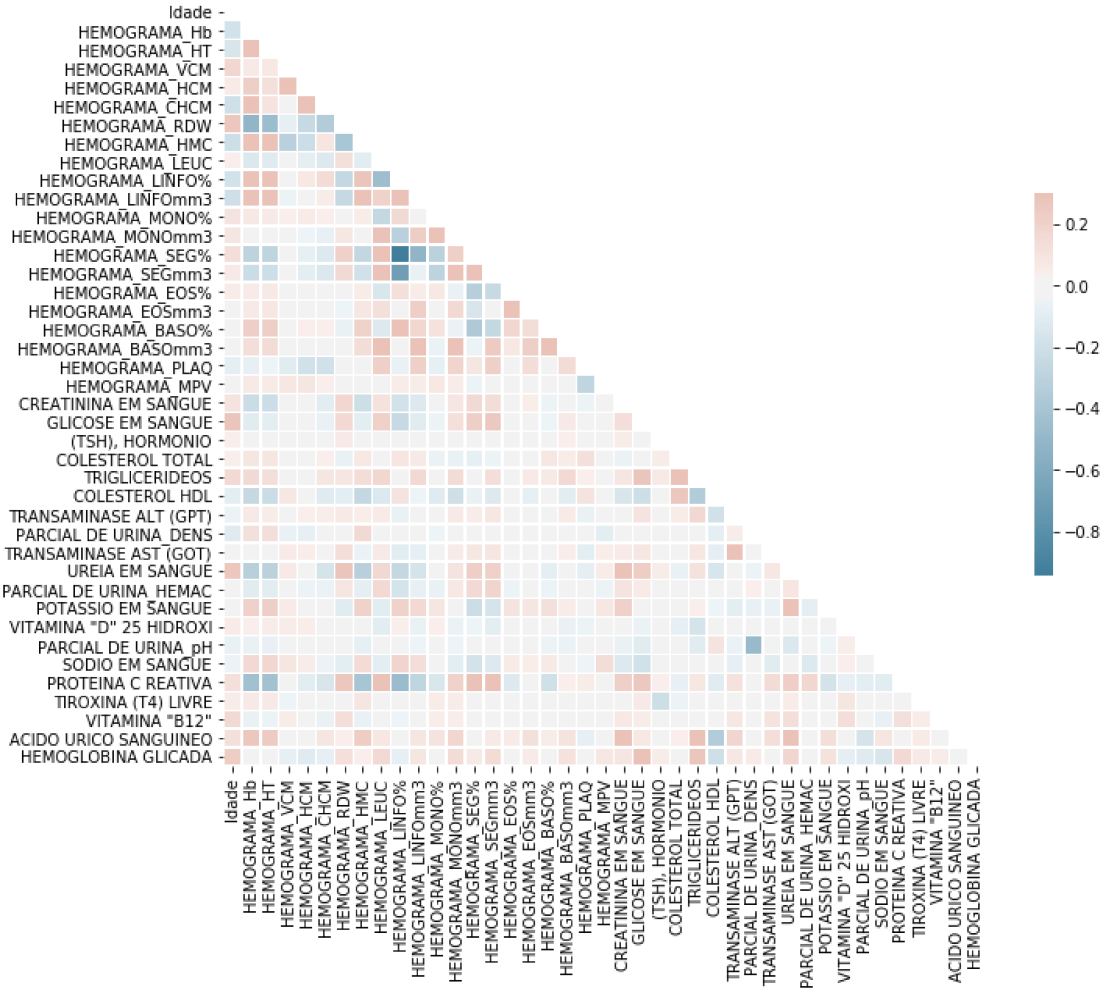
Exame (Analito)	Total	Média	SD	Min	Q1	Q2	Q3	Max
Idade	3028074	51.8	18.4	18.0	36.0	52.0	66.0	99.0
HEMOGRAMA_VCM	1977401	89.0	5.5	7.8	86.0	89.0	92.0	150.4
HEMOGRAMA_HCM	1977398	29.9	2.1	2.7	28.9	30.0	31.1	64.4
HEMOGRAMA_HT	1977396	38.3	5.8	0.3	35.5	39.0	42.1	84.5
HEMOGRAMA_CHCM	1977391	33.6	1.2	20.1	32.9	33.6	34.4	49.0
HEMOGRAMA_Hb	1977388	12.9	2.0	1.1	11.9	13.1	14.2	25.5
HEMOGRAMA_RDW	1977339	13.7	1.7	9.5	12.7	13.2	14.1	39.9
HEMOGRAMA_LINFO%	1977251	29.6	11.8	0.1	21.8	30.4	37.4	136.0
HEMOGRAMA_HMC	1977177	4.3	0.7	0.0	4.0	4.4	4.8	9.7
HEMOGRAMA_LEUC	1976697	7514.5	3874.9	10.0	5430.0	6740.0	8580.0	100000.0
HEMOGRAMA_MONO%	1976659	6.0	2.2	0.1	4.7	5.8	7.0	60.0
HEMOGRAMA_MPV	1976601	8.9	1.2	1.0	8.1	8.8	9.6	28.6
HEMOGRAMA_MONOmm3	1975131	425.4	202.0	1.0	303.0	389.0	502.0	2995.0
HEMOGRAMA_PLAQ	1974136	243.0	85.6	1.0	192.0	235.0	283.0	800.0
HEMOGRAMA_LINFOmm3	1973551	1988.3	802.3	1.0	1476.0	1945.0	2438.0	9998.0
HEMOGRAMA_SEG%	1859479	60.2	12.9	0.1	51.7	59.0	68.0	100.0
HEMOGRAMA_SEGmm3	1854139	4632.0	2912.8	1.0	2881.0	3827.0	5388.0	25000.0
HEMOGRAMA_EOS%	1836983	2.9	2.6	0.1	1.4	2.4	3.7	90.0
HEMOGRAMA_EOSmm3	1836820	188.7	147.8	1.0	91.0	152.0	243.0	999.0
HEMOGRAMA_BASO%	1816560	0.7	0.4	0.1	0.4	0.6	0.8	20.0
HEMOGRAMA_BASOmm3	1815362	46.2	33.3	1.0	27.0	40.0	57.0	700.0
CREATININA EM SANGUE	1558596	1.1	1.2	0.0	0.7	0.9	1.0	44.7
GLICOSE EM SANGUE	1417342	103.1	37.5	2.0	86.0	93.0	105.0	998.0
(TSH), HORMONIO	1135526	2.7	5.2	0.0	1.3	2.0	3.0	293.9
COLESTEROL TOTAL	1045713	190.3	41.5	36.0	161.0	187.0	216.0	598.0
TRIGLICERIDEOS	1002706	126.8	72.1	16.0	77.0	109.0	156.0	700.0
COLESTEROL HDL	995150	52.8	14.8	2.0	42.0	51.0	61.0	200.0
TRANSAMINASE ALT (GPT)	948262	29.8	29.6	0.0	17.0	22.0	32.0	500.0
PARCIAL DE URINA_DENS	878287	1015.5	7.4	1000.0	1010.0	1015.0	1020.0	1079.0
TRANSAMINASE AST (GOT)	858743	29.9	25.3	1.0	21.0	25.0	31.0	500.0
UREIA EM SANGUE	838378	47.2	36.8	1.0	28.0	36.0	49.0	500.0
PARCIAL DE URINA_HEMAC	830274	17435.7	75278.1	100.0	660.0	1320.0	6400.0	1000000.0
POTASSIO EM SANGUE	799865	4.3	0.6	1.1	3.9	4.3	4.6	10.0
VITAMINA "D" 25 HIDROXI	761135	26.6	11.1	9.0	19.7	25.2	31.5	149.6
PARCIAL DE URINA_pH	704842	6.0	0.8	5.0	5.0	6.0	6.0	9.0
SODIO EM SANGUE	703848	139.0	4.6	86.0	137.0	139.0	142.0	193.0
PROTEINA C REATIVA	591453	35.9	66.2	0.0	1.2	6.2	36.9	971.7
TIROXINA (T4) LIVRE	583772	1.2	0.3	0.1	1.0	1.1	1.3	11.5
VITAMINA "B12"	569179	490.4	286.5	61.0	336.0	420.0	539.0	2001.0
ACIDO URICO SANGUINEO	497132	5.4	1.6	0.1	4.3	5.3	6.4	41.7
HEMOGLOBINA GLICADA	489406	6.1	1.4	2.5	5.3	5.6	6.3	18.3

Na tabela temos que:

- Total é o número de registros na base que contém o analito.
- Média é o valor médio do analito.
- SD é o desvio padrão do analito.
- Min é o menor valor para aquele analito.
- Q1 é o primeiro quartil para aquele analito.
- Q2 é o segundo quartil ou mediana para aquele analito.
- Q3 é o terceiro quartil para aquele analito.
- Max é o maior valor para aquele analito.

A Figura 5.1 mostra uma análise de correlação entre os diferentes exames (analitos) previamente selecionados. De modo geral, os analitos possuem baixa correlação entre eles. A exceção é para o *HEMOGRAMA_SEG* e o *HEMOGRAMA_LINFO*.

Figura 5.1 – Correlação entre os exames previamente selecionados.



No **Error! Reference source not found.** é apresentado o histograma com a distribuição de cada parâmetro assim como o *boxplot*, a fim de observar a distribuição

dos dados e os outliers dos exames previamente selecionados. Optou-se por manter os *outlier*, uma vez que eles podem representar indivíduos portadores de alguma patologia, alvos desta pesquisa. O único parâmetro que teve os outliers removidos foi a idade (Figura 5.2). Em consequência dos valores elevados de *outliers*, outros analitos aparecem fortemente achatados nos gráficos (Figura 5.3, Figura 5.4).

De modo geral, todos os analitos plotados, possuem uma distribuição gaussiana. No entanto, devido ao intervalo de valores, alguns apresentam esta distribuição de forma discreta ou pulsada.

Figura 5.2 – Histograma e *Boxplot* da idade dos pacientes.

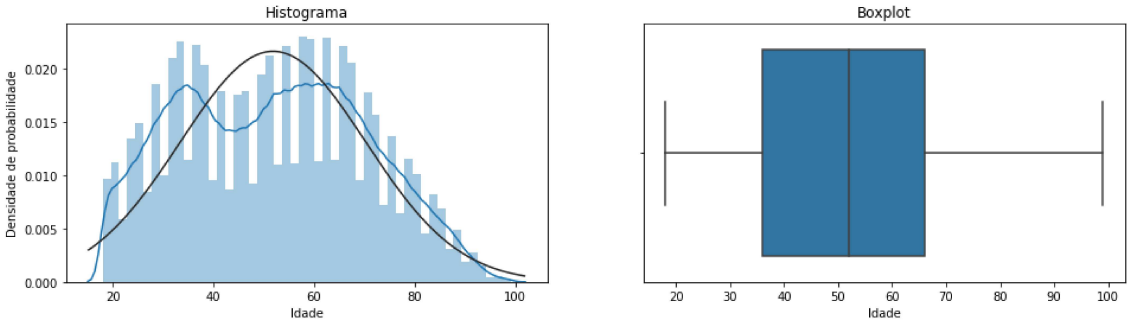


Figura 5.3 – Histograma e Boxplot do analito GLICOSE EM SANGUE.

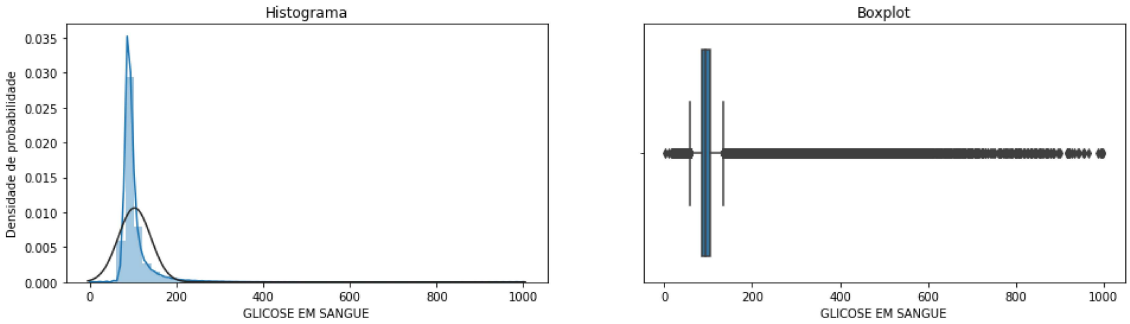
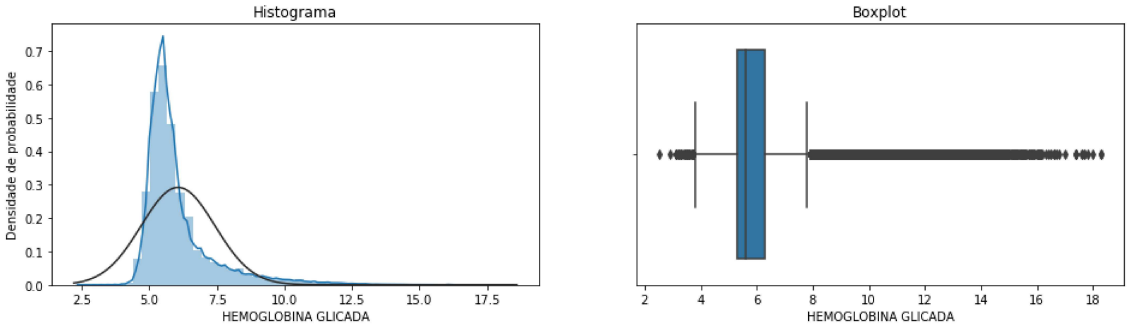


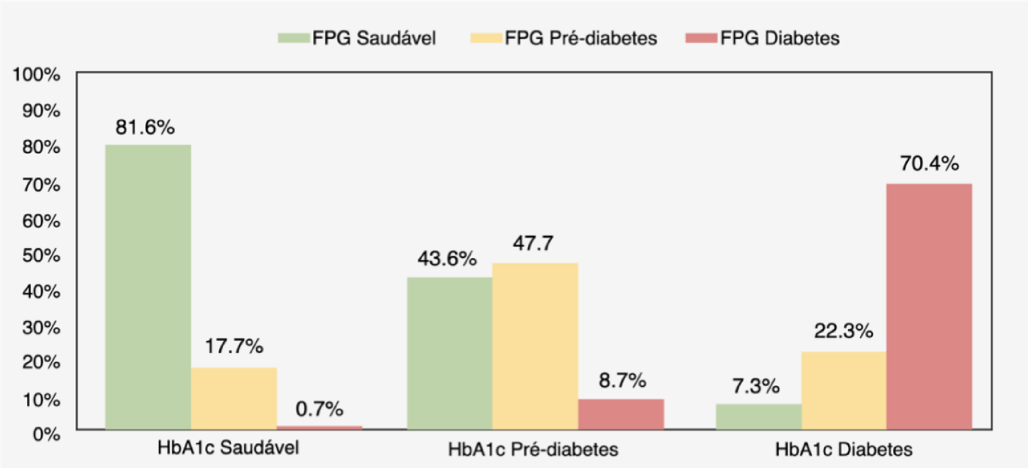
Figura 5.4 – Histograma e Boxplot do analito HEMOGLOBINA GLICADA.



5.2. COMPARAÇÃO DOS DIAGNÓSTICOS

Finalizando a análise dos dados, comparamos o diagnóstico realizado com o exame de HbA1c com o diagnóstico realizado com o exame de FPG. Considerando o diagnóstico com o exame de HbA1c como padrão ouro, analisou-se a concordância do exame de FPG. Para cada classe do diagnóstico com o HbA1c, analisamos o diagnóstico dos mesmos pacientes com o exame de FPG. Esta análise revelou certa discrepância entre o resultado dos diferentes exames, com é apresentado na Figura 5.5.

Figura 5.5 – Comparação das proporções da classificação do exame FPG em relação a classificação do exame HbA1c.



Esta comparação de resultados chama atenção para o percentual de FPG classificado como falso negativo em relação ao HbA1c. A mesma relação se mantém quando analisamos os mesmos dados isoladamente a cada ano e mesmo quando analisamos apenas indivíduos que não fazem uso de medicamentos.

O falso negativo é observado no conjunto de exames classificados pelo HbA1c como diabetes. Neste conjunto, quando o diagnóstico é realizado com o exame de FPG, 22.3% são classificados como pré-diabéticos e 7.3% são classificados como saudáveis. Estes valores vem de encontro aos estudos apontados por David (SACKS, 2011), onde cerca de 30% dos testes de FPG podem apresentar falsos negativos. O mesmo ocorre no diagnóstico de pré-diabetes, onde 43.6% são classificados como saudáveis segundo a classificação do exame de FPG.

Obviamente que neste tipo de análise é preciso considerar que alguns resultados podem estar muito próximo a valores na margem das classificações. No entanto, fica evidente a discrepância entre os exames quando observamos mais de sete porcento de

pacientes classificados como saudáveis pelo FPG quando o HbA1c aponta para o diagnóstico de diabetes.

Estes resultados de falsos negativos podem representar um risco ao paciente, levando o mesmo a não se tratar, uma vez que a doença se apresenta de forma assintomática.

6. MATERIAIS E MÉTODOS

O estudo foi devidamente aprovado junto ao comitê de ética em pesquisa da Universidade Federal de Santa Catarina sob o número de registro CAAE 02203918.0.0000.0121. Todas as simulações foram realizadas com a linguagem de programação Python utilizando o ambiente de desenvolvimento Jupyter, que é próprio para simulações científicas com uso do Python, juntamente com a biblioteca *scikit learn,* que provê uma série de recursos e métodos para se trabalhar com *machine learning*.

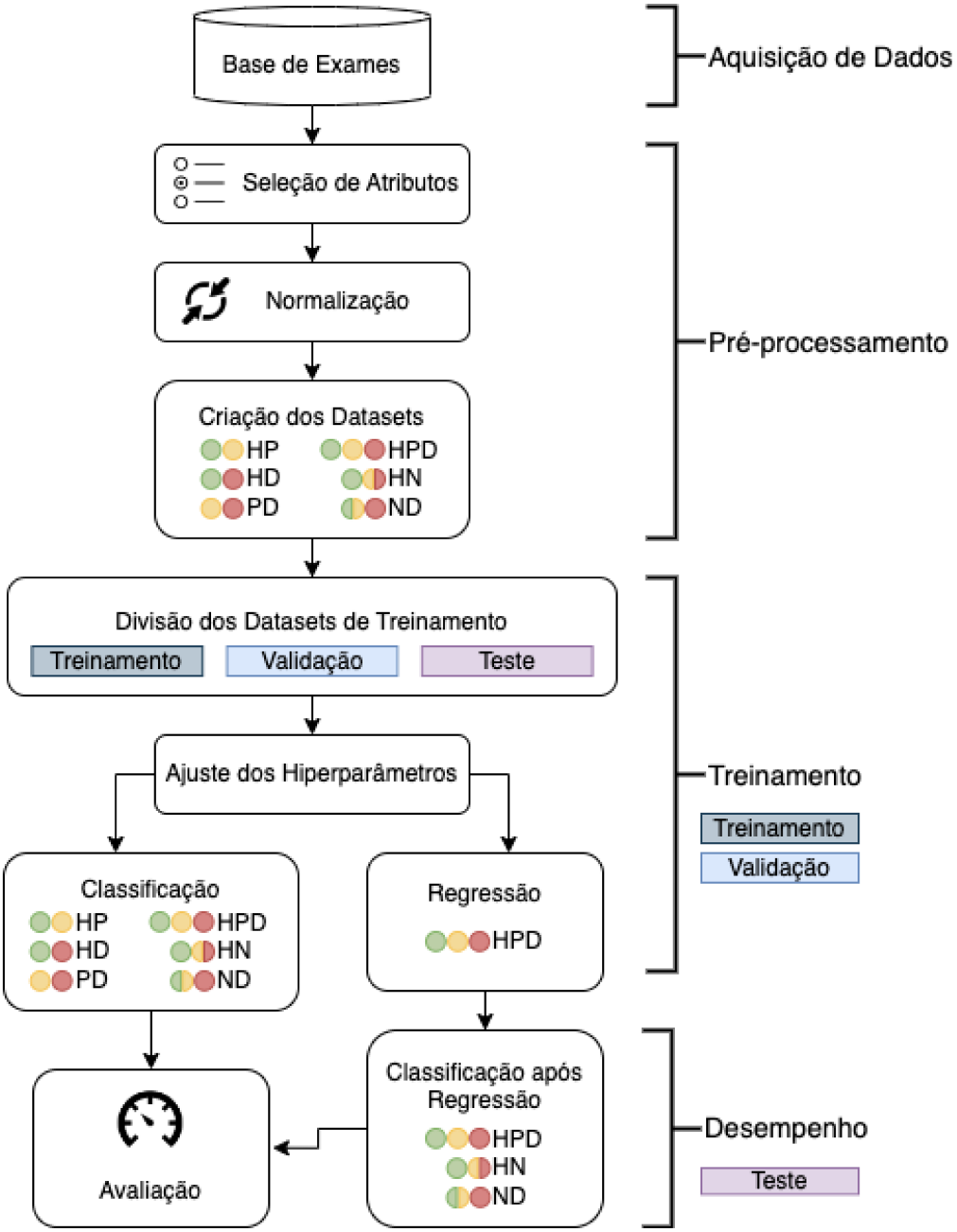
A metodologia adotada pode ser dividida em dois momentos distintos. Inicialmente realizou-se uma análise exploratória, onde foram testados diferentes modelos de machine learning para predição de exames utilizados no diagnóstico de DM. Na sequência buscou-se a construção de um método que auxiliasse na identificação de exames de FPG com resultados falso negativo, a fim de auxiliar no diagnóstico da doença e evitar que pacientes assintomáticos fiquem sem tratamento adequado.

6.1. PREDIÇÃO DO HbA1c

Esta fase teve como objetivo analisar diferentes modelos de machine learning buscando explorar as características de cada um. Para a predição do HbA1c foram testados tanto modelos de classificação como modelos de regressão, sendo que estes últimos tiveram seus resultados classificados posteriormente, de acordo com os limites apresentados na Tabela 3.1. Para a realização da análise, os dados foram divididos em diferentes combinações de datasets de acordo com cada classe do diagnóstico de diabetes.

Na Figura 6.1 é apresentado o diagrama com o fluxo da metodologia adotada nesta fase da pesquisa. Além da aquisição dos dados, esta primeira fase também é constituída por uma etapa de pré-processamento dos dados, construção dos datasets e treinamento dos modelos, finalizando com a testagem e avaliação de desempenho.

Figura 6.1 – Apresentação de diagrama com as quatro principais etapas adotadas na metodologia: Aquisição de dados, pré-processamento de dados, treinamento dos modelos e avaliação de desempenho



6.1.1 Seleção dos Atributos

Após testar diferentes técnicas de redução de dimensionalidade, optou-se por utilizar a técnica de seleção de variáveis (ou atributos) por análise fatorial. Neste processo foram utilizados os 41 parâmetros previamente selecionados.

Desconsiderando a avaliação do gráfico de ScreePlot, onde geralmente o cotovelo da curva é usado como ponto de corte para determinar o número de fatores,

realizou-se a análise fatorial para o total de 41 parâmetros aplicando uma taxa de corte de 75% de influência sobre a saída. Por fim escolheu-se os parâmetros com maior influência em cada fator.

Na Tabela 6.1 é apresentado o resultado da análise fatorial com os fatores e parâmetros com mais de 75% de influência na saída.

Tabela 6.1 – Grupo de fatores com parâmetros de maior influência sobre a saída.

Parâmetro	Item	MR1	MR2	MR3	MR29	MR5	MR6	MR9	MR11	MR8
CREATININA EM SANGUE	5	**-0.90**								
HEMOGRAMA_HT	16		**0.97**							
HEMOGRAMA_Hb	17		0.94							
HEMOGRAMA_HMC	15		0.90							
HEMOGRAMA_SEG%	26			**0.96**						
HEMOGRAMA_LINFO%	19			-0.95						
HEMOGRAMA_LEUC	18				**0.97**					
HEMOGRAMA_VCM	28					**0.97**				
HEMOGRAMA_HCM	14					0.92				
GLICOSE EM SANGUE	7						**0.93**			
HEMOGRAMA_BASO%	9							**0.95**		
HEMOGRAMA_BASOmm3	10							0.88		
TRANSAMINASE AST (GOT)	38								**0.88**	
TRANSAMINASE ALT (GPT)	37								0.85	
HEMOGRAMA_MONO%	21									**0.98**

Avaliando os valores da análise fatorial, foram selecionados 9 parâmetros principais (FACTOR-ANALYZER · PYPI, [s.d.]) com maior influência sobre a saída.

Tabela 6.2 – Lista de parâmetros de entrada.

Atributos
IDADE
CREATININA EM SANGUE
HEMOGRAMA_HT
HEMOGRAMA_SEG%
HEMOGRAMA_LEUC
HEMOGRAMA_VCM
GLICOSE EM SANGUE
TRANSAMINASE AST (GOT)
HEMOGRAMA_MONO%

Seguindo com o pré-processamento, após a seleção dos atributos, a base resultante teve os registros com dados faltantes removidos, já que nem todos os pacientes realizaram todos os exames selecionados. Isso foi necessário para ter uma base robusta e íntegra com todos os registros completos.

Como resultado desta interseção, a seleção dos atributos reduziu a quantidade de registros da base para 201338 pacientes. Analisando este novo conjunto de dados, tem-se que, de acordo com a classificação do exame de hemoglobina glicada, HbA1c (Tabela

3.1), a base está dividida com 58,26% de indivíduos saudáveis, 26,94% de indivíduos com pré-diabetes e 14,80% de indivíduos saudáveis.

6.1.2 Conjuntos de Dados (Datasets)

Buscando explorar os dados sob diferentes aspectos a fim de entender melhor a relação dos mesmos, dividiu-se a base em seis Conjuntos distintos de dados, de acordo com a classificação dos mesmo em relação ao diagnóstico da Diabetes:

 HP – Conjunto de dados com pacientes **Saudáveis e Pré-diabetes**. Neste grupo removemos os indivíduos Diabéticos.

 HD - Conjunto de dados com pacientes **Saudáveis e Diabéticos.** Neste grupo removemos os indivíduos Pré-diabéticos.

 **PD** - Conjunto de dados com pacientes **Pré-diabéticos e Diabéticos.** Neste grupo removemos os indivíduos saudáveis.

 HPD - Conjunto de dados com pacientes **Saudáveis**, **Pré-diabéticos** e **Diabéticos**.

 HN - Conjunto de dados com pacientes **Saudáveis e Não Saudáveis**. Neste grupo, a categoria 'Não Saudáveis' é composta por pacientes Pré-diabéticos e Diabéticos.

 **ND** - Conjunto de dados com pacientes **Não Diabéticos e Diabéticos**. Neste grupo, a categoria 'Não Diabéticos' é composta por pacientes Saudáveis e Pré-Diabéticos.

Na sequência, os dados foram normalizados, tendo em vista que alguns modelos exigem tal prática. O método utilizado foi a padronização de escala ou escore-z, resultando em dados com uma distribuição gaussiana, com média zero e desvio padrão igual a um. Os cálculos foram realizados com o método *StandardScaler* da biblioteca *scikit-learn* (PREPROCESSING DATA — SCIKIT-LEARN 0.22.2 DOCUMENTATION, [s.d.]).

Continuando a etapa de pré-processamento, os datasets foram divididos de forma aleatória em dois grupos distintos, sendo o primeiro formado por 60% do total e destinado

ao treinamento e validação dos modelos. Já o segundo grupo, com os 40% restantes, foi destinado aos testes finais e avaliação de desempenho dos modelos.

Para o treinamento com os modelos de classificação, as saídas foram categorizadas de forma binária e com multi-classes de acordo com cada grupo de dados, conforme é apresentado na Tabela 6.3. Já para o modelo de redes neurais, a saída multi-classe do grupo HPD, foi categorizada utilizando o método *One Hot Encode*, onde três novas variáveis de saída foram criadas, sendo uma para cada classe, conforme Tabela 6.4.

Tabela 6.3 – Categorização binária das saídas de acordo com cada grupo de dados.

Grupo	0	1	2
HP	Saudáveis	Pré-diabéticos	
HD	Saudáveis	Diabéticos	
PD	Pré-diabéticos	Diabéticos	
HN	Saudáveis	Não Saudáveis	
ND	Não Diabéticos	Diabéticos	
HPD	Saudáveis	Pré-diabéticos	Diabéticos

Tabela 6.4 – Categorização da saída multi-classe para o grupo HPD e modelo de rede neural com a utilização do método One Hot Encode.

Saudáveis	Pré-diabéticos	Diabéticos
1	0	0
0	1	0
0	0	1

6.1.3 Treinamento

Nesta etapa foi utilizado o grupo de dados para treinamento e validação, definido anteriormente e formado por 60% dos dados totais. Este grupo de dados foi novamente dividido em outros dois grupos com a proporção de 70% destinado ao treinamento dos modelos e os 30% restantes destinados ao processo de validação e ajuste dos hiperparâmetros. Neste processo foi utilizado um método de otimização (*GridSearchCV* do *sklearn*) que realizou inúmeros testes com diferentes conjuntos de hiperparâmetros até encontrar aquele que configurasse o melhor resultado. No entanto, após esta etapa, alguns

hiperparâmetros foram ajustados manualmente, buscando assim reduzir o *overfitting* e melhorar o desempenho.

Para o processo de treinamento foram avaliados cinco diferentes modelos de aprendizado de máquina, buscando assim, explorar as características particulares de cada um. Todos os cinco modelos foram testados como classificadores assim como regressores.

A seguir segue a lista dos modelos testados:

K-Nearest Neighbors (KNN)

O primeiro modelo testado foi o KNN. Como mencionado anteriormente, ele se baseia nas características dos vizinhos para realizar a classificação de um elemento.

Seguindo as definições do método KNeighborsClassifier (versão 1.1.3), os seguintes parâmetros foram configurados:

- Número de vizinhos: Foram utilizados 8 vizinhos na análise de cada elemento. Observamos que aumentando o número de vizinhos a sensibilidade diminui, mas em contrapartida a precisão aumenta.
- Peso: Aplica um peso dependendo da distância do vizinho. Neste parâmetro usamos o valor "*uniform*", que trata todos os vizinhos com o mesmo peso. Observamos que desta forma os modelos se tornam mais generalistas.
- Algoritmos: Usamos a função "*ball_tree*", que particiona a busca dos vizinhos em bolhas.

Os demais parâmetros foram mantidos conforme a configuração padrão.

Support Vector Machine (SVM)

Para classificação com SVM usamos o método SVC (versão 1.1.3) da biblioteca *Scikit-learn* e configuramos os seguintes parâmetros:

- Regularização: Utilizamos o valor de 0.8 para uma regularização do tipo L2, buscando melhorar a generalização do modelo.
- Kernel: Utilizamos a função padrão "rbf".

- Função de Decisão: Usamos a função “ovr”, que aplica o método “um versus o restante”.

Os demais parâmetros foram mantidos como padrão.

Naïve Bayes (NB)

Neste caso manteve-se todos os parâmetros padrão do método GaussianNB (versão 1.1.3), disponibilizado pela biblioteca *scikit-learn* (NAIVE BAYES — SCIKIT-LEARN 0.21.3 DOCUMENTATION, [s.d.]).

Random Forest (RF)

No caso do modelo Floresta Randômica, utilizou-se o método RandomForestClassifier (versão 1.1.3) sendo que os seguintes parâmetros foram configurados.

- Estimadores: Representa o número de árvores na estrutura. Neste caso forma utilizados 10.
- Profundidade máxima: Representa a profundidade máxima da árvore. Utilizou-se 5.
- Random State: Controla a aleatoriedade na construção das árvores. Aqui utilizou-se o valor 10.

Artificial Neural Network (ANN)

No caso da Rede Neural Artificial, utilizou-se o modelo Perceptron multicamadas fazendo uso do método MLPClassifier (versão 1.1.3).

Foram utilizadas duas camadas ocultas, sendo a primeira com 20 neurônios e segunda com 50 neurônios, ambas com função de ativação “relu”. Na otimização foi utilizado o método “*adam*” com um aprendizado adaptativo e uma taxa de inicial de 0,001. Também foi utilizado um alfa de 0,01, batch_size “auto” e tol de 0,001.

No caso da regressão, os hiperparâmetros foram os mesmos do modelo de classificação. Os modelos de regressão receberam os mesmos nomes dados aos de classificação, seguidos pela letra “r” (**KNNr**, **SVMr**, **NBr**, **RFr** e **ANNr**).

No caso da rede neural, também foi realizada uma classificação após o processo de regressão. Neste modelo, após a predição das saídas, os valores foram classificados de

acordo com os grupos, “Saudáveis e Não Saudáveis” (HN), “Não Diabéticos e Diabéticos” (ND) e “Saudáveis, Pré-diabéticos e Diabéticos” (HPD), seguindo os limites de referência conforme a Tabela 3.1.

6.1.4 Avaliação de Desempenho

Como métrica para avaliação dos modelos testados, além da acurácia, que neste caso não é recomendada devido ao desbalanceamento da base, utilizamos a Escore-F1, a área sob a curva com o gráfico Precision-Recall e a Matriz de Confusão, onde avaliamos a Sensibilidade (SN), Especificidade (SP), Precisão (PR) e Predição Negativa (NPR), sempre utilizando a base de teste final, que foi separada inicialmente.

Devido a característica de desbalanceamento da base, o recomendado é que seja utilizada a Escore-F1 e o valor da área sob a curva do gráfico Precision-Recall, no entanto, a métrica que melhor permite avaliar a capacidade preditiva dos modelos é análise conjunta da Sensibilidade (SN) e da Precisão (PR).

Para os modelos de regressão, foram utilizados o erro médio absoluto (MAE), erro médio quadrático (MSE) e o a raiz do erro médio quadrático (RMSE).

Após a regressão, os valores preditos foram também classificados, sendo que o resultado deste processo foi também avaliado com a mesmas ferramentas utilizadas para os modelos de classificação. Isso permitiu comparar o resultado dos modelos de classificação com as classificações realizadas após a predição dos modelos de regressão.

