UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
CURSO DE GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Gabriel Frank Simonetto

Estudo de possíveis discriminantes de TEAF utilizando dados do DATASUS

Florianópolis
[2022]

Gabriel Frank Simonetto

Estudo de possíveis discriminantes de TEAF utilizando dados do DATASUS

Trabalho de Conclusão de Curso do Curso de Graduação em Ciências da Computação do Campus Florianópolis da Universidade Federal de Santa Catarina para a obtenção do título de bacharel em Ciências da Computação.
Orientador: Prof. Dr. Mateus Grellert
Coorientador: Prof. Dr. Jônata Tyska Carvalho

Florianópolis
[2022]

Gabriel Frank Simonetto

Estudo de possíveis discriminantes de TEAF utilizando dados do DATASUS

Florianópolis, 25 de julho de 2022.

Coordenador do Curso

Banca Examinadora:

Prof. Dr. Mateus Grellert
Orientador

Prof. Dr. Jônata Tyska Carvalho
Coorientador

Profª. Drª. Carina Friedrich Dorneles
Membro da banca

Profª Drª. Patricia Della Méa Plentz
Membro da banca

RESUMO

Mesmo após inúmeros avanços nas Ciências da Saúde, diversos tipos de anomalias congênitas ainda são consideradas subdiagnosticadas devido à dificuldade do diagnóstico e à baixa visibilidade dos casos. A falta de diagnóstico impede tanto o tratamento de recém-nascidos e crianças quanto a busca por políticas públicas que diminuam a incidência de casos. Neste cenário, podemos destacar o Transtorno do Espectro Alcoólico Fetal (TEAF) como um caso importante de anomalia congênita subdiagnosticada. Esse transtorno é ocasionado principalmente pelo consumo de bebida alcoólica pela gestante durante a gravidez, causando microcefalia, dismorfias faciais e déficit neurocognitivo. O diagnóstico de TEAF é desafiador, uma vez que ele é multifatorial e que os sintomas apresentados por esse transtorno são comuns a outros transtornos. A literatura aponta um crescimento no uso de técnicas de Inteligência Artificial (IA) para apoiar profissionais da saúde que atuam na área de diagnóstico e tratamento de doenças e transtornos complexos. Portanto, este trabalho propõe o uso de mineração de dados e aprendizado de máquina para promover o diagnóstico assistido de TEAF, utilizando a base dados do DATASUS com informações do nascimento de crianças e fatores sociodemográficos. Até onde sabemos, este é o primeiro trabalho que busca estudar SAF usando aprendizado de máquina. Inicialmente uma análise de dados é realizada, a fim de ampliar o conhecimento sobre o transtorno. Os principais achados desta etapa são uma análise das variáveis mais relevantes segundo a teoria atual, na qual boa parte das variáveis falha em ser apoiada pelos dados como discriminante entre indivíduos SAF positivos e negativos. Em seguida, definimos e usamos técnicas de amostragem de dados para garantir o balanceamento de classes, assim como aumentar a confiança estatística dos achados realizados neste trabalho. Por último, é realizada uma modelagem preditiva para detectar casos prováveis de TEAF logo após o parto. Os modelos treinados apresentam métricas modestas, e levantam feature importances com poucas variáveis de alta importância para o valor preditivo do modelo. Este trabalho busca contribuir para a detecção dos casos de alto risco TEAF, identificando seus biomarcadores mais proeminentes, assim como fatores regionais que interfiram na incidência deste transtorno, permitindo a criação de políticas de intervenção adequadas para cada localidade.

Palavras-chave: Inteligência artificial, mineração de dados, aprendizado de máquina, TEAF.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As anomalias congênitas são um grupo de alterações estruturais ou funcionais que ocorrem durante a vida intrauterina e que podem ser detectadas antes, durante ou após o nascimento (MENDES *et al.*, 2018). Elas podem afetar diversos órgãos e sistemas do corpo humano e são causadas por um ou mais fatores genéticos, infecciosos, nutricionais e ambientais, podendo ser resultado de uma combinação desses fatores.

Exemplos de anomalias congênitas envolvem os Transtornos do Espectro Alcoólico Fetal (TEAF). Estes transtornos incluem uma vasta gama de condições patológicas relacionadas ao consumo de álcool durante a gestação. O espectro de distúrbios pode incluir defeitos congênitos, distúrbios neurológicos relacionados ao álcool, assim como a manifestação mais grave, a Síndrome Alcoólica Fetal (SAF) (COOK *et al.*, 2016).

Não há cura para os TEAF, portanto a forma mais efetiva de combater esses transtornos é evitando sua ocorrência durante a gestação. Uma estratégia para isso é a conscientização dos riscos de consumo de álcool durante a assistência pré-natal. Porém, a assistência pré-natal no Brasil ainda carece do desenvolvimento de rotinas e instrumentos confiáveis que auxiliem os profissionais de saúde nas ações de prevenção e diagnóstico precoce para esses problemas (LOPES *et al.*, 2016).

Mesmo sem cura definitiva, o tratamento dos sintomas de TEAF é fundamental para garantir o desenvolvimento sadio de crianças afetadas por esse transtorno. Devido ao cenário adverso de incapacidades cognitivas e comportamentais, os indivíduos com SAF se beneficiam muito de acesso a serviços de apoio, abrandando assim, as manifestações adversas do espectro (QUEIROZ, 2016). Quando se tem acesso ao diagnóstico, torna-se possível procurar o amparo necessário para a criança.

O diagnóstico médico é amparado pela presença ou ausência de biomarcadores(discriminantes biológicos), que são medidas ou indicadores médicos que avaliam a presença progressão ou resposta a uma determinada condição ou doença. O principal desafio para diagnóstico precoce de TEAF vem do fato de serem transtornos multifatoriais, afetados por biomarcadores e fatores socioeconômicos que são compartilhados por outros transtornos (RILEY; INFANTE; WARREN, 2011). Portanto, soluções que buscam identificar recém-nascidos afetados por TEAF são de extrema importância.

A literatura demonstra que técnicas de Inteligência Artificial (IA) podem ser fortes aliadas nessa linha, gerando modelos capazes de auxiliar no diagnóstico e no tratamento destes transtornos (HEERINGEN; MANN, 2014; ARSENAULT-LAPIERRE; KIM; TURECKI, 2004). Portanto, o desenvolvimento de ferramentas e tecnologias de IA capazes de apoiar os profissionais de saúde que atuam na área, assim como a democratização ao acesso à informação para a população geral, são recursos vitais para o combate de problemas da saúde mental.

O desenvolvimento de soluções em IA tem sido amparado pela disponibilidade de

dados, e o Brasil também tem trabalhado em ações nessa direção. Como exemplo, podemos citar as recentes iniciativas para acesso aberto a dados de saúde por parte do governo brasileiro, providenciando um amplo espaço amostral de dados, que possibilitam a realização de um trabalho de análise e busca por conhecimento orientado pelos processos de *data mining*. O Departamento de Informática do Sistema Único de Saúde (DATASUS) fornece diferentes dados digitais sobre a saúde através de suas APIs (*Application Programming Interface*)(SOUZA; AUTRAN *et al.*, 2019). Em 2020, o DATASUS coordenou a criação da Rede Nacional de Dados em Saúde (RNDS), com uma API que disponibiliza acesso a vários dados de saúde (informações sobre pacientes, dados de exame, do examinador, etc). Infelizmente, essa API só é acessível a estabelecimentos de saúde.

Esse trabalho procura encontrar discriminantes elucidativas sobre quais fenômenos, acontecimentos, e correlações estão envolvidos com o surgimento de um diagnóstico de TEAF assim como das outras anomalias a serem estudadas. Até onde sabemos, este é o primeiro trabalho que busca estudar SAF usando aprendizado de máquina.

O link para o código fonte usado no desenvolvimento desse trabalho pode ser encontrado no apêndice C.

1.1 OBJETIVOS

O foco principal do trabalho é o desenvolvimento de um sistema de mineração de dados aplicado à base de dados do DATASUS, procurando o enriquecimento dos dados a partir dos preceitos de *data mining*. Uma vez criada essa base de dados enriquecida, esse trabalho também usa técnicas de aprendizado de máquina para a criação de modelos preditivos que são usados tanto para a extração de conhecimento, entendendo as variáveis de maior significância para o modelo, quanto para o uso posterior de maneira prática, procurando identificar casos de risco na população.

1.1.1 Objetivos Específicos

1. Extrair os dados relevantes da plataforma do DATASUS e fazer uma análise exploratória sobre as variáveis relevantes de acordo com a literatura;

2. Realizar o pré-processamento dos dados coletados no objetivo anterior. Verificar balanceamento dos conjuntos de dados, dados faltantes, dados incorretos, dados discrepantes e então aplicar as respectivas técnicas de correção para cada problema;

3. Treinamento de modelos de aprendizado de máquina com base nos dados processados;

4. Validação e avaliação do desempenho dos modelos treinados;

5. Divulgação e publicação do Trabalho de Conclusão de Curso e da apresentação de slides sobre o mesmo;

1.2 JUSTIFICATIVA

O transtorno do Espectro Alcoólico Fetal (TEAF) inclui a vasta gama de condições patológicas que podem ocorrer quando o álcool é consumido durante diferentes períodos da gravidez. Essas condições são geralmente acompanhadas por danos encefálicos estruturais e funcionais. As crianças afetadas têm problemas de aprendizagem, memória, atenção, linguagem, comportamento e dificuldade de se relacionarem com os outros.

Apesar disso, atualmente não há políticas públicas de prevenção e protocolos para diagnóstico precoce reconhecidos para indivíduos com TEAF no Brasil. Considerando o grande fardo emocional deste transtorno, além dos altos custos financeiros destes indivíduos para a sociedade, justifica-se o estudo da incidência do consumo de álcool entre gestantes e a associação com partos prematuros.

O desenvolvimento de um modelo que usando os dados da gestante e do recém-nascido avalie o risco de desenvolver o TEAF usando aprendizado de máquina poderá auxiliar no desenvolvimento de políticas públicas e entendimento científico, uma vez que o uso de estatística massificado em projetos de aprendizado de máquina é capaz de desvendar e definir correlações que antes não haviam sido percebidas.

Assim como a análise massificada de dados é capaz de derivar os alicerces principais da análise, também pode-se estudar como a estratificação de certos grupos altera o perfil de análise do problema. A título de exemplo, é possível traçar como diferentes regiões são afetadas pela TEAF, e se existem diferenças entre quais fatores são mais relevantes em diferentes áreas, o que imediatamente demonstra que a proposta de intervenção, deve ser, também, adaptada.

No Brasil, a incidência e prevalência dos TEAF são subnotificados e ainda há muito desconhecimento sobre o tema entre profissionais de saúde. Os dados coletados das gestantes e dos recém-nascidos ainda são muito insipientes e o diagnóstico precoce muitas vezes é negligenciado. Com o auxílio da plataforma proposta, os profissionais da saúde poderão ser direcionados a ter um novo olhar para os múltiplos diagnósticos ao qual os recém nascidos podem possuir, auxiliando o profissional numa tarefa de alta complexidade. O que beneficia tanto os indivíduos que podem contar com ajuda especializada, quanto o sistema informacional dos órgãos públicos, que recebem dados mais assertivos.

O trabalho também trará contribuições para a comunidade científica, já que fará o uso de técnicas de aprendizado de máquina através de uma metodologia customizada para os temas deste projeto. Além de criar novos conjuntos de dados com dados agregados de fontes diferentes. Essa contribuição é importante, pois serve como inspiração, ou referência, para outros projetos similares que aplicam conceitos de computação na área da saúde.

1.3 ESTRUTURA DO TRABALHO

O trabalho a seguir está organizado da seguinte forma:

- Capítulo 2: explica alguns conceitos que são usados no trabalho e faz uma revisão da literatura científica sobre SAF;
- Capítulo 3: apresenta a solução proposta e explica brevemente os passos seguidos no trabalho;
- Capítulo 4: faz uma análise exploratória dos dados a serem utilizados, baseado em discriminantes proeminentes de SAF segundo a literatura;
- Capítulo 5: discorre extensivamente sobre as decisões técnicas encontradas ao trabalhar com os dados de acordo com cada etapa apresentada na solução proposta;
- Capítulo 6: inicia a conclusão do texto ao apresentar as métricas e discriminantes obtidas na execução dos modelos;
- Capítulo 7: consolida as conclusões do trabalho e apresenta trabalhos futuros;

2 REFERENCIAL TEÓRICO

O objetivo deste capítulo é elucidar os principais conceitos relativos a esse trabalho, assim como discutir trabalhos relacionados que servirão como base para seu desenvolvimento. Primeiro, serão apresentados os conceitos básicos na seção 2.1. Já a seção 2.2 apresenta alguns dos trabalhos que exploraram a análise de SAF recentemente.

2.1 CONCEITOS BÁSICOS

Aqui exploram-se alguns conceitos importantes para o entendimento completo das seções posteriores desse trabalho.

2.1.1 SAF e TEAF

Escolhida como a anomalia congênita (uma alteração estrutural ou funcional que ocorrem durante a vida intrauterina) que será usada de modelo nesse trabalho, o SAF (Síndrome Alcoólica Fetal) é uma condição decorrente do consumo de álcool durante a gravidez pela mãe.

Segundo (QUEIROZ, 2016), "A Síndrome Alcoólica Fetal se caracteriza pela tríade microcefalia-dismorfias faciais-déficit neurocognitivo.". O SAF foi a primeira manifestação documentada do TEAF(Transtorno do Espectro Alcoólico Fetal), devido ao fato de ser o caso mais extremo do espectro, no qual o diagnóstico a partir da dismorfia facial é possível, entretanto, os déficits associados ao consumo de álcool podem tambem se manifestar em graus de intensidade menor.

Ao longo do tempo, o TEAF também foi estudado e definido pela comunidade científica como uma fonte de problemas para o portador em diversas áreas da vida como a saúde mental, educação, comportamento criminal e independência dos indivíduos (RILEY; INFANTE; WARREN, 2011).

Não existe ainda uma definição da comunidade científica acerca de uma quantidade segura de ingestão de álcool durante a gravidez, e, boa parte das pesquisas chama a atenção para o fato de que SAF, é uma das poucas anomalias congênitas com possibilidade total de intervenção, uma vez que existem somente o fator ambiental de consumo de álcool envolvido. (QUEIROZ, 2016)

2.1.2 Plataformas de dados de saúde

Nesse trabalho, usaremos diversas ferramentas relacionadas ao setor público de saúde no Brasil, boa parte dos sistemas coletores de dados, como o DATASUS, estão ativos há um tempo considerável. Segundo o site do DATASUS, o Departamento de Informática do Sistema Único de Saúde (DATASUS) em 1991, e desde então, foram desenvolvidos

mais de 200 sistemas que auxiliam a construção e fortalecimento do Sistema Único de Saúde (SUS) (SOBREDATASUS2022. . . , s.d.).,

Se por um lado, iniciativas como o DATASUS tem certa história de sucesso, foi somente recentemente, através da lei de dados abertos (LEIDADOS. . . , s.d.), e do decreto da política de dados abertos (POLITICADADOS. . . , s.d.) que começaram as iniciativas para que esses dados se tornassem acessíveis para a comunidade científica e a população em geral. Através dessas iniciativas, as bases de dados dos inúmeros serviços do DATASUS tornaram-se disponíveis pela comunidade. Foi dessa forma que durante a pandemia de COVID-19 tiveram-se notícias de diversas iniciativas independentes que puderam usar os dados públicos do Brasil para criar infogramas públicos. O que criou uma onda de ferramentas que buscavam fazer previsões e visualizações de dados acerca da pandemia. (COELHO; ARAUJO MORAIS; SILVA ROSA *et al.*, 2020) (VALENTIM *et al.*, 2021)

Neste trabalho, usaremos o *PySUS* (COELHO; BARON *et al.*, 2021) como API (*Application Programming Interface*) que permite o acesso aos dados do DATASUS. Ela foi criada inicialmente para auxiliar nas pesquisas envolvendo os casos de dengue no Brasil através de análises sobre os dados do banco Sistema de Informações de Agravos de Notificação (SINAN). A versão atual da biblioteca *PySUS* possui acesso a diversos outros bancos, permitindo que ela seja utilizada em diversas pesquisas voltadas à saúde.

