UNIVERSIDADE FEDERAL DE SANTA CATARINA CAMPUS ARARANGUÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

Rafael Damiani Alves

PREDIÇÃO DA FORÇA DE REAÇÃO DO SOLO DURANTE A CAMINHADA E CORRIDA NA ÁGUA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Araranguá
2022

Rafael Damiani Alves

PREDIÇÃO DA FORÇA DE REAÇÃO DO SOLO DURANTE A CAMINHADA E CORRIDA NA ÁGUA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

Dissertação submetida ao Programa de Pós Graduação em Tecnologias da Informação e
Comunicação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Tecnologias da Informação e Comunicação.
Orientador: Prof. Cristian Cechinel, Dr. Coorientador: Prof. Alessandro Haupenthal , Dr.

Araranguá
2022

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

Alves, Rafael
PREDIÇÃO DA FORÇA DE REAÇÃO DO SOLO DURANTE A CAMINHADA E CORRIDA NA ÁGUA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS / Rafael Alves ; orientador, Cristian Cechinel, coorientador, Alessandro Haupenthal, 2022.
127 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Araranguá, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2022.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Mineração de Dados, Modelo preditivo, Força de reação do solo. I. Cechinel, Cristian . II. Haupenthal, Alessandro. III. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Tecnologias da Informação e Comunicação. IV. Título.

Rafael Damiani Alves

PREDIÇÃO DA FORÇA DE REAÇÃO DO SOLO DURANTE A CAMINHADA E CORRIDA NA ÁGUA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

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

Prof.(a) Heiliane de Brito Fontana, Dr.(a)
Universidade Federal de Santa Catarina

Prof.(a) Merisandra Côrtes de Mattos, Dr.(a)
Universidade do Extremo Sul Catarinense

Prof. Roderval Marcelino, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Tecnologias da informação e comunicação.

Documento assinado digitalmente
Fernando Jose Spanhol
Data: 10/03/2022 17:41:50-0300
CPF: 642.656.419-20
Verifique as assinaturas em https://v.ufsc.br

Prof. Fernando José Spanhol
Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Cristian Cechinel
Data: 10/03/2022 18:11:55-0300
CPF: 454.628.449-72
Verifique as assinaturas em https://v.ufsc.br

Prof. Cristian Cechinel, Dr.
Orientador

Araranguá, 2022.

Este trabalho é dedicado a minha namorada e aos meus queridos pais.

AGRADECIMENTOS

Primordialmente agradeço a Deus por me dar força, fé e sabedoria para a produção desta pesquisa e por permitir que aos poucos eu alcance meus objetivos, como é o caso da obtenção do Grau de mestre em Tecnologias da Informação e Comunicação.

Agradeço imensamente também aos meus pais Sander e Luciani, por todo amor, confiança e atenção passados para mim. Agradeço ainda os mesmos por terem me proporcionado todas as oportunidades possíveis para a realização do mestrado e consequentemente deste trabalho, deixo claro que sem os mesmos nada disso seria possível. Necessito salientar minha gratidão a minha namorada, que nos momentos de cansaço e esgotamento me proporcionou alegria e descontração, contribuindo então para a retomada do foco na conquista dos meus objetivos.

Em relação ao ambiente acadêmico demonstro aqui meus honestos agradecimentos ao meu orientador Dr. Cristian Cechinel e também ao meu coorientador Dr. Alessandro Haupenthal, por esclarecerem todas as minhas dúvidas e também por em nenhum momento desistirem dessa pesquisa.

Escreva algo que valha a pena ler ou faça algo que valha a pena escrever.

(Benjamin Franklin)