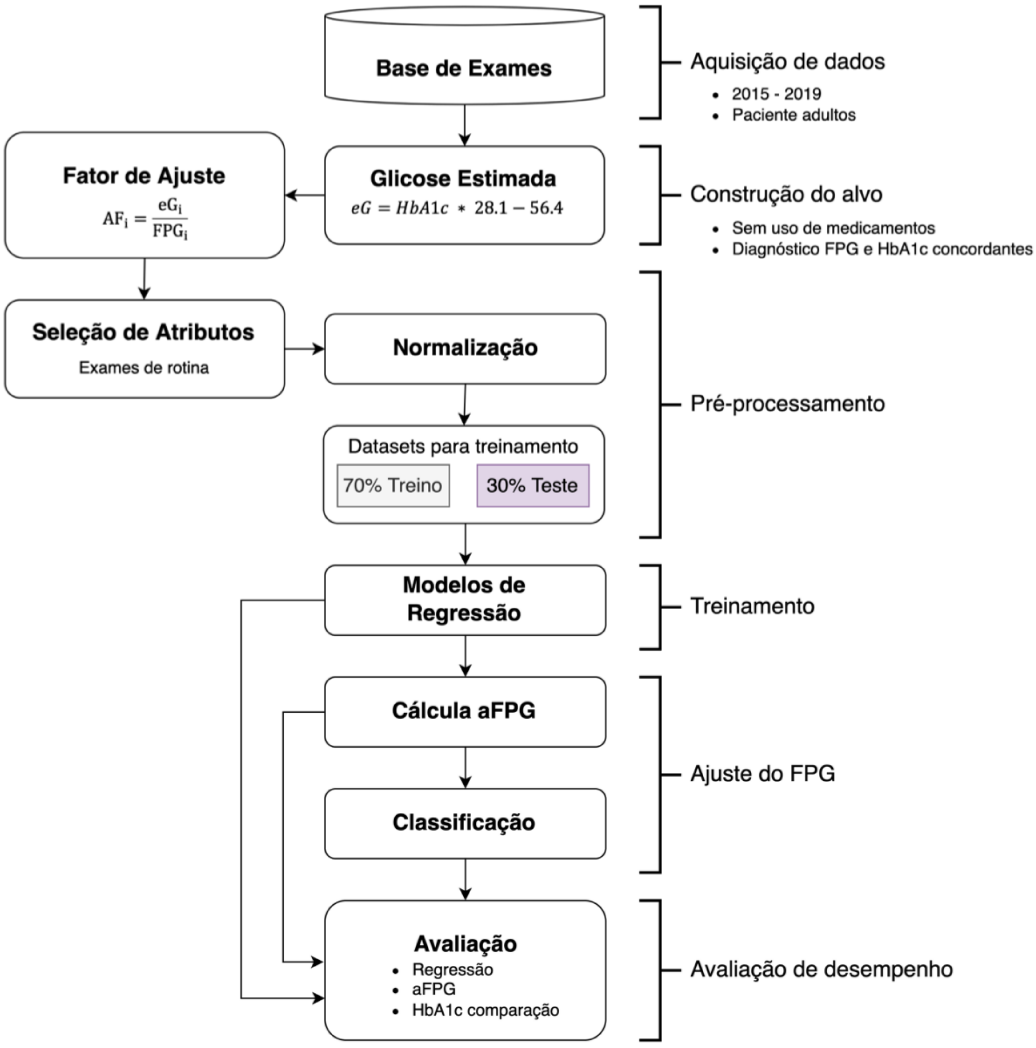
6.2. IDENTIFICAÇÃO DE FALSOS NEGATIVOS

Buscando aperfeiçoar a metodologia adotada até então, assim como melhorar a contribuição para a triagem e apoio ao diagnóstico da Diabetes, esta segunda etapa teve como meta a identificação de exames de FPG com resultados falso negativo.

Para isso foi desenvolvido um método para calcular um valor de “FPG Ajustado”, cuja classificação do diagnóstico da diabetes possua maior concordância em comparação aos resultados obtidos com o exame de HbA1c. O objetivo deste fator de ajuste não é alterar o valor do FPG, mas sim possibilitar uma comparação entre diferentes métodos de diagnóstico.

A metodologia seguiu seis etapas distintas, sendo composta pela (1) aquisição dos dados, (2) construção do alvo, (3) pré-processamento, (4) treinamento dos modelos, (5) ajuste do FPG e (6) avaliação de desempenho, como apresentado na Figura 6.2.

Figura 6.2 – Fluxograma proposto para a metodologia adotada: (1) aquisição dos dados, (2) construção do alvo, (3) pré-processamento de dados, (4) treinamentos dos modelos, (5) Ajuste do FPG e (6) avaliação de desempenho.



Inicialmente o alvo ou variável de saída dos modelos, assim como as variáveis de entrada ou atributos receberam tratamento adequado. Na sequência, após o pré-processamento necessário, os modelos foram treinados. Como resultado do treinamento um valor ajustado de FPG foi calculado. Por fim, os resultados encontrados foram testados e analisados.

Possuindo forte correlação com FPG (r = 0.84, p < 0.0001) (RITA BETTENCOURT SILVA et al., [s.d.]), valores médios de glicose (AG) seriam mais estáveis e adequados ao diagnóstico, amortizando erros e oscilações ocorridas ao longo do dia. De acordo com o A1c-Derived Average Glucose study (NATHAN et al., 2008), HbA1c pode ser utilizado para calcular a Glicose Média Estimada (Equação 6.1). Desta forma, utilizando os valores de HbA1c, calculamos a eAG para a base de dados analisada (251814 pacientes).

$$eAG = HbA1c\ .\ 28{,}7 - 46{,}7 \qquad \text{Equação 6.1}$$

No entanto, quando se compara a classificação do diagnóstico de diabetes referente aos valores calculados de eAG com a classificação do HbA1c, também pode-se observar certa discrepância entre os resultados. Nesta comparação, aproximadamente 67% dos pacientes, classificados como saudáveis pelo HbA1c, são classificados com pré-diabetes pelo eAG, gerando falsos positivos. O mesmo ocorre com a classificação de pré-diabetes, onde cerca de 30% são classificados pelo eAG com diabetes.

Diante da pouca concordância entre o diagnóstico do exame de HbA1c e a glicose média estimada (eAG), buscou-se a construção de uma nova equação que possuísse maior concordância entre a classificação de HbA1c e a de FPG. Neste processo, utilizou-se apenas os registros cujos pacientes não faziam uso de medicamentos e cuja classificação dos exames de HbA1c concordava com a classificação dos exames de FPG. Este novo grupo, com 32555 registro, foi utilizado para construir uma nova equação para o calculo de um valor de glicose estimada (eG). Neste processo, utilizando regressão linear (SKLEARN.LINEAR_MODEL.LINEARREGRESSION — SCIKIT-LEARN 0.24.2 DOCUMENTATION, [s.d.]), chegamos a seguinte equação de primeira ordem.

$$eG = HbA1c\ .\ 28{,}1 - 56{,}4 \qquad \text{Equação 6.2}$$

Esta equação foi utilizada para o cálculo de um valor de glicose estimada (eG) sobre toda base de dados (251814 pacientes). A comparação da classificação dos valores calculados com essa nova equação e a classificação do HbA1c pode ser observado na Tabela 6.5. Diferentemente do FPG e da eAG, é possível observar maior concordância entre a classificação dos valores de glicose estimada (eG) e a classificação com HbA1c.

Tabela 6.5 - Comparação da classificação dos valores de HbA1c com FPG, Glicose Média Estimada (eAG= $HbA1c\,.28{,}7 - 46{,}7$) e Glicose Estimada (eG= $HbA1c\,.\ 28{,}1 - 56{,}4$).

HbA1c	H - Saudável			P – Pré-diabetes			D - Diabetes		
Classificação	H	P	D	H	P	D	H	P	D
FPG	81.6%	17.7%	0.7%	43.6%	47.7%	8.7%	7.3%	22.3%	70.4%
eAG	32.6%	67.4%	0%	0%	70.4%	29.6%	0%	0%	100%
eG	**91.1%**	**8.9%**	**0%**	**0%**	**100%**	**0%**	**0%**	**0%**	**100%**

6.2.1 Construção do Alvo

Utilizando o valor de eG calculado e o valor de FPG, de acordo com a Equação 6.3, foi proposto um fator de ajuste (AF) para aproximar a classificação do exame de FPG da classificação do exame de HbA1c.

$$\mathrm{AF_i} = \frac{\mathrm{eG_i}}{\mathrm{FPG_i}} \qquad \text{Equação 6.3}$$

Estes valores de fator de ajuste, são utilizados como o alvo do treinamento dos modelos de machine learning. Com o objetivo de construir modelos que consigam predizer novos valores do fator de ajuste, estes valores preditos do fator de ajuste (pAF) são aplicados sobre os valores dos exames de FPG, retornando assim um FPG ajustado (aFPG), conforme é apresentado na Equação 6.4. Nesta equação, a variável a representa um índice de correção que poderá ser aplicado com objetivo de realizar um ajuste fino no cálculo.

$$\mathrm{aFPG_i} = \mathrm{FPG_i}\,.\,\mathrm{pAF_i} + a \qquad \text{Equação 6.4}$$

6.2.2 Pré-processamento

Diferentemente do processo anterior, onde foi utilizado a técnica estatística de análise fatorial (DA et al., [s.d.]) (BARTHOLOMEW, 1984) para seleção dos atributos. Agora foram selecionados apenas os analitos que compõem o hemograma completo e a glicose em sangue, já que o objetivo é o ajuste da glicose estimada.

Esta etapa de pré-processamento resultou em 15 diferentes variáveis de entrada, como apresentado na Tabela 6.6

Tabela 6.6 – Atributos utilizados no processo de treinamento dos modelos para o ajuste da glicose estimada.

Atributo	Descrição	Min	Max	Mean	SD
AF	**Fator de Ajuste**	**0,53**	**1,89**	**1,01**	**0,12**
Gender	Gênero	-	-	-	-
Age	Idade (ano)	20	99	46,22	14,85
FPG	Glicose sanguínea (mg/dl – milligram/decalitr)	38,00	583,00	97,92	28,26
Baso%	Basófilos (%)	0,10	6,80	0,73	0,32
MCHC	Concentração de hemoglobina corpuscular (g/dl – gram/decalitre)	24,2	37,9	33,83	1,04
MCH	Hemoglobina corpuscular (pg - picogram)	15,50	43,20	29,97	1,75
HT	Eritrócitos (%)	13,20	62,30	41,82	3,78
LC	Leucócitos (unidades/mm2)	1640,0	25280,0	6486,9	1754,6
Linfo%	Linfócitos (%)	2,40	79,50	34,93	7,81
Mono%	Monócitos (%)	0,80	22,10	6,15	1,49
MPV	Volume médio das plaquetas (%)	5,60	19,10	8,98	1,09
PLT	Contagem de plaquetas (unidades/mm2)	8,00	796,00	240,96	60,71
RDW	Distribuição de células vermelhas (%)	10,50	27,10	13,05	0,83
SEG%	Netrófilos segmentado (%)	13,10	92,80	54,97	8,39
MVC	Volume médio corpuscular (ft - fentoliter)	55,20	123,00	88,59	4,64

Como os hyperparâmetros já foram ajustados durante o treinamento apresentado na etapa anterior da metodologia, aqui não se precisou do grupo de validação.

Desta forma, a base de dados foi dividida em apenas dois datasets, sendo um destinado ao treinamento dos modelos, com 70% dos dados e outro destinado ao teste dos modelos, com os 30% restantes. Esta divisão foi feita de forma aleatória utilizando o método *train_test_split* da biblioteca sklearn (SKLEARN.LINEAR_MODEL.LINEARREGRESSION — SCIKIT-LEARN 0.24.2 DOCUMENTATION, [s.d.]). Por fim os dados de entrada foram também normalizados com média 0 e desvio padrão 1.

6.2.3 Treinamento

Diferentemente do processo de construção da equação de glicose estimada (eG), onde foram utilizados apenas exames concordantes entre FPG e HbA1c e cujos pacientes não faziam uso de medicamento, aqui foram utilizados todos os dados da base sem

distinção. Os outliers existentes também foram mantidos por entender que eles possam representar indivíduos doentes.

Para a predição do fator de ajuste (AF) foram utilizadas técnicas de regressão, sendo testados os modelos de machine learning treinados na etapa anterior:

- K-nearest Neighbors Regressor (KNNr)
- Support Vector Machine Regressor (SVMr)
- Naïve Bayes Regressor (NBr)
- Random Forest Regressor (RFr)
- Artificial Neural Networks Regressor (ANNr)

Além dos valores padrão definidos pela biblioteca, alguns valores de hiperparâmetros foram configurados conforme é apresentado a seguir:

- **KNN Function**: KNeighborsRegressor()

  n_neighbors=8,
  weights='distance',
  algorithm='ball_tree',
  leaf_size=30,
  p=2,
  metric='minkowski',
  metric_params=None,

- **SVM Function**: LinearSVR()

  epsilon=0.1,
  tol=1e-5,
  C=1.0,
  loss='epsilon_insensitive',
  fit_intercept=True,
  intercept_scaling=1.0,
  dual=True,
  random_state=None,
  max_iter=1000

- **NB Function**: BayesianRidge()

  n_iter=300,
  tol=0.001,
  alpha_1=1e-06,
  alpha_2=1e-06,

lambda_1=1e-06,
lambda_2=1e-06,
alpha_init=None,
lambda_init=None,
compute_score=False,
fit_intercept=True,
normalize=False,
copy_X=True,

- **RF Function**: RandomForestRegressor()

  n_estimators=50,
  criterion='mse',
  max_depth=5,
  min_samples_split=2,
  min_samples_leaf=1,
  min_weight_fraction_leaf=0.0,
  max_features='auto',
  max_leaf_nodes=None,
  min_impurity_decrease=0.0,
  min_impurity_split=None,
  bootstrap=True,
  oob_score=False,
  random_state=10,
  warm_start=False,
  ccp_alpha=0.0,
  max_samples=None

- **ANN Function**: MLPRegressor()

  hidden_layer_sizes=[20,50],
  activation='relu',
  solver='adam',
  alpha=0.1,
  batch_size='auto',
  learning_rate='adaptive',
  learning_rate_init=0.001,
  power_t=0.5,
  max_iter=150,
  shuffle=True,
  random_state=10,
  tol=0.00001,
  warm_start=False,
  momentum=0.9,

```
nesterovs_momentum=True,
early_stopping=False,
validation_fraction=0.1,
beta_1=0.9,
beta_2=0.999,
epsilon=1e-08,
n_iter_no_change=10,
max_fun=15000
```

6.2.4 Avaliação de desempenho

Como métrica de desempenho para o resultado da regressão nós utilizamos o erro médio absoluto (MAE), erro médio quadrático (MSE) e raiz do erro médio quadrático (RMSE), para avaliar os erros do modelo sobre a predição do fator de ajuste.

Com os valore de AF preditos (pAF), nós calculamos os valores de FPG ajustado (aFPG) e com os valores originais de glicose estimada (eG), plotamos um gráfico de distribuição a fim de observarmos o comportamento dos modelos.

Após esta etapa de regressão, os valores de aFPG foram classificados seguindo os critérios adotados para o diagnóstico de diabetes.

Como mencionado, uma métrica bastante comum, mas não recomendada para bases desbalanceadas, é a acurácia (ACC). Ela nos fornece uma ideia geral sobre o desempenho do modelo e pode ser calculada também sobre cada classe de diagnóstico. Apear da base estudada ser desbalanceada, nossos modelos são de regressão com classificação posterior, o que se entende não influenciar no resultado. No entanto, buscando uma melhor análise, utilizou-se como métrica de desempenho para a classificação do aFPG a matriz de confusão e o Escore-F1 (EVALUATION: FROM PRECISION, RECALL AND F-FACTOR TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION, [s.d.]) (REHMAN et al., 2018).

7. RESULTADOS

Os resultados são apresentados na mesma sequência adotada com a metodologia, sendo construídos ao longo do processo. Neste sentido, inicialmente são apresentados os resultados dos modelos de classificação para diferentes grupos de dados (datasets) criados e analisados. Para cada grupo de dados são plotadas as matrizes de confusão tanto com os valores totais de pacientes como com os percentuais. Os valores resultantes da matriz de confusão, são então utilizados para calcular a sensibilidade, a especificidade, a precisão, o valor preditivo negativo e o Escore-F1. Um gráfico com a curva (AUC) Precision-Recall também é apresentado, assim como os valores da área sob a curva para cada modelo testado.

Os valores da área sob a curva Precision-Recall, assim como os valores da matriz de confusão tem como alvo a classe que faz referencia a patologia ou a falta de saúde. No entanto, os valores de

Na sequência, utilizando os mesmos grupos de dados, são apresentados os resultados para os modelos de regressão e consequentemente a classificação dos respectivos resultados. Por fim, para os diferentes grupos de dados, compara-se os resultados dos modelos de classificação e classificação após regressão, concluindo assim a primeira etapa da metodologia.

Posteriormente são apresentados os resultados para o processo de ajuste do FPG, comparando a classificação dos resultados obtidos com a classificação do exame de FPG.

7.1. MODELOS DE CLASSIFICAÇÃO

Seguindo a metodologia para exploração dos modelos e predição do HbA1c, inicialmente foram testados apenas os modelos de classificação para os diferentes grupos de dados criados (HP, HD, PD, NH, ND e HPD). Para cada grupo e modelo testado, avaliou-se a Acurácia o Escore-F1 e a Matriz de Confusão que fornece a Sensibilidade (SN), Especificidade (SP), Precisão (PR) e o Valor Preditivo Negativo (NPR).

7.1.1 Grupo HP – Saudável e Pré-diabetes

Figura 7.1 – Matriz de confusão do modelo de classificação KNN para o grupo HP.

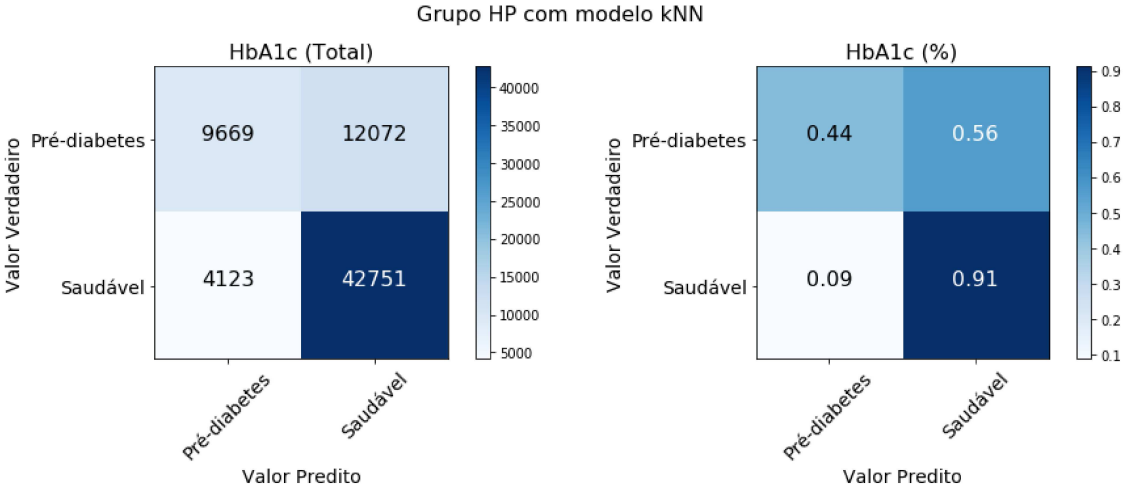


Figura 7.2 - Matriz de confusão do modelo de classificação SVM para o grupo HP.

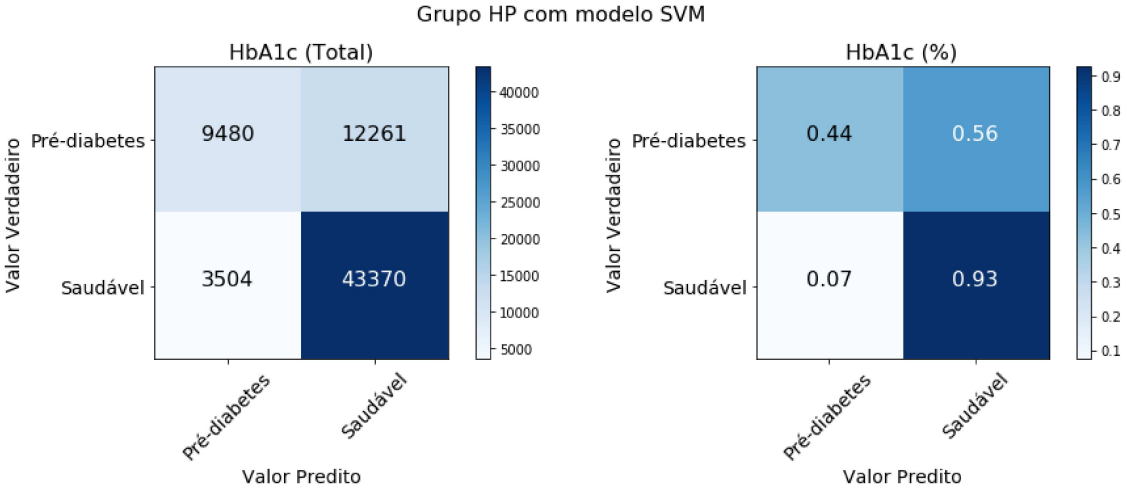


Figura 7.3 - Matriz de confusão do modelo de classificação GNB para o grupo HP.

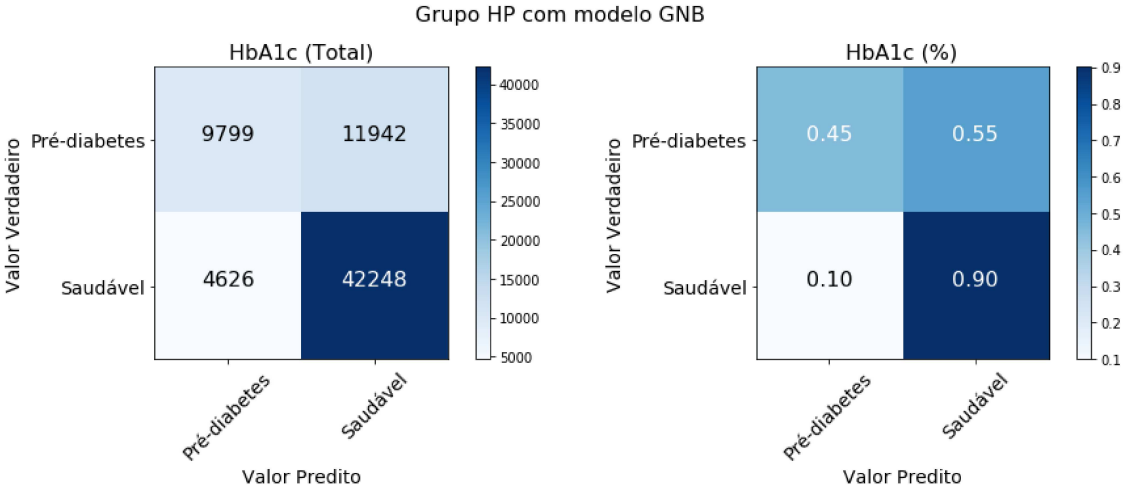


Figura 7.4 - Matriz de confusão do modelo de classificação RF para o grupo HP.

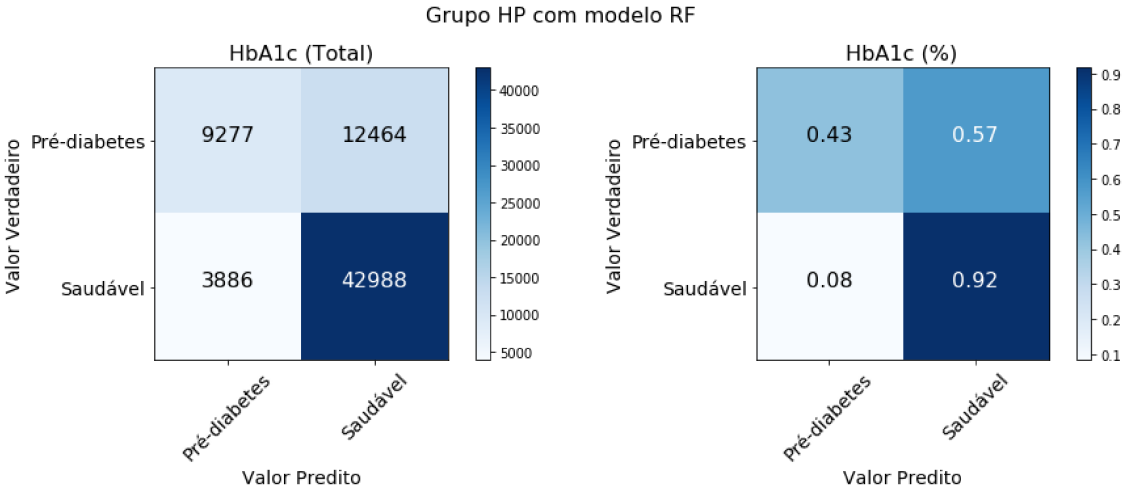
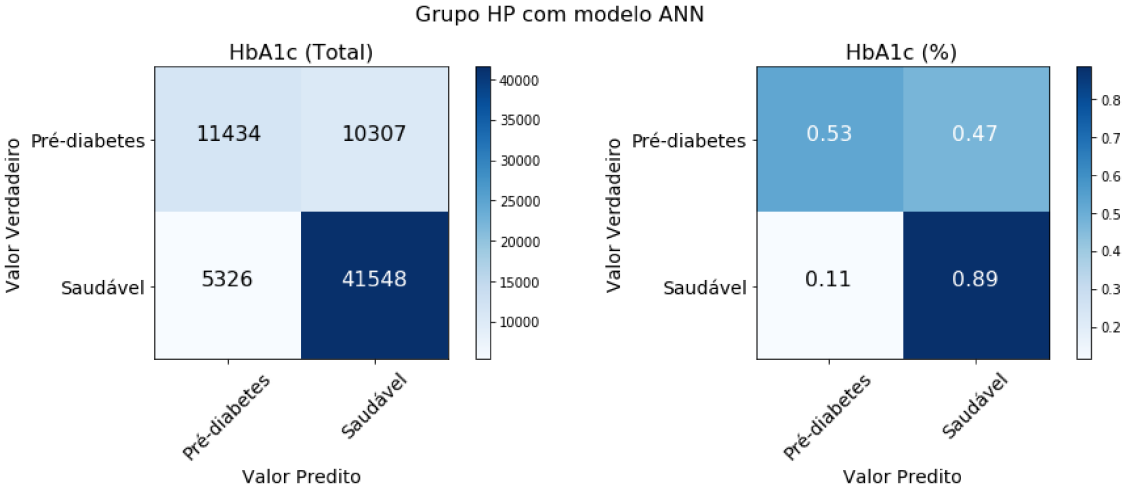
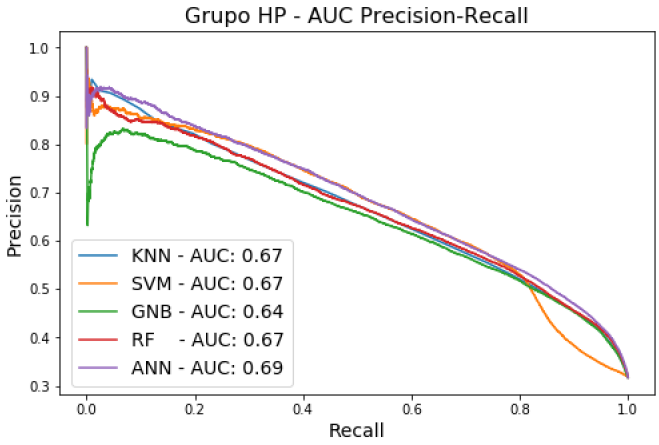


Figura 7.5 - Matriz de confusão do modelo de classificação ANN para o grupo HP.



Pelo fato de a base ser desbalanceada, optou-se por utilizar o gráfico Precision-Recall. Na Figura 7.6 – Área sob a curva para o gráfico Precision-Recall tendo com alvo a classe Pré-diabetes no grupo HP.é apresentado também os valores da área sob a curva do gráfico (AUC-PR) para cada modelo testado, tomando como alvo a Pré-diabetes.

Figura 7.6 – Área sob a curva para o gráfico Precision-Recall tendo com alvo a classe Pré-diabetes no grupo HP.



Na tabela é apresentado os resultados das métricas de avaliação para cada um dos modelos testados para o grupo HP (Saudável e Pré-diabetes), sobre cada classe do diagnóstico. Os valores de acurácia não são confiáveis como métricas para avaliação do poder de predição dos modelos, mas comparando a acurácia de treinamento com a de teste é possível observar a ocorrência ou não de overfitting.

Tabela 7.1 – Métrica de avaliação dos modelos de classificação para o grupo HP.

Classes	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	78,0	77,8	76,1	76,5	77,4
Acurácia Teste	76,4	77,0	75,9	76,2	77,2
AUC Precision-Recall	**67,0**	**67,0**	**64,0**	**67,0**	**69,0**
Pré-diabetes					
Sensibilidade ou Recall (SN)	**44,5**	**43,6**	**45,1**	**42,7**	**52,6**
Especificidade (SP)	91,2	92,5	90,1	91,7	88,6
Precisão (PR) ou Valor Preditivo Positivo	**70,1**	**73,0**	**67,9**	**70,5**	**68,2**
Valor Preditivo Negativo (NPV)	78,0	78,0	78,0	77,5	80,1
Escore-F1	**54,4**	**54,6**	**54,2**	**53,2**	**59,4**
Saudável					
Sensibilidade ou Recall (SN)	91,2	92,5	90,1	91,7	88,6
Especificidade (SP)	44,5	43,6	45,1	42,7	52,6
Precisão (PR) ou Valor Preditivo Positivo	78,0	78,0	78,0	77,5	80,1
Valor Preditivo Negativo (NPV)	70,1	73,0	67,9	70,5	68,2
Escore-F1	84,1	84,6	83,6	84,0	84,2

7.1.2 Grupo HD – Saudável e Diabetes

Figura 7.7 – Matriz de confusão do modelo de classificação KNN para o grupo HD.

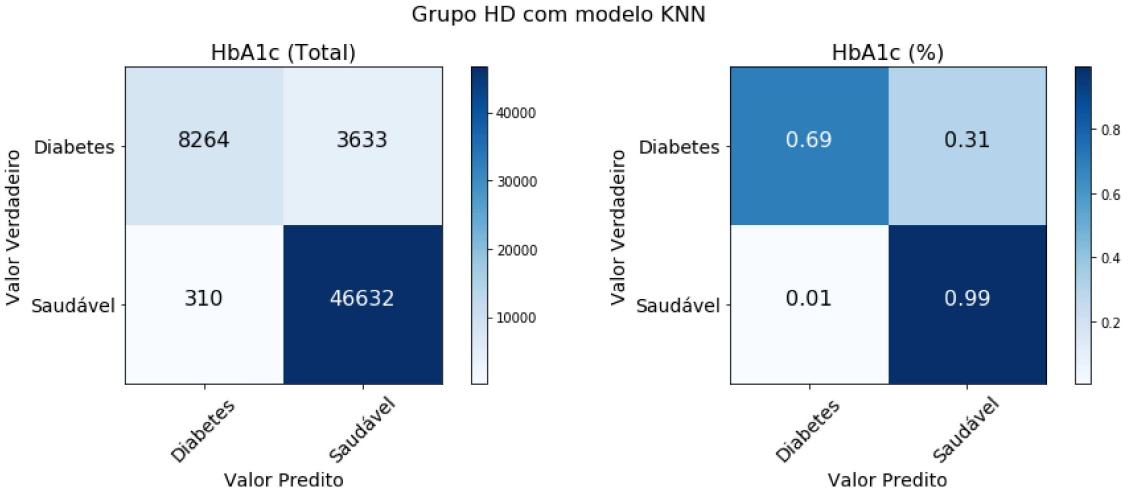


Figura 7.8 - Matriz de confusão do modelo de classificação SVM para o grupo HD.

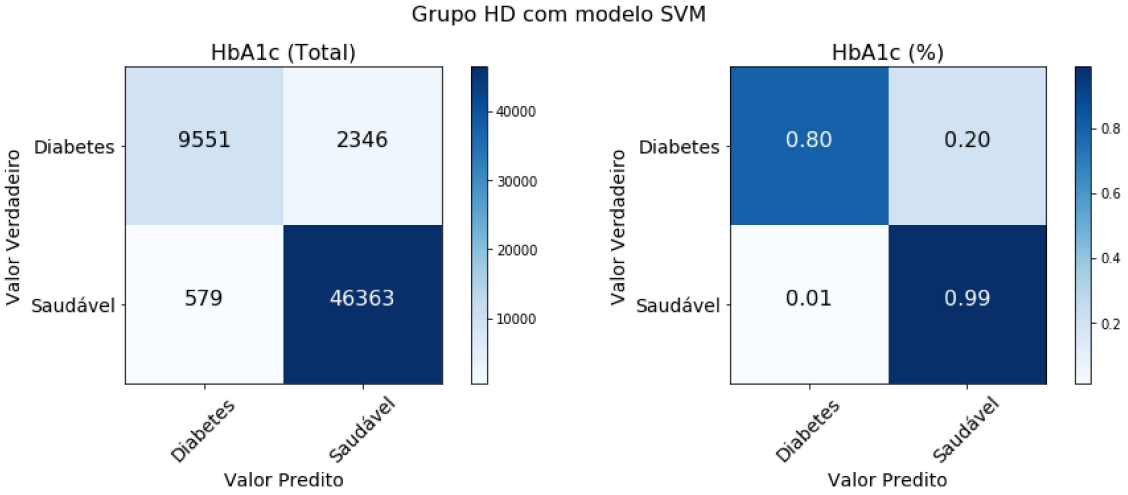


Figura 7.9 - Matriz de confusão do modelo de classificação GNB para o grupo HD.

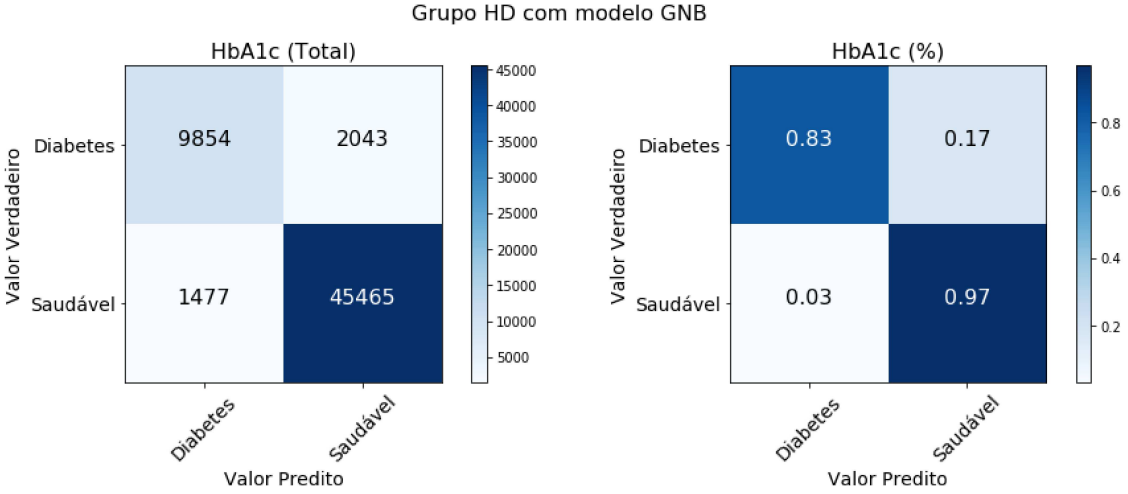


Figura 7.10 - Matriz de confusão do modelo de classificação RF para o grupo HD.