A figura 1 representa a estrutura do DATASUS e módulos, de diferentes setores da saúde que estão ligados a ele.

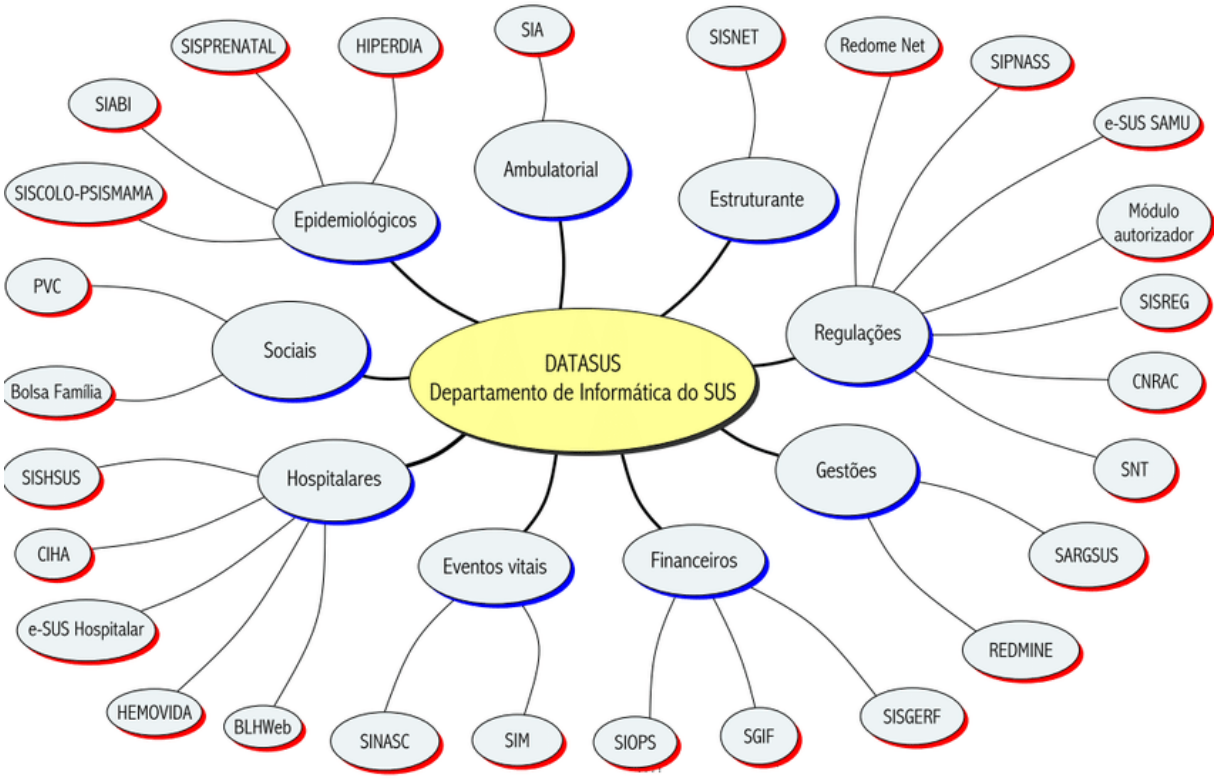


Figura 1 – Mapa Conceitual - Alimentação do DATASUS (SOUZA; AUTRAN *et al.*, 2019)

Mais em específico para nosso caso, estamos interessados em particular no Sistema de Informações sobre Nascidos Vivos (SINASC) (SITESINASC..., s.d.) que é responsável por registrar variadas informações sobre o trabalho de parto, e condições da mãe e da criança. Este sistema é instrumental para a construção de indicadores úteis para o planejamento de gestão dos serviços de saúde.

2.1.3 Aprendizado de máquina

Segundo (MITCHELL, 1999), a redução do custo de armazenamento de grandes volumes de dados, a facilidade de coletar dados através das redes, a redução do custo de poder computacional, e o desenvolvimento de algoritmos de aprendizado de máquina eficientes; fizeram com que fosse possível o uso de dados históricos extensivos para a análise de dados, previsão de cenários futuros, e detecção de irregularidades.

Alguns exemplos de aplicações que podem ser criadas através da análise de dados históricos seriam a detecção de fraude em cartão de crédito, prever o comportamento de consumidores, detecção de mensagens de *spam* em e-mail. Entretanto, mais recentemente, os fatores apontados por Mitchell apenas se acentuaram, e campos mais e mais complexos podem ser atacados, usando os mesmos princípios e algoritmos, ainda que em versões mais evoluídas.

Um algoritmo de aprendizado de máquina, ou aprendizado de máquina, é um algoritmo capaz de derivar conhecimento a partir de observações (GRELLERT, 2018). Existem diversos mecanismos matemáticos e estatísticos capazes de compilar as correlações e efeitos das variáveis entre si. Exploram-se nesse capítulo alguns desses mecanismos, quando aplicados em um algoritmo, também chamados de "modelo".

Esse trabalho irá aplicar técnicas de aprendizado de máquina supervisionado, pois está disponível se um bebê foi diagnosticado com SAF, ou outra anomalia congênita, mas também é possível usar-se técnicas com definições menos sólidas do que se busca no *dataset*. Por exemplo, os modelos não supervisionados, que não usam um gabarito para a definição de sucesso de um problema.

Além disso, existem os modelos de aprendizado por reforço, que procuram desempenhar uma tarefa do melhor jeito possível ao atribuir uma pontuação as suas tentativas prévias, buscando sempre aumentar essa pontuação.

Além de ser possível classificar um modelo entre supervisionado, não supervisionado, de reforço... etc, mesmo dentro dessas categorias, encontram-se mais categorias ainda. O problema de identificação de SAF é considerado um problema de classificação, pois existe um espaço discreto de possibilidades de respostas possíveis, nesse caso, duas, ou a criança tem um diagnóstico de SAF, ou não tem. Isso poderia ser comparado com, por exemplo, o problema de descobrir o preço de uma casa, que possuiriam um espaço contínuo de respostas. Problemas como esse são chamados de problemas de regressão.

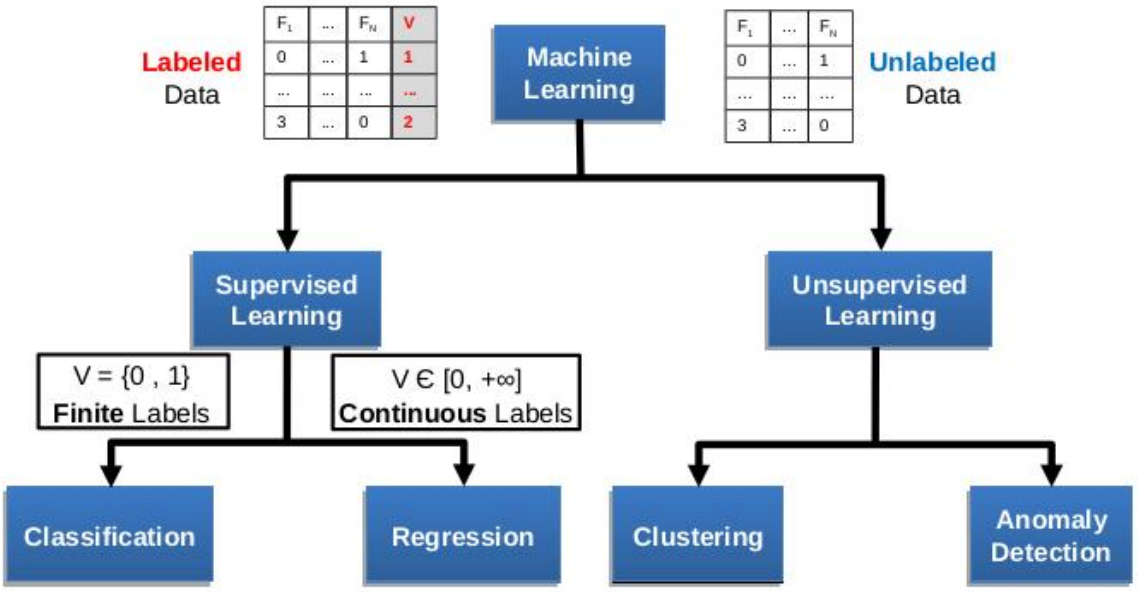


Figura 2 – Subdivisões de Aprendizado de Máquina (GRELLERT, 2018)

Nesse trabalho, usam-se inicialmente os modelos de árvore de decisão e *random forest*, os quais serão explicados nas seções a seguir

2.1.3.1 Árvore de decisão

O modelo da árvore de decisão funciona através uma sequência de *branches*, que, em seu modo mais básico, usam de condicionais para separar os dados entre dados que testam positivamente, e negativamente para tal condicional. Após uma série de condicionais para cada entrada, espera-se que tenha-se adquirido informação suficiente para definir de qual classe essa entrada deve ser. A árvore procurará criar as condições que mais diminuem a incerteza geral do sistema, assim, conquistando o máximo de informação sobre quais classes vão ser mais prováveis nos próximos *branches*.

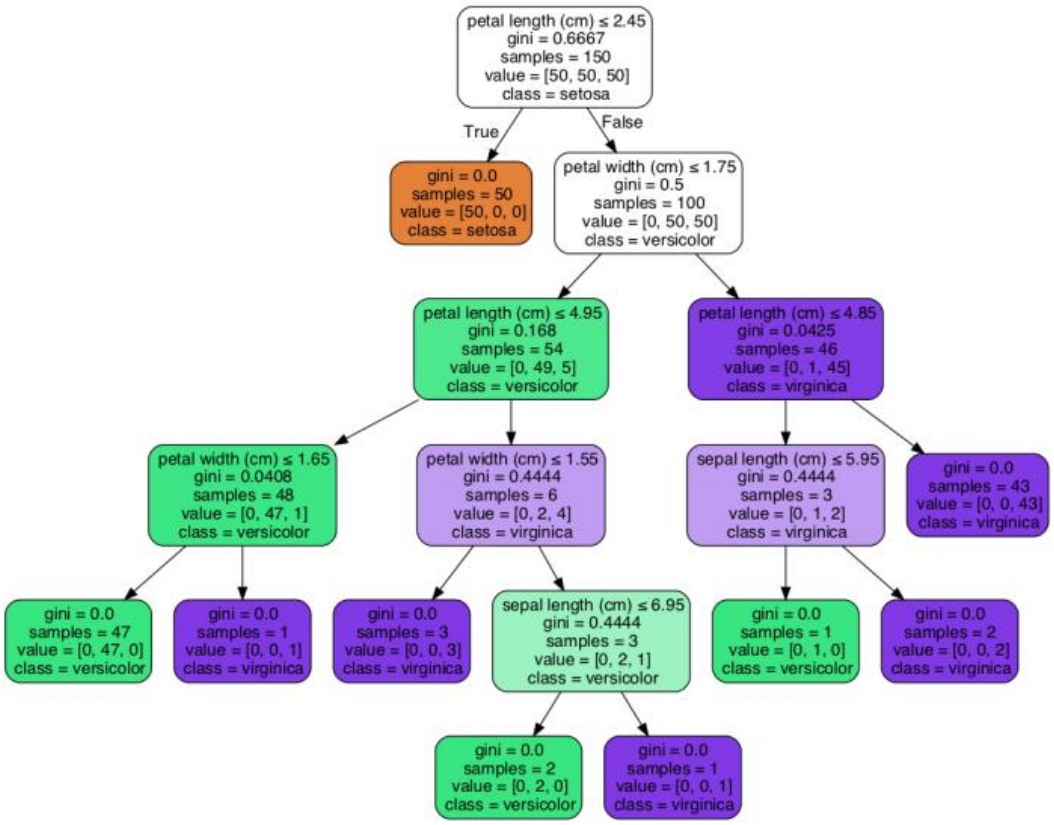


Figura 3 – Exemplo de Visualização da Árvore de Decisão (GRELLERT, 2018)

2.1.3.2 Random Forest

Uma *random forest*, é, basicamente, uma agregação de várias árvores de decisão, que usam de algum mecanismo de consenso para decretar a previsão final. Esse conceito de "agrupamento"de determinados modelos é conhecido na literatura como Ensemble, e, no caso das *random forests*, vem aliado a estratégia de gerar árvores de maneira randômica, de modo a introduzir variabilidade na construção das árvores, o que torna o consenso póstumo dessas variadas árvores, mais poderoso.

O algoritmo de uma *random forest* consiste, então, tanto da junção das predições das árvores de decisão (*tree bagging*), quanto a construção randômica dessas árvores (*randomized decision tree training*). A junção das predições pode ser realizada por voto de maioria, para modelos de classificação, ou usando média simples, para modelos de regressão (GRELLERT, 2018).

2.1.3.3 XGBoost

O *XGBoost* é um algoritmo de aprendizado de máquina que pertence à família de métodos Ensemble. Assim como as *random forests*, o *XGBoost* também utiliza o conceito de agregar vários modelos para obter uma previsão final. No entanto, o *XGBoost* adota uma abordagem diferente em relação à construção desses modelos (XGBOOST. . . , s.d.).

O algoritmo do *XGBoost* combina árvores de decisão, de forma que cada árvore é construída de maneira sequencial, buscando minimizar uma função de perda específica. A construção das árvores é realizada de forma iterativa, adicionando uma nova árvore a cada iteração para corrigir os erros cometidos pelas árvores anteriores. Esse processo de construção iterativa e corretiva é conhecido como *boosting*.

Ao final do processo de treinamento, o *XGBoost* combina as previsões de todas as árvores para obter uma previsão final. Para problemas de classificação, o consenso pode ser alcançado através de voto de maioria, enquanto para problemas de regressão, a média simples é utilizada.

2.1.4 Amostragem

Segundo Barbetta (BARBETTA, 2012) quando se deseja conhecer uma população (conjunto de elementos que se deseja estudar), também é possível analisar uma amostra, para se aprender de maneira aproximada as características da população. Entretanto, esse é um processo que somente obtém sucesso se realizado com uma metodologia de seleção de elementos, de maneira que sejam representativos da população como um todo.

Neste trabalho, temos uma dificuldade devido a grande quantidade de dados, e portanto, será necessário o uso de amostras para realizar a análise da população. Além disso, há um segundo problema a ser considerado, os portadores de SAF são um grupo sub-representado na população, portanto, uma mera amostragem aleatória contaria com poucos indivíduos para serem analisados em cada amostra.

Para lidarmos com isso, usamos uma estratégia de amostragem estratificada. Esta, tem como objetivo, reduzir o viés que seria criado ao analisar-se uma amostra contendo poucas entradas do grupo de interesse sub representado (BARBETTA, 2012). Em outras palavras, a analise estatística, ou o modelo, falhariam em produzir conhecimento confiável se não tivessem um espaço amostral significativo dentro de cada classe a ser analisada.

No nosso caso em específico, precisamos garantir que cada amostra usada para treino de um modelo tenha uma representatividade significativa para SAF. Além da estratificação de SAF em relação a pessoas sem SAF, também estratificamos os valores por estado, de maneira que as diferenças regionais possam ser contempladas na análise.