RESUMO

O avanço das tecnologias da informação e comunicação tem gerado progresso nas técnicas de mineração de dados. Os processos e técnicas de mineração de dados são utilizados para procurar padrões que proporcionam vantagens e aperfeiçoamentos em diversas áreas, tais como: marketing, detecção de fraude, investimentos financeiros, área da saúde entre outras. Com o objetivo de auxiliar na prescrição dos exercícios de caminhada e corrida na água, no que refere-se ao conhecimento e controle das forças de reação do solo durante exercícios subaquáticos, esta pesquisa procurou gerar modelos preditivos para o pico das componentes vertical, ântero-posterior e da resultante da força de reação do solo durante a caminhada e corrida na água, por meio das técnicas de mineração de dados. Os dados utilizados neste estudo, contam com informações como sexo, idade, massa corporal, cintura, coxa e estatura de 143 pessoas. No que se refere aos exercícios subaquáticos, os dados contam com informações sobre exercícios de caminhada e corrida na água em 3 diferentes tipos de velocidades: normal, rápida e lenta. Os níveis de imersão presentes nos dados são 0,75, 0,90, 1,05, 1,20, e 1,35 metros. Foram utilizados 14 tipos diferentes de algoritmos de regressão para predição das variáveis FR, Fy e Fx, por meio de uma biblioteca python denominada de scikit-learn. Através do uso das técnicas de mineração de dados, como cross-validation K-Fold, foram gerados modelos de predição que conseguiram prever até 93% das amostras presentes na base de dados usada. Os algoritmos Support Vector Regression e Tweedie Regressor foram os algoritmos que demonstraram maior potencial na predição das forças de reação do solo durante os exercícios aquáticos em questão.

Palavras-chave: Mineração de Dados. Modelo preditivo. Força de reação do solo. Caminhada. Corrida.

ABSTRACT

The advancement of information and communication technologies has generated progress in data mining techniques. Data mining processes and techniques are used to look for patterns that provide advantages and improvements in several areas, such as: marketing, fraud detection, financial investments, healthcare, among others. In order to assist in the prescription of walking and running exercises in the water, with regard to the knowledge and control of ground reaction forces during underwater exercises, this research sought to generate predictive models for the peak of the vertical, antero- and the resulting ground reaction force during walking and running in the water, using data mining techniques. The data used in this study have information such as sex, age, body mass, waist, thigh and height of 143 people. As far as underwater exercises are concerned, the data has information about walking and running exercises in the water at 3 different types of speeds: normal, fast and slow. The immersion levels present in the data are 0.75, 0.90, 1.05, 1.20, and 1.35 meters. Fourteen different types of regression algorithms were used to predict the variables FR, Fy and Fx, using a python library called scikit-learn. Through the use of data mining techniques, such as cross-validation K-Fold, prediction models were generated that were able to predict up to 93% of the samples present in the database used. The Support Vector Regression and Tweedie Regressor algorithms were the algorithms that showed the greatest potential in predicting the ground reaction forces during the aquatic exercises in question.

Keywords: Data Mining. Predictive model. Ground reaction force. Walk. Race.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CSV - Comma-separated values

DDL - Linguagem de definição de dados DF - Data Frame

FR - valor máximo da resultante das componentes da FRS durante a realização do contato com a plataforma.

FRS - Força de reação do solo.

Fx - Componente ântero-posterior da força de reação do solo. Fy - Componente vertical da força de reação do solo.

GB - Gradient Boosting

KDD - Processo de Descoberta do Conhecimento LARS - regressão de ângulo mínimo

MAE - Mean absolute error

MD - Mineração de dados

MSE - Erro quadrático médio

RMSE - Raiz quadrada do erro-médio

R^2 - Coeficiente de determinação

SQL - Linguagem de Consulta Estruturada

SVM - Support Vector Machines

SVR - Support Vector Regression

SCRIPT - Linguagem de script (Algoritmos de computadores)

WEKA - Waikato Environment for Knowledge Analysis

LISTA DE FÓRMULAS

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA E JUSTIFICATIVA DA PESQUISA

A Para as pessoas que buscam o aperfeiçoamento ou manutenção da saúde, a água fornece um ambiente seguro, agradável e com uma probabilidade menor de lesão e acidentes (TAKESHIMA et al., 2002).

As vantagens provenientes da prática de atividade física na água são diversas, tais como o aumento da força muscular, do condicionamento aeróbico, melhora do equilíbrio e da flexibilidade corporal (TAKESHIMA et al., 2002). Além do mais, a sensação de bem estar causada pela temperatura da água e a minimização da sudorese e da exposição do corpo no decorrer do exercício da experiência proporcionam vantagens ligadas ao avanço da qualidade de vida (TAKESHIMA et al., 2002).