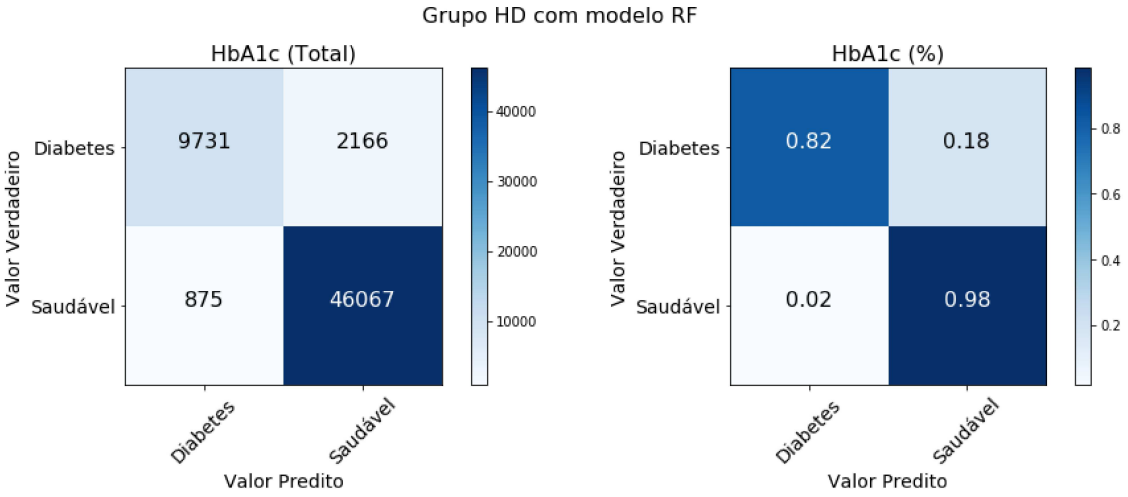


Figura 7.11 - Matriz de confusão do modelo de classificação ANN para o grupo HD.

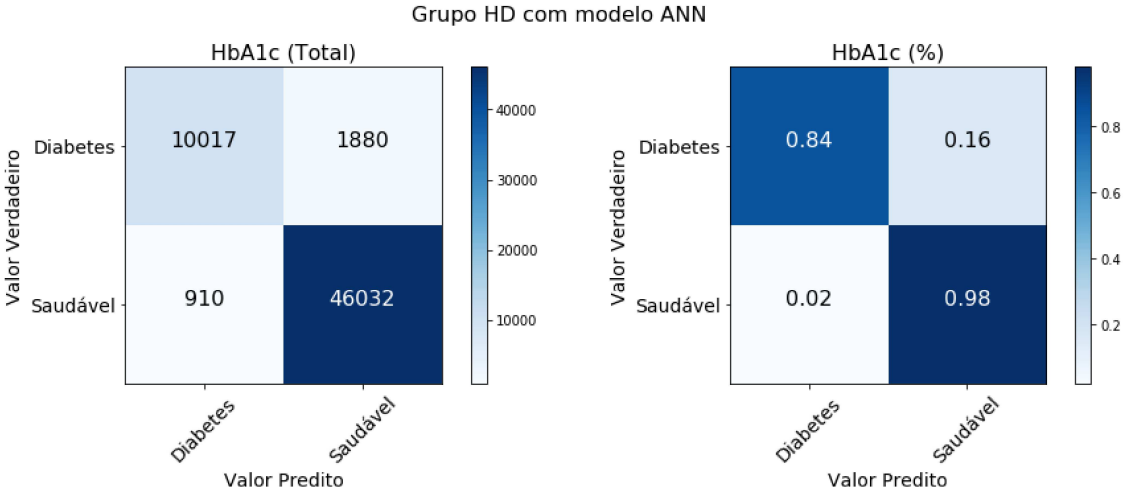


Figura 7.12 - Área sob a curva para o gráfico Precision-Recall tendo com alvo a classe Diabetes no grupo HD.

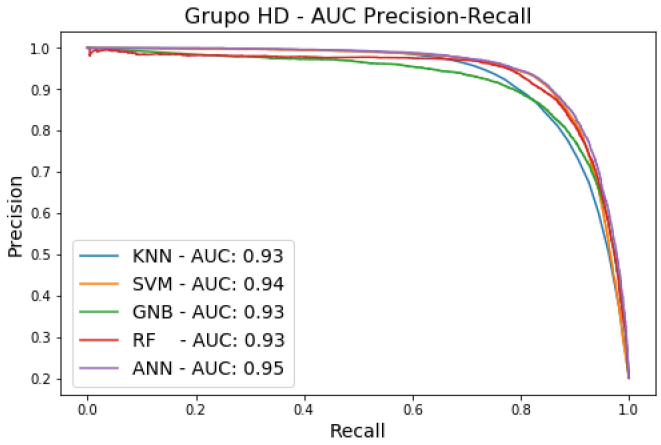


Tabela 7.2 – Métrica de avaliação dos modelos de classificação para o grupo HD.

Classes **Métricas**	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	93,5	95,3	93,9	94,8	95,2
Acurácia Teste	93,3	95,0	94,0	94,8	95,2
AUC Precision-Recall	**93,0**	**94,0**	**93,0**	**93,0**	**95,0**
Diabetes					
Sensibilidade ou Recall (SN)	**69,5**	**80,3**	**82,8**	**81,8**	**84,2**
Especificidade (SP)	99,3	98,8	96,9	98,1	98,1
Precisão (PR) ou Valor Preditivo Positivo	**96,4**	**94,3**	87,0	**91,7**	**91,7**
Valor Preditivo Negativo (NPV)	92,8	95,2	**95,7**	95,5	96,1
Escore-F1	**80,7**	**86,7**	84,8	**86,5**	**87,8**
Saudável					
Sensibilidade ou Recall (SN)	99,3	98,8	96,9	98,1	98,1
Especificidade (SP)	69,5	80,3	82,8	81,8	84,2
Precisão (PR) ou Valor Preditivo Positivo	92,8	95,3	95,7	95,5	96,1
Valor Preditivo Negativo (NPV)	96,4	94,3	87,0	91,7	91,7
Escore-F1	95,9	96,9	96,3	96,8	97,1

7.1.3 Grupo PD – Pré-diabetes e Diabetes

Figura 7.13 – Matriz de confusão do modelo de classificação KNN para o grupo PD.

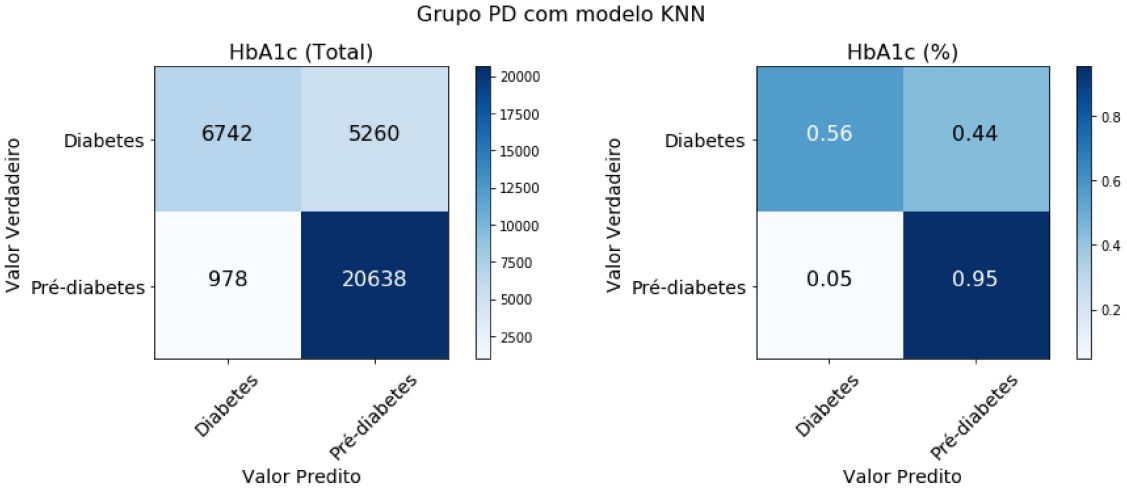


Figura 7.14 - Matriz de confusão do modelo de classificação SVM para o grupo PD.

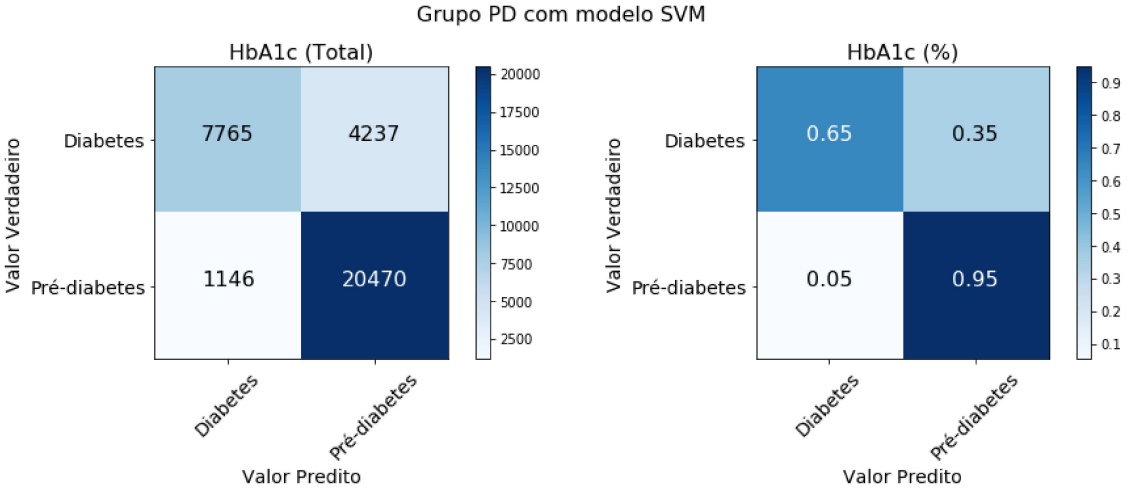


Figura 7.15 - Matriz de confusão do modelo de classificação GNB para o grupo PD.

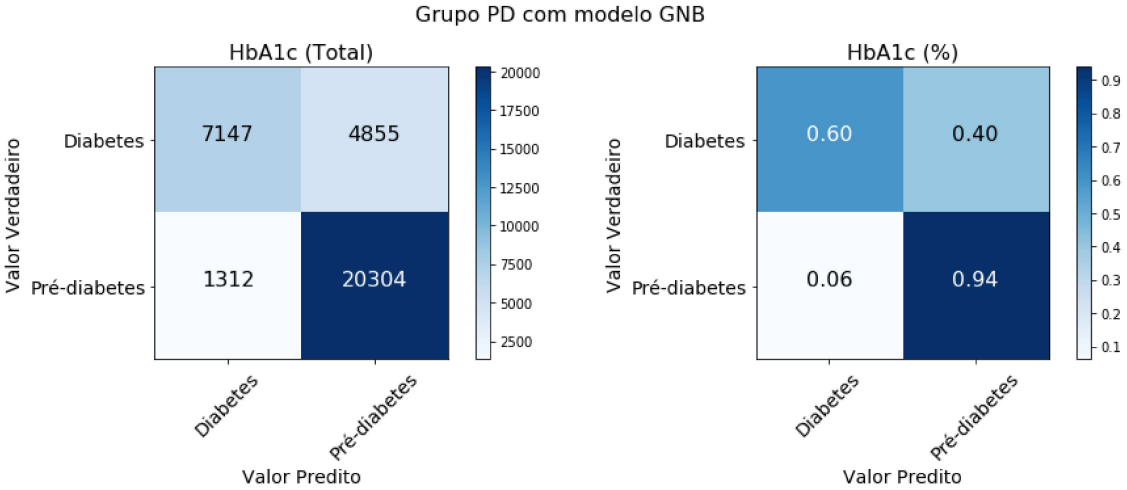


Figura 7.16 - Matriz de confusão do modelo de classificação RF para o grupo PD.

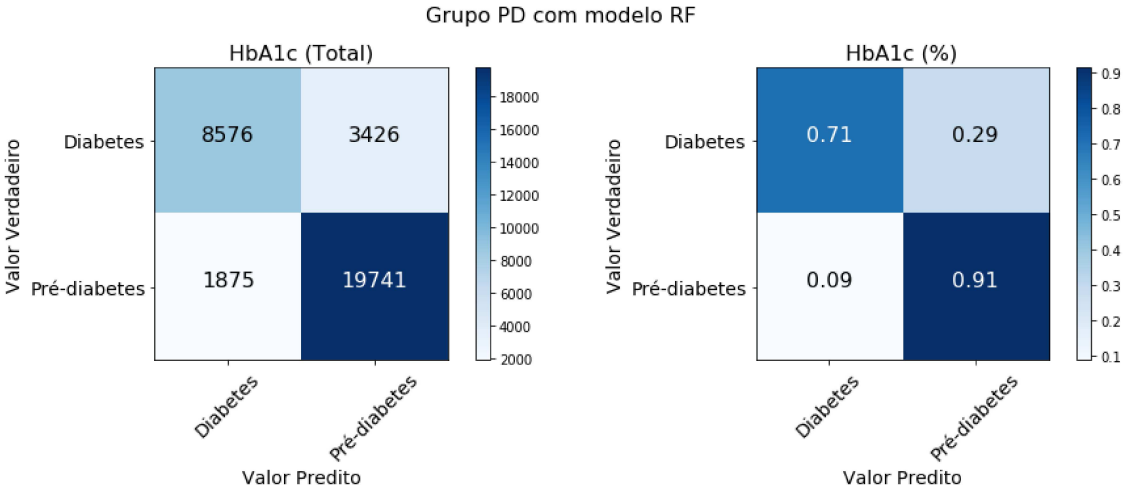


Figura 7.17 - Matriz de confusão do modelo de classificação ANN para o grupo PD.

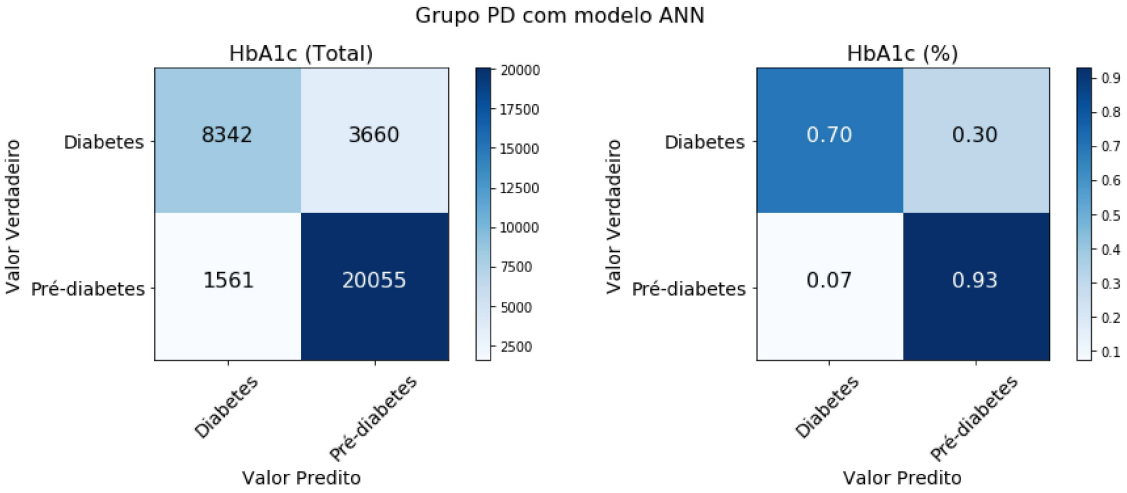


Figura 7.18 - Área sob a curva para o gráfico Precision-Recall tendo com alvo a classe Diabetes no grupo PD.

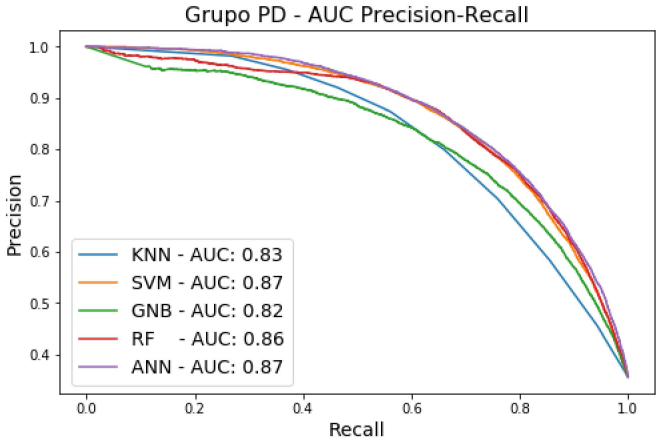


Tabela 7.3 – Métrica de avaliação dos modelos de classificação para o grupo PD.

Classes **Métricas**	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	84,1	84,8	81,9	84,5	84,8
Acurácia Teste	81,4	83,4	81,6	84,2	84,5
AUC Precision-Recall	**82,7**	**86,6**	**82,5**	**86,0**	**87,2**
Diabetes					
Sensibilidade ou Recall (SN)	**56,2**	**64,7**	**59,5**	**71,5**	**69,5**
Especificidade (SP)	95,5	94,7	93,9	91,3	92,8
Precisão (PR) ou Valor Preditivo Positivo	**87,3**	**87,1**	**84,5**	**82,1**	**84,2**
Valor Preditivo Negativo (NPV)	79,7	82,9	80,7	85,2	84,6
Escore-F1	**68,4**	**74,3**	**69,9**	**76,4**	**76,2**
Pré-diabetes					
Sensibilidade ou Recall (SN)	95,5	94,7	93,9	91,3	92,8
Especificidade (SP)	56,2	64,7	59,5	71,5	69,5
Precisão (PR) ou Valor Preditivo Positivo	79,7	82,9	80,7	85,2	84,6
Valor Preditivo Negativo (NPV)	87,3	87,1	84,5	82,1	84,2
Escore-F1	86,9	88,4	86,8	88,2	88,5

7.1.4 Grupo HN – Saudável e Não Saudável

Figura 7.19 – Matriz de confusão do modelo de classificação KNN para o grupo HN.

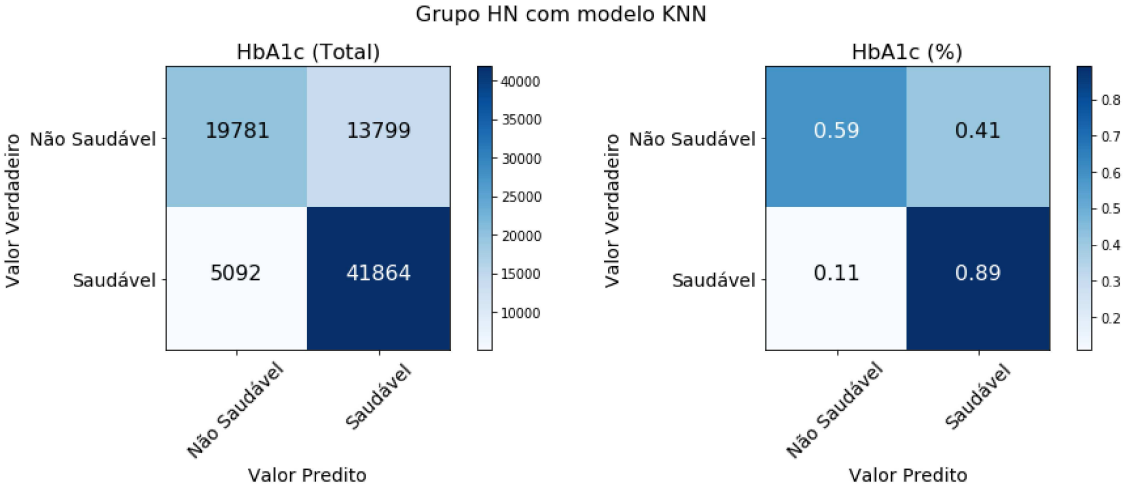


Figura 7.20 - Matriz de confusão do modelo de classificação SVM para o grupo HN.

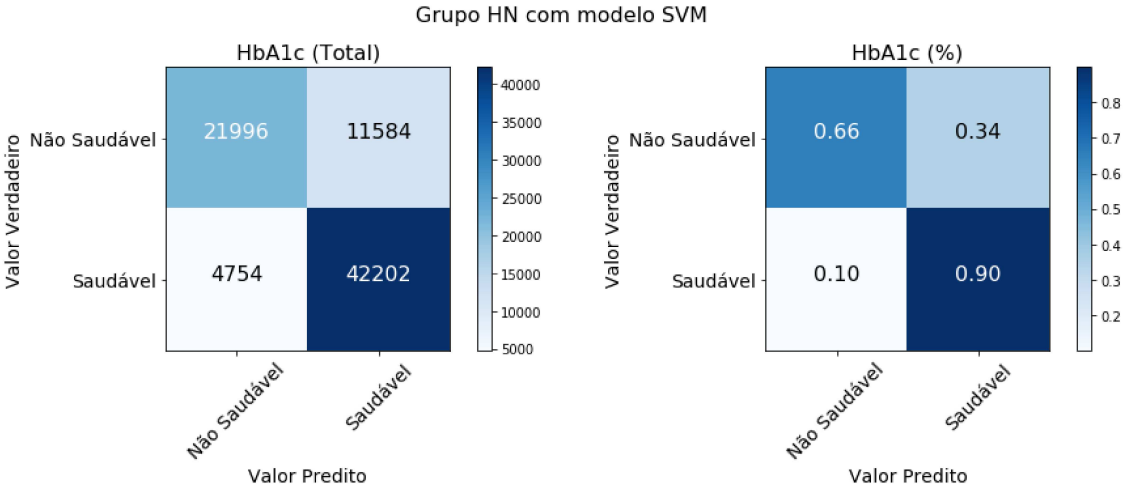


Figura 7.21 - Matriz de confusão do modelo de classificação GNB para o grupo HN.

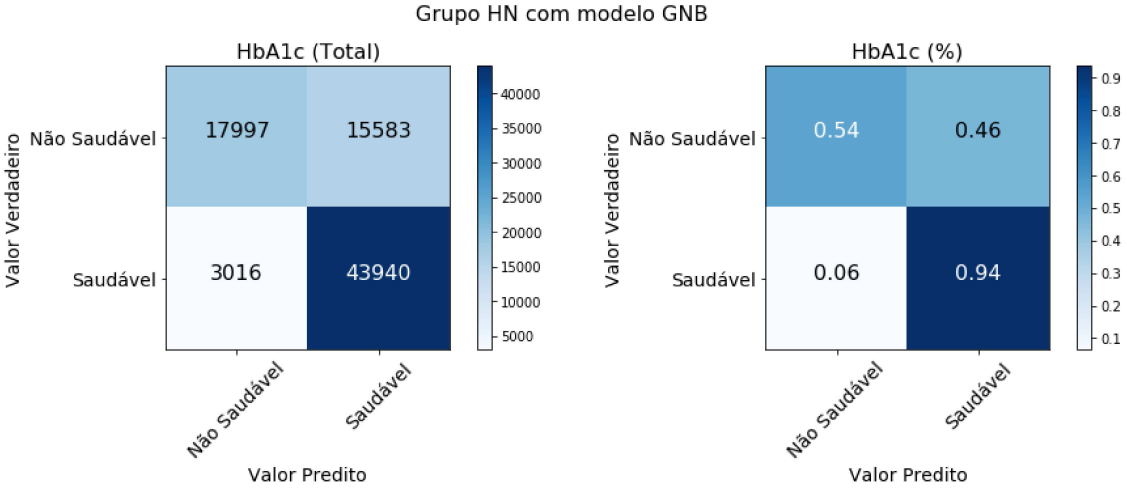


Figura 7.22 - Matriz de confusão do modelo de classificação RF para o grupo HN.

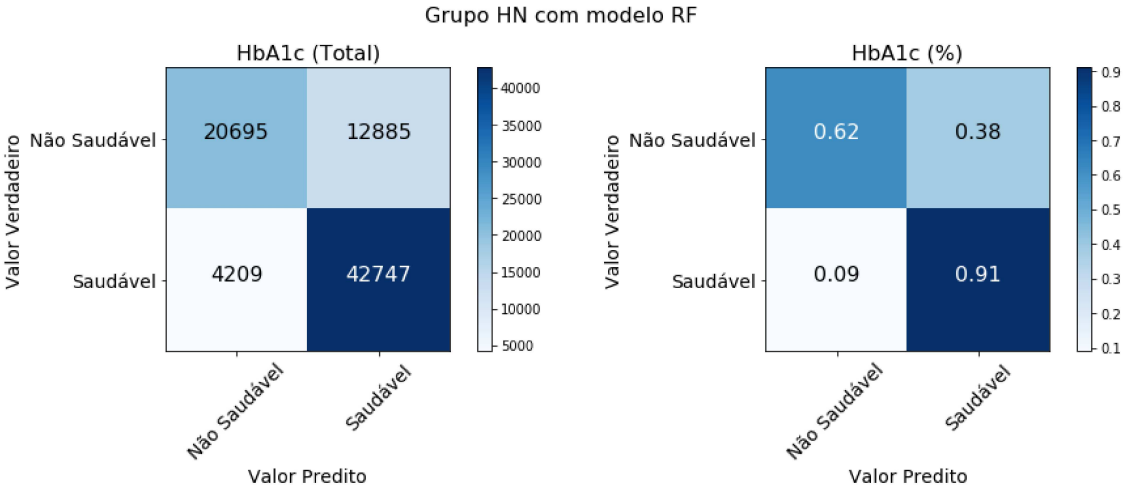


Figura 7.23 - Matriz de confusão do modelo de classificação ANN para o grupo HN.

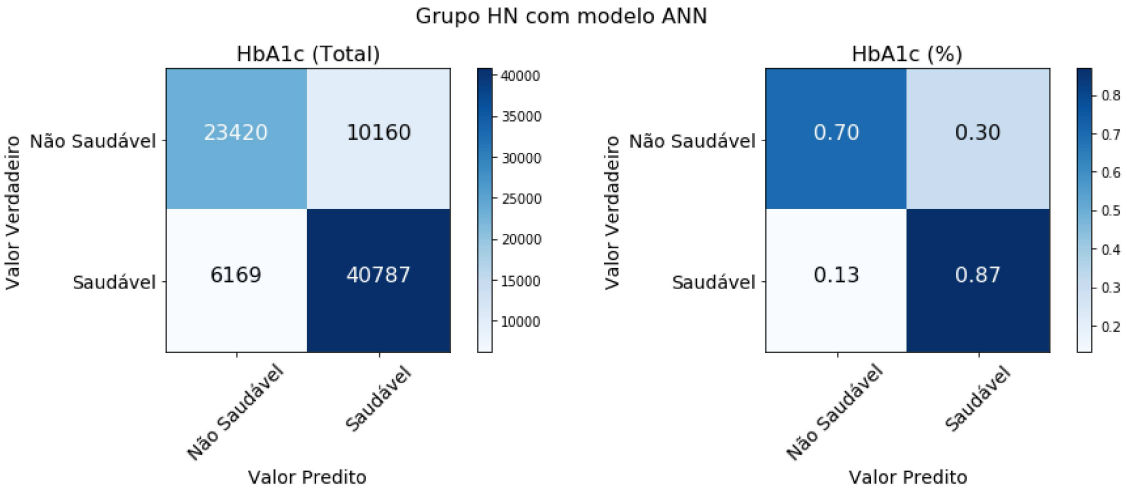


Figura 7.24 - Área sob a curva para o gráfico Precision-Recall tendo com alvo a classe Não Saudável no grupo HN.

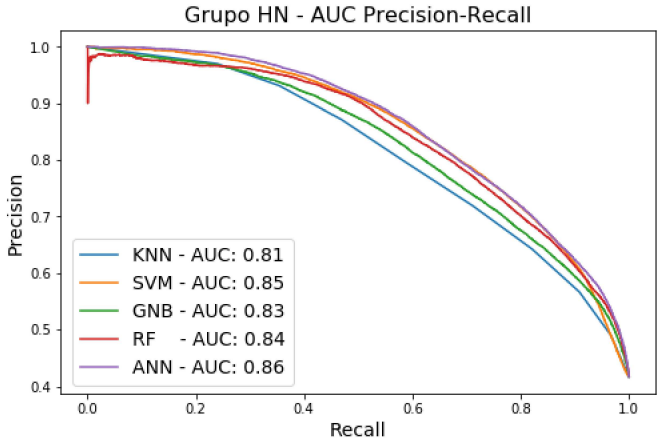


Tabela 7.4 – Métrica de avaliação dos modelos de classificação para o grupo HN.

Classes **Métricas**	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	81,2	79,9	76,7	78,7	79,5
Acurácia Teste	76,5	79,7	76,9	78,8	79,7
AUC Precision-Recall	**81,5**	**85,2**	**82,8**	**84,3**	**85,8**
Não Saudável (Pré-diabetes e Diabetes)					
Sensibilidade ou Recall (SN)	**58,9**	**65,5**	**53,6**	**61,6**	**69,7**
Especificidade (SP)	89,2	89,9	93,6	91,0	86,9
Precisão (PR) ou Valor Preditivo Positivo	**79,5**	**82,2**	**85,6**	**83,1**	**79,2**
Valor Preditivo Negativo (NPV)	75,2	78,5	73,8	76,8	80,1
Escore-F1	**67,7**	**72,9**	**65,9**	**70,8**	**74,2**
Saudável					
Sensibilidade ou Recall (SN)	89,2	89,9	93,6	91,0	86,9
Especificidade (SP)	58,9	65,5	53,6	61,6	69,7
Precisão (PR) ou Valor Preditivo Positivo	75,2	78,5	73,8	76,8	80,1
Valor Preditivo Negativo (NPV)	79,5	82,2	85,6	83,1	79,2
Escore-F1	81,6	83,8	82,5	83,3	83,3

7.1.5 Grupo ND – Não Diabetes e Diabetes

Figura 7.25 – Matriz de confusão do modelo de classificação KNN para o grupo ND.

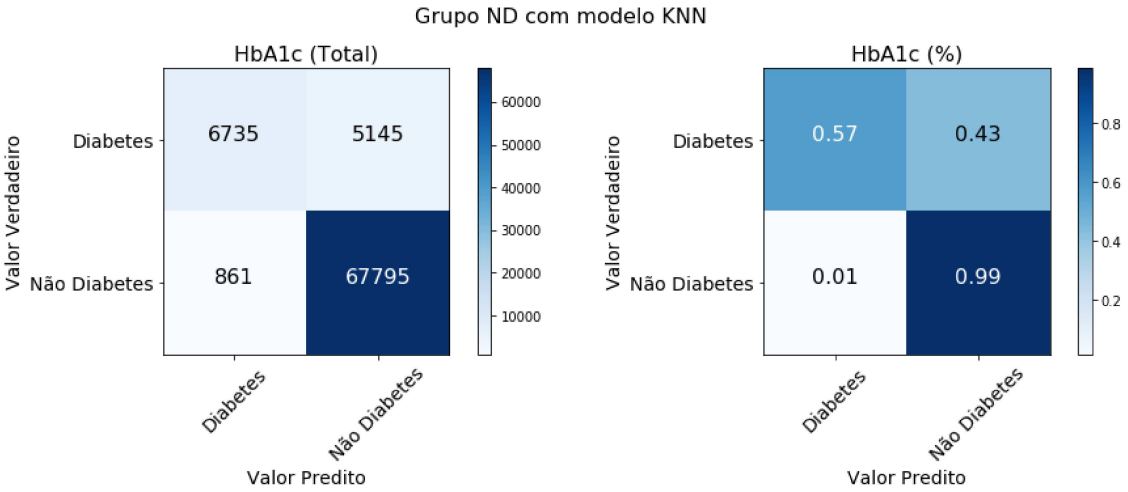


Figura 7.26 - Matriz de confusão do modelo de classificação SVM para o grupo ND.

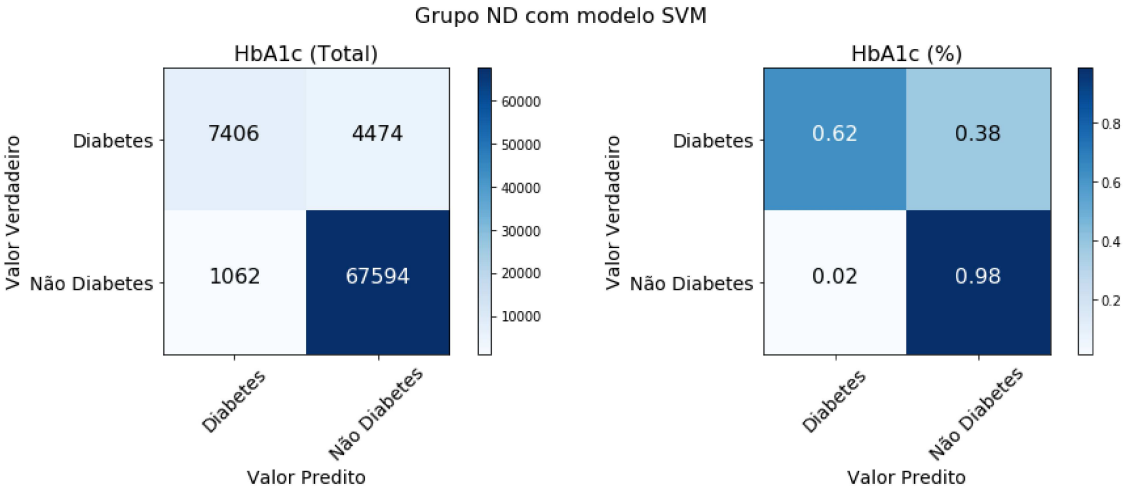


Figura 7.27 - Matriz de confusão do modelo de classificação GNB para o grupo ND.