Mesmo com esses conceitos definidos, ainda existe um problema, a confiança da análise em cima de uma única amostragem seria gravemente comprometida pela mera aleatoriedade do processo. Uma amostra qualquer poderia produzir conclusões muito diferentes de uma outra amostra qualquer.

Para lidar com isso, realizamos o processo de amostragem, treino e teste, múltiplas vezes, e agrupamos os resultados de todas as amostras, para cada parâmetro sendo analisado. Este protocolo é apoiado pelo teorema central do limite.

O Teorema central do limite diz que as médias e variâncias de um espaço amostral de variáveis tende a uma distribuição normal se tirarmos uma medida de agrupamento

para varias amostras de uma mesma população (BUSSAB; MORETTIN, 2002)

Assim, a estratégia de amostragem estratificada e a aplicação do Teorema Central do Limite fornecem uma base sólida para lidar com a dificuldade de grandes volumes de dados, garantindo representatividade e confiabilidade nas análises estatísticas realizadas neste trabalho.

2.2 TRABALHOS RELACIONADOS

A seguir discutimos parte da literatura teórica sobre SAF e TEAF.

2.2.1 *Prevalência e Fatores Associados com o Consumo de Álcool Durante a Gravidez* (BAPTISTA *et al.*, 2017)

Esse artigo foca-se menos em analizar os efeitos do consumo de álcool do que os outros artigos mencionados nessa seção, focando-se mais na prevalência do consumo de álcool, e os fatores que levariam a tal.

A metodologia usa de um estudo *cross-sectional* em um espaço amostral de puérperas após o trabalho de parto, recrutadas ao longo de um espaço de 6 meses na cidade de São Carlos - SP. Dessas mulheres, coletou-se informações demográficas e reprodutivas, assim como a aplicação de um questionário de T-ACE para definir o padrão de uso de álcool das mesmas. As mulheres então são divididas entre o grupo de consumidoras de álcool, e não consumidoras de álcool, no limiar de score de T-ACE maior ou menor que 2. As comparações nesses grupos então serão realizadas usando um T test não pareado, um teste de qui-quadrado, ou teste exato de Fischer, dependendo da variável sendo analisada, sendo o nível de significância estatística definido em 5% ($p = 0.05$).

O estudo contou com 88.4% de participação (818 de 925 mulheres requisitadas). Dentre essas, 7.3% (60 mulheres) foram qualificadas como consumidoras de álcool. Dos fatores analisados, os únicos 2 fatores com diferença estatística significativa foram: ausência de parceiro fixo ($p=0.010$), e peso menor nas crianças com mães T-ACE positivas (3,045g$\pm$71.0 vs 3,192g$\pm$19.2; $p=0.040$). Nenhuma variável reprodutiva apresentou diferença estatística significante. O estudo então conclui que identificar mulheres suscetíveis ao consumo de álcool na gravidez pode contribuir para desenvolver políticas publicas de saúde mais eficientes. Usaremos em nosso trabalho o conjunto de variáveis levantadas nesse estudo como ponto de partida para a exploração de dados do DATASUS.

2.2.2 *Transtorno do Espectro Alcoólico Fetal: Uma Visão Geral* (RILEY; INFANTE; WARREN, 2011)

Publicado pela springer, uma das revistas de ciências *peer-reviewed* mais conhecidas na comunidade científica, esse trabalho é um *overview* que busca analisar vários trabalhos

na área e evidenciar a TEAF como um todo, chamando atenção para o fato que, enquanto o SAF costumeiramente é diagnosticado através de parâmetros físicos, a ingestão de álcool na gravidez também se manifesta na criança através de traços comportamentais e intelectuais em todo o espectro alcoólico fetal.
Ao longo do artigo os autores guiam os leitores através do histórico de desenvolvimentos na área, e das diferentes descobertas que vários artigos proeminentes encontraram ao longo do tempo, comentando sobre epidemiologia, demografia, os biomarcadores, o custo associado, e os efeitos relacionados ao SAF.
O artigo conclui que os efeitos da exposição a álcool na gestação produzem implicações vitalícias na saúde mental, educação, comportamento criminal e independência dos indivíduos acometidos por SAF. Também é apontado que existe o desafio de determinar-se a manifestação de TEAF em um individuo através de sintomas, quando pode se haver uma intersecção de sintomas relacionados a outras condições. E por último, comenta-se que enquanto a pesquisa elucidou vários mecanismos envolvidos na manifestação de SAF, é necessário trabalhar-se em traduzir essas descobertas em maneiras de prevenir e intervir nas consequências de exposição a álcool.

2.2.3 Síndrome Alcoólica Fetal – Revisão Sistematizada (SANTANA; ALMEIDA; MONTEIRO, 2014)

Esse *overview* busca expandir o entendimento que temos sobre SAF, para quais possíveis efeitos o uso de etanol durante a gravidez pode causar no embrião humano.
O trabalho usa Medline, SciELO, LILACS e Cochrane; como bases de dados para analizar tanto a literatura nacional como internacional. A busca inicial por artigos resulta em 118 resultados, que através de vários filtros chegam ao numero final de 24 artigos sendo estudados no *overview*.
O estudo identifica que as principais complicações dos fetos, cujas mães fizeram uso do álcool na gravidez são: baixo peso ao nascer (BPN), crescimento intrauterino restrito (CIUR), prematuridade, retardo no neurodesenvolvimento e microcefalia. Além disso, falhando em encontrar um consenso sobre a dose mínima que cause efeitos no feto, o estudo indica que nenhuma quantidade de ingestão de álcool é segura, e que é necessário a aplicação de medidas educativas para gestantes, visando prevenir as manifestações de TEAF.

2.2.4 A Síndrome Alcoólica Fetal: Revisão sistemática (QUEIROZ, 2016)

Essa monografia busca realizar a revisão sistemxática da literatura de SAF focando em fatores de risco e prevenção.
O trabalho usará trabalhos publicados nas bases de dados Pubmed e SciELO. Foram

selecionados os descritores referentes à síndrome alcoólica fetal. Os filtros escolhidos foram: artigos em língua portuguesa, inglesa ou espanhola, estudos realizados em seres humanos e publicados entre os anos de 2000 e 2016 (a data de publicação da monografia). Após a filtragem dos trabalhos, termina-se com um número de 8 deles que serão usados para compôr o compilado de conhecimento da monografia.
É então concluído que os fatores de risco para alcoolismo materno estão sendo bem documentados na literatura na forma de: não habitar com o cônjuge, gravidez inesperada, uso de outras drogas lícitas ou ilícitas e residir com outros consumidores de bebida alcoólica. Esta revisão também aponta que a discussão sobre a prevenção de SAF possui pouco amparo na literatura, assim como as discussões de tratamento e acompanhamento do portador de SAF.

2.2.5 Sumário dos trabalhos relacionados

Os artigos descritos resumem os esforços da academia em descrever os aspectos físicos, comportamentais, biológicos e psicológicos envolvidos no diagnóstico e manifestação da SAF e TEAF, assim como também é descrito em alguns deles os fatores sociais que desencadeariam um quadro de SAF. Entretanto, ainda não existem artigos suficientes discutindo como realizar a intervenção e detecção prévia dos casos de risco. Além disso, ao comparar-se com o que será feito nesse trabalho, são estudos clássicos de psicologia. A grande diferença entre os trabalhos descritos e este, é a metodologia utilizada. Os trabalhos descritos em um geral criaram seus próprios dados ao entrevistar as mães diretamente, conseguindo uma base de dados customizada para a análise proposta em seus contextos. Este trabalho, entretanto, terá o desafio de usar bases de dados públicas criadas de maneira genérica, o que faz com que as relações das bases com o problema de SAF e TEAF precisem ser cuidadosamente criadas, mas obtendo a vantagem de conseguir um grande montante de dados para a análise.

A tabela abaixo apresenta um resumo comparativo dos trabalho relacionados listados anteriormente.

Artigo	Objetivo	Ano	Fonte de Dados	Resultados
(BAPTISTA *et al.*, 2017)	Analisar prevalência do uso de álcool e fatores relacionados.	2017	Estudo próprio com 818 puérperas.	Fatores identificados: ausência de parceiro fixo e peso menor nos bebês.
(RILEY; INFANTE; WARREN, 2011)	Overview de TEAF como um fenômeno.	2011	Não mencionado	Aponta implicações do SAF no dia-a-dia dos indivíduos.
(SANTANA; ALMEIDA; MONTEIRO, 2014)	Entender os possíveis efeitos de ingestão de álcool para o feto humano.	2014	24 artigos de Medline, SciELO, LILACS e Cochrane;	Encontra principais complicações dos fetos.
(QUEIROZ, 2016)	Revisão bibliográfica focada em fatores de risco e prevenção	2016	8 artigos de Pubmed e SciELO	Identifica alguns fatores, chama a atenção para a ausência de propostas de intervenção

Tabela 1 – Resumo comparativo dos trabalhos relacionados

3 SOLUÇÃO PROPOSTA

O fluxo da solução proposta por este trabalho é apresentado na Figura 4.

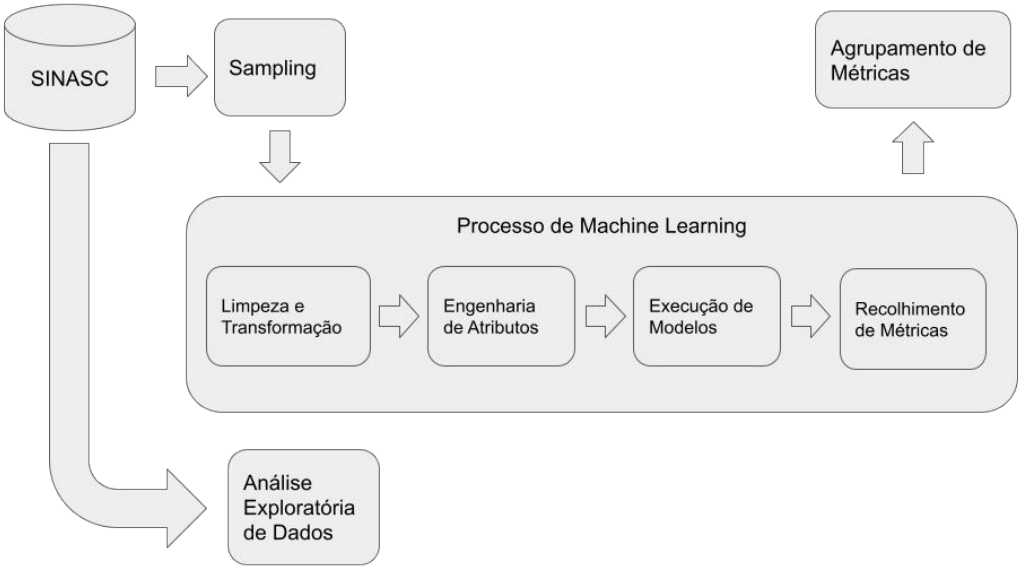


Figura 4 – Metodologia global

Primeiramente, faremos uma análise exploratória de dados com o intuito de estabelecer uma intuição sobre o dado, checar as interações com dados faltantes, e verificar as distribuições das variáveis mais relevantes de acordo com a literatura.

A base de dados do SINASC oferecida pelo DATASUS contém dados de todos os estados do país entre os anos de 1994 a 2019, totalizando 14,087 GB. Para esse trabalho, usaremos somente os dados de 2010 a 2019, devido tanto a modernização do sistema ao longo dos anos, como para representar a realidade atual do país melhor. Além disso, devido ao grande volume de dados, e ao percentual pequeno de entradas de SAF, precisaremos usar uma estratégia de estratificação de dados.

A estratégia escolhida atualmente envolve 100 iterações de *sampling* com o quíntuplo de dados neutros, cada um desses *datasets* estratificados então passará pelas mesmas etapas de limpeza e enriquecimento até finalmente ser treinada pelo modelo. Após essas etapas, colheremos as métricas de cada modelo, e uma etapa final de agregação dos resultados de cada um dos 100 modelos expressará a realidade estatística da resolução do problema pelo modelo, e expressividade dos discriminantes como descritores do problema.

Para realizar essa estratificação, primeiro, extraímos todas as entradas SAF positivas entre 2010 e 2019. Isso é realizado através da coluna “CODANOMAL”, que representa o “Código da anomalia (CID 10)”. (DICTSINASC..., s.d.) da criança nascida, assim, filtramos essa coluna usando:

- Q86 - Síndromes com malformações congênitas devidas a causas exógenas conhecidas,

não classificadas em outra parte (FIOCRUZCID10..., s.d.)

- Q870 - Síndromes com malformações congênitas afetando predominantemente o aspecto da face (CIDQ870..., s.d.)

O *dataset* final contém 1262 entradas. Isso faz com que nossa estratificação precise de 6310 valores selecionados aleatoriamente, totalizando *datasets* de trabalho de 7572 entradas por iteração.

A base de dados então passa por uma série de processos de limpeza e enriquecimento: imputação de dados faltantes, normalização de valores contínuos, binarização de valores categóricos, tradução de codificação para linguagem natural, dentre outros

ESTCIVMAE	Estado Civil	ESTCIV.CNV	caracter	1	Situação conjugal da mãe: 1– Solteira; 2– Casada; 3– Viúva; 4– Separada judicialmente/divorciada; 5– União estável; 9– Ignorada.

Figura 5 – Exemplo de codificação - Coluna ESTCIVMAE - Fonte: DANTPS - CGIAE

Roda-se então o modelo selecionado para o sample atual, colhem-se as métricas, e no fim das 100 iterações, serão agrupadas as métricas usando média e desvio padrão.

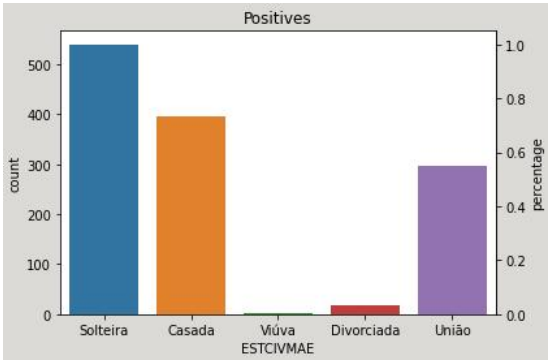
4 ANÁLISE EXPLORATÓRIA DE DADOS

Como guia desta análise, usaremos alguns dos fatores que são teorizados como discriminantes proeminentes de SAF segundo a literatura (BAPTISTA *et al.*, 2017)

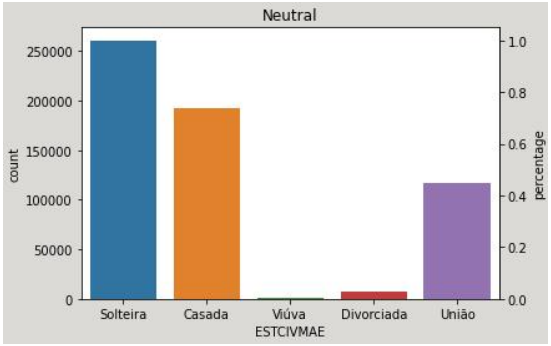
- Estado Civil (ausência de parceiro fixo)
- Menor peso do bebê ao nascer
- Bebês do sexo feminino podem ter maior perda de peso devido a TEAF
- Baixa escolaridade da mãe
- Baixo nível socioeconômico da mãe
- Idade superior a 30 anos
- Cor de pele não branca
- Desemprego

Como apontamos anteriormente, existe uma dificuldade no manuseio da grande quantidade de dados disponíveis. Para essa análise em específico, produziu-se um *dataset* de todas as entradas SAF positivas entre 2010 e 2019, e, para atuar como base de comparação, usou-se um *dataset* de dados neutros com 2% dos dados entre 2010 e 2019, o *dataset* de positivos contém 1262 entradas positivas, enquanto o *dataset* neutro possui 584269 entradas.

Uma análise inicial do estado civil de mães com crianças SAF positivas não parece ter um padrão divergente do estado civil de mães com crianças SAF negativas, como mostram as figuras 6a e 6b.

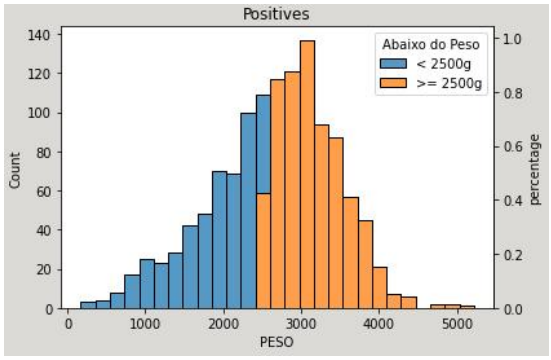


(a) Gráfico ESTCIVMAE para *dataset* positivo

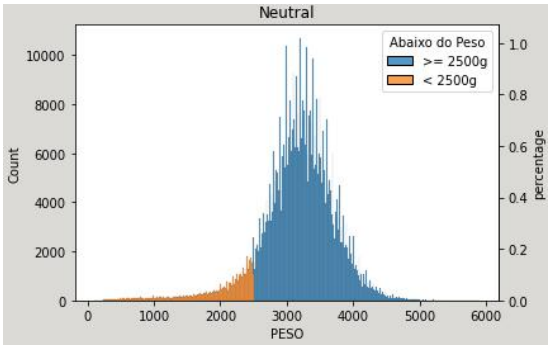


(b) Gráfico ESTCIVMAE para *dataset* neutro

Já o peso das crianças ao nascer claramente é afetado pelo consumo de álcool na gravidez, como demonstram as figuras 7a e 7b.

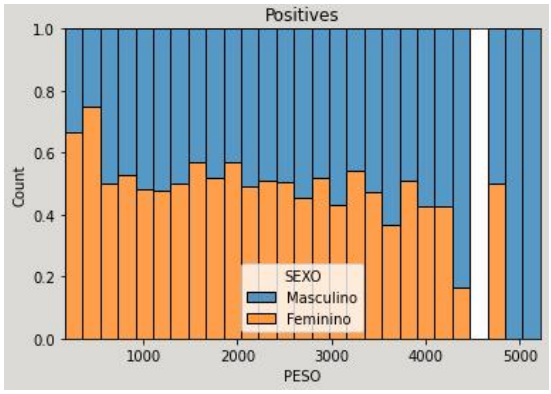


(a) Gráfico PESO para *dataset* positivo

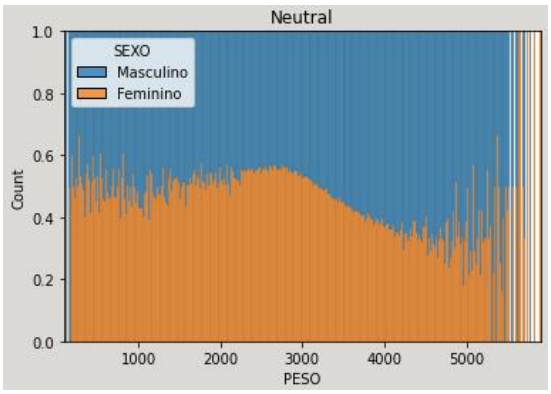


(b) Gráfico de peso para *dataset* neutro

Por outro lado, a exploração falha em encontrar alguma correlação entre uma perda de peso acentuada nas crianças do sexo feminino, como mostram as figuras 8a e 8b.

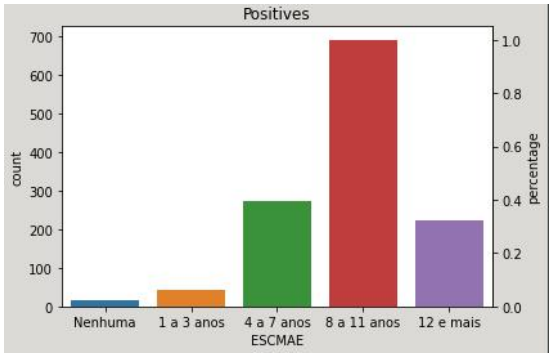


(a) Gráfico de peso por sexo para *dataset* positivo

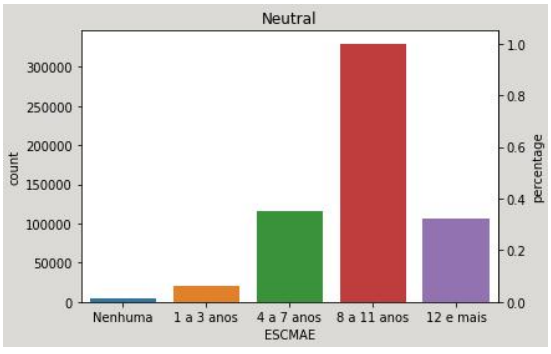


(b) Gráfico de peso por sexo para *dataset* neutro

As figuras 9a e 9b mostram que assim como o estado civil, a escolaridade da mãe não parece influenciar a manifestação de SAF na criança.



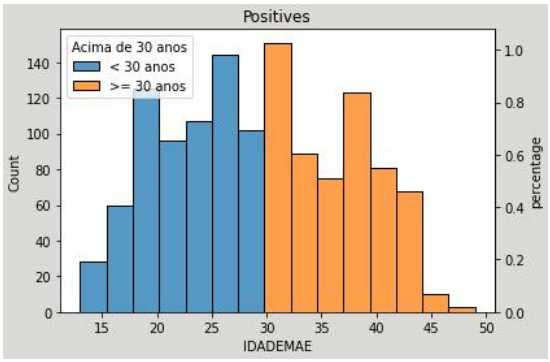
(a) Gráfico de escolaridade para *dataset* positivo



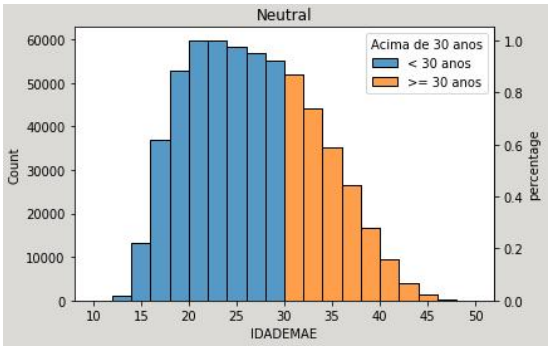
(b) Gráfico de escolaridade para *dataset* neutro

Não foi possível encontrar informações socioeconômicas no *dataset* do SINASC para averiguar se um baixo nível socioeconômico influencia na manifestação de SAF na criança.

A exploração incentiva uma análise mais cuidadosa a respeito da idade da mãe, que, num primeiro momento parecem ter alguma divergência entre os 2 grupos avaliados, como demonstrado nas figuras 10a e 10b.

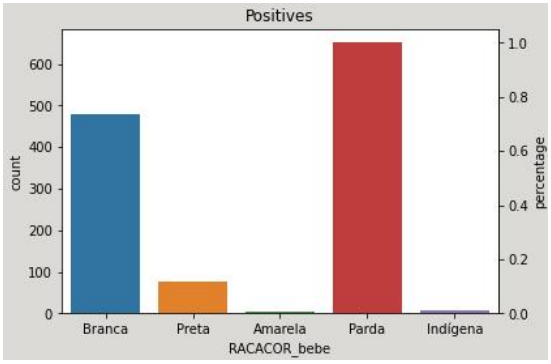


(a) Gráfico de idade para *dataset* positivo

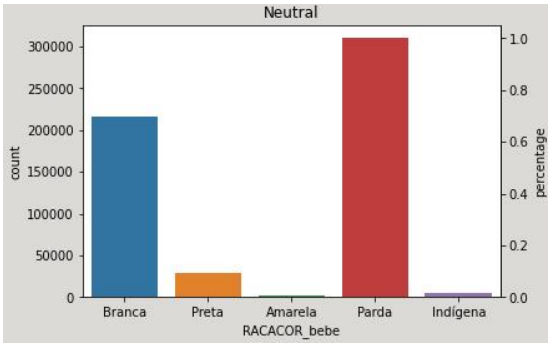


(b) Gráfico de idade para *dataset* neutro

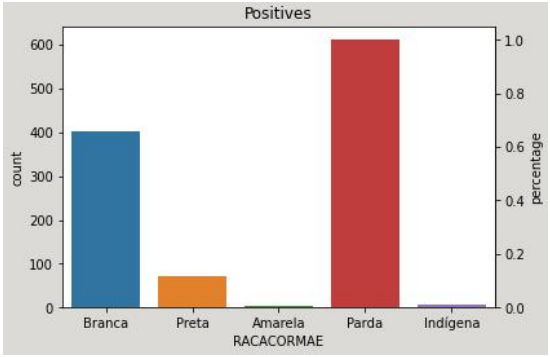
Tanto a raça da mãe quanto do bebê parecem não influenciar o diagnóstico de SAF, como demonstrado nas figuras 11a,11b, 12a e 12b.



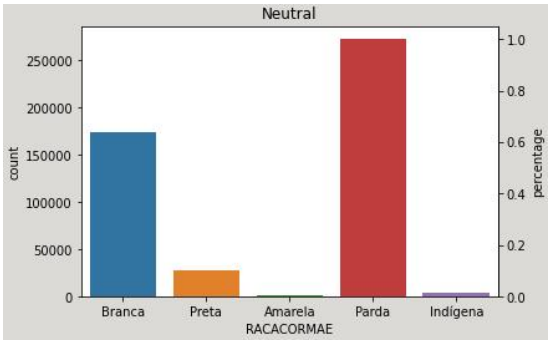
(a) Gráfico de raça do bebê para *dataset* positivo



(b) Gráfico de raça do bebê para *dataset* neutro



(a) Gráfico de raça da mãe para *dataset* positivo

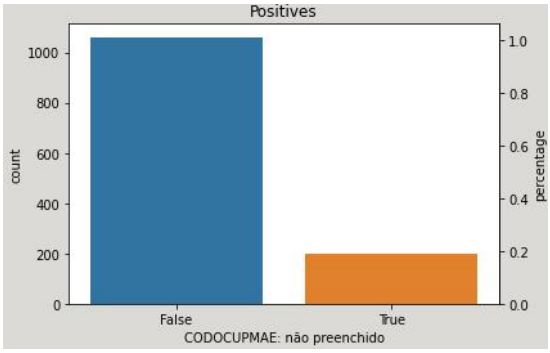


(b) Gráfico de raça da mãe para *dataset* neutro

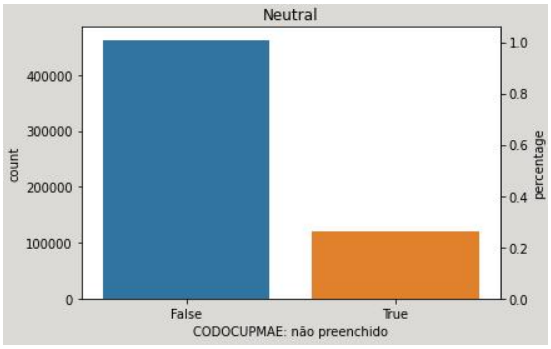
Sobre desemprego, também não existem informações satisfatórias no SINASC. O melhor esforço que poderia ser feito seria verificar a nulidade de valores na coluna "CODOCUPMAE", que representa um código de: "Ocupação, conforme a Classificação

Brasileira de Ocupações (CBO-2002)" (DICTSINASC. . . , s.d.). A CBO, entretanto, não possui um valor destinado a pessoas desempregadas.

A ausência de um valor não deve receber um significado semântico numa análise, uma vez que a noção de um dado faltante é justamente a ausência de informação. Entretanto, mesmo ao fazer a contagem de dados faltantes em ambos os grupos, não existe diferença significativa na proporção, como mostram as figuras 13a e 14b.

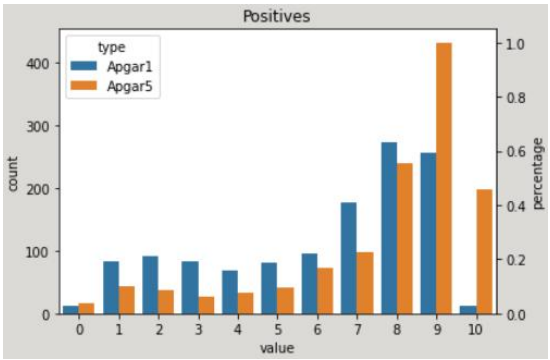


(a) Dados faltantes em CODOCUPMAE para *dataset* positivo

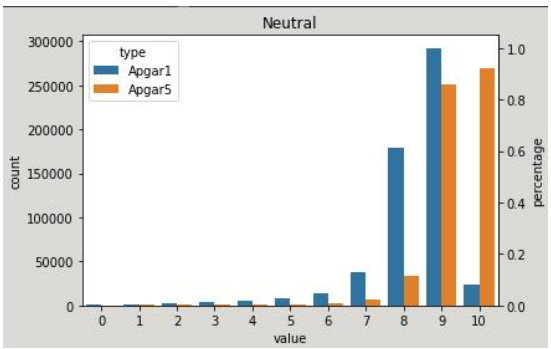


(b) Dados faltantes em CODOCUPMAE para *dataset* neutro

Por último, encontramos uma forte correlação na variável de APGAR, que não está presente no estudo mencionado no inicio desse capítulo. APGAR sendo um score de 0 a 10, com os componentes: Aparência, Pulso, Gesticulação, Atividade e Respiração indo de 0 a 2. Também se destaca que APGAR1 é referente ao APGAR medido no primeiro minuto de vida, e APGAR5, no quinto minuto de vida. As figuras 13a e 14b mostram como a distribuição tende a números mais altos em indivíduos saudáveis, e mais baixos em casos de SAF positivo.



(a) Distribuição APGAR para indivíduos SAF positivo



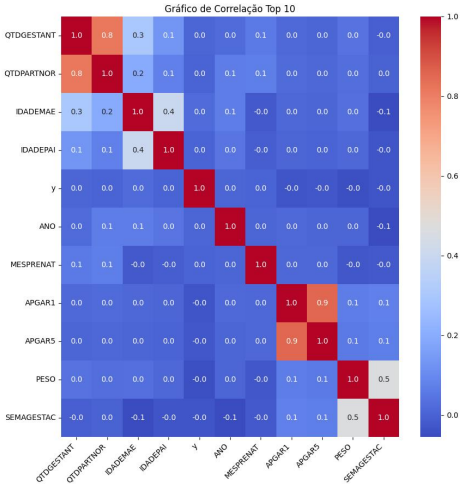
(b) Distribuição APGAR para indivíduos neutros

4.1 CORRELAÇÃO ENTRE VARIÁVEIS

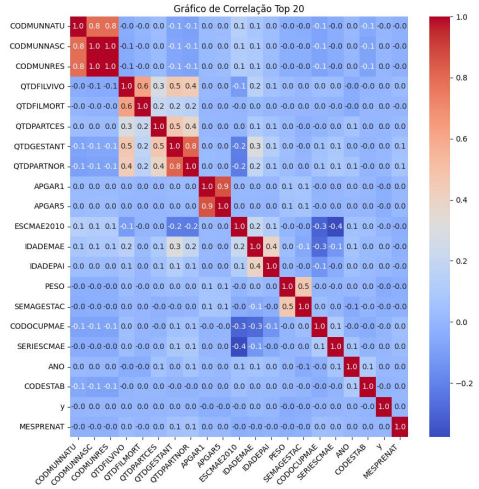
Para a seguinte análise, selecionamos as 10 e 20 variáveis que tem a maior correlação absoluta (calculada usando o método de pearson) com o diagnóstico positivo ou negativo de SAF. Além disso, estamos usando a biblioteca scipy para fazer o clustering hierárquico

das variáveis apresentadas, de forma a criar "ilhas"de itens similares baseado na semelhança de correlação que estes têm entre si.