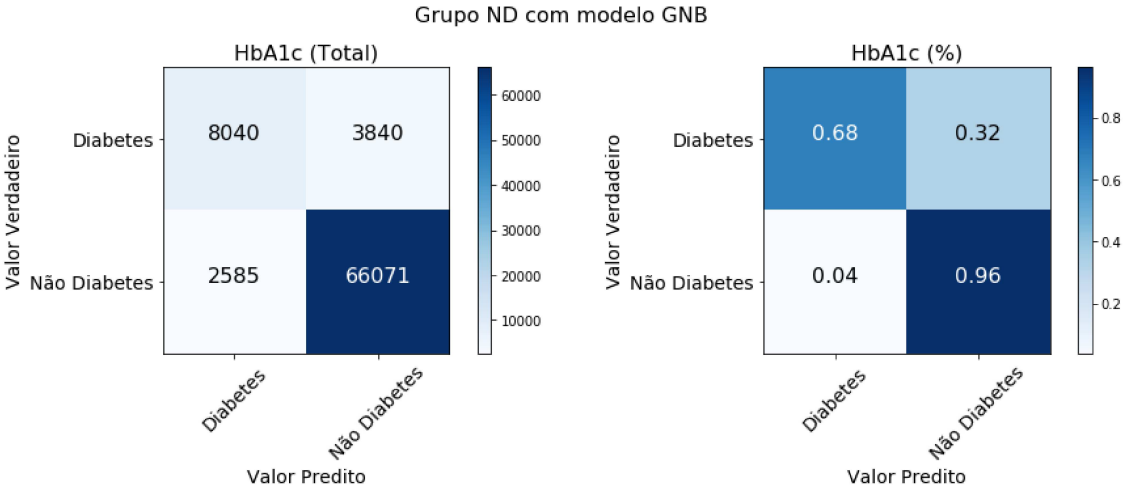


Figura 7.28 - Matriz de confusão do modelo de classificação RF para o grupo ND.

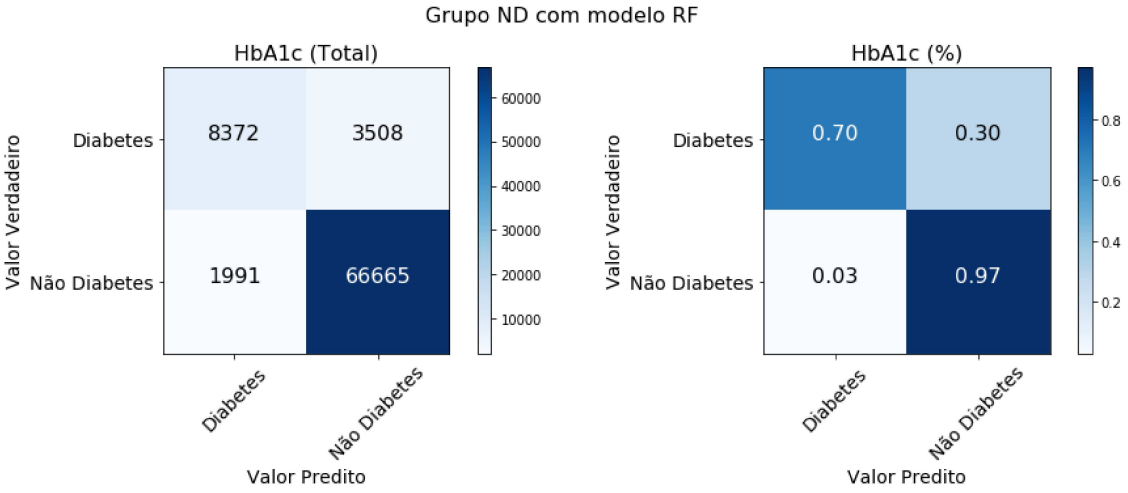


Figura 7.29 - Matriz de confusão do modelo de classificação ANN para o grupo ND.

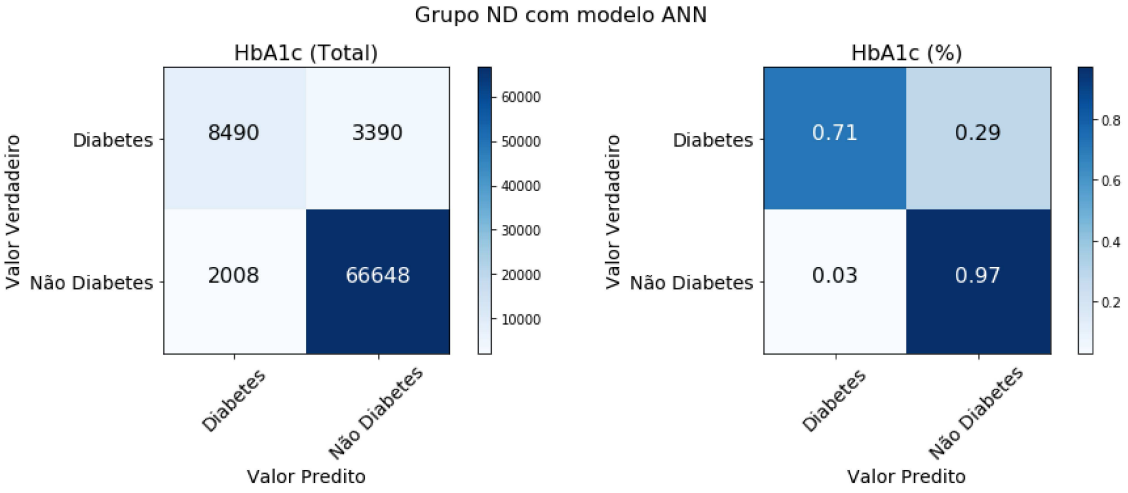


Figura 7.30 - Área sob a curva para o gráfico Precision-Recall tendo como alvo a classe Diabetes no grupo ND.

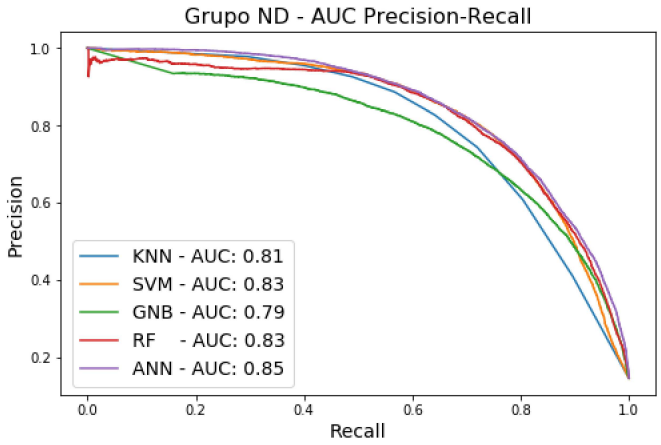


Tabela 7.5 – Métricas de avaliação dos modelos de classificação para o grupo ND.

Classes **Métricas**	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	93,3	93,3	91,9	93,2	93,3
Acurácia Teste	92,4	93,1	92,0	93,2	93,3
AUC Precision-Recall	**80,6**	**83,5**	**78,6**	**82,8**	**81,0**
Diabetes					
Sensibilidade ou Recall (SN)	**56,7**	**62,3**	**67,7**	**70,5**	**71,5**
Especificidade (SP)	98,7	98,5	96,2	97,1	97,1
Precisão (PR) ou Valor Preditivo Positivo	**88,7**	**87,5**	**75,7**	**80,8**	**80,9**
Valor Preditivo Negativo (NPV)	92,9	93,8	94,5	95,9	95,2
Escore-F1	**69,2**	**72,8**	**71,5**	**75,3**	**75,9**
Não Diabetes (Saudável e Pré-diabetes)					
Sensibilidade ou Recall (SN)	98,7	98,5	96,2	97,1	97,1
Especificidade (SP)	56,7	62,3	67,7	70,5	71,5
Precisão (PR) ou Valor Preditivo Positivo	92,9	93,8	94,5	95,0	95,2
Valor Preditivo Negativo (NPV)	88,7	87,5	75,7	80,8	80,9
Escore-F1	95,8	96,1	95,4	96,0	96,1

7.1.6 Grupo HPD – Saudável, Pré-diabete e Diabetes

Figura 7.31 – Matriz de confusão do modelo de classificação KNN para o grupo HPD.

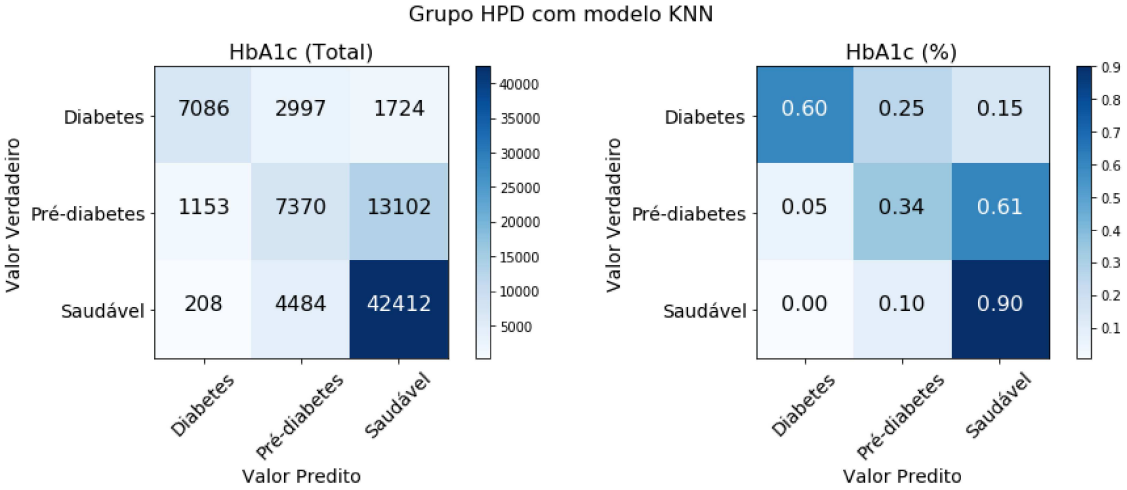


Figura 7.32 - Matriz de confusão do modelo de classificação SVM para o grupo HPD.

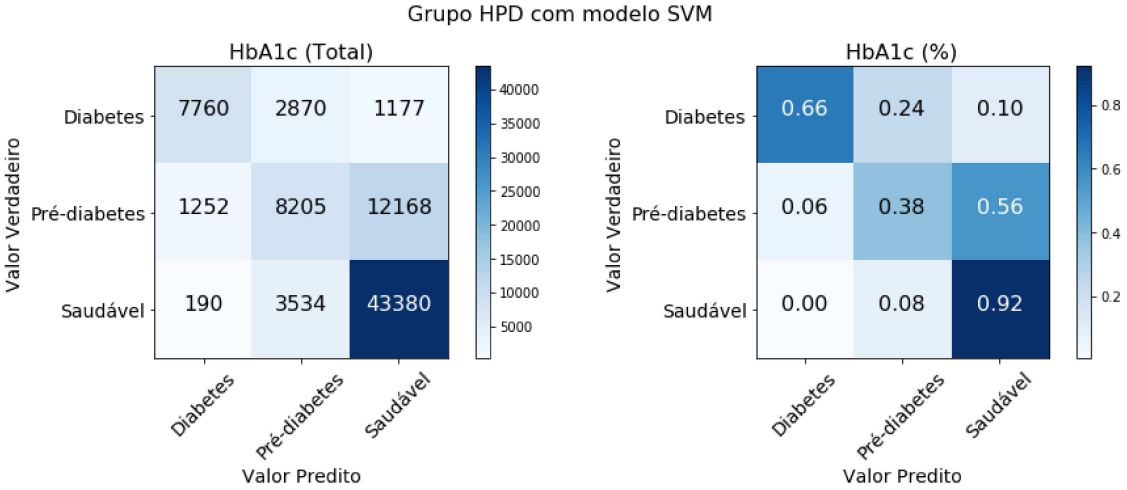


Figura 7.33 - Matriz de confusão do modelo de classificação NB para o grupo HPD.

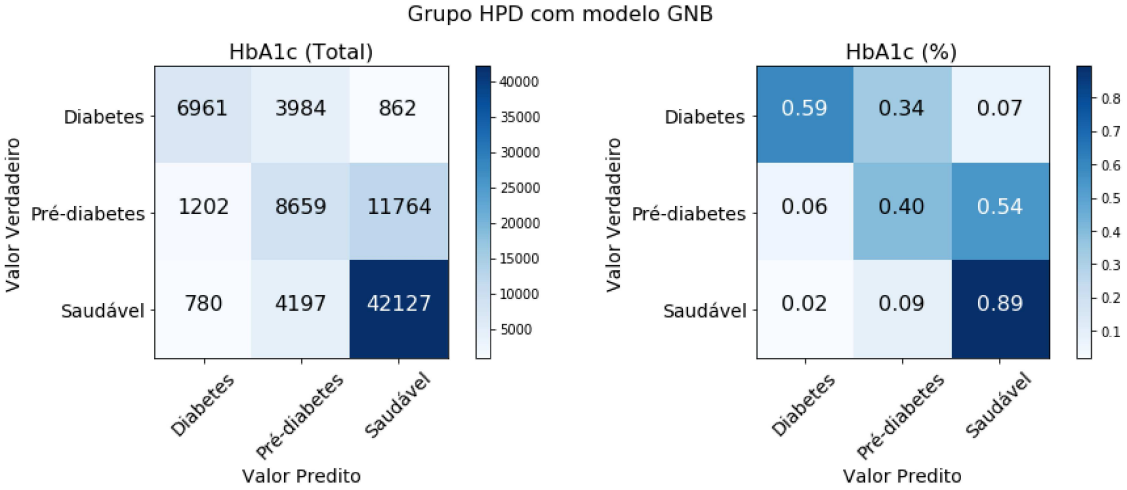


Figura 7.34 - Matriz de confusão do models de classificação RF para o grupo HPD.

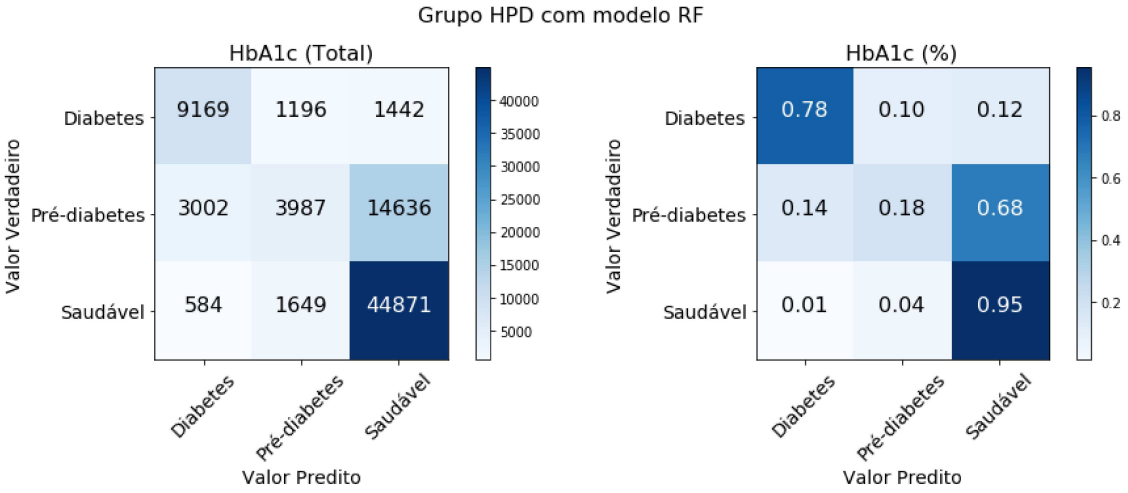


Figura 7.35 - Matriz de confusão do modelo ANN para o grupo HPD.

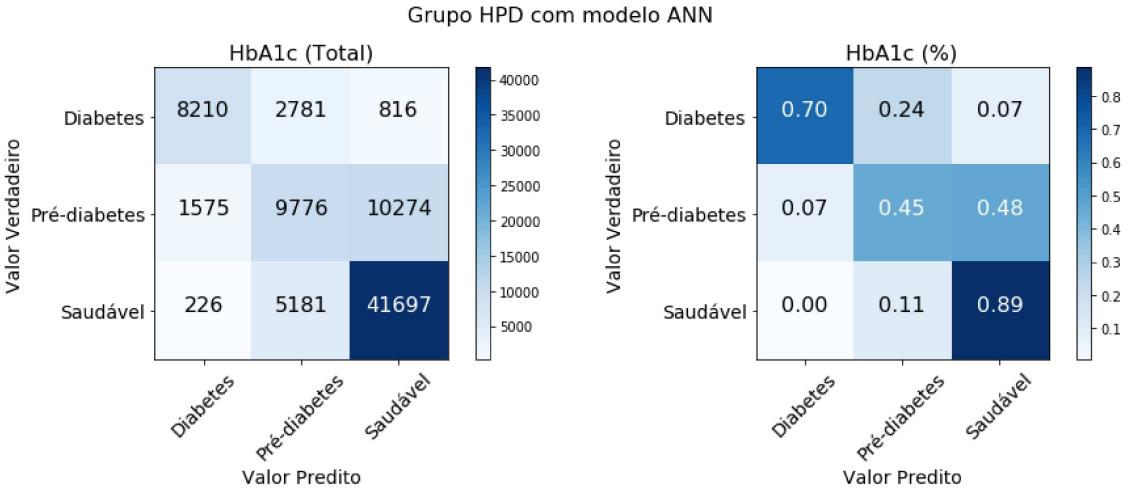


Figura 7.36 - Área sob a curva para o gráfico Precision-Recall tendo como alvo a classe Diabetes no grupo HPD.

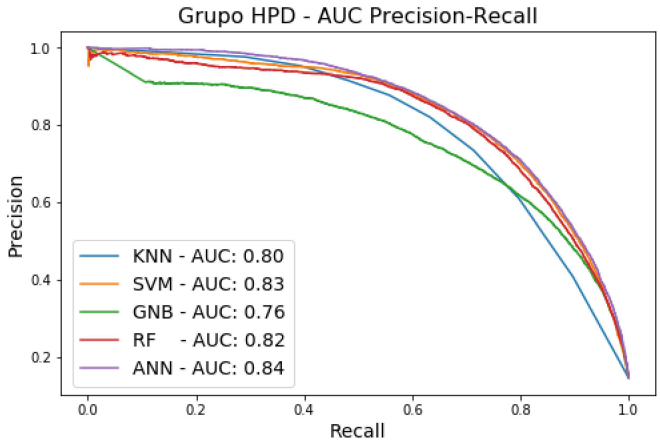


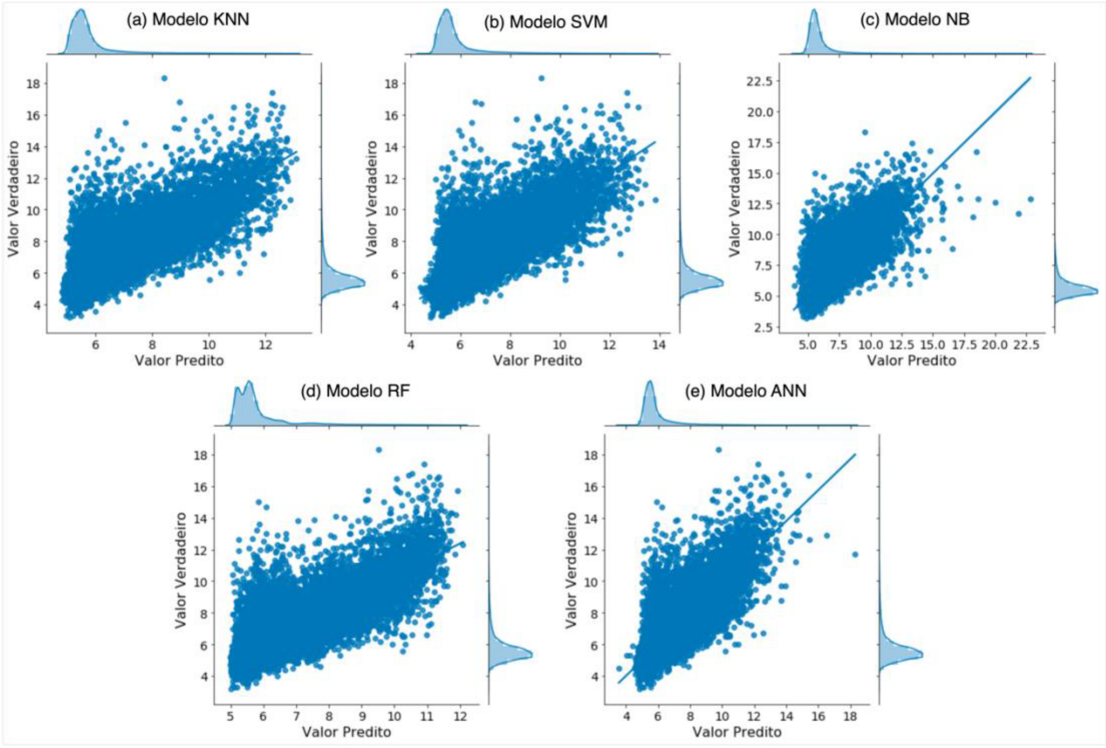
Tabela 7.6 – Métricas de avaliação dos modelos de classificação para o grupo HPD.

Classes / **Métricas**	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Acurácia Treinamento	76,2	74,9	71,9	72,2	74,5
Acurácia Teste	70,6	73,7	71,7	72,0	74,1
AUC Precision-Recall	**79,9**	**83,4**	**76,3**	**82,3**	**84,5**
Diabetes					
Sensibilidade ou Recall (SN)	**60,0**	**65,7**	**59,0**	**65,7**	**69,5**
Especificidade (SP)	98,0	97,9	97,1	97.9	97,4
Precisão (PR) ou Valor Preditivo Positivo	**83,9**	**84,3**	**77,8**	**84,3**	**82,0**
Valor Preditivo Negativo (NPV)	93,5	94,3	93,2	94,3	94,9
Escore-F1	**70,0**	**73,9**	**67,1**	**73,9**	**75,3**
Pré-diabetes					
Sensibilidade ou Recall (SN)	34,1	37,9	40,0	37,9	45,2
Especificidade (SP)	87,3	89,1	86,1	89,1	86,5
Precisão (PR) ou Valor Preditivo Positivo	49,6	56,2	51,4	56,2	55,1
Valor Preditivo Negativo (NPV)	78,3	79,6	79,6	79,6	81,1
Escore-F1	40,4	45,3	45,0	45,3	49,7
Saudável					
Sensibilidade ou Recall (SN)	90,0	92,1	89,4	92,1	88,5
Especificidade (SP)	55,7	60,1	62,2	60,1	66,8
Precisão (PR) ou Valor Preditivo Positivo	74,1	76,5	76,9	76,5	79,0
Valor Preditivo Negativo (NPV)	79,9	84,4	80,7	84,4	80,5
Escore-F1	81,3	83,6	82,7	83,6	83,5

7.2. MODELOS DE REGRESSÃO

Os mesmos conjuntos de modelos formam testados para regressão, claro que utilizando as respectivas funções e parâmetros para cada caso. Na Figura 7.37 é apresentado o gráfico de dispersão comparando os valores preditos com os valores verdadeiros para cada um dos modelos testados na predição do HbA1c.

Figura 7.37 - Gráfico de dispersão comparando os valores preditos com os valores verdadeiros para os modelos (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN.



De modo geral os modelos de regressão conseguiram predizer os valores de HbA1c, como pode ser observado no gráfico da Figura 7.37. No entanto, buscando compreender melhor o poder de cada modelo, é apresentado na Tabela 7.7 os erros e a correlação para cada um dos modelos testados

Tabela 7.7 – Comparação da métricas de avaliação para os modelos de regressão.

Métricas	**KNN**	**SVM**	**NB**	**RF**	**ANN**
Erro Médio Absoluto (MAE)	**0,38**	**0,37**	**0,39**	**0,39**	**0,36**
Erro Médio Quadrático (MSE)	0,63	0,62	0,64	0,62	0,60
Raiz do Erro Médio Quadrático (RMSE)	0,39	0,38	0,41	0,38	0,35
Correlação (Pearson)	0,84	0,84	0,83	0,84	0,85

7.2.1 Classificação da Regressão

Utilizando os valores preditos dos modelos de regressão, foi realizada a classificação dos resultados obtidos para posteriormente serem comparados com os resultados dos modelos de classificação. Este método foi adotado para as classes dos grupos HPD, HN e ND.

Figura 7.38 – Matriz de confusão da classificação dos resultados da regressão para os modelos (a) KNNr, (b) SVMr, (c) NBr, (d) RFr e (e) ANNr. O resultado foi classificado de acordo com o grupo HPD – possuindo as classes Saudável, Pré-diabetes e Diabetes.

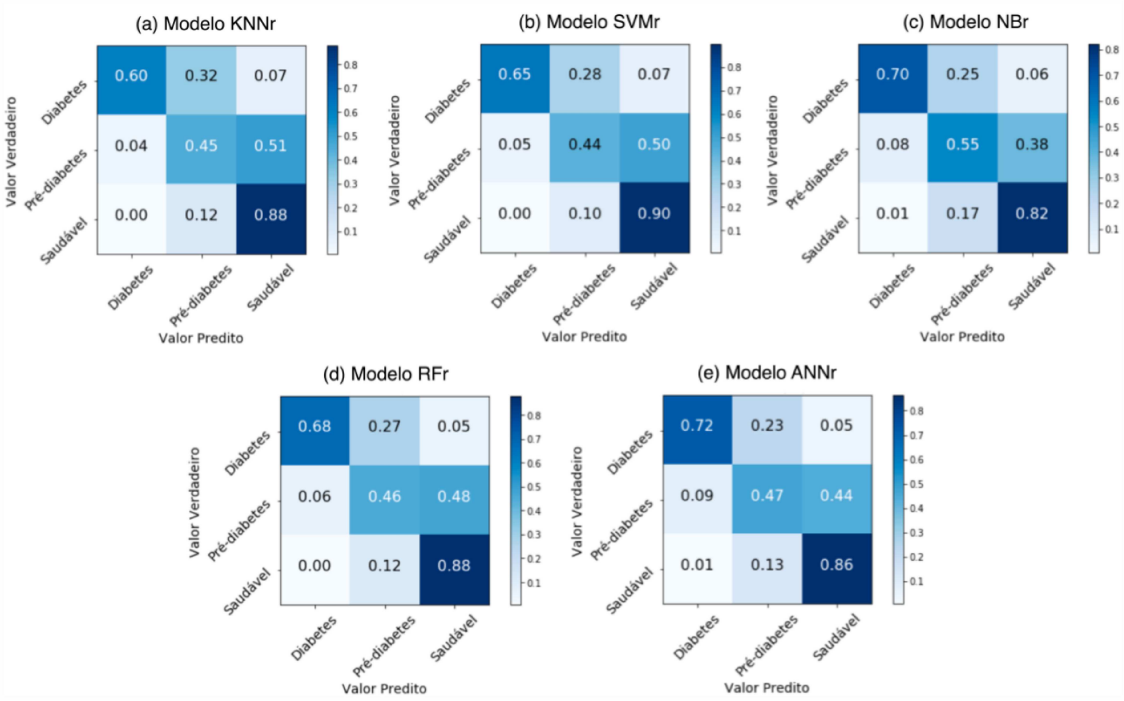


Tabela 7.8 - Métricas de avaliação dos modelos para classificação após regressão.

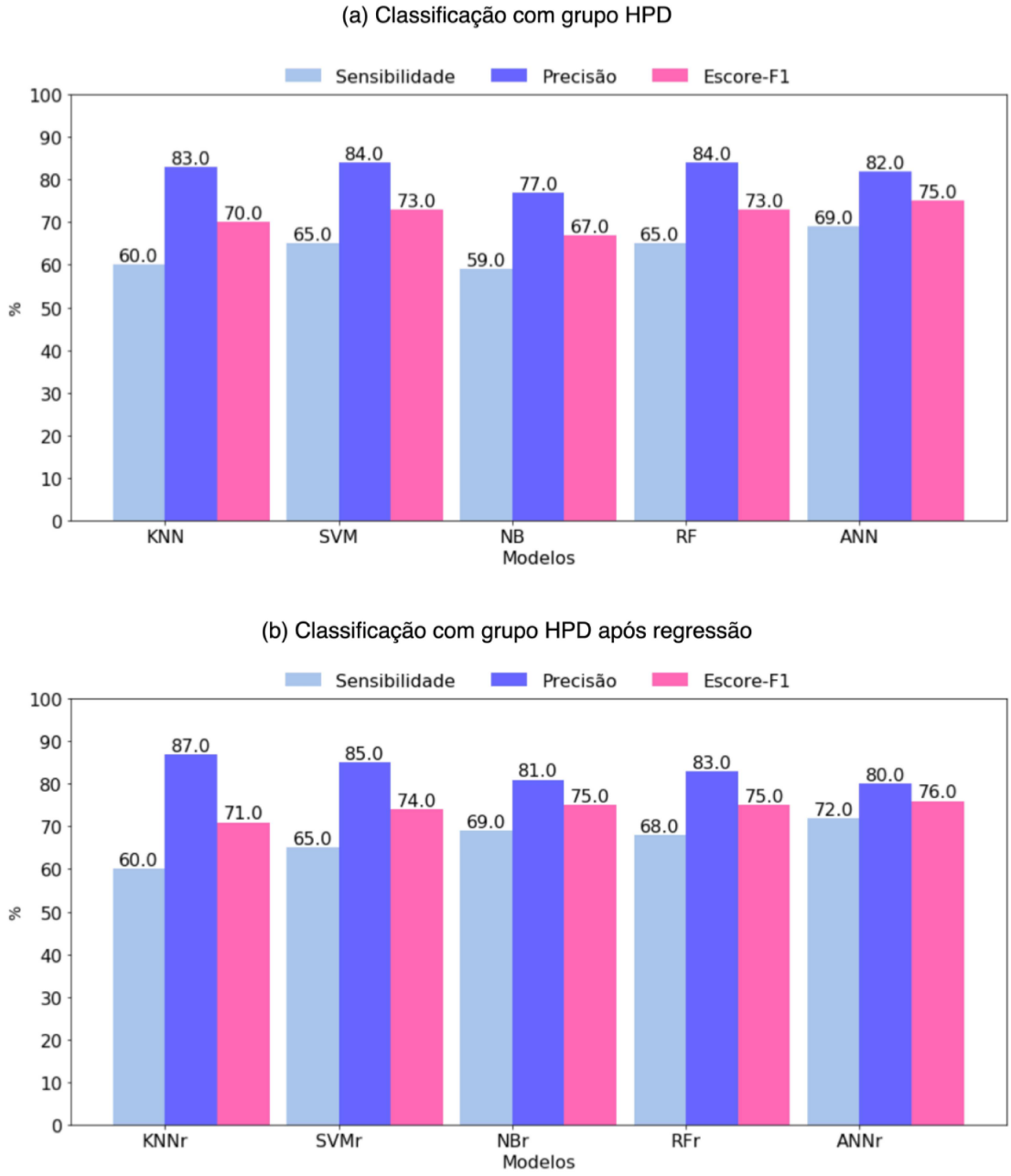
Classes Métricas	KNNr	SVMr	NBr	RFr	ANNr
Diabetes					
Sensibilidade ou Recall (SN)	**60,4**	**65,4**	**69,8**	**68,3**	**72,3**
Especificidade (SP)	98,5	98,1	97,3	97,7	96,9
Precisão (PR) ou Valor Preditivo Positivo	**87,4**	**85,7**	**81,5**	**83,6**	**80,3**
Valor Preditivo Negativo (NPV)	93,5	94,2	94,9	94,7	95,3
Escore-F1	**71,5**	**74,2**	**75,2**	**75,2**	**76,1**
Pré-diabetes					
Sensibilidade ou Recall (SN)	45,1	44,3	54,9	45,8	47,3
Especificidade (SP)	83,9	86,4	81,1	85,0	84,8
Precisão (PR) ou Valor Preditivo Positivo	51,0	54,8	51,9	53,2	53,5
Valor Preditivo Negativo (NPV)	80,5	80,7	82,9	80,9	81,3
Escore-F1	47,9	49,0	53,4	49,2	50,2
Saudável					
Sensibilidade ou Recall (SN)	87,7	89,8	82,0	87,6	86,2
Especificidade (SP)	64,6	65,1	73,7	67,4	69,7
Precisão (PR) ou Valor Preditivo Positivo	77,5	78,1	81,3	78,9	79,8
Valor Preditivo Negativo (NPV)	79,1	82,1	74,7	79,7	78,4
Escore-F1	82,3	83,6	81,6	83,0	82,9

Como a classificação é feita sobre os valores da regressão, não se faz necessário repetir as métricas da tabela para os demais grupos de classificação, uma vez que os resultados seriam os mesmos. No entanto é interessante a comparação dos modelos de classificação com os de regressão.

7.2.2 Comparação dos resultados

Na classificação dos resultados e triagem de indivíduos doentes, a predição dos valores de HbA1c e posterior classificação dos mesmos, também se apresenta como um processo bastante interessante. Buscando analisar melhor essa possibilidade, na Figura 7.39 é apresentado os resultados das métrica Sensibilidade, Precisão e Escore-F1 referentes aos modelos de classificação do HbA1c em comparação a classificação após a regressão.

Figura 7.39 – Comparação das métricas de Sensibilidade, Precisão e Escore-F1 com os modelos de classificação (KNN, SVM, NB, RF, ANN) e classificação após regressão (KNNr, SVMr, NBr, RFr, ANNr), sobre os grupos HN e ND.

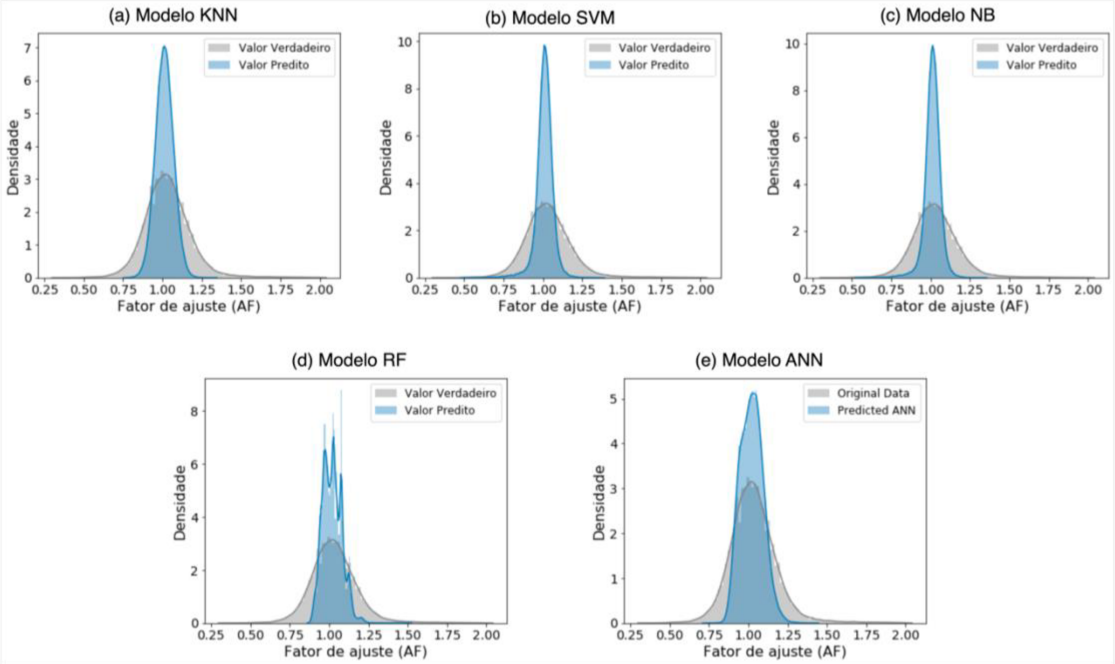


7.3. IDENTIFICAÇÃO DE FALSOS NEGATIVOS

Com os resultados obtidos na etapa de predição de novos exames, definiu-se os caminhos a serem tomados para identificação de possíveis falsos negativos no exame de FPG. Assim, utilizou-se os modelos de machine learning para predizer o fator de ajuste (pAF).

Como o alvo dos modelos desta etapa é o fator de ajustes (AF), na Figura 7.40 é apresentado o gráfico com a distribuição dos valores de fator de ajuste preditos (pAF) juntamente com os valores originais para cada um dos modelos testados: (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN. Nos gráficos é possível observar o comportamento do resultado da predição em relação ao valor original. O modelo cujo resultados mais se aproximaram dos valores verdadeiros foi o da rede neural artificial (ANN).

Figura 7.40 - Comparação do fator de ajuste predito (pAF) com o fator de ajuste original (AF) calculado



Buscando uma melhor análise dos modelos de predição, na Tabela 7.9, além das métricas de erros, também são aprestados dados estatísticos dos valores preditos pelos modelos de machine learning. Estes dados são importantes para conhecer o desempenho de predição dos modelos utilizados. Além dos erros, que apresenta de forma direta a capacidade de predição, os dados estatísticos nos mostram o quão próximo dos dados reais os modelos podem chegar.

Tabela 7.9 – Comparação estatística do valor original do fator de ajuste com os valores preditos pelos modelos de machine learning.

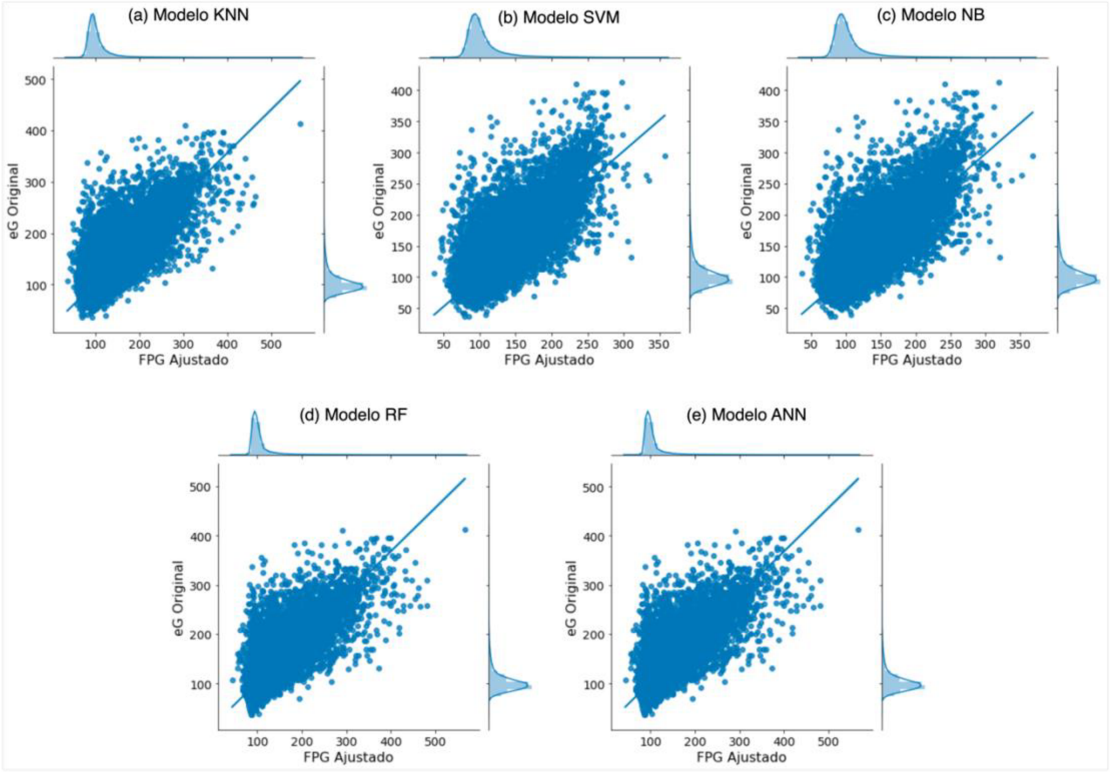
Métrica	AF Original	KNN	SVM	NB	RF	ANN
Mean	**1.034**	**1.013**	**1.011**	**1.012**	**1.021**	**1.031**
SD	**0.178**	**0.057**	**0.053**	**0.051**	**0.061**	**0.070**
Min	0.340	0.077	0.527	0.529	0.881	0.741
Max	4.523	1.289	1.370	1.343	1.503	1.433
MAE	**-**	**0.105**	**0.104**	**0.104**	**0.100**	**0.096**
MSE	-	0.028	0.028	0.028	0.026	0.025
RMSE	-	0.169	0.167	0.167	0.161	0.157

Analisando a Tabela 7.9, observamos que SD dos valores preditos ficou abaixo do valor original, provavelmente devido aos outliers existentes nos dados originais. O maior valor (Max) do dado original é 4.523. Já nas predições, esse valor não passa de 1.503 (modelo RF).

Na

Figura 7.41 são apresentados os gráficos de dispersão das amostras de FPG ajustados (aFPG) em relação a glicose estimada (eG) verdadeira, para cada um dos modelos testados, (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN. Nos gráficos é possível observar o comportamento dos modelos após o cálculo do FPG ajustado (aFPG). De modo geral, todos os modelos se comportaram de maneira semelhante, apresentando forte relação entre os valores calculados e os originais.

Figura 7.41 - Distribuição dos valores ajustados de FPG em relação aos valores verdadeiros de glicose estimada (eG), para cada um dos modelos testados, (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN



Já na são apresentadas as métricas de erros e os dados estatísticos dos valores de FPG ajustados, calculados com os valores de AF preditos pelos modelos de machine learning. De acordo com os valores apresentados na Tabela 7.10, o modelo ANN teve o menor erro entre os modelos testados. Da mesma forma, o valor médio (Mean) de FPG ajustado, calculados com o modelo ANN, foi o que mais se aproximou do dado original.

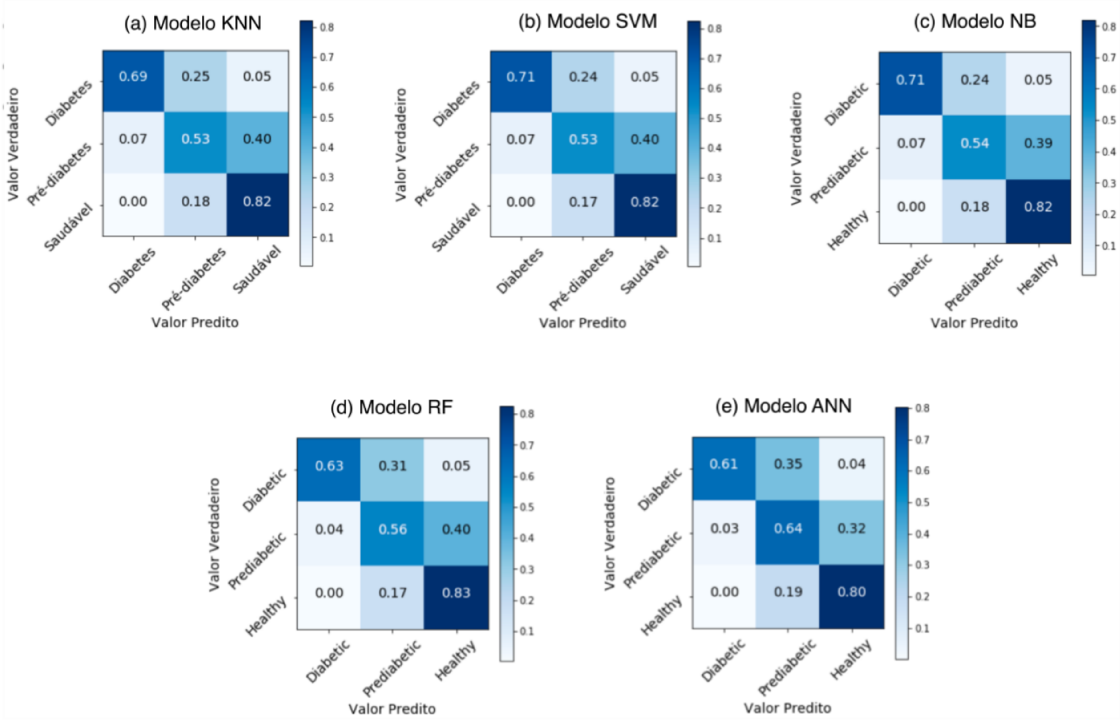
Tabela 7.10 - Comparação estatística do valor original do da glicose estimada com os valores preditos pelos modelos de machine learning.

Métrica	eG original	KNN	SVM	NB	RF	ANN
Mean	**107.838**	**106.089**	**105.405**	**105.533**	**106.383**	**107.390**
SD	31.548	30.693	26.773	26.650	29.223	29.338
Min	36.330	35.918	37.134	36.882	43.639	45.381
Max	412.870	560.877	373.248	368.393	566.350	574.564
MAE	-	**11.385**	**11.169**	**11.152**	**10.943**	**10.517**
MSE	-	347.140	324.019	323.151	337.668	320.934
RMSE	-	18.632	18.000	17.976	18.376	17.914

Após o calculo dos valores de FPG ajustado (aFPG), calculados com os valores de AF preditos pelos modelos de machine learning. Os valores calculados foram classificados para o diagnosticado de Diabetes.

Para avaliar a classificação dos valores de aFPG para diagnóstico de DM, utilizamos a matriz de confusão. Na Figura 7.38 são apresentadas as matrizes de confusão para cada um dos modelos testados, (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN.

Figura 7.42 - Matriz de confusão da classificação de diagnóstico de diabetes com os valores de FPG ajustado, para cada um dos modelos testados, (a) KNN, (b) SVM, (c) NB, (d) RF e (e) ANN.



Com a matriz de confusão é apresentado mais detalhadamente as características de cada modelo diante da classificação no diagnostico de Diabetes. De modo geral, assim como nos modelos anteriores, tem-se maior dificuldade na classificação dos pacientes com pré-diabetes.

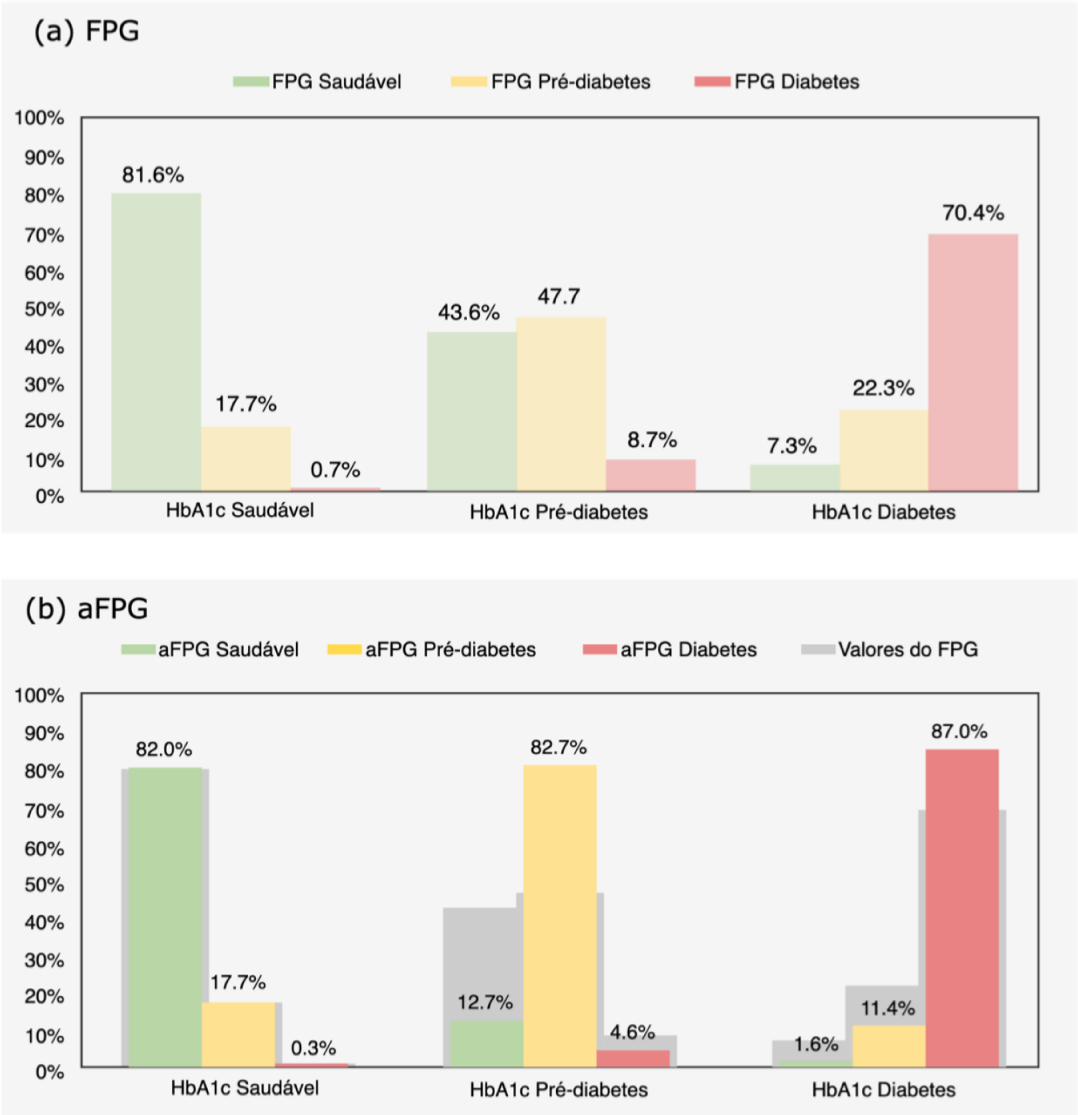
Na Tabela 7.11 é apresentado os detalhes referentes ao desempenho dos modelos na classificação dos valores de FPG ajustado (aFPG). Diante do conjunto de informações extraída da matriz de confusão, buscamos modelos com maior valor de Sensibilidade (SN) e Precisão (PR). Desta forma focamos em modelos com maior capacidade de identificar pacientes doentes assim como maior certeza dentre os pacientes identificados.

Tabela 7.11 - Métricas de avaliação dos modelos para classificação do FPG Ajustado (aFPG)

Classes Métricas	KNN	SVM	NB	RF	ANN
Diabetes					
Sensibilidade ou Recall (SN)	**69.8%**	**71.2%**	**71.3%**	**63.4%**	**64.0%**
Especificidade (SP)	97.0%	96.9%	96.8%	98.1%	98.2%
Precisão (PR) ou Valor Preditivo Positivo	**80.5%**	**80.1%**	**80.0%**	**85.8%**	**86.1%**
Valor Preditivo Negativo (NPV)	94.8%	95.0%	95.0%	93.8%	93.9%
Escore-F1	**74.8%**	**75.4%**	**75.4%**	**72.9%**	**73.4%**
Pré-diabetes					
Sensibilidade ou Recall (SN)	53.3%	53.0%	53.6%	55.8%	66.3%
Especificidade (SP)	80.8%	81.3%	81.0%	79.4%	75.9%
Precisão (PR) ou Valor Preditivo Positivo	60.4%	60.9%	60.8%	59.8%	60.1%
Valor Preditivo Negativo (NPV)	75.9%	75.9%	76.1%	76.6%	80.4%
Escore-F1	56.6%	56.7%	57.0%	57.7%	63.1%
Healthy					
Sensibilidade ou Recall (SN)	82.1%	82.4%	82.0%	82.5%	78.0%
Especificidade (SP)	70.2%	70.4%	70.8%	70.3%	78.0%
Precisão (PR) ou Valor Preditivo Positivo	73.0%	73.2%	70.3%	73.1%	77.7%
Valor Preditivo Negativo (NPV)	80.0%	80.4%	80.1%	80.4%	78.4%
Escore-F1	77.3%	77.5%	77.4%	77.5%	77.9%

Na Figura 7.43 é apresentado a classificação dos valores de FPG ajustados em comparação com a HbA1c utilizando o modelo de ANN. No gráfico a classificação do diagnóstico de Diabetes com o FPG ajustado (b) possui maior concordância com a classificação do HbA1c em comparação aos valores FPG originais (a).

Figura 7.43 - Comparação da classificação com os valores de FPG original (a) e FPG ajustado (b) em relação ao HbA1c.



Para melhor observar os resultados, na Tabela 7.12 comparamos a concordância do FPG e aFPG em relação a HbA1c, assim como também calculamos o ganho obtido em cada classe de diagnóstico.

Tabela 7.12- Comparação e ganho das porcentagens no diagnóstico de Diabetes com o FPG e aFPG

Classes Métrica	FPG original	FPG ajustado	Ganho
HbA1c Diabetes			
FPG Diabetes	70.4%	87.0%	**16.6%**
FPG Pré- diabetes	22.3%	11.4%	- 10.9%
FPG Healthy	7.3%	1.6%	- 5.7%
HbA1c Pré-diabetes			
FPG Diabetes	8.7%	4.6%	- 4.1%
FPG Pré-diabetes	47.7%	82.7%	**35%**
FPG Healthy	43.6%	12.7%	- 30.9%
HbA1c Saudável			
FPG Diabetes	0.7%	0.3%	- 0.4%
FPG Pré-diabetes	17.7%	17.7%	- 0%
FPG Saudável	81.6%	82.0%	**0.4%**

8. DISCUSSÃO

Nos últimos anos, o uso de testes laboratoriais para prever resultados e apoiar o diagnóstico de diferentes patologias tem sido cada vez mais explorado. Como relatado na Revisão Bibliográfica, vários estudos vem utilizando exames laboratoriais para prever novos resultados e buscar o diagnóstico de doenças que não são alvo do teste, sendo que as publicações mais recentes (ALJAME et al., 2021; CAMPAGNER; CAROBENE; CABITZA, 2021; MYARI; PAPAPETROU; TSAOUSI, 2021; RAHMAN et al., 2021) tiveram como objetivo o apoio ao diagnóstico do SARS-CoV-2.

Na tentativa de auxiliar a previsão de Diabetes Mellitus, vários estudos têm utilizado diferentes conjuntos de dados e técnicas de aprendizado de máquina com bons resultados. Em seu estudo, Zheng et al. (ZHENG et al., 2017) obteve 100% de sensibilidade e precisão acima de 90% em vários modelos usando um conjunto de dados de 300 amostras com diferentes atributos de entrada, incluindo notas de autorrelato do paciente e medicação. Oliveira et al. (OLIVERA et al., 2017b) obteve 68% de sensibilidade e 68% de especificidade após usar um conjunto de dados menor e variáveis categóricas obtidas por meio de entrevistas. Lei et al. (LAI et al., 2019) obteve 71,6% de sensibilidade e 73,4% de especificidade com um conjunto de dados de 13.309 amostras usando características laboratoriais e clínicas.

Esses estudos obtiveram bons resultados, mas utilizaram dados clínicos ou de diagnósticos. Essas informações são geradas através da análise de um médico, ao contrário da maioria dos exames laboratoriais, como o hemograma completo, que segue um processo analítico automatizado sem a intervenção de fatores humanos.

Desta forma, os estudos citados não podem ser diretamente comparados com este trabalho, pois seguem metodologias com características diferentes. Este trabalho utilizou apenas dados quantitativos de exames laboratoriais de rotina para treinar diferentes modelos de classificação e regressão, bem como diferentes arranjos de conjuntos de dados, tendo como objetivo a triagem de indivíduos potencialmente diabéticos assim como a identificação de exames de FPG com resultado falso negativo.

Está metodologia se mostra inédita e capaz de contribuir para a triagem de indivíduos diabéticos assintomáticos e não identificados, como são casos de falsos negativos com uso do exame de FPG, auxiliando assim no diagnóstico de diabetes e no

tratamento precoce. O método pode aprimorar o processo de diagnóstico dos laboratórios médicos, possibilitando uma análise automática e de baixo custo, sem comprometer o fluxo de trabalho atualmente adotado.

Buscando analisar a possibilidade de predição do exame HbA1c sobre diferentes aspectos, a base foi dividida em 5 diferentes grupos (datasets). Esta divisão buscou avaliar as características de cada classe no processo de diagnóstico da Diabetes Mellitus.

Inicialmente as classes foram divididas aos pares, buscando assim identificar o comportamento dos modelos de classificação com os diferentes arranjos. Uma matriz de confusão foi escolhida para a avaliação dos modelos, uma vez temos uma base desbalanceada. Nesses casos, o Escore-F1 e a área sob a curva Precision-Reacll (AUC PR) são as métricas de avaliação mais recomendadas. O Escore-F1 representa uma média consoante entre sensibilidade e precisão e é uma maneira simples de avaliar modelos com bancos de dados desbalanceados. No entanto, ao avaliar um modelo de classificação na busca de um alvo preferimos a análise conjunta de sensibilidade e precisão. Isso nos possibilita conhecer as características de cada modelo.

Na Tabela 7.1 é apresentado de forma detalhada as métricas de avaliação do processo de classificação do grupo Saudável e Pré-diabete (HP) para todos os modelos testados. Aqui o alvo é a pré-diabetes. Analisando os valores de sensibilidade deste grupo, vemos a dificuldade dos modelos em identificar o alvo, sendo que todos tiveram resultado em torno de 50%. Já quando o alvo é o indivíduo saudável, a sensibilidade aumenta significativamente, ficando em torno de 90%. Isto se deve ao fato de a base ser desbalanceada, possuindo o dobro de indivíduos saudáveis em relação aos com pré-diabetes. Analisando os valores de precisão, vê-se que estes são ligeiramente melhores, ficando em torno 70% para a pré-diabetes, mas ainda sendo pouco relevante para o bom desempenho do modelo. Isso é confirmado quando observamos os valores do Escore-F1 e a área sob a curva Precision-recall (AUC PR), que também se mostram relativamente baixos. Nesta tabela também foi plotados os valores de acurácia. Não com o objetivo de avaliar o desempenho do modelo, mas com o intuito de verificar a ocorrência ou não de overfitting durante o treinamento. Neste caso é possível afirma que todos os modelos tiveram um treinamento sem overfitting.

Já quando analisamos os resultados da classificação sobre o grupo Saudável e Diabetes (HD) (Tabela 7.2), tendo a Diabetes como alvo, observamos um aumento na sensibilidade de todos os modelos, sendo que o modelo de ANN teve o melhor resultado

(84,2%). O mesmo pode-se dizer pra a precisão, onde todos os modelos alcançaram valores em torno dos 90%, sendo que o modelo KNN teve o melhor resultado, com 96,4% Esta melhora já era esperada, uma vez que a classe de pré-diabetes foi suprimida, deixando o conjunto de dados mais heterogêneo e consequentemente facilitando a classificação. O mesmo ocorreu com os valores de Escore-F1 e AUC PR, ficando em torno de 85% e 93%. Este resultado demostra a capacidade dos modelos em identificar e classificar indivíduos com Diabetes em relação a indivíduos saudáveis.

Na análise da classificação do grupo Pré-diabetes e Diabetes (PD) (Tabela 7.3), observamos um resultado inferior ao do grupo HD, mas superior ao grupo HP, sendo que o alvo aqui é a Diabetes. Em geral os valores de sensibilidade ficaram entre 56,2% (KNN) e 71,5% (RF). Os valores de precisão também ficaram um pouco melhores que no grupo HP, ficando entre 82,1% (RF) e 87,3% (KNN). Os valores de Escore-F1 e AUC PR também ficaram melhores que o grupo HP, sendo que os melhores resultados foram 76,4% para o modelo RF (Escore-F1) e 87,2% para o modelo ANN (AUC PR).

Comparando os resultados do grupo Saudável e Pré-diabetes (HP) com os resultados do grupo Pré-diabetes e Diabetes (PD), verifica-se que os modelos possuem maior capacidade de separar o Pré-diabetes do Diabetes do que do Saudável.

Avançando com a análise unimos a classe Pré-diabetes tanto a classe Diabetes, para formar o grupo Saudável e Não Saudável (HN), assim como a classe Saudável, formando o grupo Não diabetes e Diabetes (ND).

Na Tabela 7.4 é apresentado os valores referentes ao grupo HN. O modelo com maior sensibilidade foi o ANN com 69,7%. Já o modelo com maior precisão foi NB com 85,6%. Os melhores valores de Escore-F1 e de AUC PR foram respectivamente 74,2% e 85,8%, ambos para o modelo ANN.

Já na Tabela 7.5 temos os valores da análise para o grupo ND. Aqui o modelo com maior sensibilidade foi o ANN com 71,5% e o de maior precisão foi o KNN com 88,7%. No caso do Escore-F1, o maior valor foi 75,9% com o modelo KNN e o maior valor de AUC PR foi 83,5% com o modelo SVM.

De modo geral, os grupos HN e ND obtiveram resultados semelhantes. No entanto podemos afirma que o grupo ND teve um desempenho ligeiramente melhor que o grupo HN.

Por fim, analisamos o grupo HPD formado pelas classes Saudável, Pré-diabetes e Diabetes. Diferentemente dos grupos anteriores, este é formado pelas três classes que compõem o diagnóstico de Diabetes Mellitus. Novamente foram testados todos o cinco modelo de machine learning. O resultado detalhado para cada um dos modelos e para cada classe do diagnóstico pode ser visto na Tabela 7.6. Neste grupo, os valores da classe Diabetes são muito próximos, quando não o mesmo, da classe Diabetes do Grupo ND. Isso por que em ambos os grupos a classe Diabetes é formada pelo mesmo conjunto de dados. Analisando os valores, vê-se que sensibilidade varia entre 60% (KNN) e 69,5% (ANN). O maio valor de precisão, no entanto, foi de 84,3% (SVM e RF). A precisão do modelo de ANN foi de 82%, tonando esse modelo o mais indicado para detecção da doença neste grupo de dados. Isto é confirmado com os valores do Escore-F1 (75,3) e AUC PR (84,5), o maior na comparação com os demais modelos.

Analisando os resultados obtidos com os diferentes modelos e grupos de dados, percebe-se características específicas em cada grupo, assim como no desempenho dos modelos. Desta forma, dependendo do que se busque como resultado final, será mais adequado utilizar um grupo e modelo do que outro específico, ou até uma combinação deles. Por exemplo, na triagem de indivíduos Saudáveis, com Pré-diabetes e com Diabetes, o indicado será a utilização do modelo ANN treinada com o grupo HPD, sendo que o resultado desta predição, será uma das três classes do diagnóstico. No entanto, caso se queira identificar apenas indivíduos não saudáveis, o mais indicado é a utilização do modelo ANN treinado com o grupo HN. Já na identificação apenas de indivíduos com Diabetes, tanto o modelo ANN quanto o RF, treinados com o grupo ND, poderá ser utilizado.

A maior dificuldade no processo de classificação foi em relação a classe de Pré-diabetes. De modo geral todos os modelos e grupos de dados tiveram mais dificuldade na identificação correta desta classe. Analisando os resultados da matriz de confusão no grupo HPD para cada modelo testado (Figura 7.31 a Figura 7.34), observa-se que os modelos classificam praticamente 50% dos indivíduos com Pré-diabetes como se fossem saudáveis. Isso nos mostra, que para os modelos, as classes Saudável e Pré-diabetes são mais homogêneas, de forma que os indivíduos com Pré-diabetes se assemelham mais aos indivíduos saudáveis que aos indivíduos com Diabetes.

Esta característica pode ser observada também quando analisamos os modelos treinados com os grupos HN e ND, sendo que o grupo HN (Figura 7.19 a Figura 7.23)

obteve menor sensibilidade que o grupo ND (Figura 7.25 a Figura 7.29). Da mesma forma quando analisamos a matriz de confusão do treinamento com os grupos HP, HD e PD. Os grupos HD (Figura 7.7 a Figura 7.11) e PD (Figura 7.13 a Figura 7.17) possuem maior sensibilidade em comparação ao grupo HP (Figura 7.1 a Figura 7.5), confirmando a maior dificuldade dos modelos de classificação em separar os indivíduos com Pré-diabetes dos indivíduos Saudáveis.

Continuando a análise dos modelos de machine learning na predição do HbA1c, cada um dos modelos foi testado como regressor, buscando assim encontrar um valor para o exame de HbA1c e não mais uma classe para o diagnóstico de Diabetes Mellitus.

Na Figura 7.37 foi apresentada o gráfico de dispersão comparando os valore verdadeiros de HbA1c com aqueles preditos pelos modelos de regressão. De modo geral, observa-se certa similaridade estre os valores na dispersão dos dados, sendo que o modelo ANN obteve o melhor comportamento. Isto pode ser melhor observado, analisando as métricas apresentadas na Tabela 7.7. Comparando os valores de correlação, tem-se que o modelo ANN é o que, cuja predição, mais se aproxima dos valores verdadeiros. O mesmo pode ser observado em relação aos erros de predição, de forma que o modelo ANN obteve o melhor resultado com erro médio absoluto de (MAE) de 0,36 e raiz do erro médio quadrático de (RMSE) de 0,35. Estes resultados afirmam a capacidade dos modelos, em especial o ANN, em conseguir predizer valores de HbA1c com base em outros exames de rotina.

A fim de comparam os resultados obtidos com a classificação com os da regressão, realizou-se a classificação dos valores preditos com os modelos de regressão. Para isso utilizamos as mesmas regras de classificação para Diabetes Mellitus apresentadas na Tabela 3.1. Os resultados desta classificação foram inicialmente apresentados na matriz de confusão, conforme Figura 7.38. Já na Figura 7.39, é apresentado uma comparação entre os resultados dos modelos de classificação e a classificação dos modelos de regressão. De modo geral observa-se que os resultados são semelhantes aos obtidos com os modelos de classificação. No entanto, a classificação dos modelos de regressão tem valores de Sensibilidade, Precisão e consequentemente Escaore-F1 um pouco melhores que os valores obtidos com os valores de classificação. Isso nos faz pensar que neste caso talvez seja melhor trabalhar com modelos de regressão do que de classificação.

Por outro lado, poderia se fazer o seguinte questionamento: “Se o exame FPG já é utilizado no diagnóstico da diabetes, por que utilizar ele como parâmetro de entrada de um modelo na busca de um resultado ao invés de utiliza-lo diretamente no diagnóstico?”

A resposta para este questionamento está na observação da classificação da FPG em relação a HbA1c na própria base de dados estudada (Figura 5.5), onde verificou-se que aproximadamente de 30% dos pacientes classificados pelo exame de HbA1c com Diabetes não foram diagnosticados corretamente pelo exame de FPG. Em uma análise mais específica, observa-se que 7,3% de diabéticos e 43,6% de pré-diabéticos foram classificados como saudáveis pelo exame de FPG.

Esse tipo de discrepância não é visto com tanta gravidade nos casos de falso positivo, considerando que muito provavelmente o paciente irá ser levado a repetir o exame antes de um possível tratamento. No entanto nos casos de falso negativo a situação se agrava, uma vez que ao descartar a presença da doença, o paciente permanecerá sem o tratamento adequado promovendo o agravamento do quadro.

No caso dos exames FPG, seria interessante ter um alerta para os casos de falso negativo. Desta forma, sempre que o exame FPG classificar o paciente como saudável, aplicaríamos um modelo de *machine learning* com o intuito de verificar a possibilidade de falso negativo, sugerindo ao paciente a realização do exame HbA1c se necessário.

Na criação de um alerta para casos não diagnosticados de Diabetes, quanto mais alta a sensibilidade mais casos serão identificados pelo modelo. No entanto isso não é tão importante quanto a precisão, já que mesmo que o modelo não identifique muitos casos, aqueles classificados como positivo estarão em sua maior parte corretos. Neste caso, o ideal é que tenhamos modelos com alta precisão, o que fará com que os resultados do modelo em questão tenham poucos falsos positivos.

Com base nos resultados obtidos, desenvolveu-se um método para identificar os valores do exame de FPG que representam falsos negativos. Este método prediz valores ajustados de FPG, onde a classificação do resultado é fortemente concordante com a classificação do exame de HbA1c.

Conforme apresentado na Tabela 6.5, o diagnóstico de diabetes com base no exame de FPG possui baixa concordância com o diagnóstico realizado com o exame de HbA1c. Apesar do exame de HbA1c ser mais recomendado e possuir vantagens em relação ao FPG (ASSOCIATION, 2017; MALKANI; DESILVA, 2012), este último

continua sendo o exame mais realizado. Desta forma, a taxa de falso negativo apresentada pelo FPG, podem trazer complicações para a vida do paciente, uma vez que 60% dos indivíduos não apresentam sintomas na fase inicial da doença (DIABETES UK, 2020).

A falta de concordância provocada por variações nos valores de FPG, sugerem ser mais adequado o uso do exame de HbA1c, uma vez que este representa uma média estimada do nível de glicose no sangue [8, 26]. De maneira semelhante, os valores de glicose média estimada (eAG), calculados com a equação: $eAG = HbA1c * 28{,}7 - 46{,}7$, também apresenta discrepância no diagnóstico quando comparados com o HbA1c, nos levando assim a construção de uma nova equação de glicose estimada (eG) (Equação 6.2).

O objetivo de usar um fator de ajuste (AF) para os valores de FPG, seria o de aproximar estes valores de um valor de glicose média, evitando assim que oscilações neste exame possam atrapalhar no diagnóstico.