Abaixo podemos ver a correlação entre as variáveis 15a e 15b.



(a) Correlação top 10 variáveis



(b) Correlação top 20 variáveis

Podemos observar que as variáveis não possuem uma correlação alta com o y, o que é de se esperar, uma vez que SAF é uma condição rara e deve acontecer a partir de uma sequência complexa de fatores, ao invés de ser resumida a variáveis simples.

Outras correlações observadas também não apresentam um ganho de conhecimento fora do comum: peso tem alguma correlação com semanas de gestação, quantidade de gestações anteriores correlaciona com quantidade de partos normais, e ambas as medidas de APGAR tem forte correlação entre si.

4.2 CONCLUSÃO DA ANÁLISE DE DADOS

Falhamos em encontrar as mesmas correlações indicadas pela literatura de estudo clínico.As variáveis mais notáveis como discriminantes entre os grupos SAF positivos e SAF negativos são: o peso da criança (sem distinção de gênero), e num grau menor, a idade da mãe acima de 30 anos. A teoria a respeito das variáveis de estado civil, baixa escolaridade da mãe, baixo nível socioeconômico da mãe, cor de pele não branca e desemprego como potenciais discriminantes de SAF não foi amparada pela análise exploratória.

5 DESENVOLVIMENTO

5.1 SAMPLING

Devido a dimensão dos dados, não é possivel carregar em memória todo o dataset histórico dos anos 1994 a 2019, que juntos pesam um total de 14,087 GB.

Dessa forma, esse trabalho usa uma metodologia de sampling para que seja possível computar diversos espaços amostrais dos dados, e ao agrupá-las, pode-se ter uma representação do espaço total de dados.

Para esse fim, primeiro, definimos que usaremos apesar informações dos anos de 2010 a 2019, excluindo os anos de 1994 a 2010, uma vez que dados mais recentes vão ser mais capazes de expressar o estado atual dos fatores que influenciam SAF.

Depois disso, usaremos o conceito de balanceamento de classes para definir um alfa de multiplicação de maneira que tenhamos 5x mais entradas negativas do que positivas para SAF.

Dessa forma, chegamos ao número de 1262 entradas positivas em todo o país, e por consequência, escolhemos 6310 entradas para completar o dataset de treino para cada amostra, completando um dataset com um total de 7572 entradas.

Essas 6310 entradas selecionadas randomicamente serão estratificadas por estado, para que seja possível representar cada estado do país no treinamento dos modelos. Na tabela abaixo encontram-se, tanto a distribuição encontrada para a presença de SAF ao longo dos anos de 2010 a 2019, quanto o número de amostras neutras correspondentes por estado 2.

Estado	Indivíduos com SAF	Indivíduos neutros
AC	19	95
AL	23	115
AM	37	185
AP	2	10
BA	66	330
CE	51	255
DF	20	100
ES	22	110
GO	47	235
MA	51	255
MG	98	490
MS	10	50
MT	19	95
PA	25	125
PB	30	150

Continued on next page

Tabela 2 – *Continued from previous page*

Estado	Indivíduos com SAF	Indivíduos neutros
PE	88	440
PI	16	80
PR	65	325
RJ	84	420
RN	27	135
RO	7	35
RR	12	60
RS	51	255
SC	44	220
SE	10	50
SP	330	1650
TO	8	40

Tabela 2 – Distribuição da estratificação de cada amostra

Com esses valores conseguimos ter tanto uma distribuição equilibrada entre indivíduos com e sem SAF, quanto uma representação equiparada de cada estado do país.

5.2 DATA CLEANING

Uma vez carregados os dados criados no processo de *sampling*, realiza-se uma série de procedimentos para que o dado esteja pronto para a ingestão do modelo de aprendizagem de máquina.

5.2.1 Remoção de Colunas Excedentes

Para dados faltantes, definiremos que se uma coluna possui abaixo de 40% de dados preenchidos, essa coluna será removida da nossa análise.

Devido às limitações de memória, usa-se um sample de 2% dos dados para a análise de dados faltantes.

Abaixo podemos ver os dados faltantes no dataset 16.

Figura 16 – Gráfico de dados faltantes

Isso se aplica as colunas:

- RACACORN
- CODANOMAL
- NUMREGCART
- DTREGCART
- CODCART
- RACACOR_RN
- DTRECORIGA
- CODMUNCART
- DTRECORIG
- IDADEPAI

Ultimamente, checa-se a semântica dessas colunas, e se existe alguma outra coluna que é capaz de expressar a mesma informação.

Faremos aqui entretanto uma exceção pra coluna "IDADEPAI", pois a falta de informação pode ter o valor semântico de que o pai não estaria presente na gravidez da mãe. Essa é uma escolha que possui seus contras, já que em meio a pais ausentes também haverão simplesmente erros de imputação por isso deve-se exercer cautela ao se tirar conclusões a respeito desse atributo.

Por fim, fazem-se algumas análises sobre a semântica das colunas remanescentes, e se estas demonstrarão alguma utilidade ou não. No apêndice A é possível ver uma tabela de variáveis com todas as colunas que de fato são utilizadas no treinamento.

5.3 EXECUÇÃO DOS MODELOS

O treino acontece com 3 tipos de modelos de *Machine Learning*, *Decision Tree*, *Random Forest*, e *XGBoost* (XGB). Escolhemos 3 modelos baseados em árvore pois o escopo do nosso problema gira em torno da capacidade de expressar quais variávies são as mais relevantes para a decisão do modelo. Precisamos de modelos interpretáveis e expressivos para que seja possível a extração de conhecimento dos modelos.

Cada definição de modelo será rodada para cada uma das N samples de dados definidas.

Os modelos não fazem ajuste de hiperparâmetros, sendo rodados apenas com a configuração básica conforme mostra a Tabela 3

Modelo	**Hiperparâmetros**
RF	random_state=0
XGB	random_state=0
	learning_rate=1,
	max_depth=2,
	n_estimators=2,
	objective='binary:logistic'
DT	random_state=0

Tabela 3 – Tabela de Hyperparâmetros

Todos os modelos usam um *random_state=0* como semente de aleatoriedade para garantir a reprodutibilidade do modelo. Os valores específicos para XGB são recomendados pela documentação da biblioteca, e significam, respectivamente, a taxa de aprendizado do modelo, a profundidade máxima de uma árvore, o número de estimadores do modelo e o objetivo como regressão logística para classificação binária.

Os modelos rodam sua pipeline de treinamento, e então, são salvos os metadados sobre o treinamento desses modelos na classe *ModelEvaluator*, que é genérica para todos os modelos.

5.4 RECOLHIMENTO DE MÉTRICAS

A classe *ModelEvaluator* é responsável por guardar as métricas de cada rodada de treinamento por sample. Isso faz com que seja possível a análise de como uma classe de modelo se comporta no dataset do SINASC, baseado em como cada instância de modelo em específico performou em sua sample.

A metodologia de treino e teste de modelo se dá por validação cruzada com 10 rodadas. São recolhidas as métricas:

- F1
- ROC_AUC
- Feature Importance

O F1 score é uma métrica de avaliação de desempenho de um modelo de classificação que combina duas medidas importantes: precisão e *recall*. A precisão mede a proporção de verdadeiros positivos em relação a todos os exemplos classificados como positivos pelo modelo. O *recall*, por sua vez, mede a proporção de verdadeiros positivos em relação a todos os exemplos que realmente são positivos. O F1 Score é a média harmônica dessas

duas medidas, fornecendo uma medida agregada do desempenho do modelo em equilibrar a precisão e o *recall*. Valores mais altos de F1 Score indica um melhor desempenho do modelo na classificação correta da classe positiva e negativa.

A sigla da métrica ROC-AUC significa *Receiver Operating Characteristic Area Under the Curve*. A curva ROC é uma representação gráfica que ilustra a relação entre a taxa de verdadeiros positivos (*True Positive Rate* - TPR) e a taxa de falsos positivos (*False Positive Rate* - FPR) do modelo em diferentes pontos de corte de classificação.

A ROC-AUC é a área sob essa curva ROC e fornece uma medida da capacidade do modelo de distinguir entre a classe positiva e negativa. Um valor de ROC-AUC próximo de 1 indica um modelo com excelente capacidade de classificação, enquanto um valor próximo de 0.5 indica um desempenho aleatório do modelo.

Já a importância dos atributos (em inglês, *feature importance*) é uma representação numérica da importância relativa de cada atributo para determinar o veredito de uma previsão feita pelo modelo. Enquanto números mais altos indicam um maior peso da variável na análise, e números menores indicam um peso menor, em geral, a parte de maior destaque na análise de importância de atributos se dá na distância relativa entre cada uma das variáveis.

Com o uso dessas métricas, torna-se possível avaliar o desempenho e comportamento de cada classe de modelo.

Essas métricas nos fornecem informações valiosas sobre a capacidade do modelo em lidar com as amostras específicas, permitindo-nos analisar seu desempenho em cada rodada de treinamento.

Com base nessas informações, podemos tomar decisões informadas sobre a escolha e otimização dos modelos para atingir os melhores resultados possíveis.

Uma vez criados todos os *ModelEvaluator*, inicia-se a etapa de agregação de métricas para que seja possível analisar a performance global de um único tipo de modelo.

6 ANÁLISE DE RESULTADOS DOS MODELOS

Este capítulo apresenta a análise dos resultados obtidos por meio das métricas e modelos desenvolvidos ao longo deste estudo. Nosso objetivo é compreender como as informações encontradas se relacionam com o escopo global deste trabalho, considerando também o caso de estudo específico.

Olharemos o comparativo modelo a modelo das métricas F1 e *roc_auc*, assim como as *feature_importances* obtidas, para analisar como as diferentes estratégias de cada modelo acabam por gerar perspectivas diferentes em relação ao mesmo caso de estudo.

6.1 ANÁLISE DE MÉTRICAS

A tabela abaixo apresenta a média das métricas obtidas para cada modelo ao serem rodados para 100 amostras de dados.

Modelo	**Hiperparâmetros**	**Média F1**	**Média ROC_AUC**
RF	random_state=0	0.27	0.60
XGB	random_state=0	0.41	0.73
	learning_rate=1,		
	max_depth=2,		
	n_estimators=2,		
	objective='binary:logistic'		
DT	random_state=0	0.23	0.47

Tabela 4 – Tabela de Resultados

Ambos os modelos de *Random Forest* e *Decision Tree* tiveram uma baixa performance em resumir o SAF como uma soma das variáveis fornecidas no treinamento dos modelos.

O modelo XGB teve uma performance consideravelmente maior em ambas as métricas. É difícil dizer se esses números estão próximos do máximo teórico com as atributos sendo utilizadas nesse modelo em específico, ou se existem ganhos possíveis a serem feitos durante o pipeline de treinamento.

6.2 IMPORTÂNCIA DOS ATRIBUTOS

Abaixo podemos ver a importância de atributos dos modelos escolhidos para nosso projeto: DT 17, RF 18 e XGB 19.

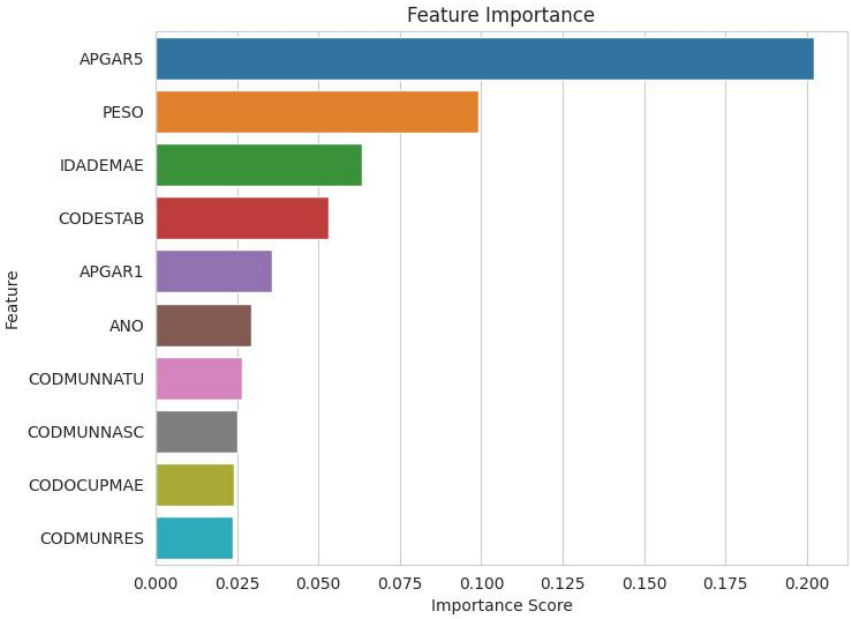


Figura 17 – Gráfico de importância de atributos para DT

O modelo de *Decision Tree* dá grande importância ao APGAR5, ao peso do bebê e a idade da mãe. Algumas variáveis curiosas aparecem no gráfico, como o código do estabelecimento, o município de naturalidade, nascimento e residência, essas variáveis poderiam apontar uma região localizada na qual existe maior incidência de SAF, ou possivelmente, uma região localizada na qual se diagnostica mais SAF.

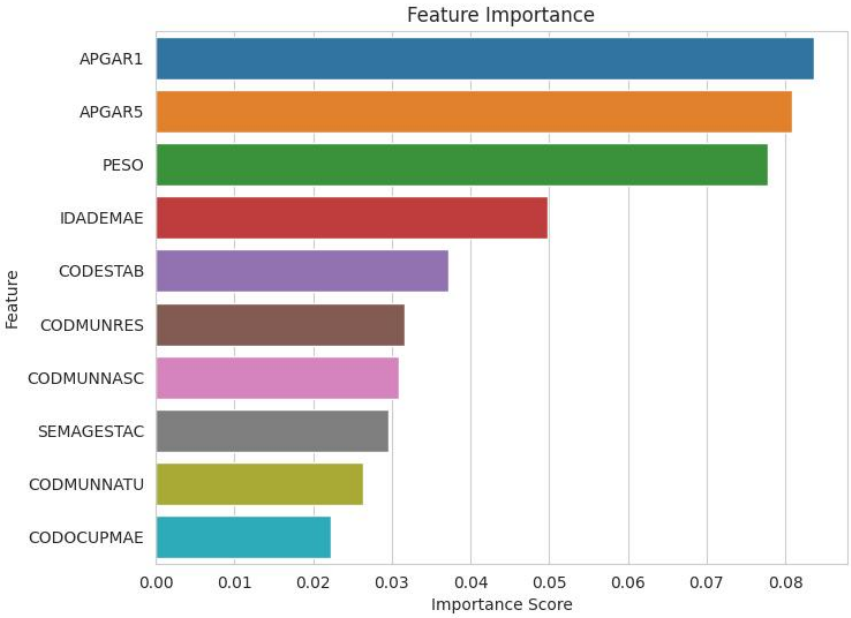


Figura 18 – Gráfico de importância de atributos para RF

O modelo de *Random Forest* aponta boa parte das mesmas variáveis que a *Decision*

Tree, mas observa-se que existe menos disparidade entre a importância do APGAR5 e o resto das atributos.

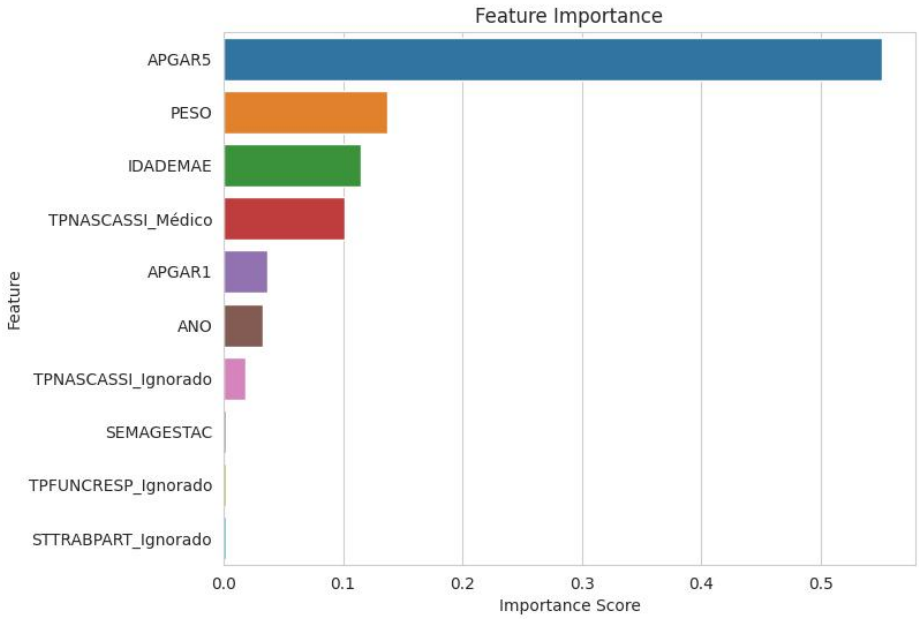


Figura 19 – Gráfico de importância de atributos para XGB

Já a importância de atributos do XGB diferem um pouco dos outros modelos. Não é mencionado coisas como a localidade, e existe uma estratégia de priorização do APGAR5 como fator de maior relevância.

Todos os modelos apontaram APGAR5 e APGAR1 como fatores bastante relevantes no diagnóstico de SAF, assim como o peso observado da criança e a idade da mãe.

7 CONCLUSÃO

A Síndrome Alcoólica Fetal (SAF) é uma condição grave que afeta indivíduos expostos ao álcool durante a gestação. É uma preocupação de saúde pública devido aos efeitos adversos que pode causar no desenvolvimento físico, cognitivo e comportamental das crianças afetadas.

O fato de SAF ser uma doença congênita com potencial de 100% de redução faz com que o ímpeto da intervenção clínica surja, somente é necessário que identifiquem-se os casos de risco com antecedência, e então, as devidas providências sejam tomadas.

Este estudo teve como objetivo principal analisar possíveis discriminantes de SAF através de um pipeline de dados e uso de machine learning utilizando dados do Sistema de Informações sobre Nascidos Vivos (SINASC).

Também foi feita uma releitura das conclusões acadêmicas históricas sobre o problema de SAF, agora, a partir de uma perspectiva de dados massivos, ao invés de uma perspectiva de análise e entrevista clínica.

Essas descobertas são de grande importância para profissionais de saúde e pesquisadores que lidam com a prevenção e o tratamento da SAF. A identificação precoce de gestações de alto risco permite a adoção de medidas preventivas e intervenções adequadas, visando reduzir os impactos da exposição ao álcool durante a gestação. Além disso, a abordagem metodológica proposta neste estudo pode servir como base para futuras pesquisas e aplicações relacionadas à predição de eventos em outros contextos de saúde pública.

Para trabalhos futuros, uma das sugestões de melhorias é a incorporação de novas bases de dados para complementarem as informações trazidas pelo SINASC. Além disso, observa-se que o uso de classes SAF vs. neutralidade não resolve os problemas de diagnóstico diferencial, e portanto, é interessante trazer modelos que classifiquem também entre diferentes anomalias e distúrbios.

REFERÊNCIAS

ARSENAULT-LAPIERRE, Geneviève; KIM, Caroline; TURECKI, Gustavo. Psychiatric diagnoses in 3275 suicides: a meta-analysis. **BMC psychiatry**, Springer, v. 4, n. 1, p. 1–11, 2004.

BAPTISTA, Flavia Hashizume *et al.* Prevalência e fatores associados ao consumo de álcool durante a gravidez. **Revista Brasileira de Saúde Materno Infantil**, SciELO Brasil, v. 17, p. 271–279, 2017.

BARBETTA, Pedro Alberto. **Estatistica Aplicada as Ciencias Sociais**. 8th. [*S.l.*]: Editora Atlas, 2012. ISBN 9788532806048.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística Básica**. 5th. São Paulo: Saraiva, 2002. P. 272. ISBN 85-02-03497-9.

CIDQ870. [*S.l.*: *s.n.*]. Disponível em: https://www.medicinanet.com.br/cid10/2374/q87_outras_sindromes_com_malformacoes_congenitas_que_acometem_multiplos_sistemas.htm.

COELHO, Akeni Lobo; ARAUJO MORAIS, Indyara de;
SILVA ROSA, Weverton Vieira da *et al.* A utilização de tecnologias da informação em saúde para o enfrentamento da pandemia do Covid-19 no Brasil. **Cadernos Ibero-Americanos de Direito Sanitário**, v. 9, n. 3, p. 183–199, 2020.

COELHO, Flávio Codeço; BARON, Bernardo Chrispim *et al.* **AlertaDengue/PySUS: Vaccine**. [*S.l.*]: Zenodo, mai. 2021. DOI: 10.5281/zenodo.4883502. Disponível em: https://doi.org/10.5281/zenodo.4883502.

COOK, Jocelynn L. *et al.* Fetal alcohol spectrum disorder: a guideline for diagnosis across the lifespan. Edição: **CMAJ**, CMAJ, v. 188, n. 3, p. 191–197, 2016. ISSN 0820-3946. DOI: 10.1503/cmaj.141593. eprint: https://www.cmaj.ca/content/188/3/191.full.pdf. Disponível em: https://www.cmaj.ca/content/188/3/191.

DICIONÁRIO de variáveis. [*S.l.*: *s.n.*]. Disponível em: https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacao-sobre-nascidos-vivos/dicionario-de-variaveis/.

DICTSINASC. [*S.l.*: *s.n.*]. Disponível em: http://svs.aids.gov.br/dantps/cgiae/sinasc/documentacao/dicionario_de_dados_SINASC_tabela_DN.pdf.

FIOCRUZCID10. [*S.l.*: *s.n.*]. Disponível em: https://github.com/bigdata-icict/ETL-Dataiku-DSS/blob/master/SIM/CID-10-CATEGORIAS.CSV.utf8#L1287.

GRELLERT, Mateus. Machine learning mode decision for complexity reduction and scaling in video applications, 2018.

HEERINGEN, Kees van; MANN, J John. The neurobiology of suicide. **The Lancet Psychiatry**, Elsevier, v. 1, n. 1, p. 63–72, 2014.

LEIDADOS. [*S.l.*: *s.n.*]. Disponível em: `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`.

LOPES, Claudia S *et al.* ERICA: prevalência de transtornos mentais comuns em adolescentes brasileiros. **Revista de Saúde Pública**, SciELO Brasil, v. 50, 2016.

MENDES, Isadora Cristina *et al.* Anomalias congênitas e suas principais causas evitáveis: uma revisão. **Revista Médica de Minas Gerais**, v. 28, n. 1, p. 1–6, 2018.

MITCHELL, Tom M. Machine learning and data mining. **Communications of the ACM**, ACM New York, NY, USA, v. 42, n. 11, p. 30–36, 1999.

POLITICADADOS. [*S.l.*: *s.n.*]. Disponível em: `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm`.

QUEIROZ, Marise Rosas. A Siêndrome Alcoólica Fetal: Revisão Sistemática. Faculdade de Medicina da UFBA, 2016.

RILEY, Edward P; INFANTE, M Alejandra; WARREN, Kenneth R. Fetal alcohol spectrum disorders: an overview. **Neuropsychology review**, Springer, v. 21, n. 2, p. 73–80, 2011.

SANTANA, Rogério A; ALMEIDA, Leonardo FJL; MONTEIRO, Denise LM. Siêndrome alcoólica fetal–revisão sistematizada. **Revista Hospital Universitário Pedro Ernesto (TÍTULO NÃO-CORRENTE)**, v. 13, n. 3, 2014.

SITESINASC. [*S.l.*: *s.n.*]. Disponível em: `http://svs.aids.gov.br/dantps/cgiae/sinasc/`.

SOBREDATASUS2022. [*S.l.*: *s.n.*]. Disponível em: `https://datasus.saude.gov.br/sobre-o-datasus/`.

SOUZA, Pollianna Marys de; AUTRAN, Marynice Medeiros Matos de *et al.* Repositório datasus: organização e relevância dos dados abertos em saúde para a vigilância epidemiológica. **P2P E INOVAÇÃO**, v. 6, p. 50–59, 2019.

VALENTIM, Ricardo Alexsandro de Medeiros *et al.* A relevância de um ecossistema tecnológico no enfrentamento à Covid-19 no Sistema Único de Saúde: o caso do Rio

Grande do Norte, Brasil. **Ciência & Saúde Coletiva**, SciELO Brasil, v. 26, p. 2035–2052, 2021.

XGBOOST. [*S.l.*: *s.n.*]. Disponível em: `https://xgboost.readthedocs.io/en/stable/index.html`.

APÊNDICE A – DICIONÁRIO DE VARIÁVEIS

Ao final de todas as etapas de limpeza e criação de colunas, temos as seguintes colunas entrando no modelo:

Tabela 5 – Dicionário de Variáveis

Coluna	Tipo	Descrição
CODESTAB	text	Código de estabelecimento
CODMUNNASC	int8	Município de ocorrência, em codificação idêntica a de CODMUNRES, conforme tabela TABMUN
LOCNASC	int8	Local de ocorrência do nascimento, conforme a tabela: 9: Ignorado; 1: Hospital; 2: Outro Estab Saúde; 3: Domicílio; 4: Outros
IDADEMAE	int8	Idade da mãe em anos
ESTCIVMAE	int8	Estado civil, conforme a tabela: 1: Solteira; 2: Casada; 3: Viúva; 4: Separado judicialmente/Divorciado; 5: União consensual (versões anteriores); 9: Ignorado
ESCMAE	int8	Escolaridade, anos de estudo concluídos: 1: Nenhuma; 2: 1 a 3 anos; 3: 4 a 7 anos; 4: 8 a 11 anos; 5: 12 e mais; 9: Ignorado
CODOCUPMAE	text	Ocupação, conforme a Classificação Brasileira de Ocupações (CBO-2002)
QTDFILVIVO	text	Número de filhos vivos
QTDFILMORT	text	Número de filhos mortos
CODMUNNASC	int8	Município de ocorrência, em codificação idêntica a de CODMUNRES, conforme tabela TABMUN
GESTACAO	int8	Semanas de gestação, conforme a tabela: 9: Ignorado; 1: Menos de 22 semanas; 2: 22 a 27 semanas; 3: 28 a 31 semanas; 4: 32 a 36 semanas; 5: 37 a 41 semanas; 6: 42 semanas e mais
GRAVIDEZ	int8	Tipo de gravidez, conforme a tabela: 9: Ignorado; 1: Única; 2: Dupla; 3: Tripla e mais
PARTO	int8	Tipo de parto, conforme a tabela: 9: Ignorado; 1: Vaginal; 2: Cesáreo

CONSULTAS	int8	Número de consultas de pré-natal: 1: Nenhuma; 2: de 1 a 3; 3: de 4 a 6; 4: 7 e mais; 9: Ignorado
SEXO	int8	Sexo, conforme a tabela: 0: Ignorado, não informado; 1: Masculino; 2: Feminino
APGAR1	text	Apgar no primeiro minuto 00 a 10
APGAR5	text	Apgar no quinto minuto 00 a 10
RACACOR	int8	Raça/Cor: 1: Branca; 2: Preta; 3: Amarela; 4: Parda; 5: Indígena
PESO	text	Peso ao nascer, em gramas
CODMUNNATU	int8	Código do município de naturalidade da mãe
SERIESCMAE	int8	Série escolar da mãe. Valores de 1 a 8.
RACACORMAE	int8	Raça/cor da mãe
QTDGESTANT	text	Número de gestações anteriores
QTDPARTNOR	text	Número de partos vaginais
QTDPARTCES	text	Número de partos cesáreos
IDADEPAI	int8	Idade do pai
SEMAGESTAC	int8	Número de semanas de gestação.
TPMETESTIM	int8	Método utilizado. Valores: 1– Exame físico; 2– Outro método; 9– Ignorado.
CONSPRENAT	text	Número de consultas pré-natal
MESPRENAT	text	Mês de gestação em que iniciou o pré-natal
TPAPRESENT	int8	Tipo de apresentação do RN. Valores: 1– Cefálico; 2– Pélvica ou podálica; 3– Transversa; 9– Ignorado.
STTRABPART	int8	Trabalho de parto induzido? Valores: 1– Sim; 2– Não; 3– Não se aplica; 9– Ignorado.
STCESPARTO	int8	Cesárea ocorreu antes do trabalho de parto iniciar? Valores: 1–Sim; 2– Não; 3– Não se aplica; 9– Ignorado.
TPROBSON	text	Código do Grupo de Robson, gerado pelo sistema
STDNEPIDEM	int8	Status de DO Epidemiológica. Valores: 1 – SIM; 0 – NÃO.
STDNNOVA	int8	Status de DO Nova. Valores: 1 – SIM; 0 – NÃO.
IDANOMAL	int8	Anomalia congênita: 9: Ignorado; 1: Sim; 2: Não

ESCMAE2010	int8	Escolaridade 2010. Valores: 0 – Sem escolaridade; 1 –Fundamental I (1a a 4a série); 2 – Fundamental II (5a a 8a série); 3 – Médio (antigo 2o Grau); 4 – Superior incompleto; 5 –Superior completo; 9 – Ignorado.
CODUFNATU	int8	Código da UF de naturalidade da mãe
TPNASCASSI	int8	Nascimento foi assistido por? Valores: 1– Médico; 2–Enfermeira/obstetriz; 3– Parteira; 4– Outros; 9– Ignorado.
TPFUNCRESP	int8	Tipo de função do responsável pelo preenchimento. Valores:1– Médico; 2– Enfermeiro; 3– Parteira; 4– Funcionário docartório; 5– Outros.
TPDOCRESP	int8	Tipo do documento do responsável. Valores: 1-CNES; 2-CRM; 3-COREN; 4-RG; 5-CPF.
ESCMAEAGR1	text	Escolaridade 2010 agregada. Valores: 00 – Sem Escolaridade; 01 – Fundamental I Incompleto; 02 – Fundamental I Completo; 03 – Fundamental II Incompleto; 04 – Fundamental II Completo; 05 – Ensino Médio Incompleto; 06 – Ensino Médio Completo; 07 – Superior Incompleto; 08 – Superior Completo; 09 – Ignorado; 10 – Fundamental I Incompleto ou Inespecífico; 11 – Fundamental II Incompleto ou Inespecífico; 12 – Ensino Médio Incompleto ou Inespecífico.
CODMUNCART	int8	Descrição faltante
ESTADO	text	Estado de nascimento da criança
is_equal_CODMUNRES_ and_CODMUNNASC	bool	Se as colunas CODMUNRES e CODMUNNASC sao iguais
is_equal_CODMUNRES_ and_CODMUNNATU	bool	Se as colunas CODMUNRES e CODMUNNATU sao iguais
is_equal_CODMUNNASC_ and_CODMUNNATU	bool	Se as colunas CODMUNNAS e CODMUNNATU sao iguais
is_missing_IDADEPAI	bool	Se o campo IDADEPAI está preenchido ou não

Tabela adaptada do dado original. Informações adicionais sobre as colunas podem ser encontradas em (DICIONÁRIO. . . , s.d.).

Referência: Tabela da Fiocruz com informações referentes às colunas presentes no dado original.