Na Figura 7.40 temos os gráficos de distribuição referente aos valores do fator de ajuste preditos (pAF) com os modelos de machine learning. Nos gráficos é possível observar certa semelhança entre os resultados dos modelos testados e os valores originais. Todos apresentaram média aproximada igual a 1. Já o desvio padrão, se difere do original (0,178), mas se assemelham entre os diferentes modelos de predição, com valores em torno de 0,05. Entendemos que essa diferença em relação ao valor original, ocorre por que os modelos tem dificuldade em prever os outliers (por serem a grande minoria na base de dados), mantendo os valores mais perto da média. Estes gráficos nos mostram a capacidade dos modelos em predizer os valores do Fator de Ajuste, o que é reforçado observando os erros de predição e dados estatísticos apresentados na Tabela 7.9. Tanto no gráfico como na tabela, observamos que o modelo ANN foi o que mais se assemelha ao original.

Já na

Figura 7.41, temos os gráficos de dispersão dos valores de FPG ajustados (aFPG), calculado com o FPG original e o fator de ajuste predito (pAF) conforme Equação 6.4, juntamente com o valor da glicose estimada (eG) original, calculado conforme Equação 6.2. De modo geral, todos os modelos se comportaram de forma semelhante. Visualmente, todos os modelos apresentam uma distribuição linear correspondente ao valor da glicose estimada (eG) original. Analisando os dados da Tabela 7.10, vemos que o modelo ANN e KNN, respectivamente, tiveram valores de média e desvio padrão mais próximos do original. Já os menores valores de erro (RMSE = 17,9, MSE = 320,9 e MAE = 10,5), foram alcançados com o modelo ANN.

Na sequência, realizamos a classificação dos valores de FPG ajustado (aFPG) de acordo com o diagnóstico de Diabetes Mellitus. O resultado desta classificação foi plotado na forma de uma matriz de confusão, apresentada na Figura 7.42. Analisando as matrizes, vemos que todos os modelos possuem maior dificuldade na classificação dos indivíduos com pré-diabetes. Sendo que aproximadamente 40% dos pré-diabéticos foram classificados como saudáveis. Parte dessa dificuldade pode-se atribuir as regiões de transição entre as classes. Essa característica fuzzy no processo de classificação tende a prejudicar o desempenho de classificação dos modelos. De modo geral, os modelos KNN, SVM e NB, apresentaram os melhores resultados na classificação do aFPG. No entanto o modelo ANN obteve melhor regularidade entre as classes.

Na busca pela identificação de indivíduos com a doença, buscamos modelos com alta sensibilidade e alta precisão, uma vez que queremos ter certeza do resultado do modelo. No entanto, quando buscamos por falsos negativos, mesmo um valor de sensibilidade menor é interessante, já que qualquer identificação diminui o erro no processo. No entanto o valor da precisão deve ser alto, já que precisamos ter certeza do resultado apontado pelo modelo. Neste senário de ajuste do FPG e identificação de falsos negativos, o modelo que apresentou maiores valores de precisão, tanto para a classe de Diabetes quanto para a de Pré-diabetes, foi o ANN. Para a classificação dos indivíduos com diabetes, o modelo apresentou valores de sensitividade e precisão igual a 64,0% e 86,1%, respectivamente. Já para a classificação de indivíduos com pré-diabetes, o modelo apresentou valores de sensitividade e precisão igual a 66,3% e 60,1%, respectivamente. Analisando os valores de Escore-F1, que representa uma média harmônica de precisão e sensibilidade, os modelos SVM e NB tiveram melhor resultado para a classificação da diabetes, com 75,4%. Já no caso dos indivíduos com pré-diabetes, o melhor resultado foi também do modelo ANN com 63,1%.

Já a classificação de indivíduos saudáveis nos é menos relevante, uma vez que não possui falsos negativos. Na Figura 7.43, plotamos o gráfico da classificação dos valores de FPG ajustado (aFPG) (Figura 7.43 - b) em comparação aos valores originais de HbA1c, juntamente com os antigos valores de FPG (Figura 7.43 - a). Nele podemos observar os ganhos no ajuste dos resultados apresentados pelo modelo em relação aos antigos valores de FPG, assim como a maior concordância com os valores de HbA1c.

Na Tabela 7.12 apresentamos o ganho na concordância do diagnóstico de Diabetes, em relação ao HbA1c, utilizando o FPG ajustado (aFPG). A classificação dos indivíduos com diabetes teve um ganho de 16,6% e a classificação dos indivíduos com pré-diabetes teve um ganho de 35%. Já a classificação dos saudáveis se manteve com um leve aumento de 0,4%.

Neste caso, vale ressaltar que os mesmos pacientes presentes em cada classe de diagnóstico da HbA1c são mantidos no respectivo dataset. Assim podemos afirmar que não houve reclassificação de pacientes entre classes de HbA1c, de forma que o ganho obtido com o aFPG é absoluto em cada classe de diagnóstico.

9. CONCLUSÃO

Pacientes com Diabetes Mellitus podem ser assintomáticos e passar despercebidos em diagnósticos baseados apenas em exames de FPG. Esses exames podem sofrer variações, sendo suscetíveis a metodologias não padronizadas, assim como a preparação do paciente antes do exame e o uso de medicamentos.

A possibilidade de um sistema computacional encontrar automaticamente informações ocultas em dados de testes laboratoriais é altamente vantajosa para o processo de diagnóstico em laboratórios médicos. Esses sistemas poderiam realizar triagem de pacientes para descobrir doenças precoces, gerar alertas e recomendar exames complementares como contraprova de exames com resultados falso negativo. Muitas vezes, esses exames poderiam ser realizados com a mesma amostra de sangue do paciente, gerando agilidade e economia.

Este trabalho demonstrou que modelos de aprendizado de máquina, em especial o ANN, podem auxiliar na triagem de Diabetes Mellitus usando dados de exames laboratoriais realizados rotineiramente, incluindo hemogramas, fornecendo evidências para encaminhar um paciente para testes adicionais (por exemplo, HbA1c). O sistema proposto pode operar em conjunto com métodos tradicionais e não interromper o processo de fluxo normal de exames.

Após a análise dos resultados apresentados na etapa de identificação de falsos negativos, podemos afirmar a capacidade do método proposto em efetuar um ajuste nos valores de FPG, aumentando a concordância com o HbA1c e diminuindo a ocorrência de falsos negativos.

De modo geral, o modelo apresentou um ganho de 52% sobre a classificação dos valores de FPG ajustados. Desta forma, durante a rotina de exames laboratoriais, o modelo poderia ser utilizado na triagem de possíveis falsos negativos e consequentemente sugerir a realização do exame de HbA1c para confirmação do diagnóstico de Diabetes Mellitus.

9.1. TRABALHOS FUTUROS

Dando continuidade aos estudos iniciados neste trabalho, sugere-se o cruzamento e análise de dados de fotoplestimografia (PPG) com os exames de hemograma completo. A ideia principal constrói a tese de que se possa predizer exames de rotina com base em dados de PPG.

Desta forma, podería-se realizar triagem de pacientes de forma não invasiva, identificando possível alterações nos exames sanguíneos. Assim, após uma indicação positiva, o paciente seria orientado a realização dos exames em questão, a fim de confirmar a suspeita indica pelo sistema.

Esse processo, seria extremamente vantajoso para a identificação de doenças de forma precoce. Permitindo que muitos pacientes possam se tratar de forma preditiva e com maior chance de sucesso. Da mesma forma, esse processo poderia acarretar em economia aos sistemas de saúde, pois evitaria o agravamento de doenças e gastos com tratamentos mais complexos.

REFERÊNCIAS

About Chronic Diseases | CDC. [s.d.]. Disponível em: https://www.cdc.gov/chronicdisease/about/index.htm. Acesso em: 17 nov. 2019.

ABU-MOSTAFA, Yaser S.; MAGDON-ISMAIL, M.; LIN, H. T. **Learning from Data: A Short Course**. [s.l.] : AMLBook.com, 2012. Disponível em: https://books.google.com.br/books?id=iZUzMwEACAAJ.

AIKENS, Rachael C.; BALASUBRAMANIAN, Santhosh; CHEN, Jonathan H. A Machine Learning Approach to Predicting the Stability of Inpatient Lab Test Results. **AMIA Summits on Translational Science Proceedings**, *[S. l.]*, v. 2019, p. 515, 2019. Disponível em: /pmc/articles/PMC6568078/. Acesso em: 5 fev. 2022.

ALJAME, Maryam; IMTIAZ, Ayyub; AHMAD, Imtiaz; MOHAMMED, Ameer. Deep forest model for diagnosing COVID-19 from routine blood tests. **Scientific reports**, *[S. l.]*, v. 11, n. 1, 2021. DOI: 10.1038/S41598-021-95957-W. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34404838/. Acesso em: 18 nov. 2021.

AMERICAN DIABETES ASSOCIATION STANDARDS OF MEDICAL CARE IN DIABETES-2017. *[S. l.]*, [s.d.]. Disponível em: http://www.epi.uff.br/wp-content/uploads/2013/10/dc_40_s1_final.pdf. Acesso em: 15 abr. 2019.

ASSOCIATION, American Diabetes. Classification and diagnosis of diabetes. **Diabetes Care**, *[S. l.]*, v. 40, n. Supplement 1, p. S11–S24, 2017. DOI: 10.2337/dc17-S005.

BABAEI RIKAN, Samin; SORAYAIE AZAR, Amir; GHAFARI, Ali; BAGHERZADEH MOHASEFI, Jamshid; PIRNEJAD, Habibollah. COVID-19 Diagnosis from Routine Blood Tests using Artificial Intelligence Techniques. **Biomedical signal processing and control**, *[S. l.]*, v. 72, 2021. DOI: 10.1016/J.BSPC.2021.103263. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34745318/. Acesso em: 5 fev. 2022.

BANERJEE, Abhirup; RAY, Surajit; VORSELAARS, Bart; KITSON, Joanne; MAMALAKIS, Michail; WEEKS, Simonne; BAKER, Mark; MACKENZIE, Louise S. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. **International immunopharmacology**, *[S. l.]*,

v. 86, 2020. DOI: 10.1016/J.INTIMP.2020.106705. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32652499/. Acesso em: 5 fev. 2022.

BARON, Jason M.; DIGHE, Anand S. **The role of informatics and decision support in utilization management**. **Clinica Chimica Acta**, 2014. DOI: 10.1016/j.cca.2013.09.027. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/24084507. Acesso em: 20 abr. 2020.

BARON, Jason M.; MERMEL, Craig H.; LEWANDROWSKI, Kent B.; DIGHE, Anand S. Detection of Preanalytic Laboratory Testing Errors Using a Statistically Guided Protocol. **American Journal of Clinical Pathology**, *[S. l.]*, v. 138, n. 3, p. 406–413, 2012. DOI: 10.1309/AJCPQIRIB3CT1EJV. Disponível em: https://academic.oup.com/ajcp/article-lookup/doi/10.1309/AJCPQIRIB3CT1EJV. Acesso em: 20 abr. 2020.

BARTHOLOMEW, D. J. The foundations of factor analysis. **Biometrika**, *[S. l.]*, v. 71, n. 2, p. 221–232, 1984. DOI: 10.1093/BIOMET/71.2.221. Disponível em: https://academic-oup-com.eres.qnl.qa/biomet/article/71/2/221/233220. Acesso em: 11 jul. 2021.

BELLMAN, Richard. **Adaptive Control Processes: A Guided Tour**. [s.l.] : Princeton University Press, 1961. Disponível em: http://www.jstor.org/stable/j.ctt183ph6v. Acesso em: 23 abr. 2020.

BERNARDINI, Michele; MORETTINI, Micaela; ROMEO, Luca; FRONTONI, Emanuele; BURATTINI, L. TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic Health Records. **Computers in Biology and Medicine**, *[S. l.]*, v. 112, 2019. Disponível em: http://dx.doi.org/10.1016/j.compbiomed.2019.103358.

BIRKS, Jacqueline; BANKHEAD, Clare; HOLT, Tim A.; FULLER, Alice; PATNICK, Julietta. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. **Cancer medicine**, *[S. l.]*, v. 6, n. 10, p. 2453–2460, 2017. DOI: 10.1002/CAM4.1183. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28941187/. Acesso em: 5 fev. 2022.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [s.l.] : Springer New

York, 2016. Disponível em: https://books.google.com.br/books?id=kOXDtAEACAAJ.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. A training algorithm for optimal margin classifiers. *In*: PROCEEDINGS OF THE FIFTH ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY - COLT '92 1992, New York, New York, USA. **Anais** [...]. New York, New York, USA: ACM Press, 1992. p. 144–152. DOI: 10.1145/130385.130401. Disponível em: http://portal.acm.org/citation.cfm?doid=130385.130401. Acesso em: 25 abr. 2020.

BREIMAN, Leo. Random forests. **Machine Learning**, *[S. l.]*, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324.

BRINATI, Davide; CAMPAGNER, Andrea; FERRARI, Davide; LOCATELLI, Massimo; BANFI, Giuseppe; CABITZA, Federico. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. **Journal of medical systems**, *[S. l.]*, v. 44, n. 8, 2020. DOI: 10.1007/S10916-020-01597-4. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32607737/. Acesso em: 5 fev. 2022.

BROWNLEE, Jason. **Develop k-Nearest Neighbors in Python From Scratch**. [s.d.]. Disponível em: https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/. Acesso em: 25 abr. 2020.

BROWNLEE, Jason. **Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End**. 2016. Disponível em: https://machinelearningmastery.com/machine-learning-with-python/. Acesso em: 24 nov. 2022.

CABITZA, Federico et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. **Clinical chemistry and laboratory medicine**, *[S. l.]*, v. 59, n. 2, p. 421–431, 2020. DOI: 10.1515/CCLM-2020-1294. Disponível em: https://pubmed.ncbi.nlm.nih.gov/33079698/. Acesso em: 5 fev. 2022.

CAMARGO, Joíza Lins; GROSS, Jorge Luiz. **Glycohemoglobin (GHb): clinical and analytical aspects**. **Arquivos brasileiros de endocrinologia e metabologia**Sociedade Brasileira de Endocrinologia e Metabologia, , 2004. DOI: 10.1590/s0004-27302004000400005.

CAMPAGNER, Andrea; CAROBENE, Anna; CABITZA, Federico. External validation of Machine Learning models for COVID-19 detection based on Complete Blood Count. **Health information science and systems**, *[S. l.]*, v. 9, n. 1, 2021. DOI: 10.1007/S13755-021-00167-3. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34721844/. Acesso em: 18 nov. 2021.

CASTRILLÓN, Omar D.; SARACHE, William; CASTAÑO, Eduardo. Sistema Bayesiano para la Predicción de la Diabetes. **Información tecnológica**, *[S. l.]*, v. 28, n. 6, p. 161–168, 2017. DOI: 10.4067/S0718-07642017000600017. Disponível em: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642017000600017&lng=en&nrm=iso&tlng=en. Acesso em: 23 abr. 2018.

CHAUHAN, Nagesh Singh. **Random Forest — A powerful Ensemble Learning algorithm**. [s.d.]. Disponível em: https://towardsdatascience.com/random-forest-a-powerful-ensemble-learning-algorithm-2bf132ba639d. Acesso em: 26 abr. 2020.

CHEN, Min; HAO, Yixue; HWANG, Kai; WANG, Lin; WANG, Lu. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. **IEEE Access**, *[S. l.]*, v. 5, p. 8869–8879, 2017. DOI: 10.1109/ACCESS.2017.2694446.

CHICCO, Davide; JURMAN, Giuseppe. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. **BMC Medical Informatics and Decision Making**, *[S. l.]*, v. 20, n. 1, p. 16, 2020. DOI: 10.1186/s12911-020-1023-5. Disponível em: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5.

CLSI Standards & Guidelines: Shop for CLSI Standards. [s.d.]. Disponível em: https://clsi.org/standards/. Acesso em: 19 nov. 2019.

CORTES, CORINNA; VAPNIK, VLADIMIR. Support-Vector Networks. *In*: Boston, MA. v. 20p. 273–297.

DA, André; PEREIRA, Silva; PALUDO, Berenice; VIEIRA, Manoel; CERBARO, Rodolfo Henrique. APOSTILA ANÁLISE FATORIAL. *[S. l.]*, [s.d.].

DE SILVA, Kushan; MATHEWS, Noel; TEEDE, Helena; FORBES, Andrew;

JÖNSSON, Daniel; DEMMER, Ryan T.; ENTICOTT, Joanne. Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: A retrospective cohort analysis using machine learning and unstructured big data. **Computers in Biology and Medicine**, *[S. l.]*, v. 132, p. 104305, 2021. DOI: 10.1016/J.COMPBIOMED.2021.104305.

DEMIRCI, Ferhat; AKAN, Pinar; KUME, Tuncay; SISMAN, Ali Riza; ERBAYRAKTAR, Zubeyde; SEVINC, Suleyman. Artificial Neural Network Approach in Laboratory Test Reporting Learning Algorithms. **American Journal of Clinical Pathology**, *[S. l.]*, v. 146, n. 2, p. 227–237, 2016. DOI: 10.1093/ajcp/aqw104.

DIABETES UK. **Diabetes UK - Know diabetes. Fight diabetes | Diabetes UK**. , 2018. Disponível em: https://www.diabetes.org.uk/. Acesso em: 20 abr. 2020.

DIABETES UK. **Prediabetes | Diabetes UK**. 2020. Disponível em: https://www.diabetes.org.uk/preventing-type-2-diabetes/prediabetes. Acesso em: 20 abr. 2020.

DIGHE, AnandS et al. The 2013 symposium on pathology data integration and clinical decision support and the current state of field. **Journal of Pathology Informatics**, *[S. l.]*, v. 5, n. 1, p. 2, 2014. DOI: 10.4103/2153-3539.126145.

DIRETRIZES DA SOCIEDADE BRASILEIRA DE DIABETES 2017-2018. [s.l: s.n.]. Disponível em: www.editoraclannad.com.br. Acesso em: 15 abr. 2019.

DU, Y.; FANG, Z.; JIAO, J.; XI, G.; ZHU, C.; REN, Y.; GUO, Y.; WANG, Y. Application of ultrasound-based radiomics technology in fetal-lung-texture analysis in pregnancies complicated by gestational diabetes and/or pre-eclampsia. **ULTRASOUND IN OBSTETRICS & GYNECOLOGY**, *[S. l.]*, v. 57, n. 5, p. 804–812, 2021. DOI: 10.1002/uog.22037.

DUAN, Kai-Bo; KEERTHI, S. Sathiya. Which Is the Best Multiclass SVM Method? An Empirical Study. *In*: [s.l.] : Springer, Berlin, Heidelberg, 2005. p. 278–285. DOI: 10.1007/11494683_28. Disponível em: http://link.springer.com/10.1007/11494683_28. Acesso em: 25 abr. 2020.

Ensemble methods — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em:

https://scikit-learn.org/stable/modules/ensemble.html#ensemble. Acesso em: 25 abr. 2020.

Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. [s.d.]. Disponível em: https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation. Acesso em: 17 jul. 2021.

factor-analyzer · PyPI. [s.d.]. Disponível em: https://pypi.org/project/factor-analyzer/. Acesso em: 18 abr. 2020.

FACURE, Matheus. **Funções de ativação**. 2017. Disponível em: https://matheusfacure.github.io/2017/07/12/activ-func/. Acesso em: 21 out. 2019.

Feature Selection in Python with Scikit-Learn. [s.d.]. Disponível em: https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/. Acesso em: 23 out. 2018.

FERNÁNDEZ-LLATAS, Carlos; GARCÍA-GÓMEZ, Juan Miguel. **Data Mining in Clinical Medicine**. New York, NY: Springer New York, 2015. v. 1246 DOI: 10.1007/978-1-4939-1985-7.

GLADDING, Patrick A. et al. A machine learning PROGRAM to identify COVID-19 and other diseases from hematology data. **Future science OA**, *[S. l.]*, v. 7, n. 7, 2021. DOI: 10.2144/FSOA-2020-0207. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34254032/. Acesso em: 5 fev. 2022.

GLOBAL REPORT ON DIABETES WHO Library Cataloguing-in-Publication Data Global report on diabetes. [s.l: s.n.]. Disponível em: http://www.who.int/about/licensing/copyright_form/index.html. Acesso em: 26 abr. 2020.

GOLDSTEIN, Matthew. Kn- Nearest Neighbor Classification. **IEEE Transactions on Information Theory**, *[S. l.]*, v. 18, n. 5, p. 627–630, 1972. DOI: 10.1109/TIT.1972.1054888.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] : MIT Press, 2016. Disponível em: https://books.google.com.br/books?id=Np9SDQAAQBAJ.

GUNČAR, Gregor; KUKAR, Matjaž; NOTAR, Mateja; BRVAR, Miran; ČERNELČ, Peter; NOTAR, Manca; NOTAR, Marko. An application of machine learning to haematological diagnosis. **Scientific reports**, *[S. l.]*, v. 8, n. 1, p. 411, 2018. DOI: 10.1038/s41598-017-18564-8. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/29323142. Acesso em: 4 abr. 2019.

HALL, Patrick; PHAN, Wen; WHITSON, Katie. **Opportunities and Challenges for Machine Learning in Business**. [s.l: s.n.].

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. [s.l.] : Elsevier Inc., 2012. DOI: 10.1016/C2009-0-61819-5.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2. ed. [s.l.] : Artmed, 2007. Disponível em: https://books.google.com.br/books?id=bhMwDwAAQBAJ.

HERNANDEZ, Bernard; HERRERO, Pau; RAWSON, Timothy Miles; MOORE, Luke S. P.; EVANS, Benjamin; TOUMAZOU, Christofer; HOLMES, Alison H.; GEORGIOU, Pantelis. Supervised learning for infection risk inference using pathology data. **BMC Medical Informatics and Decision Making**, *[S. l.]*, v. 17, n. 1, p. 168, 2017. DOI: 10.1186/s12911-017-0550-1. Disponível em: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0550-1.

HO, Tin Kam. Random decision forests. *In*: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, ICDAR 1995, **Anais** [...]. : IEEE Computer Society, 1995. p. 278–282. DOI: 10.1109/ICDAR.1995.598994.

HOLSBACH, Nicole; FOGLIATTO, Flávio Sanson; ANZANELLO, Michel Jose. Método de mineração de dados para identificação de câncer de mama baseado na seleção de variáveis. **Ciencia e Saude Coletiva**, *[S. l.]*, v. 19, n. 4, p. 1295–1304, 2014. DOI: 10.1590/1413-81232014194.01722013.

HOSSAIN, Md. Ekramul; KHAN, Arif; MONI, Mohammad Ali; UDDIN, Shahadat.

Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, *[S. l.]*, v. 18, n. 2, p. 745–758, 2021. DOI: 10.1109/TCBB.2019.2937862. Disponível em: https://ieeexplore.ieee.org/document/8815739/.

HU, Lufeng; YANG, Panxin; WANG, Xianqin; LIN, Feiyan; CHEN, Huiling; CAO, Hongcui; LI, Haiying. Using Biochemical Indexes to Prognose Paraquat-Poisoned Patients: An Extreme Learning Machine-Based Approach. **IEEE Access**, *[S. l.]*, v. 7, p. 42148–42155, 2019. Disponível em: http://dx.doi.org/10.1109/ACCESS.2019.2907272.

Imbalanced data : How to handle Imbalanced Classification Problems. [s.d.]. Disponível em: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/. Acesso em: 25 abr. 2020.

INTERNATIONAL DIABETES FEDERATION. **IDF Atlas 9th edition and other resources**. 2019. Disponível em: https://www.diabetesatlas.org/en/resources/?gclid=CjwKCAjwkPX0BRBKEiwA7THxiGGu4-2c0TG2qc5kNjW4IN3fZxM4g-QbZUX8CeHB_z8ccB_cet-KeBoC-6wQAvD_BwE. Acesso em: 20 abr. 2020.

JOSHI, Rohan P. et al. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. **Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology**, *[S. l.]*, v. 129, 2020. DOI: 10.1016/J.JCV.2020.104502. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32544861/. Acesso em: 5 fev. 2022.

KAFTAN, Ahmed N.; HUSSAIN, Majid K.; ALGENABI, Abdulhussein A.; NASER, Farah H.; ENAYA, Muslim A. Predictive Value of C-reactive Protein, Lactate Dehydrogenase, Ferritin and D-dimer Levels in Diagnosing COVID-19 Patients: a Retrospective Study. **Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH**, *[S. l.]*, v. 29, n. 1, p. 45–50, 2021. DOI: 10.5455/AIM.2021.29.45-50. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34012213/. Acesso em: 5 fev. 2022.

KAVAKIOTIS, Ioannis; TSAVE, Olga; SALIFOGLOU, Athanasios; MAGLAVERAS,

Nicos; VLAHAVAS, Ioannis; CHOUVARDA, Ioanna. Machine Learning and Data Mining Methods in Diabetes Research. **Computational and Structural Biotechnology Journal**, *[S. l.]*, v. 15, p. 104–116, 2017. DOI: 10.1016/J.CSBJ.2016.12.005. Disponível em: https://www.sciencedirect.com/science/article/pii/S2001037016300733. Acesso em: 23 set. 2019.

KINAR, Yaron; KALKSTEIN, Nir; AKIVA, Pinchas; LEVIN, Bernard; HALF, Elizabeth E.; GOLDSHTEIN, Inbal; CHODICK, Gabriel; SHALEV, Varda. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. **Journal of the American Medical Informatics Association**, *[S. l.]*, v. 23, n. 5, p. 879–890, 2016. DOI: 10.1093/jamia/ocv195. Disponível em: https://academic.oup.com/jamia/article/23/5/879/2379871.

KORB, K. B.; NICHOLSON, A. E. **Bayesian Artificial Intelligence**. [s.l.] : CRC Press, 2010. Disponível em: https://books.google.com.br/books?id=LxXOBQAAQBAJ.

KUKAR, Matjaž et al. COVID-19 diagnosis by routine blood tests using machine learning. **Scientific Reports 2021 11:1**, *[S. l.]*, v. 11, n. 1, p. 1–9, 2021. DOI: 10.1038/s41598-021-90265-9. Disponível em: https://www.nature.com/articles/s41598-021-90265-9. Acesso em: 21 fev. 2022.

LAI, Hang; HUANG, Huaxiong; KESHAVJEE, Karim; GUERGACHI, Aziz; GAO, Xin. Predictive models for diabetes mellitus using machine learning techniques. **BMC Endocrine Disorders**, *[S. l.]*, v. 19, n. 1, p. 101, 2019. DOI: 10.1186/s12902-019-0436-6. Disponível em: https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6. Acesso em: 6 jun. 2021.

LORENA, Ana Carolina; DE CARVALHO, André C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, *[S. l.]*, v. 14, n. 2, p. 43–67, 2007. DOI: 10.22456/2175-2745.5690. Disponível em: https://seer.ufrgs.br/rita/article/view/5690. Acesso em: 25 abr. 2020.

LOUIS, David N. et al. **Computational pathology: An emerging definition**. **Archives of Pathology and Laboratory Medicine**the College of American Pathologists, , 2014. DOI: 10.5858/arpa.2014-0034-ED. Disponível em:

http://www.archivesofpathology.org/doi/abs/10.5858/arpa.2014-0034-ED. Acesso em: 22 mar. 2018.

LUO, Yuan; SZOLOVITS, Peter; DIGHE, Anand S.; BARON, Jason M. Using Machine Learning to Predict Laboratory Test Results. **American journal of clinical pathology**, *[S. l.]*, v. 145, n. 6, p. 778–788, 2016. DOI: 10.1093/ajcp/aqw064. Disponível em: https://pubmed.ncbi.nlm.nih.gov/27329638/. Acesso em: 7 mar. 2018.

MA, Han; XU, Cheng Fu; SHEN, Zhe; YU, Chao Hui; LI, You Ming. Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China. **BioMed Research International**, *[S. l.]*, v. 2018, 2018. DOI: 10.1155/2018/4304376.

MALKANI, Samir; DESILVA, Taniya. Controversies on how diabetes is diagnosed. . abr. 2012, 2, p. 97–103.

MCCULLOCH, Warren S.; PITTS, Walter. **A logical calculus of the ideas immanent in nervous activityBulletin of Mothemnticnl Biology**. [s.l: s.n.].

Medical Subject Headings - Home Page. [s.d.]. Disponível em: https://www.nlm.nih.gov/mesh/meshhome.html. Acesso em: 21 fev. 2022.

METSKER, Oleg; MAGOEV, Kirill; YAKOVLEV, Alexey; YANISHEVSKIY, Stanislav; KOPANITSA, Georgy; KOVALCHUK, Sergey; KRZHIZHANOVSKAYA, Valeria V. Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study. **BMC Medical Informatics and Decision Making**, *[S. l.]*, v. 20, n. 1, p. 201, 2020. DOI: 10.1186/s12911-020-01215-w. Disponível em: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01215-w.

MINSKY, M.; PAPERT, S. A.; BOTTOU, L. **Perceptrons: An Introduction to Computational Geometry**. [s.l.] : MIT Press, 2017. Disponível em: https://books.google.com.br/books?id=PLQ5DwAAQBAJ.

MISHRA, Aditya. **Metrics to Evaluate your Machine Learning Algorithm**. **Towards Data Science**, 2018. Disponível em: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. Acesso em: 24 abr. 2020.

MITCHELL, Tom M. **Machine Learning. Annual Review Of Computer Science**. [s.l: s.n.]. v. 4

MOONEY, Ciarán; EOGAN, Maeve; NÍ ÁINLE, Fionnuala; CLEARY, Brian; GALLAGHER, Joeseph J.; O'LOUGHLIN, John; DREW, Richard J. Predicting bacteraemia in maternity patients using full blood count parameters: A supervised machine learning algorithm approach. **International Journal of Laboratory Hematology**, *[S. l.]*, v. 43, n. 4, p. 609–615, 2021. DOI: 10.1111/ijlh.13434. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/ijlh.13434.

MUCHERINO, Antonio; PAPAJORGJI, Petraq J.; PARDALOS, Panos M. k-Nearest Neighbor Classification. *In*: [s.l.] : Springer, New York, NY, 2009. p. 83–106. DOI: 10.1007/978-0-387-88615-2_4. Disponível em: http://link.springer.com/10.1007/978-0-387-88615-2_4. Acesso em: 8 out. 2019.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [s.l.] : MIT Press, 2012. Disponível em: https://books.google.com.br/books?id=NZP6AQAAQBAJ.

MYARI, Alexandra; PAPAPETROU, Evangelia; TSAOUSI, Christina. Diagnostic value of white blood cell parameters for COVID-19: Is there a role for HFLC and IG? **International Journal of Laboratory Hematology**, *[S. l.]*, 2021. DOI: 10.1111/IJLH.13728. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/ijlh.13728. Acesso em: 18 nov. 2021.

Naive Bayes — scikit-learn 0.21.3 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/naive_bayes.html. Acesso em: 13 out. 2019.

NATHAN, David M. et al. **International expert committee report on the role of the A1C assay in the diagnosis of diabetes**. **Diabetes Care**American Diabetes Association, , 2009. DOI: 10.2337/dc09-9033.

NATHAN, David M.; KUENEN, Judith; BORG, Rikke; ZHENG, Hui; SCHOENFELD, David; HEINE, Robert J. Translating the A1C assay into estimated average glucose values. **Diabetes Care**, *[S. l.]*, v. 31, n. 8, p. 1473–1478, 2008. DOI: 10.2337/dc08-0545. Disponível em: http://creativecommons. Acesso em: 17 maio. 2021.

Nearest Neighbors — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/neighbors.html. Acesso em: 25 abr. 2020.

Neural network models (supervised) — scikit-learn 0.21.3 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/neural_networks_supervised.html. Acesso em: 20 out. 2019.

NIH. **Study Quality Assessment Tools | NHLBI, NIH**. **National Heart, Lung, and Blood Institute**, 2014. Disponível em: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools. Acesso em: 21 fev. 2022.

OLIVERA, André Rodrigues et al. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. **Sao Paulo Medical Journal**, *[S. l.]*, v. 135, n. 3, p. 234–246, 2017. a. DOI: 10.1590/1516-3180.2016.0309010217. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802017000300234&lng=en&tlng=en. Acesso em: 7 mar. 2018.

OLIVERA, André Rodrigues; ROESLER, Valter; IOCHPE, Cirano; SCHMIDT, Maria Inês; VIGO, Álvaro; BARRETO, Sandhi Maria; DUNCAN, Bruce Bartholow. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. **Sao Paulo Medical Journal**, *[S. l.]*, v. 135, n. 3, p. 234–246, 2017. b. DOI: 10.1590/1516-3180.2016.0309010217.

PARK, Dong Jin; PARK, Min Woo; LEE, Homin; KIM, Young-Jin; KIM, Yeongsic; PARK, Young Hoon. Development of machine learning model for diagnostic disease prediction based on laboratory tests. **Scientific Reports**, *[S. l.]*, v. 11, n. 1, p. 7567, 2021. DOI: 10.1038/s41598-021-87171-5. Disponível em: http://www.nature.com/articles/s41598-021-87171-5.

PARSIAN, Mahmoud. **Data algorithms : recipes for scaling up with Hadoop and Spark**. [s.l: s.n.]. Disponível em: https://learning.oreilly.com/library/view/data-algorithms/9781491906170/. Acesso em: 8 out. 2019.

PEARL, J. **Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning**. [s.l.] : UCLA, Computer Science Department, 1985. Disponível em: https://books.google.com.br/books?id=1sfMOgAACAAJ.

PEEK, Niels; COMBI, Carlo; MARIN, Roque; BELLAZZI, Riccardo. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. **Artificial Intelligence in Medicine**, *[S. l.]*, v. 65, n. 1, p. 61–73, 2015. DOI: 10.1016/j.artmed.2015.07.003. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0933365715000871. Acesso em: 22 abr. 2020.

PLANTE, Timothy B.; BLAU, Aaron M.; BERG, Adrian N.; WEINBERG, Aaron S.; JUN, Ik C.; TAPSON, Victor F.; KANIGAN, Tanya S.; ADIB, Artur B. Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study. **Journal of Medical Internet Research**, *[S. l.]*, v. 22, n. 12, p. e24048, 2020. DOI: 10.2196/24048. Disponível em: https://www.jmir.org/2020/12/e24048.

Preprocessing data — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing. Acesso em: 26 abr. 2020.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. An Empirical Study of Chronic Diseases in the United States: A Visual Analytics Approach to Public Health. **International Journal of Environmental Research and Public Health**, *[S. l.]*, v. 15, n. 3, p. 431, 2018. DOI: 10.3390/ijerph15030431. Disponível em: http://www.mdpi.com/1660-4601/15/3/431. Acesso em: 26 abr. 2020.

RAHMAN, Tawsifur et al. Mortality prediction utilizing blood biomarkers to predict the severity of COVID-19 using machine learning technique. **Diagnostics**, *[S. l.]*, v. 11, n. 9, 2021. DOI: 10.3390/DIAGNOSTICS11091582/S1. Disponível em: /pmc/articles/PMC8469072/. Acesso em: 18 nov. 2021.

RAO, P. V. Type 2 diabetes in children: Clinical aspects and risk factors. **Indian Journal of Endocrinology and Metabolism**, *[S. l.]*, v. 19, n. 7, p. S47–S50, 2015. DOI: 10.4103/2230-8210.155401. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/25941651. Acesso em: 22 set. 2019.

RAVAUT, Mathieu; HARISH, Vinyas; SADEGHI, Hamed; LEUNG, Kin Kwan;

VOLKOVS, Maksims; KORNAS, Kathy; WATSON, Tristan; POUTANEN, Tomi; ROSELLA, Laura C. Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. **JAMA Network Open**, *[S. l.]*, v. 4, n. 5, p. e2111315, 2021. DOI: 10.1001/jamanetworkopen.2021.11315. Disponível em: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2780137.

RAWSON, T. M. et al. Supervised machine learning for the prediction of infection on admission to hospital: A prospective observational cohort study. **Journal of Antimicrobial Chemotherapy**, *[S. l.]*, v. 74, n. 4, p. 1108–1115, 2019. DOI: 10.1093/jac/dky514. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063012496&doi=10.1093%2Fjac%2Fdky514&partnerID=40&md5=b6885d5b7bcd6954020f0b6cf42e070e.

RAZAVIAN, Narges; MARCUS, Jake; SONTAG, David. **Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests**. PMLR, , 2016. Disponível em: https://proceedings.mlr.press/v56/Razavian16.html. Acesso em: 5 fev. 2022.

REHMAN, A.; ABBAS, N.; SABA, T.; MAHMOOD, T.; KOLIVAND, H. Rouleaux red blood cells splitting in microscopic thin blood smear images via local maxima, circles drawing, and mapping with original RBCs. **MICROSCOPY RESEARCH AND TECHNIQUE**, *[S. l.]*, v. 81, n. 7, p. 737–744, 2018. DOI: 10.1002/jemt.23030.

RICHARDSON, Alice M.; LIDBURY, Brett A. Infection status outcome, machine learning method and virus type interact to affect the optimised prediction of hepatitis virus immunoassay results from routine pathology laboratory assays in unbalanced data. **BMC Bioinformatics**, *[S. l.]*, v. 14, n. 1, 2013. DOI: 10.1186/1471-2105-14-206.

RICHARDSON, Alice M.; LIDBURY, Brett A. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. **BMC Medical Informatics and Decision Making**, *[S. l.]*, v. 17, n. 1, 2017. DOI: 10.1186/s12911-017-0522-5. Disponível em: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85027404321&doi=10.1186%2Fs12911-017-0522-

5&partnerID=40&md5=cbe8bbc0867b0ef4593b54c41d027c90. Acesso em: 6 jun. 2021.

RIPLEY, B. D. **Pattern Recognition and Neural Networks**. [s.l.] : Cambridge University Press, 2007. Disponível em: https://books.google.com.br/books?id=m12UR8QmLqoC.

RITA BETTENCOURT SILVA; E. COSTA; MARIA JOSÉ TELES; M. FARIA; M.F. ALMEIDA; JOANA QUEIROS; DAVIDE CARVALHO; JOÃO TIAGO GUIMARÃES. **Associação Entre Glicose Média Estimada e Glicose Plasmática em Jejum em Adultos em Seguimento Ambulatório**. [s.d.]. Disponível em: https://www.researchgate.net/publication/325380768_Associacao_Entre_Glicose_Media_Estimada_e_Glicose_Plasmatica_em_Jejum_em_Adultos_em_Seguimento_Ambulatorio. Acesso em: 4 jul. 2021.

ROBERTO NUNES GUEDES SECRETÁRIO ESPECIAL DE FAZENDA WALDERY RODRIGUES JUNIOR, Paulo; SUSANA CORDEIRO GUERRA DIRETORA-EXECUTIVA MARISE MARIA FERREIRA, Presidente; LUIZ RIOS NETO, Eduardo G.; RENATO PEREIRA COTOVIO, Carlos; DANIELLE LINS MENDES MACEDO, Carmen; LUCIA FRANÇA PONTES VIEIRA PRESIDENTE DA REPÚBLICA JAIR MESSIAS BOLSONARO, Maria; DA SAÚDE EDUARDO PAZUELLO SECRETÁRIO-EXECUTIVO ÉLCIO FRANCO, Ministro; EDUARDO GUEDES SELLERA, Paulo. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA-IBGE. *[S. l.]*, [s.d.].

ROSENBAUM, Matthew W.; BARON, Jason M. Using Machine Learning-Based Multianalyte Delta Checks to Detect Wrong Blood in Tube Errors. **American Journal of Clinical Pathology**, *[S. l.]*, v. 150, n. 6, p. 555–566, 2018. DOI: 10.1093/ajcp/aqy085.

ROSENBLATT, Frank. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, *[S. l.]*, v. 65, n. 6, p. 386–408, 1958. DOI: 10.1037/h0042519. Disponível em: http://doi.apa.org/getdoi.cfm?doi=10.1037/h0042519. Acesso em: 26 abr. 2020.

ROSS, Kenneth A. et al. Curse of Dimensionality. *In*: **Encyclopedia of Database**

Systems. Boston, MA: Springer US, 2009. p. 545–546. DOI: 10.1007/978-0-387-39940-9_133. Disponível em: http://link.springer.com/10.1007/978-0-387-39940-9_133. Acesso em: 28 out. 2019.

ROY, Shivaal K.; HOM, Jason; MACKEY, Lester; SHAH, Neil; CHEN, Jonathan H. Predicting Low Information Laboratory Diagnostic Tests. **AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science**, *[S. l.]*, v. 2017, p. 217–226, 2018. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/29888076. Acesso em: 22 abr. 2020.

ROYAL SOCIETY. **Machine learning : the power and promise of computers that learn by example**. [s.l: s.n.]. v. 66 DOI: 10.1126/scitranslmed.3002564. Disponível em: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf%0Ahttps://www.privacyinternational.org/node/1525%0Ahttps://ico.org.uk/media/2013894/ico-feedback-request-profiling-and-automated-decision-making.pdf. Acesso em: 26 abr. 2020.

SACKS, David B. **A1C versus glucose testing: A comparison**. **Diabetes Care**American Diabetes Association, , 2011. DOI: 10.2337/dc10-1546. Disponível em: https://care.diabetesjournals.org/content/34/2/518. Acesso em: 22 abr. 2020.

SAITO, Takaya; REHMSMEIER, Marc. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. **PLOS ONE**, *[S. l.]*, v. 10, n. 3, p. e0118432, 2015. DOI: 10.1371/journal.pone.0118432. Disponível em: https://dx.plos.org/10.1371/journal.pone.0118432. Acesso em: 11 dez. 2019.

SAVAN PATEL. **Chapter 2 : SVM (Support Vector Machine) — Theory – Machine Learning 101 – Medium**. **Machine Learning 101**, 2017. Disponível em: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. Acesso em: 25 abr. 2020.

SBPC/ML | SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA/MEDICINA LABORATORIAL. [s.d.]. Disponível em: http://www.sbpc.org.br/. Acesso em: 20 abr. 2020.

SCHNEIDER, Jennifer L.; LAYEFSKY, Evan; UDALTSOVA, Natalia; LEVIN, Theodore R.; CORLEY, Douglas A. Validation of an Algorithm to Identify Patients at Risk for Colorectal Cancer Based on Laboratory Test and Demographic Data in Diverse, Community-Based Population. **Clinical Gastroenterology and Hepatology**, *[S. l.]*, v. 18, n. 12, p. 2734- 2741.e6, 2020. DOI: 10.1016/j.cgh.2020.04.054. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1542356520305991.

SCHÖLKOPF, B.; SMOLA, B. S. A. J.; SMOLA, A. J.; SCHOLKOPF, M. D. M. P. I. B. C. T. G. P. B.; BACH, F.; PRESS, M. I. T. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. [s.l.] : MIT Press, 2002. Disponível em: https://books.google.com.br/books?id=y8ORL3DWt4sC.

scikit-learn: machine learning in Python — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/. Acesso em: 26 abr. 2020.

SHARMA, Anand; MANSOTRA, Vibhakar. Emerging applications of data mining for healthcare management - A critical review. *In*: 2014 INTERNATIONAL CONFERENCE ON COMPUTING FOR SUSTAINABLE GLOBAL DEVELOPMENT, INDIACOM 2014 2014, **Anais** [...]. : IEEE Computer Society, 2014. p. 377–382. DOI: 10.1109/IndiaCom.2014.6828163.

SHUBROOK, Jay et al. Standards of Medical Care in Diabetes—2017 : Summary of Revisions. **Diabetes Care**, *[S. l.]*, v. 40, n. Supplement 1, p. S4–S5, 2017. DOI: 10.2337/dc17-S003.

sklearn.decomposition.PCA — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA. Acesso em: 24 abr. 2020.

sklearn.linear_model.LinearRegression — scikit-learn 0.24.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Acesso em: 9 jul. 2021.

SOUZA, Alexandra A. De; ALMEIDA, Danilo Candido De; BARCELOS, Thiago S.; BORTOLETTO, Rodrigo Campos; MUNOZ, Roberto; WALDMAN, Helio; GOES,

Miguel Angelo; SILVA, Leandro A. Simple hemogram to support the decision-making of COVID-19 diagnosis using clusters analysis with self-organizing maps neural network. **Soft Computing**, *[S. l.]*, 2021. DOI: 10.1007/s00500-021-05810-5. Disponível em: https://link.springer.com/10.1007/s00500-021-05810-5.

Support Vector Machines — scikit-learn 0.22.2 documentation. [s.d.]. Disponível em: https://scikit-learn.org/stable/modules/svm.html. Acesso em: 25 abr. 2020.

TAMUNE, Hidetaka; UKITA, Jumpei; HAMAMOTO, Yu; TANAKA, Hiroko; NARUSHIMA, Kenji; YAMAMOTO, Naoki. Efficient Prediction of Vitamin B Deficiencies via Machine-Learning Using Routine Blood Test Results in Patients With Intense Psychiatric Episode. **Frontiers in Psychiatry**, *[S. l.]*, v. 10, 2020. DOI: 10.3389/fpsyt.2019.01029. Disponível em: https://www.frontiersin.org/article/10.3389/fpsyt.2019.01029/full.

TESFAYE, Solomon; BOULTON, Andrew. **Diabetic Neuropathy**. [s.l.] : Oxford University Press, 2009. Disponível em: https://books.google.com.br/books?id=Os1XfgeFlCcC.

TESFAYE, Solomon; SELVARAJAH, Dinesh. **Advances in the epidemiology, pathogenesis and management of diabetic peripheral neuropathy**. **Diabetes/Metabolism Research and Reviews**, 2012. DOI: 10.1002/dmrr.2239.

Towards the Applied Hybrid Model in Decision Making: Support the Early Diagnosis of Type 2 Diabetes. [s.d.]. Disponível em: https://www.researchgate.net/publication/282648803_Towards_the_Applied_Hybrid_Model_in_Decision_Making_Support_the_Early_Diagnosis_of_Type_2_Diabetes. Acesso em: 29 maio. 2022.

TROISI, Rebecca J.; COWIE, Catherine C.; HARRIS, Maureen I. Diurnal Variation in Fasting Plasma Glucose. **JAMA**, *[S. l.]*, v. 284, n. 24, p. 3157, 2000. DOI: 10.1001/jama.284.24.3157. Disponível em: https://jamanetwork.com/. Acesso em: 4 jul. 2021.

VAN 'T RIET, Esther; ALSSEMA, Marjan; RIJKELIJKHUIZEN, Josina M.; KOSTENSE, Piet J.; NIJPELS, Giel; DEKKER, Jacqueline M. Relationship between A1C and glucose levels in the general Dutch population: The new Hoorn study.

Diabetes Care, *[S. l.]*, v. 33, n. 1, p. 61–66, 2010. DOI: 10.2337/dc09-0677.

WALJEE, Akbar K.; MUKHERJEE, Ashin; SINGAL, Amit G.; ZHANG, Yiwei; WARREN, Jeffrey; BALIS, Ulysses; MARRERO, Jorge; ZHU, Ji; HIGGINS, Peter D. R. Comparison of imputation methods for missing laboratory data in medicine. **BMJ Open**, *[S. l.]*, v. 3, n. 8, p. e002847, 2013. DOI: 10.1136/BMJOPEN-2013-002847. Disponível em: https://bmjopen.bmj.com/content/3/8/e002847. Acesso em: 14 dez. 2021.

WATT, Jeremy; BORHANI, Reza; KATSAGGELOS, Aggelos K. **Machine learning refined: Foundations, algorithms, and applications**. [s.l.] : Cambridge University Press, 2016. DOI: 10.1017/CBO9781316402276.

WENG, Stephen F.; REPS, Jenna; KAI, Joe; GARIBALDI, Jonathan M.; QURESHI, Nadeem. Can machine-learning improve cardiovascular risk prediction using routine clinical data? **PLOS ONE**, *[S. l.]*, v. 12, n. 4, p. e0174944, 2017. DOI: 10.1371/journal.pone.0174944. Disponível em: https://dx.plos.org/10.1371/journal.pone.0174944. Acesso em: 21 abr. 2020.

WILLIAMSON, M. A.; SNYDER, L. M. **Wallach - Interpretação De Exames Laboratoriais**. [s.l.] : GUANABARA, 2015. Disponível em: https://books.google.com.br/books?id=ksNPvgAACAAJ.

WONG, Jenna; MURRAY HORWITZ, Mara; ZHOU, Li; TOH, Sengwee. Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data. **Current Epidemiology Reports**, *[S. l.]*, v. 5, n. 4, p. 331–342, 2018. DOI: 10.1007/s40471-018-0165-9. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/30555773. Acesso em: 24 set. 2019.

WONG, William C. W.; CHEUNG, Catherine S. K.; HART, Graham J. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. **Emerging themes in epidemiology**, *[S. l.]*, v. 5, 2008. DOI: 10.1186/1742-7622-5-23. Disponível em: https://pubmed.ncbi.nlm.nih.gov/19014686/. Acesso em: 21 fev. 2022.

WU, Yan-Ting et al. Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning. **The Journal of Clinical Endocrinology**

& Metabolism, *[S. l.]*, v. 106, n. 3, p. e1191–e1205, 2021. DOI: 10.1210/clinem/dgaa899. Disponível em: https://academic.oup.com/jcem/article/106/3/e1191/6031346.

XAVIER, R. M.; DORA, J. M.; BARROS, E. **Laboratório na Prática Clínica - 3ed: Consulta Rápida**. [s.l.] : Artmed Editora, 2016. Disponível em: https://books.google.com.br/books?id=yE_WCwAAQBAJ.

XU, Song; HOM, Jason; BALASUBRAMANIAN, Santhosh; SCHROEDER, Lee F.; NAJAFI, Nader; ROY, Shivaal; CHEN, Jonathan H. Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests. **JAMA Network Open**, *[S. l.]*, v. 2, n. 9, 2019. DOI: 10.1001/JAMANETWORKOPEN.2019.10967.

YADAW, Arjun S.; LI, Yan-chak; BOSE, Sonali; IYENGAR, Ravi; BUNYAVANICH, Supinda; PANDEY, Gaurav. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. **The Lancet Digital Health**, *[S. l.]*, v. 2, n. 10, p. e516–e525, 2020. DOI: https://doi.org/10.1016/S2589-7500(20)30217-X. Disponível em: https://www.sciencedirect.com/science/article/pii/S258975002030217X.

YANG, He S. et al. Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning. **Clinical chemistry**, *[S. l.]*, v. 66, n. 11, p. 1396–1404, 2020. DOI: 10.1093/CLINCHEM/HVAA200. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32821907/. Acesso em: 5 fev. 2022.

YIU, Tony. **Understanding Random Forest - Towards Data Science**. **Understanding Random Forest How the Algorithm Works and Why it Is So Effective**, 2019. Disponível em: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Acesso em: 26 abr. 2020.

YU, Lishan; LI, Linda; BERNSTAM, Elmer; JIANG, Xiaoqian. A deep learning solution to recommend laboratory reduction strategies in ICU. **International Journal of Medical Informatics**, *[S. l.]*, v. 144, p. 104282, 2020. a. DOI: 10.1016/j.ijmedinf.2020.104282. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1386505620305980.

YU, Lishan; ZHANG, Qiuchen; BERNSTAM, Elmer V.; JIANG, Xiaoqian. Predict or

draw blood: An integrated method to reduce lab tests. **Journal of biomedical informatics**, *[S. l.]*, v. 104, 2020. b. DOI: 10.1016/J.JBI.2020.103394. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32113004/. Acesso em: 5 fev. 2022.

ZHENG, Tao; XIE, Wei; XU, Liling; HE, Xiaoying; ZHANG, Ya; YOU, Mingrong; YANG, Gong; CHEN, You. A machine learning-based framework to identify type 2 diabetes through electronic health records. **International Journal of Medical Informatics**, *[S. l.]*, v. 97, p. 120–127, 2017. DOI: 10.1016/j.ijmedinf.2016.09.014. Disponível em: http://www.ncbi.nlm.nih.gov/pubmed/27919371. Acesso em: 23 set. 2019.

APÊNDICE A - Histograma com a distribuição e boxplot de cada parâmetro (analito) analisado.

Figura 0.1 – Histograma e *Boxplot* da idade dos pacientes.

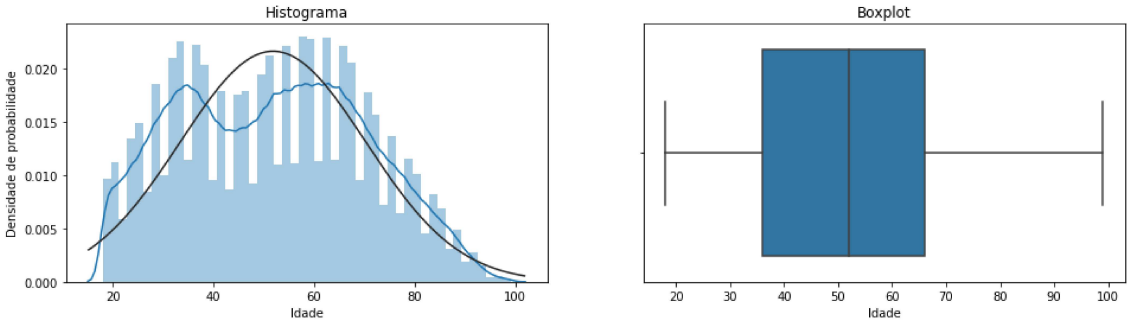


Figura 0.2 – Histograma e Boxplot do analito HEMOGRAMA_Hb.

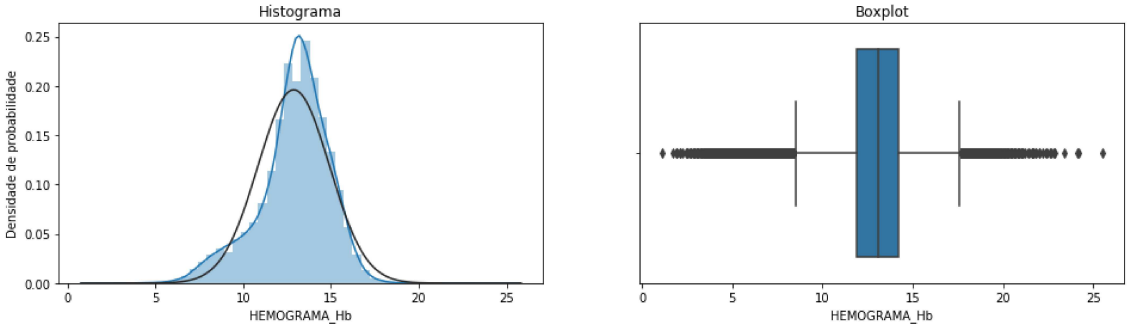


Figura 0.3 – Histograma e Boxplot do analito HEMOGRAMA_HT.

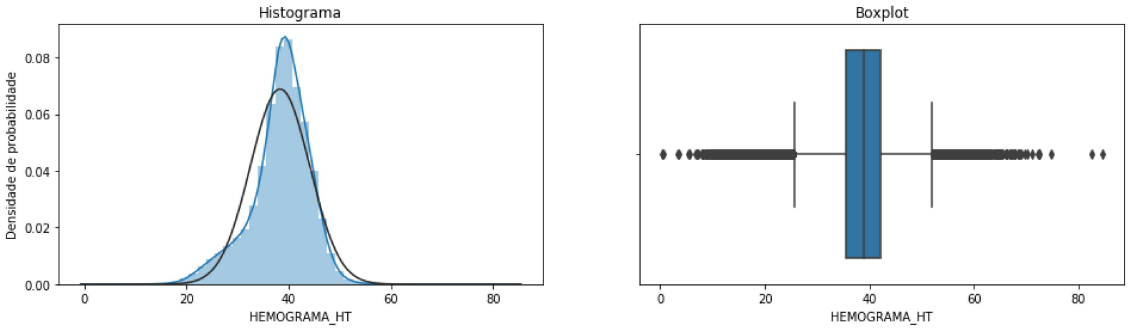


Figura 0.4 – Histograma e Boxplot do analito HEMOGRAMA_VCM.

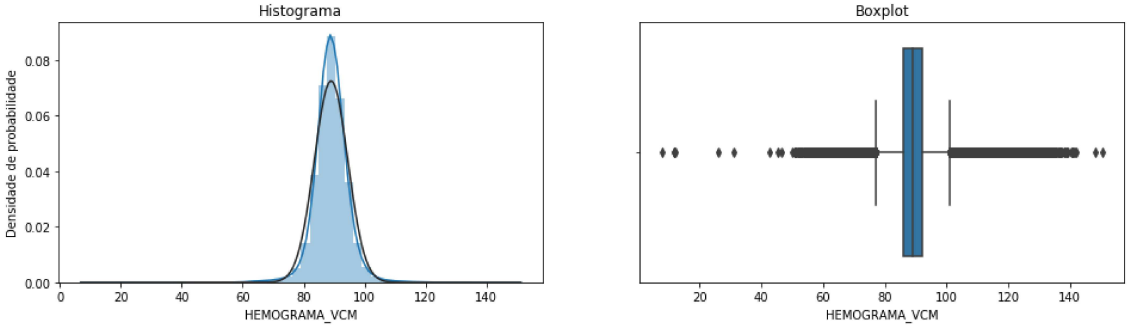


Figura 0.5 – Histograma e Boxplot do analito HEMOGRAMA_HCM.

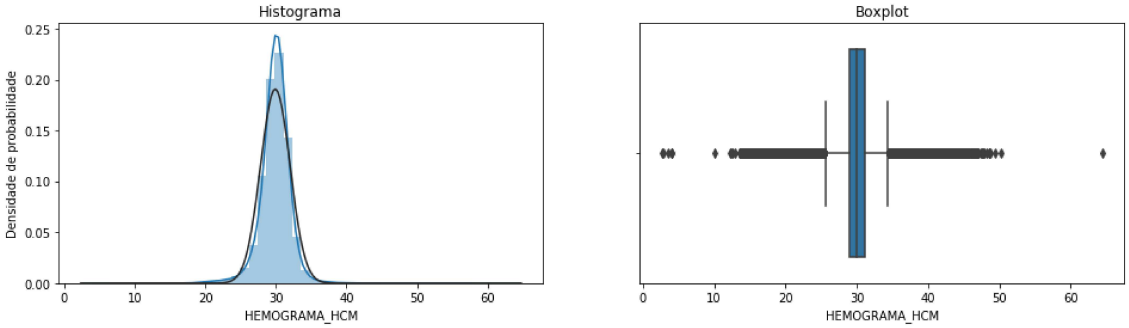


Figura 0.6 – Histograma e Boxplot do analito HEMOGRAMA_CHCM.

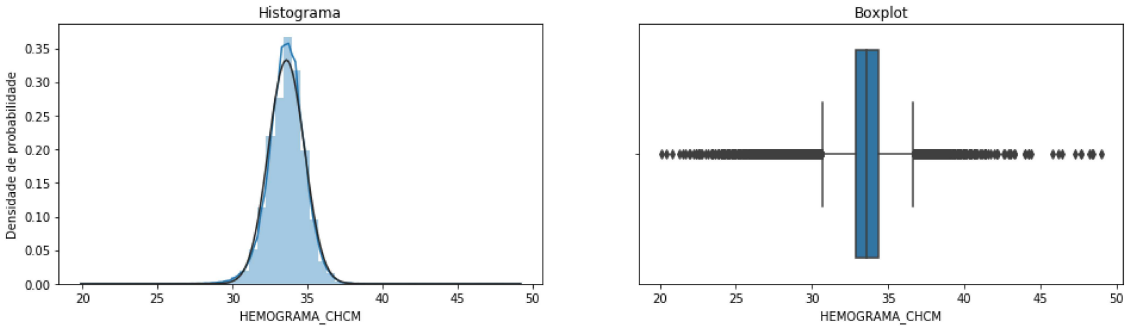


Figura 0.7 – Histograma e Boxplot do analito HEMOGRAMA_RDW.

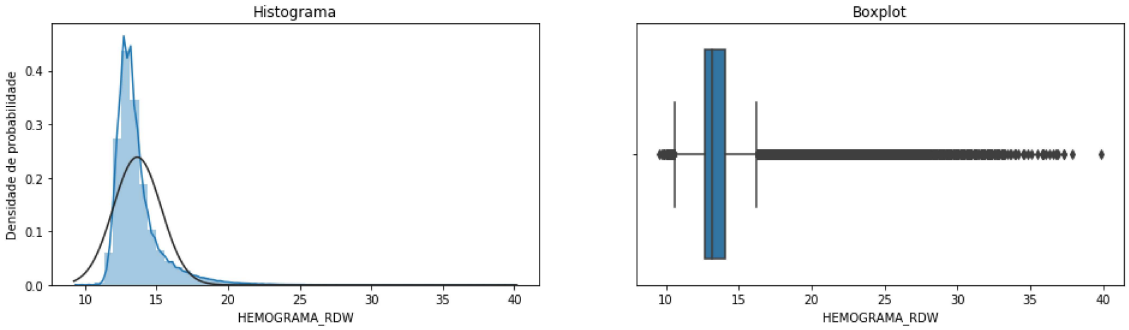


Figura 0.8 – Histograma e Boxplot do analito HEMOGRAMA_HMC.

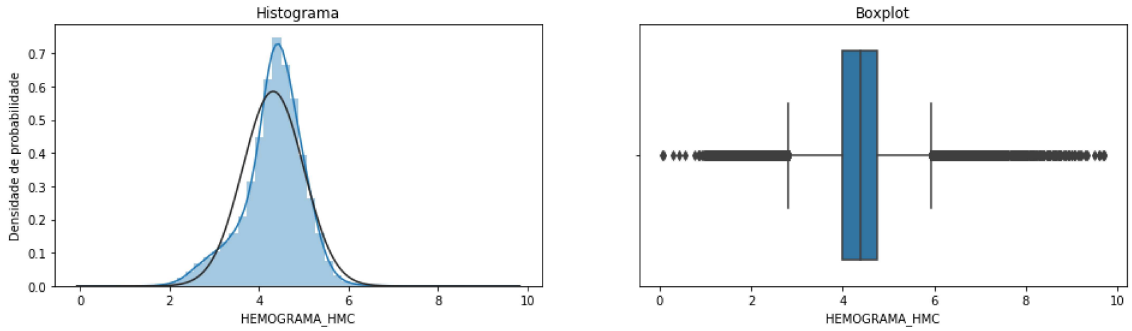


Figura 0.9 – Histograma e Boxplot do analito HEMOGRAMA_LEUC

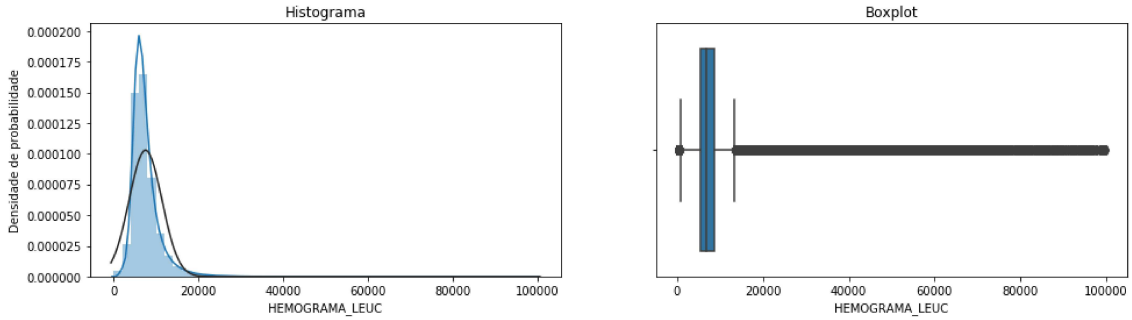


Figura 0.10 – Histograma e Boxplot do analito HEMOGRAMA_LINFO%.

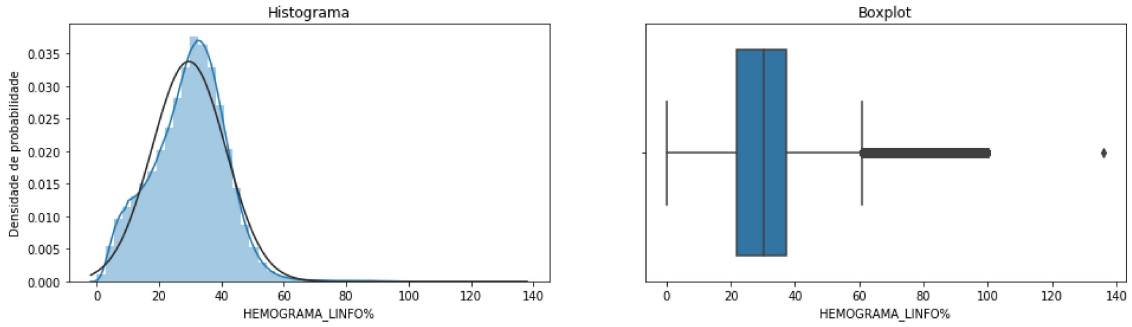


Figura 0.11 – Histograma e Boxplot do analito HEMOGRAMA_LINFOmm3.

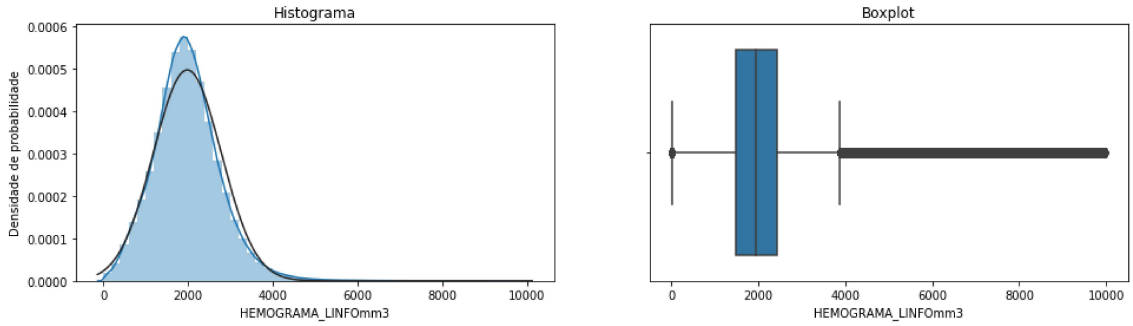


Figura 0.12 – Histograma e Boxplot do analito HEMOGRAMA_MONO%.

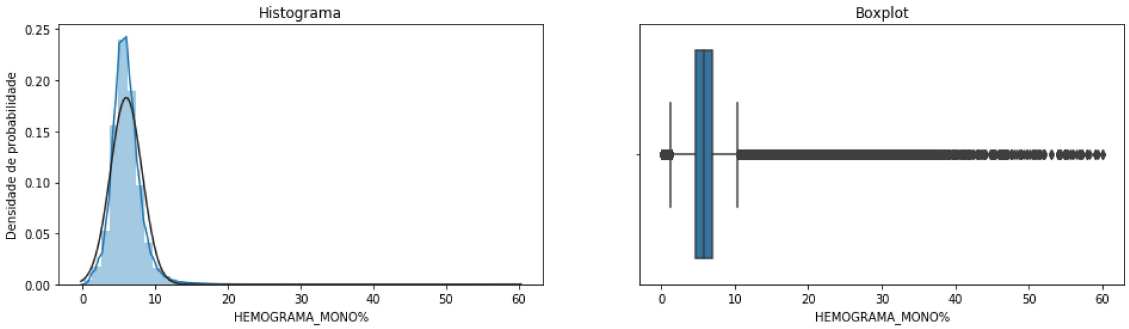


Figura 0.13 – Histograma e Boxplot do analito HEMOGRAMA_MONOmm3.

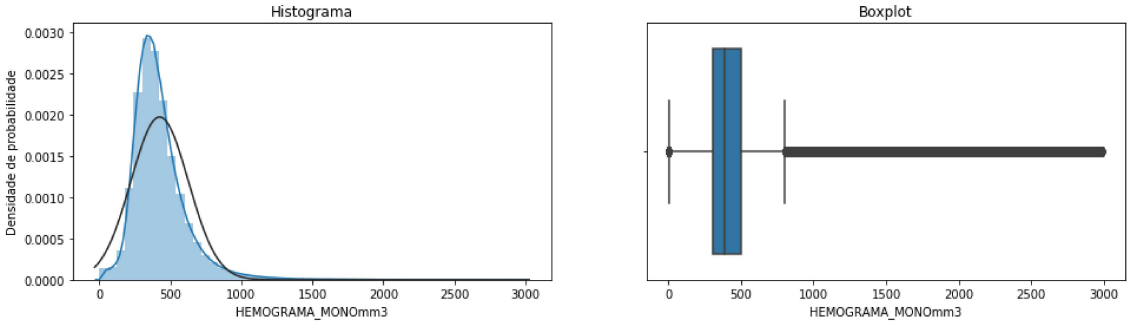


Figura 0.14 – Histograma e Boxplot do analito HEMOGRAMA_SEG%.