APÊNDICE B – CÓDIGO FONTE

O código fonte desenvolvido pelo autor e utilizado para a geração dos resultados deste trabalho está disponível no repositório público do autor na plataforma Github e pode ser acessado através do link: https://github.com/GabrielSimonetto/saf_sinasc Dentro do repositório há um arquivo README.md que contém as instruções detalhadas de como executar corretamente o código do trabalho.

APÊNDICE C – ARTIGO

Neste apêndice é apresentado o artigo sobre este trabalho seguindo o padrão da Sociedade Brasileira de Computação.

Estudo de possíveis discriminantes de TEAF utilizando dados do DATASUS

Gabriel F. Simonetto[1], **Mateus Grellert**[1], **Jônata Tyska Carvalho**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, SC – Brasil

Abstract. *Fetal Alcohol Syndrome (FAS) is a congenital anomaly caused by the consumption of alcoholic beverages by a woman during pregnancy. Due to the difficulty of diagnosis of children with FAS and low representation, an early identification of the risk by healthcare professionals and public intervention policies is required. The diagnosis of this syndrome is challenging, as there are several other conditions with similar symptoms, so solutions that are able to assist healthcare professionals in identifying children with FAS are extremely important. This work proposes the implementation of a predictive model to aid in the diagnosis of FAS based on birth data and sociodemographic factors, using machine learning to assist in diagnosis and knowledge mining of this disorder.*

Resumo. *A Síndrome Alcoólica Fetal (SAF) é uma síndrome causada pelo consumo de alcool durante a gravidez, o que faz com que seja um problema de intervenção clara. A SAF é uma anomalia congênita que devido a dificuldade do diagnóstico e a baixa representatividade caso o risco seja identificado cedo pelo profissional de saúde e pelas políticas públicas de intervenção. O diagnóstico de SAF é desafiador, uma vez existe uma série de outras condições que possuem sintomas similares apresentados como é o caso de ..., ..., ... (RILEY; INFANTE; WARREN, 2011). Assim sendo, uma ferramenta baseada em mineração de dados históricos que possa auxiliar no processo de diagnóstico é importante para a comunidade de profissionais atuando nessa área. Esse trabalho propõe a implementação de um modelo preditivo para auxílio de diagnóstico de SAF com base em dados do nascimento e fatores sociodemográficos, utilizando aprendizado de máquina para auxiliar no diagnóstico e na mineração de conhecimento sobre esse transtorno.*

1. Introdução

As anomalias congênitas são um grupo de alterações estruturais ou funcionais que ocorrem durante a vida intrauterina e que podem ser detectadas antes, durante ou após o nascimento [Mendes et al. 2018]. Elas podem afetar diversos órgãos e sistemas do corpo humano e são causadas por um ou mais fatores genéticos, infecciosos, nutricionais e ambientais, podendo ser resultado de uma combinação desses fatores.

Exemplos de anomalias congênitas envolvem os Transtornos do Espectro Alcoólico Fetal (TEAF), que incluem uma vasta gama de condições patológicas que podem ocorrer no indivíduo que foi exposto ao álcool durante sua gestação. O espectro de distúrbios pode incluir defeitos congênitos, distúrbios neurológicos relacionados

ao álcool, assim como a manifestação mais grave, a Síndrome Alcoólica Fetal (SAF) [Cook et al. 2016].

Em muitos casos, a ocorrência de anomalias congênitas pode ser mitigada com a conscientização dos riscos de consumo de álcool durante a assistência pré-natal. Porém, a assistência pré-natal no Brasil ainda carece do desenvolvimento de rotinas e instrumentos confiáveis que auxiliem os profissionais de saúde nas ações de prevenção e diagnóstico precoce para esses problemas [Lopes et al. 2016].

A literatura demonstra que técnicas de Inteligência Artificial (IA) podem ser fortes aliadas nessa linha, gerando modelos capazes de auxiliar no diagnóstico e no tratamento destes transtornos [van Heeringen and Mann 2014, Arsenault-Lapierre et al. 2004]. Portanto, o desenvolvimento de ferramentas e tecnologias de IA capazes de apoiar os profissionais de saúde que atuam na área, assim como a democratização ao acesso à informação para a população geral, são recursos vitais para o combate de problemas da saúde mental.

As recentes iniciativas para acesso aberto a dados por parte do governo brasileiro proporcionam um amplo espaço amostral de dados, que possibilitam a realização de um trabalho de análise e busca por conhecimento orientado pelos processos de mineração de dados.

Esse trabalho procura encontrar discriminantes elucidativas sobre quais fenômenos, acontecimentos, e correlações estão envolvidos com o surgimento de um diagnóstico de TEAF assim como das outras anomalias a serem estudadas.

2. Referencial Teórico

Esta seção elucida os conceitos de SAF, aprendizado de máquina e amostragem, visando facilitar o entendimento de seções posteriores deste trabalho.

2.1. SAF e TEAF

O SAF (Síndrome Alcoólica Fetal) é uma condição decorrente do consumo de álcool durante a gravidez pela mãe.

Segundo [Queiroz 2016], ”A Síndrome Alcoólica Fetal se caracteriza pela tríade microcefalia-dismorfias faciais-déficit neurocognitivo.”. O SAF foi a primeira manifestação documentada do TEAF(Transtorno do Espectro Alcoólico Fetal), devido ao fato de ser o caso mais extremo do espectro, no qual o diagnóstico a partir da dismorfia facial é possível, entretanto, os déficits associados ao consumo de álcool podem tambem se manifestar em graus de intensidade menor.

Ao longo do tempo, o TEAF também foi estudado e definido pela comunidade científica como uma fonte de problemas para o portador em diversas áreas da vida como a saúde mental, educação, comportamento criminal e independência dos indivíduos [Riley et al. 2011].

Não existe ainda uma definição da comunidade científica acerca de uma quantidade segura de ingestão de álcool durante a gravidez, e, boa parte das pesquisas chama a atenção para o fato de que SAF, é uma das poucas anomalias congênitas com possibilidade total de intervenção, uma vez que existem somente o fator ambiental de consumo de álcool envolvido. [Queiroz 2016]

2.2. Aprendizado de máquina

Um algoritmo de aprendizado de máquina, é um algoritmo capaz de derivar conhecimento a partir de observações [Grellert 2018]. Existem diversos mecanismos matemáticos e estatísticos capazes de compilar as correlações e efeitos das variáveis entre si.

Esse trabalho irá aplicar técnicas de aprendizado de máquina supervisionado, pois está disponível se um bebê foi diagnosticado com SAF, ou outra anomalia congênita.

O problema de identificação de SAF é considerado um problema de classificação, pois existe um espaço discreto de possibilidades de respostas possíveis, nesse caso, duas, ou a criança tem um diagnóstico de SAF, ou não tem.

Neste trabalho, comparamos a performance entre os modelos de árvore de decisão, *random forest*, e XGB.

2.3. Amostragem

Segundo Barbetta [Barbetta 2012] quando se deseja conhecer uma população (conjunto de elementos que se deseja estudar), também é possível analisar uma amostra, para se aprender de maneira aproximada as características da população. Entretanto, esse é um processo que somente obtém sucesso se realizado com uma metodologia de seleção de elementos, de maneira que sejam representativos da população como um todo.

Neste trabalho, temos uma dificuldade devido a grande quantidade de dados, e portanto, será necessário o uso de amostras para realizar a análise da população. Além disso, há um segundo problema a ser considerado, os portadores de SAF são um grupo sub-representado na população, portanto, uma mera amostragem aleatória contaria com poucos indivíduos para serem analisados em cada amostra.

Para lidarmos com isso, usamos uma estratégia de amostragem estratificada. Esta, tem como objetivo, reduzir o viés que seria criado ao analisar-se uma amostra contendo poucas entradas do grupo de interesse sub representado [Barbetta 2012]. Em outras palavras, a analise estatística, ou o modelo, falhariam em produzir conhecimento confiável se não tivessem um espaço amostral significativo dentro de cada classe a ser analisada.

Realizamos o processo de amostragem, treino e teste, múltiplas vezes, e agrupamos os resultados de todas as amostras, para cada parâmetro sendo analisado. Este protocolo é apoiado pelo teorema central do limite.

O Teorema central do limite diz que as médias e variâncias de um espaço amostral de variáveis tende a uma distribuição normal se tirarmos uma medida de agrupamento para varias amostras de uma mesma população [Bussab and Morettin 2002]

Assim, a estratégia de amostragem estratificada e a aplicação do Teorema Central do Limite fornecem uma base sólida para lidar com a dificuldade de grandes volumes de dados, garantindo representatividade e confiabilidade nas análises estatísticas realizadas neste trabalho.

3. Solução Proposta

O fluxo da solução proposta por este trabalho é apresentado na Figura 1.

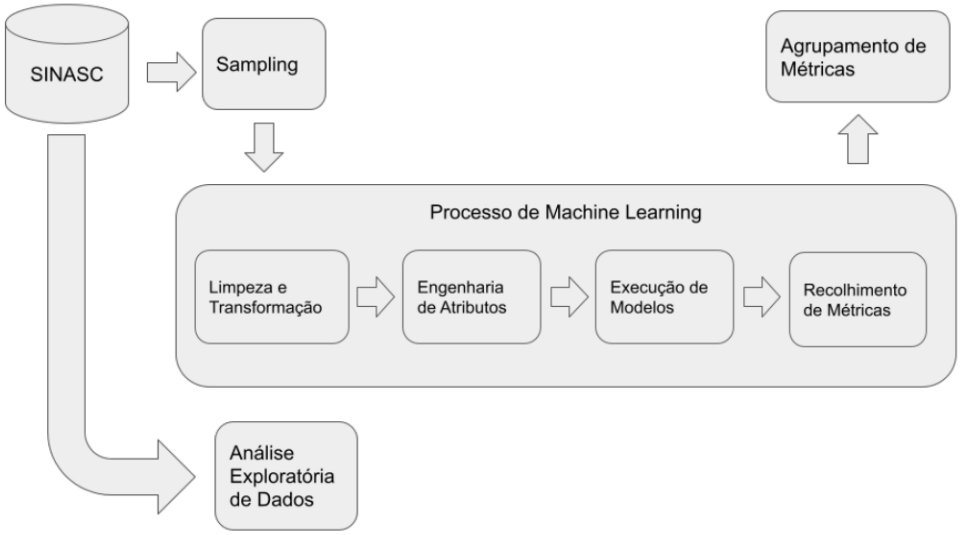


Figure 1. Metodologia global

Usaremos dados do SINASC para treinar os modelos. A base de dados do SINASC oferecida pelo DATASUS contém dados de todos os estados do país entre os anos de 1994 a 2019, totalizando 14,087 GB. Para esse trabalho, usaremos somente os dados de 2010 a 2019, devido tanto a modernização do sistema ao longo dos anos, como para representar a realidade atual do país melhor. Além disso, devido ao grande volume de dados, e ao pequeno número de entradas de SAF (1262 casos), precisamos usar uma estratégia de estratificação de dados.

Para realizar essa estratificação, primeiro, extraímos todas as entradas SAF positivas entre 2010 e 2019.

Isso é realizado através da coluna "CODANOMAL", que representa o "Código da anomalia (CID 10)" [dic] da criança nascida, assim, filtramos essa coluna usando:

- Q86 - Síndromes com malformações congênitas devidas a causas exógenas conhecidas, não classificadas em outra parte [fio]
- Q870 - Síndromes com malformações congênitas afetando predominantemente o aspecto da face [cid]

O *dataset* final contém 1262 entradas. Isso faz com que nossa estratificação precise de 6310 valores selecionados aleatoriamente, totalizando *datasets* de trabalho de 7572 entradas por iteração.

A base de dados então passa por uma série de processos de limpeza e enriquecimento: imputação de dados faltantes, normalização de valores contínuos, binarização de valores categóricos, tradução de codificação para linguagem natural, dentre outros

ESTCIVMAE	Estado Civil	ESTCIV.CNV	caracter	1	Situação conjugal da mãe: 1– Solteira; 2– Casada; 3– Viúva; 4– Separada judicialmente/divorciada; 5– União estável; 9– Ignorada.

Figure 2. Exemplo de codificação - Coluna ESTCIVMAE - Fonte: DANTPS - CGIAE

Após essas etapas, roda-se uma instância de modelo para cada sample criado, colhem-se as métricas F1 e ROC AUC de cada instância, e uma etapa final de agregação dos resultados de cada um dos 100 modelos expressará a realidade estatística da resolução do problema pelo modelo, e expressividade dos discriminantes como descritores do problema.

4. Análise Exploratória de Dados

Como guia desta análise, usa-se alguns dos fatores que são teorizados como discriminantes proeminentes de SAF segundo a literatura [Baptista et al. 2017]

- Estado Civil (ausência de parceiro fixo)
- Menor peso do bebê ao nascer
- Bebês do sexo feminino podem ter maior perda de peso devido a TEAF
- Baixa escolaridade da mãe
- Baixo nível socioeconômico da mãe
- Idade superior a 30 anos
- Cor de pele não branca
- Desemprego

Existe uma dificuldade no manuseio da grande quantidade de dados disponíveis. Para essa análise em específico, produziu-se um *dataset* de todas as entradas SAF positivas entre 2010 e 2019, e, para atuar como base de comparação, usou-se um *dataset* de dados neutros com 2% dos dados entre 2010 e 2019, o *dataset* de positivos contém 1262 entradas positivas, enquanto o *dataset* neutro possui 584269 entradas.

Nessa análise, falhamos em encontrar as mesmas correlações indicadas pela literatura de estudo clínico.As variáveis mais notáveis como discriminantes entre os grupos SAF positivos e SAF negativos são: o peso da criança (sem distinção de gênero), e num grau menor, a idade da mãe acima de 30 anos, e as marcações de APGAR. A teoria a respeito das variáveis de estado civil, baixa escolaridade da mãe, baixo nível socioeconômico da mãe, cor de pele não branca e desemprego como potenciais discriminantes de SAF não foi amparada pela análise exploratória.

4.1. Fontes de dados SINASC

Neste trabalho, usaremos o *PySUS* [Coelho et al. 2021] como API (*Application Programming Interface*) que permite o acesso aos dados do DATASUS. Ela foi criada inicialmente para auxiliar nas pesquisas envolvendo os casos de dengue no Brasil através de análises sobre os dados do banco Sistema de Informações de Agravos de Notificação (SINAN). A versão atual da biblioteca *PySUS* possui acesso a diversos outros bancos, permitindo que ela seja utilizada em diversas pesquisas voltadas à saúde.

Inicialmente, é necessário baixar individualmente cada arquivo por ano e estado, sendo necessárias a realização de quaisquer agregações a partir desses arquivos. Em específico para nosso caso, estamos interessados em particular no Sistema de Informações sobre Nascidos Vivos (SINASC) [sit] que é responsável por registrar variadas informações sobre o trabalho de parto, e condições da mãe e da criança. Este sistema é instrumental para a construção de indicadores úteis para o planejamento de gestão dos serviços de saúde.

4.2. Sampling

Devido a dimensão dos dados, não é possivel carregar em memória todo o dataset histórico dos anos 1994 a 2019, que juntos pesam um total de 14,087 GB.

Dessa forma, esse trabalho usa uma metodologia de sampling para que seja possível computar diversos espaços amostrais dos dados, e ao agrupá-las, poder ter uma representação do espaço total de dados.

Para esse fim, primeiro, definimos que usaremos apesar informações dos anos de 2010 a 2019, excluindo os anos de 1994 a 2010, uma vez que dados mais recentes vão ser mais capazes de expressar o estado atual dos fatores que influenciam SAF.

Depois disso, usaremos o conceito de balanceamento de classes para definir um alfa de multiplicação de maneira que tenhamos 5x mais entradas negativas do que positivas para SAF.

Dessa forma, chegamos ao número de 1262 entradas positivas em todo o país, e por consequência, escolhemos 6310 entradas para completar o dataset de treino para cada amostra, completando um dataset com um total de 7572 entradas.

Essas 6310 entradas selecionadas randomicamente serão estratificadas por estado, para que seja possível representar cada estado do país no treinamento dos modelos. Segue abaixo a distribuição encontrada para a presença de SAF ao longo dos anos de 2010 a 2019.

4.3. Limpeza e Transformação

Uma vez carregados os dados criados no processo de sampling, realiza-se uma série de procedimentos para que o dado esteja pronto para a ingestão do modelo de aprendizagem de máquina.

4.3.1. Remoção de Colunas Excedentes

Para dados faltantes, definiremos que se uma coluna possui abaixo de 40% de dados preenchidos, essa coluna será removida da nossa análise.

Isso se aplica as colunas:

- RACACORN
- CODANOMAL
- NUMREGCART
- DTREGCART
- CODCART
- RACACOR_RN
- DTRECORIGA
- CODMUNCART
- DTRECORIG
- IDADEPAI

Ultimamente, checa-se a semântica dessas colunas, e se existe alguma outra coluna que é capaz de expressar a mesma informação.

Faremos aqui entretanto uma exceção pra coluna “IDADEPAI“, pois a falta de informação pode ter o valor semântico de que o pai não estaria presente na gravidez da mãe. Essa é uma escolha que possui seus contras, já que em meio a pais ausentes tambem haverão simplesmente erros de inputação por isso deve se exercer cautela ao se tirar conclusões a respeito dessa feature.

Por fim, fazem-se algumas análises sobre a semântica das colunas remanescentes, e se esTas demonstrarão alguma utilidade ou não, na tabela do dicionário de variáveis **??** é possível ver todas as colunas que de fato sao utilizadas no treinamento.

4.4. Execução dos Modelos

O treino acontece com 3 tipos de modelos de *Machine Learning*, *Decision Tree*, *Random Forest*, e *XGBoost* (XGB). Escolhemos 3 modelos baseados em árvore pois o escopo do nosso problema gira em torno da capacidade de expressar quais variávies são as mais relevantes para a decisão do modelo. Precisamos de modelos interpretáveis e expressivos para que seja possível a extração de conhecimento dos modelos.

Cada definição de modelo será rodada para cada uma das N samples de dados definidas.

Os modelos não fazem ajuste de hiperparâmetros, sendo rodados apenas com a configuração básica conforme mostra a Tabela 1

Modelo	**Hiperparâmetros**
RF	random_state=0
XGB	random_state=0
	learning_rate=1,
	max_depth=2,
	n_estimators=2,
	objective=’binary:logistic’
DT	random_state=0

Table 1. Tabela de Hyperparâmetros

Os modelos rodam sua pipeline de treinamento, e então, são salvos os metadados sobre o treinamento desses modelos na classe *ModelEvaluator*, que é genérica para todos os modelos.

4.5. Recolhimento de Métricas

A classe *ModelEvaluator* é responsável por guardar as métricas de cada rodada de treinamento por sample. Isso faz com que seja possível a análise de como uma classe de modelo se comporta no dataset do SINASC, baseado em como cada instância de modelo em específico performou em sua sample.

A metodologia de treino e teste de modelo se dá por validação cruzada com 10 rodadas. São recolhidas as métricas:

- F1
- ROC_AUC
- Feature Importance

Com base nessas informações, podemos tomar decisões informadas sobre a escolha e otimização dos modelos para atingir os melhores resultados possíveis.

Uma vez criados todos os *ModelEvaluator*, inicia-se a etapa de agregação de métricas para que seja possível analisar a performance global de um único tipo de modelo.

5. Análise de Resultados dos Modelos

Apresentamos agora a performance dos modelos de acordo com as métricas escolhidas, assim como mostramos as features consideradas de maior importância para a previsão da presença de SAF.

5.1. Análise de Métricas

A tabela abaixo apresenta a média das métricas obtidas para cada modelo ao serem rodados para 100 amostras de dados.

Modelo	**Hiperparâmetros**	**Média F1**	**Média ROC_AUC**
RF	random_state=0	0.27	0.60
XGB	random_state=0 learning_rate=1, max_depth=2, n_estimators=2, objective='binary:logistic'	0.41	0.73
DT	random_state=0	0.23	0.47

Table 2. Tabela de Resultados

Ambos os modelos de *Random Forest* e *Decision Tree* tiveram uma baixa performance em resumir o SAF como uma soma das variáveis fornecidas no treinamento dos modelos.

O modelo XGB teve uma performance consideravelmente maior em ambas as métricas. É difícil dizer se esses números estão próximos do máximo teórico com as atributos sendo utilizadas nesse modelo em específico, ou se existem ganhos possíveis a serem feitos durante o pipeline de treinamento.

5.2. Importância dos atributos

Abaixo podemos ver a importância de atributos dos modelos escolhidos para nosso projeto: DT 3, RF 4 e XGB 5.

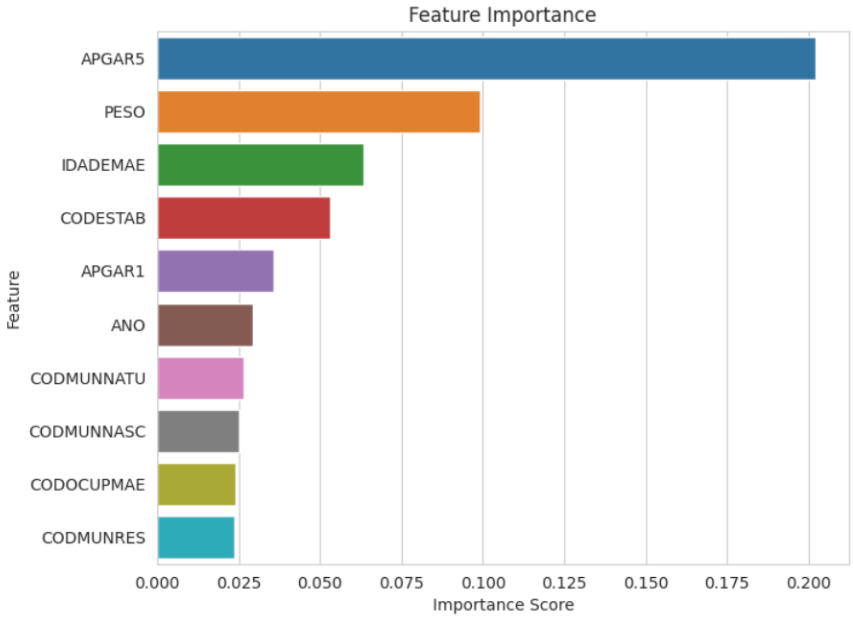


Figure 3. Gráfico de importância de atributos para DT

O modelo de *Decision Tree* dá grande importância ao APGAR5, ao peso do bebê e a idade da mãe. Algumas variáveis curiosas aparecem no gráfico, como o código do estabelecimento, o município de naturalidade, nascimento e residência, essas variáveis poderiam apontar uma região localizada na qual existe maior incidência de SAF, ou possivelmente, uma região localizada na qual se diagnostica mais SAF.

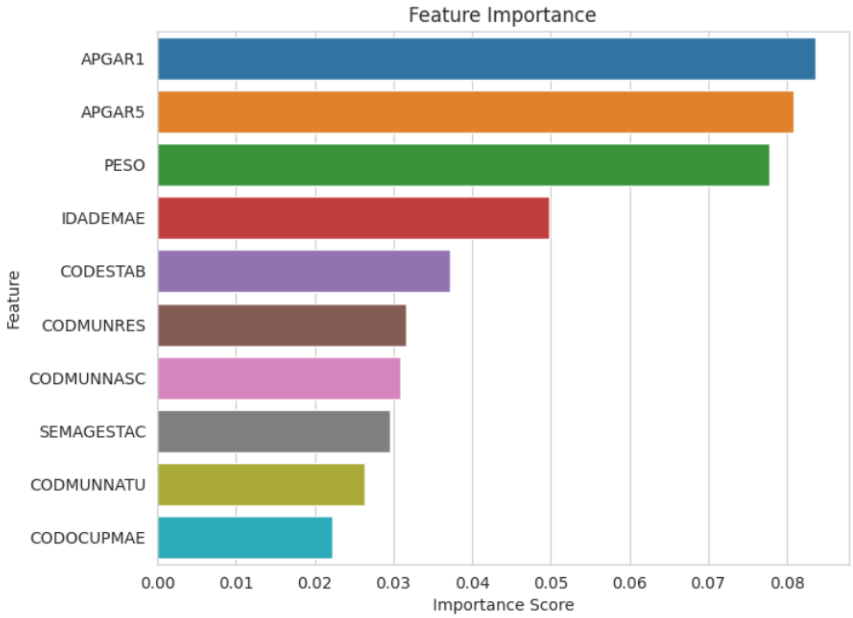


Figure 4. Gráfico de importância de atributos para RF

O modelo de *Random Forest* aponta boa parte das mesmas variáveis que a *Decision Tree*, mas observa-se que existe menos disparidade entre a importância do APGAR5 e o resto das atributos.

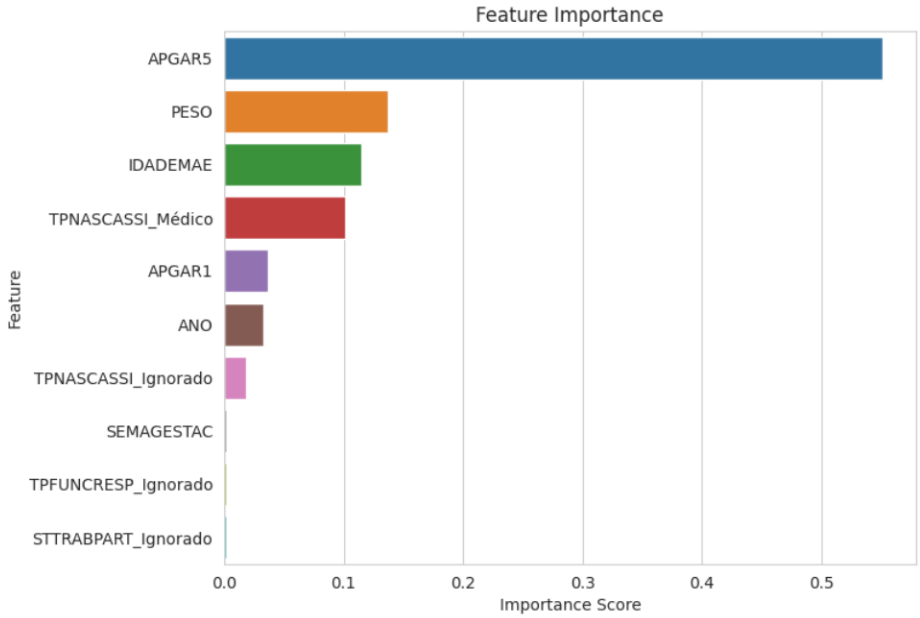


Figure 5. Gráfico de importância de atributos para XGB

Já a importância de atributos do XGB diferem um pouco dos outros modelos. Não é mencionado coisas como a localidade, e existe uma estratégia de priorização do APGAR5 como fator de maior relevância.

Todos os modelos apontaram APGAR5 e APGAR1 como fatores bastante relevantes no diagnóstico de SAF, assim como o peso observado da criança e a idade da mãe.

6. Conclusão

A Síndrome Alcoólica Fetal (SAF) é uma condição grave que afeta indivíduos expostos ao álcool durante a gestação. É uma preocupação de saúde pública devido aos efeitos adversos que pode causar no desenvolvimento físico, cognitivo e comportamental das crianças afetadas.

O fato de SAF ser uma doença congênita com potencial de 100% de redução faz com que o ímpeto da intervenção clínica surja, somente é necessário que identifiquem-se os casos de risco com antecedência, e então, as devidas providências sejam tomadas.

Este estudo teve como objetivo principal analisar possíveis discriminantes de SAF através de um pipeline de dados e uso de machine learning utilizando dados do Sistema de Informações sobre Nascidos Vivos (SINASC).

Também foi feita uma releitura das conclusões acadêmicas históricas sobre o problema de SAF, agora, a partir de uma perspectiva de dados massivos, ao invés de uma perspectiva de análise e entrevista clínica.

Essas descobertas são de grande importância para profissionais de saúde e pesquisadores que lidam com a prevenção e o tratamento da SAF. A identificação precoce de gestações de alto risco permite a adoção de medidas preventivas e intervenções adequadas, visando reduzir os impactos da exposição ao álcool durante a gestação. Além disso, a abordagem metodológica proposta neste estudo pode servir como base para futuras pesquisas e aplicações relacionadas à predição de eventos em outros contextos de saúde pública.

Para trabalhos futuros, uma das sugestões de melhorias é a incorporação de novas bases de dados para complementarem as informações trazidas pelo SINASC. Além disso, observa-se que o uso de classes SAF vs. neutralidade não resolve os problemas de diagnóstico diferencial, e portanto, é interessante trazer modelos que classifiquem também entre diferentes anomalias e distúrbios.

References

cidq870.

dictsinasc.

fiocruzcid10.

sitesinasc.

Arsenault-Lapierre, G., Kim, C., and Turecki, G. (2004). Psychiatric diagnoses in 3275 suicides: a meta-analysis. *BMC psychiatry*, 4(1):1–11.

Baptista, F. H., Rocha, K. B. B., Martinelli, J. L., Avó, L. R. d. S. d., Ferreira, R. A., Germano, C. M. R., and Melo, D. G. (2017). Prevalência e fatores associados ao consumo de álcool durante a gravidez. *Revista Brasileira de Saúde Materno Infantil*, 17:271–279.

Barbetta, P. A. (2012). *Estatistica Aplicada as Ciencias Sociais*. Editora Atlas, 8th edition.

Bussab, W. d. O. and Morettin, P. A. (2002). *Estatística Básica*. Saraiva, São Paulo, 5th edition.

Coelho, F. C., Baron, B. C., de Castro Fonseca, G. M., Reck, P., and Palumbo, D. (2021). Alertadengue/pysus: Vaccine.

Cook, J. L., Green, C. R., Lilley, C. M., Anderson, S. M., Baldwin, M. E., Chudley, A. E., Conry, J. L., LeBlanc, N., Loock, C. A., Lutke, J., Mallon, B. F., McFarlane, A. A., Temple, V. K., and Rosales, T. (2016). Fetal alcohol spectrum disorder: a guideline for diagnosis across the lifespan. *CMAJ*, 188(3):191–197.

Grellert, M. (2018). Machine learning mode decision for complexity reduction and scaling in video applications.

Lopes, C. S., Abreu, G. d. A., Santos, D. F. d., Menezes, P. R., Carvalho, K. M. B. d., Cunha, C. d. F., Vasconcellos, M. T. L. d., Bloch, K. V., and Szklo, M. (2016). Erica: prevalência de transtornos mentais comuns em adolescentes brasileiros. *Revista de Saúde Pública*, 50.

Mendes, I. C., Jesuino, R. S. A., Pinheiro, D. d. S., and Rebelo, A. C. S. (2018). Anomalias congênitas e suas principais causas evitáveis: uma revisão. *Revista Médica de Minas Gerais*, 28(1):1–6.

Queiroz, M. R. (2016). A síndrome alcoólica fetal: Revisão sistemática.

Riley, E. P., Infante, M. A., and Warren, K. R. (2011). Fetal alcohol spectrum disorders: an overview. *Neuropsychology review*, 21(2):73–80.

van Heeringen, K. and Mann, J. J. (2014). The neurobiology of suicide. *The Lancet Psychiatry*, 1(1):63–72.