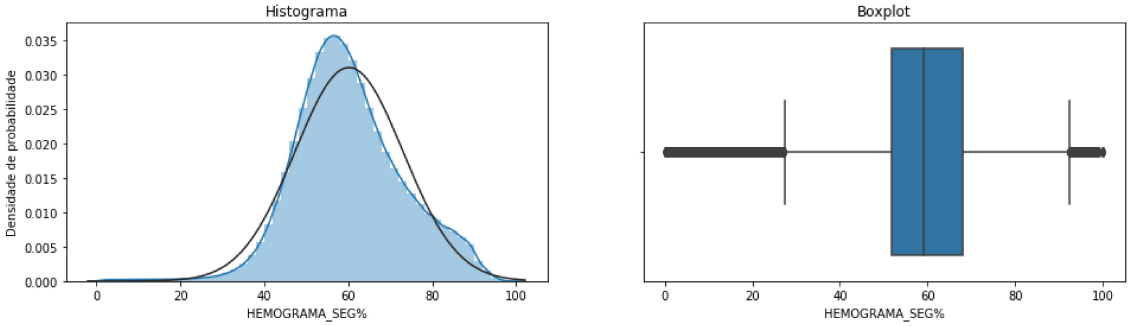


Figura 0.15 – Histograma e Boxplot do analito HEMOGRAMA_SEGmm3.

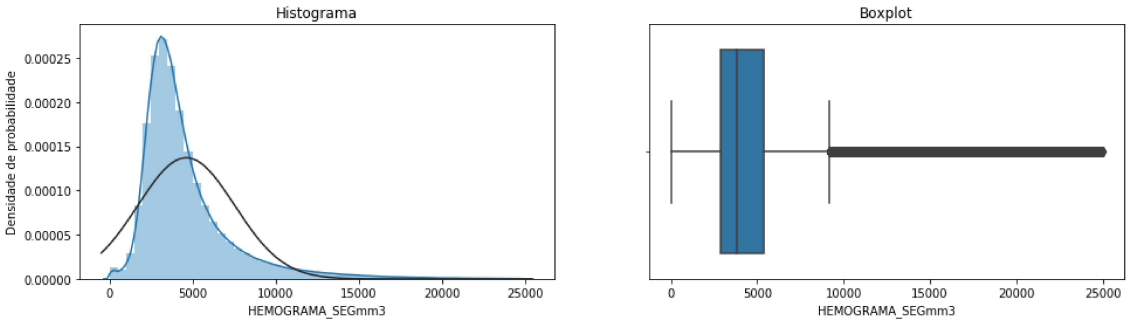


Figura 0.16 – Histograma e Boxplot do analito HEMOGRAMA_EOS%.

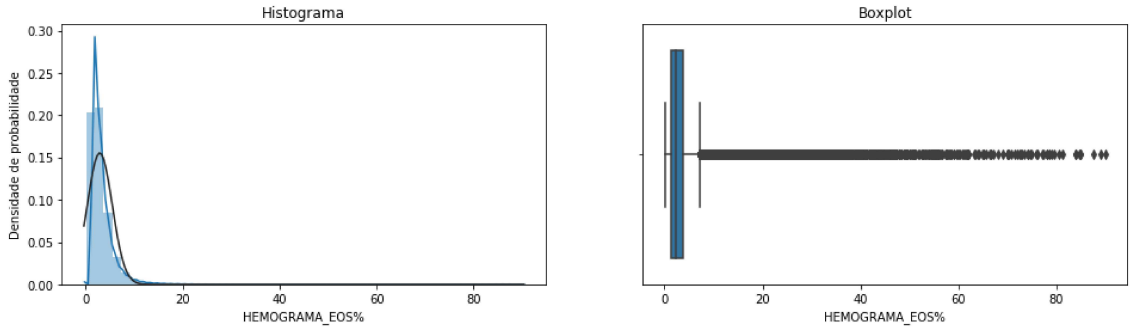


Figura 0.17 – Histograma e Boxplot do analito HEMOGRAMA_EOSmm3.

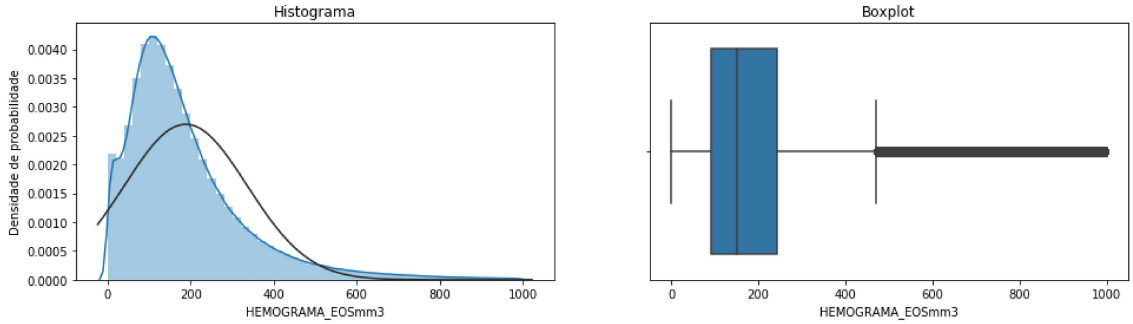


Figura 0.18 – Histograma e Boxplot do analito HEMOGRAMA_BASO%.

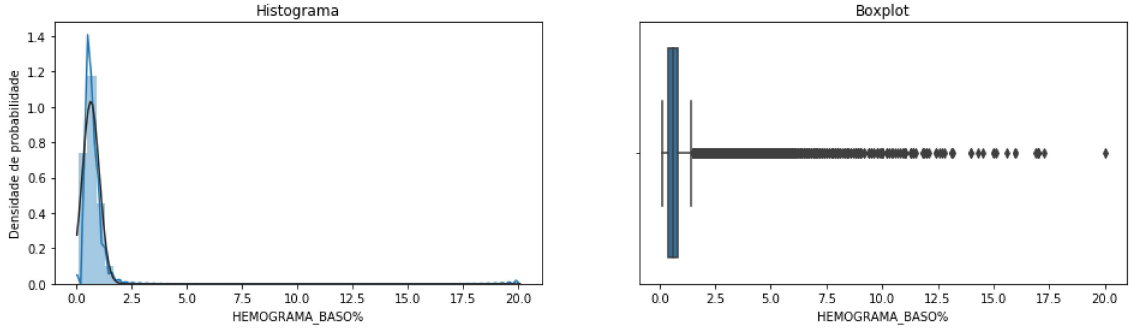


Figura 0.19 – Histograma e Boxplot do analito HEMOGRAMA_BASOmm3.

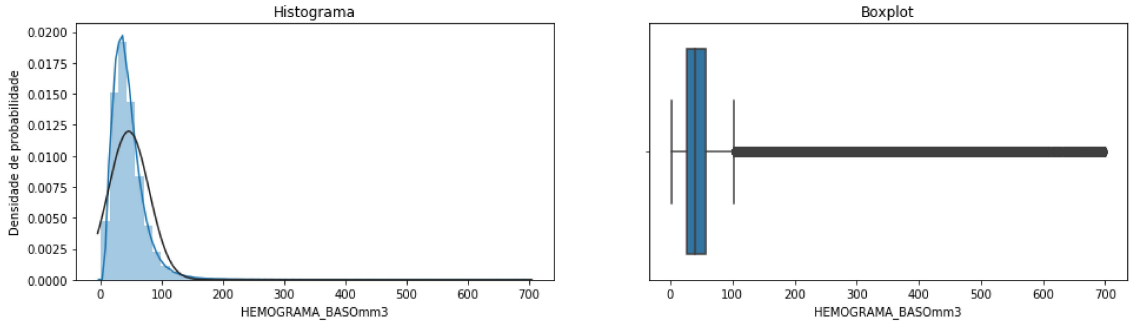


Figura 0.20 – Histograma e Boxplot do analito HEMOGRAMA_PLAQ.

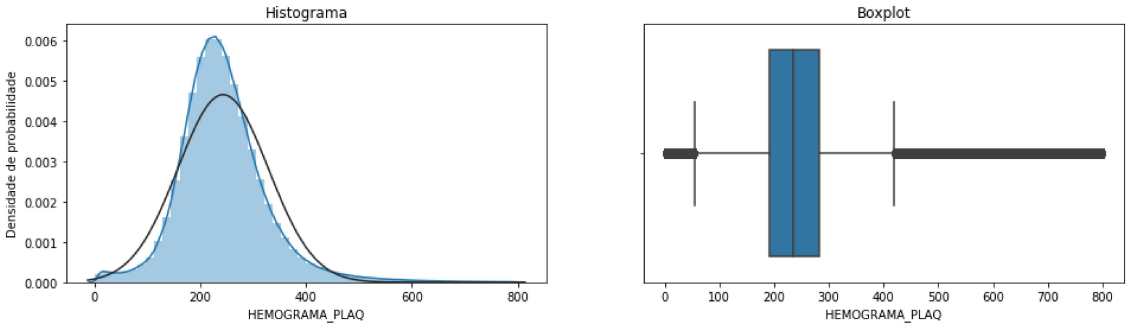


Figura 0.21 – Histograma e Boxplot do analito HEMOGRAMA_MPV.

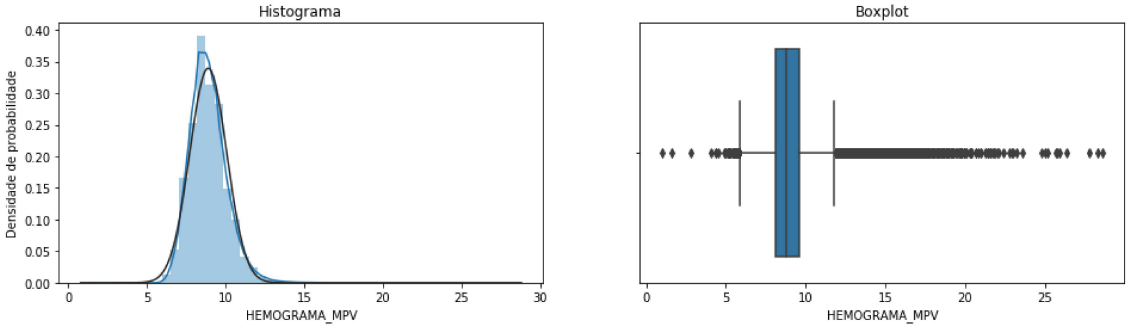


Figura 0.22 – Histograma e Boxplot do analito CREATININA EM SANGUE.

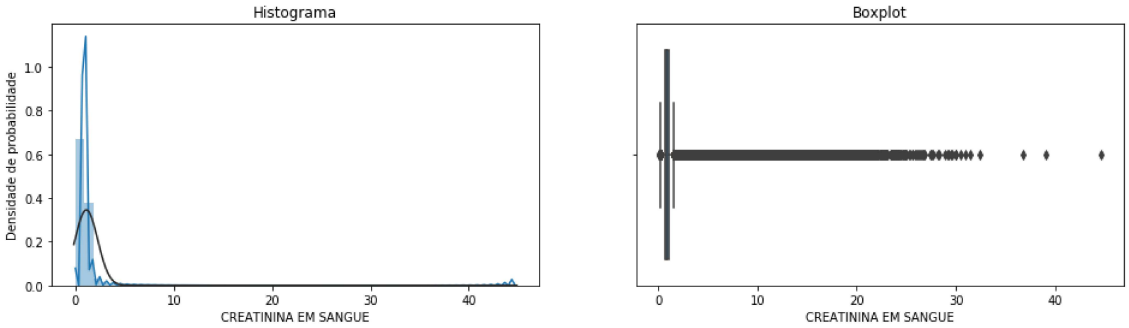


Figura 0.23 – Histograma e Boxplot do analito GLICOSE EM SANGUE.

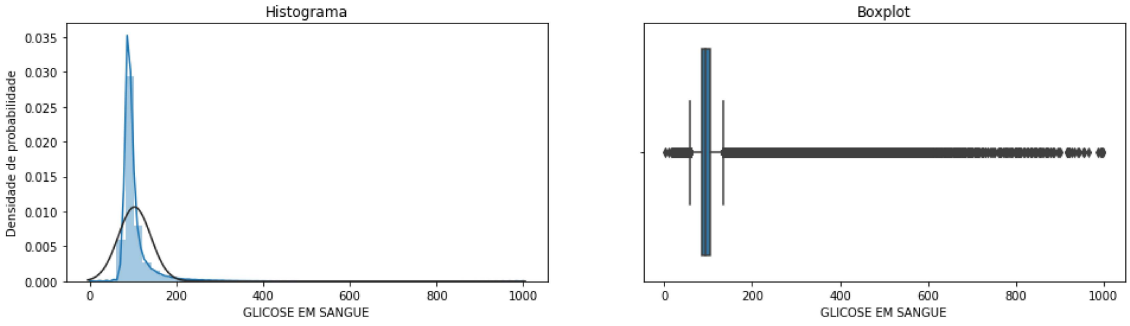


Figura 0.24 – Histograma e Boxplot do analito (TSH), HORMONIO.

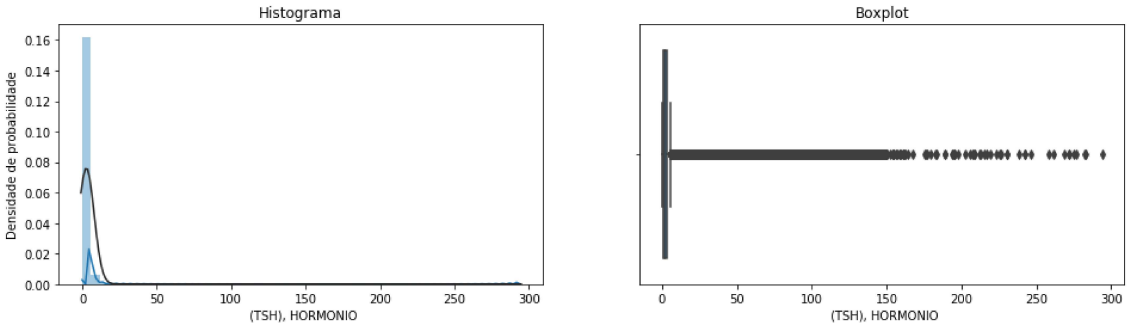


Figura 0.25 – Histograma e Boxplot do analito COLESTEROL TOTAL.

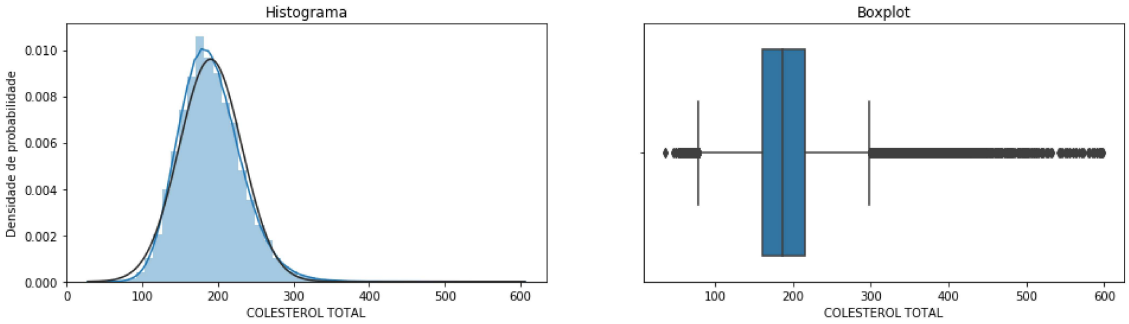


Figura 0.26 – Histograma e Boxplot do analito TRIGLICERIDEOS.

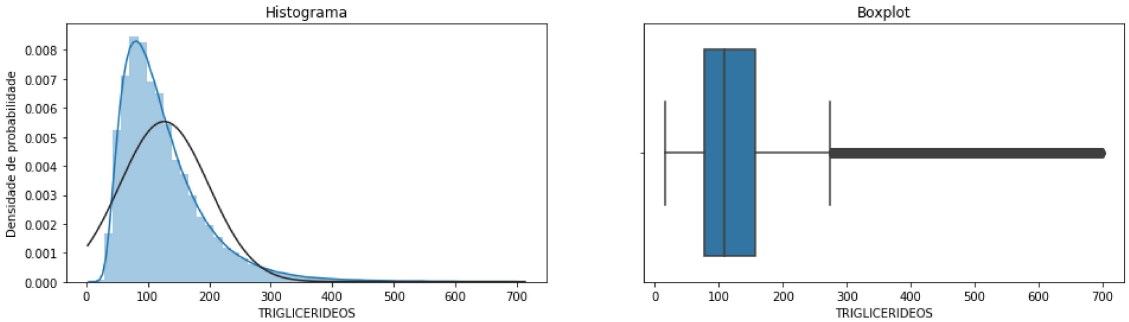


Figura 0.27 – Histograma e Boxplot do analito COLESTEROL HDL.

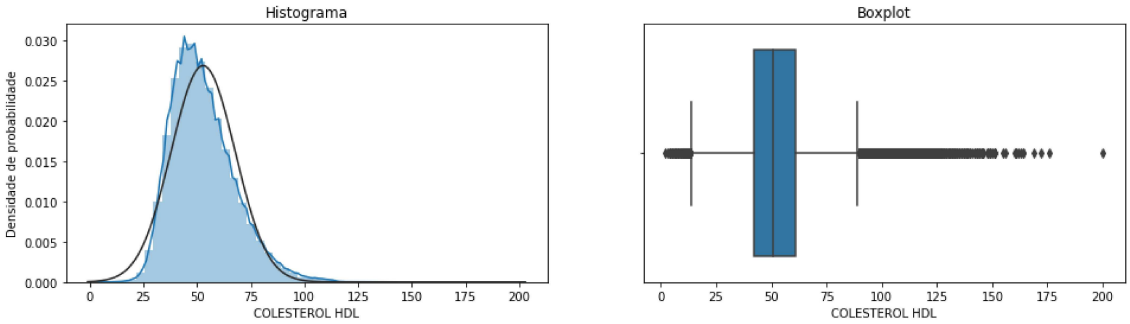


Figura 0.28 – Histograma e Boxplot do analito TRANSAMINASE ALT (GPT).

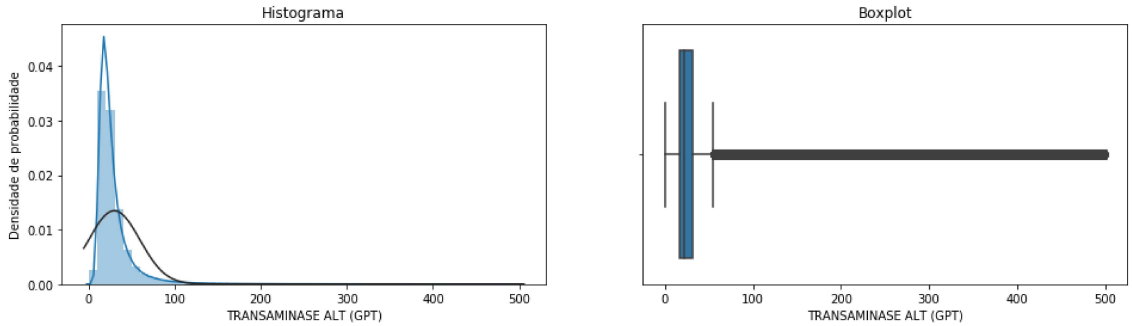


Figura 0.29 – Histograma e Boxplot do analito PARCIAL DE URINA_DENS

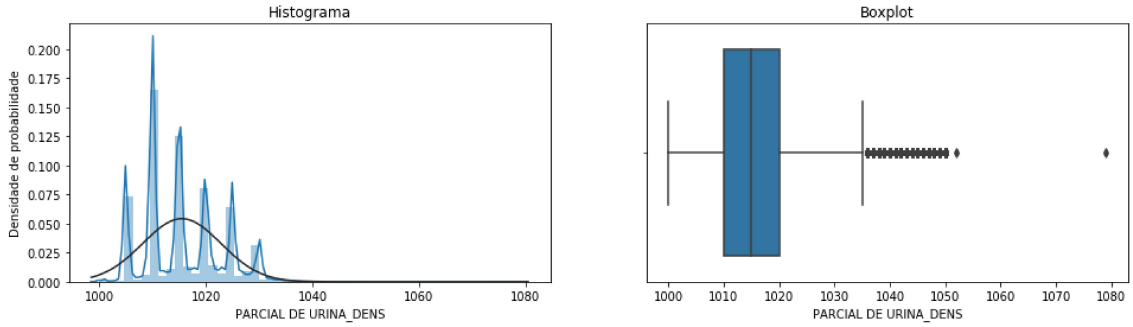


Figura 0.30 – Histograma e Boxplot do analito TRANSAMINASE AST (GOT.

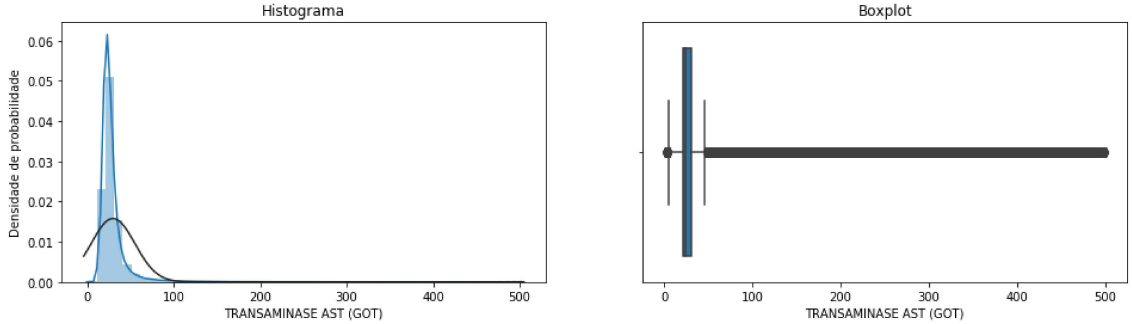


Figura 0.31 – Histograma e Boxplot do analito UREIA EM SANGUE.

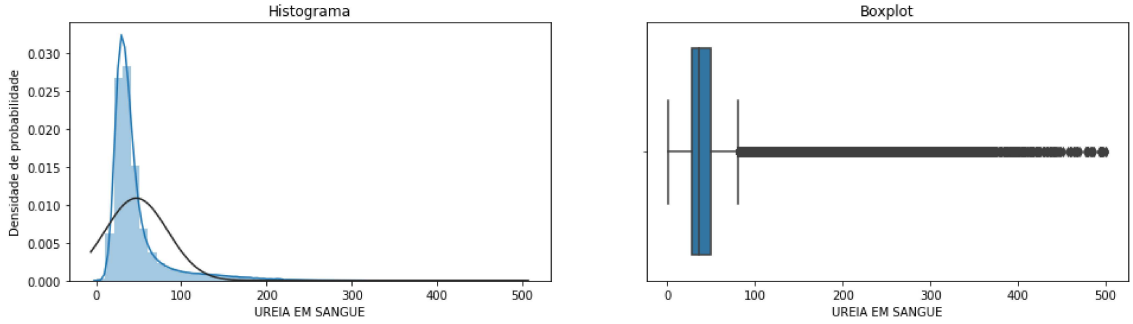


Figura 0.32 – Histograma e Boxplot do analito URINA_HEMAC.

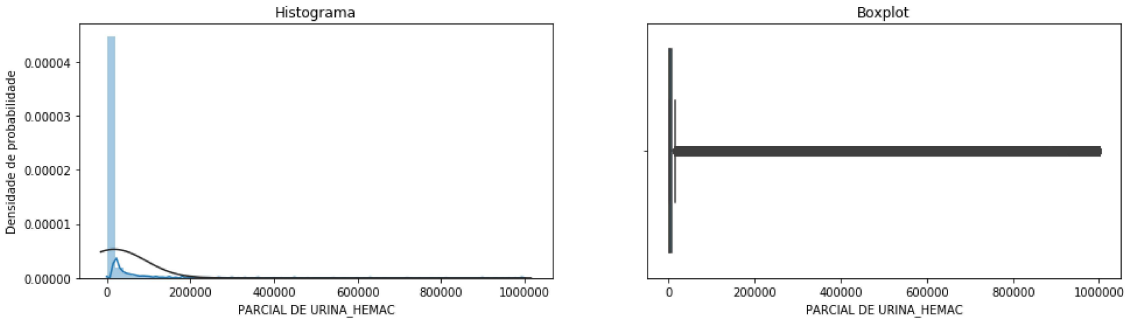


Figura 0.33 – Histograma e Boxplot do analito POTASSIO EM SANGUE.

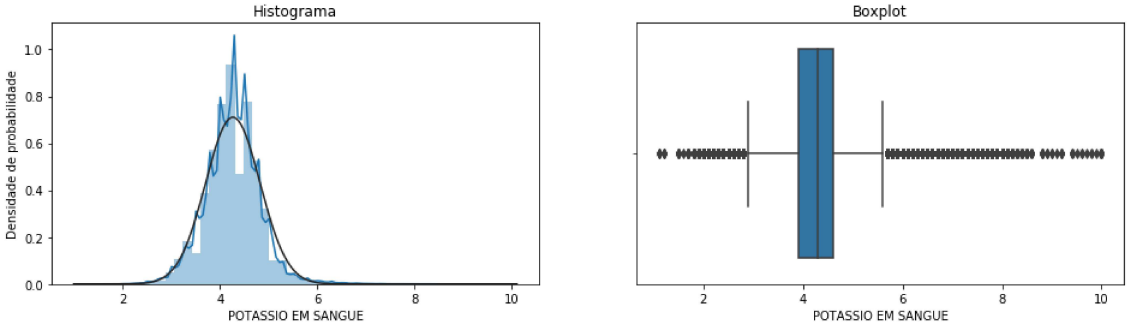


Figura 0.34 – Histograma e Boxplot do analito VITAMINA “D” 25 HIDROXI).

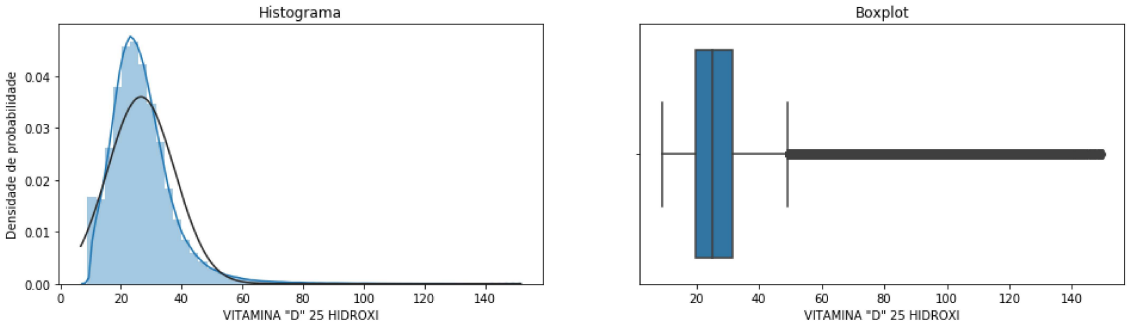


Figura 0.35 – Histograma e Boxplot do analito PARCIAL DE URINA_ph.

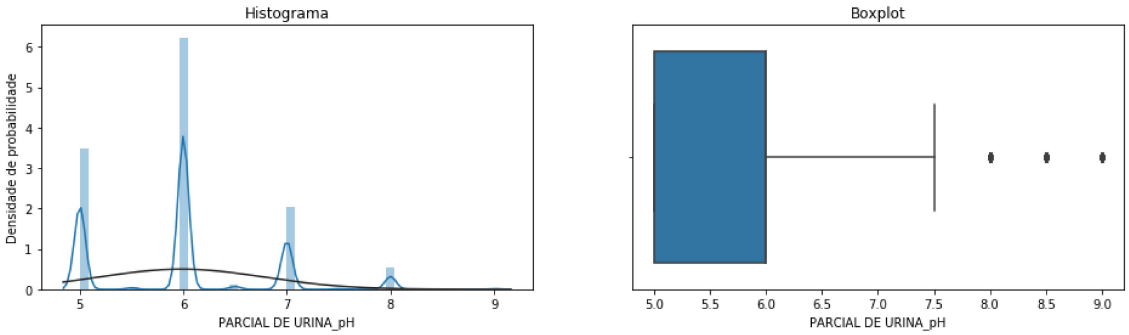


Figura 0.36 – Histograma e Boxplot do analito SODIO EM SANGUE.

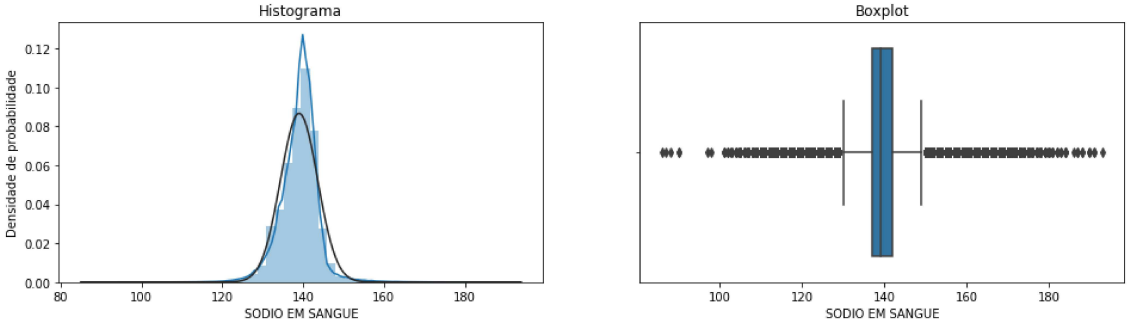


Figura 0.37 – Histograma e Boxplot do analito PROTEINA C REATIVA.

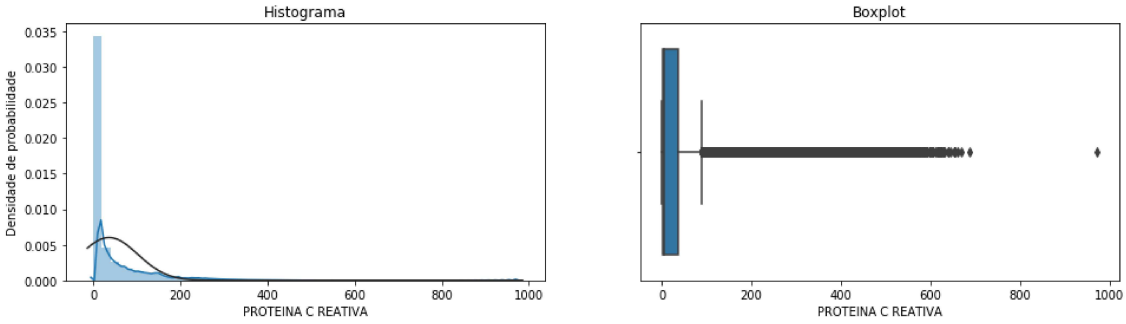


Figura 0.38 – Histograma e Boxplot do analito TIROXINA (T4) LIVRE.

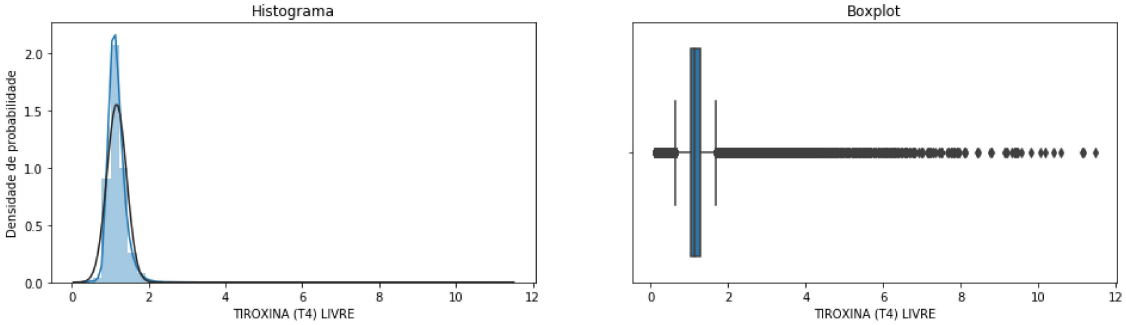


Figura 0.39 – Histograma e Boxplot do analito VITAMINA “B12”.

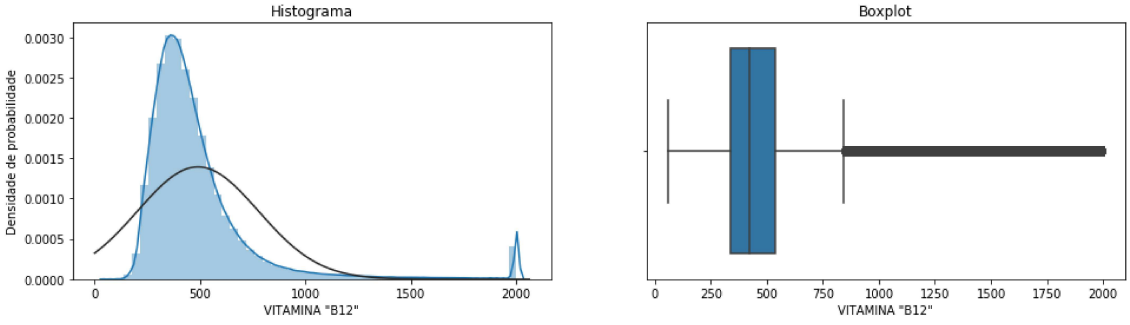


Figura 0.40 – Histograma e Boxplot do analito ACIDO URICO SANGUINEO

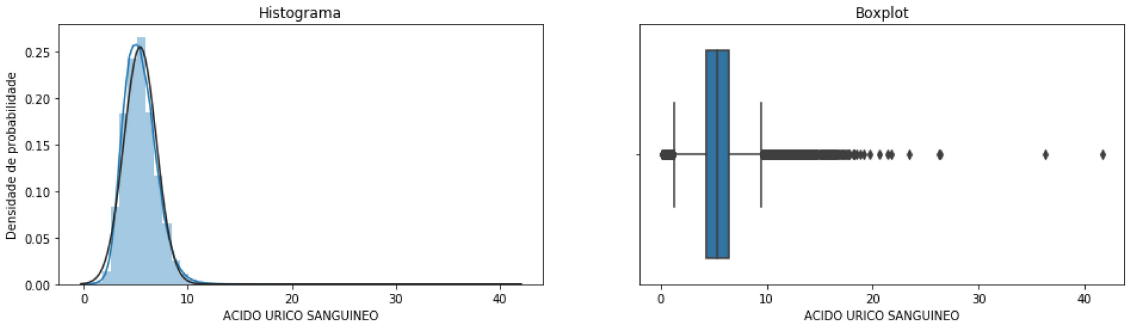


Figura 0.41 – Histograma e Boxplot do analito HEMOGLOBINA GLICADA.