Deste modo, os exercícios aquáticos estão sendo aplicados em programas de reabilitação e/ou condicionamento físico para um grande número de pessoas (MASUMOTO et al., 2004).

Conforme Haupenthal (2013), a prática da atividade física em meio aquático tem como uma das metas fundamentais a redução do peso aparente (diferença entre o peso do indivíduo e a força do empuxo). Sendo assim, o profissional incumbido pelo exercício usa o meio aquático como forma de diminuição da força de sustentação do peso corporal.

Dessa forma, constata-se que as forças de impacto intervindo sobre o sistema musculoesquelético são menores (ROESLER et al., 2006). Uma das formas de estudar a mudança gerada pela ação do empuxo durante o exercício na água é a análise do pico de força na componente vertical da força de reação do solo (FRS) (HAUPENTHAL, 2013).

A FRS vertical é a resposta ou reação gerada a contar do contato do pé do indivíduo no decorrer do exercício e é a principal representante da magnitude do impacto que irá agir nas estruturas corporais (KELLER et al., 1996). De acordo com Haupenthal (2013), às mudanças vigentes nos movimentos executados na água geram mudanças também na componente ântero-posterior.

Destaca-se que é de grande importância saber quais variáveis podem implicar na FRS. Para a prescrição de exercícios subaquáticos, é importante compreender de que forma o tipo de exercício (como caminhada e corrida) causam alterações nas componentes da FRS e

principalmente o valor que pode ser obtido para a variação da força por meio da variação de cada variável (HAUPENTHAL, 2013).

No estudo de Haupenthal (2013), o mesmo apresenta uma pesquisa a qual possuía a meta de colaborar na prescrição dos exercícios de caminhada e corrida na água, no que tange o conhecimento e controle das forças de reação do solo (FRS) durante o contato com o fundo da piscina. O trabalho de Haupenthal (2013) apresentou modelos de regressão para o pico das componentes vertical (Fy), ântero-posterior (Fx) e da resultante (FR) da FRS durante a caminhada e corrida na água.

O avanço das tecnologias da informação e comunicação tem proporcionado o avanço das técnicas de mineração de dados, as mesmas vêm sendo usadas para realizar descobertas de padrões. As técnicas de mineração de dados podem ser empregadas em quaisquer áreas, normalmente gerando diversos benefícios para a área que utiliza a mesma (ALVES et al., 2018).

Baseado nesse contexto, nota-se que é possível utilizar as técnicas de mineração de dados na área da saúde e da fisioterapia. Dessa maneira pode-se buscar predizer as forças de reação do solo (FRS), para procurar algum padrão que ajude os profissionais da saúde a prescrever exercícios subaquáticos de forma ainda mais segura e efetiva. Dessa maneira observa-se que modelos preditivos obtidos por meio da mineração de dados e aprendizado de máquina podem ser utilizados como ferramenta de apoio pelos profissionais da área para a prescrição dos exercícios subaquáticos.

Diante deste contexto, surge a seguinte pergunta de pesquisa: Com qual exatidão é possível predizer a força de reação do solo durante a caminhada e corrida na água?

1.2 OBJETIVOS

Essa dissertação busca gerar modelos para predição da força de reação do solo durante a caminhada e corrida na água utilizando técnicas de mineração de dados. A predição da FRS de maneira automática pode ajudar os gestores e especialistas da área da saúde a melhorar seus trabalhos de prescrição dos exercícios de caminhada e corrida na água. Dessa forma, por intermédio dos modelos de predição e consequentemente da descoberta de conhecimento, busca-se oferecer contribuições para a criação de iniciativas que visem melhorias na reabilitação e no setor da saúde de modo geral.

1.2.1 Objetivo Geral

Gerar modelos para predição da força de reação do solo durante a caminhada e corrida na água por meio de algoritmos de regressão.

1.2.2 Objetivos Específicos

- Desenvolver modelos preditivos para os valores de pico das componentes vertical e ântero-posterior e a resultante da FRS para a caminhada e corrida dentro da água;
- Verificar a eficácia do modelo para estimar os valores de pico para as componentes vertical e ântero-posterior e a resultante da FRS para a caminhada e corrida dentro da água;
- Implementar scripts para a fase de transformação;
- Utilizar diversos algoritmos de regressão usando Python.

1.3 ADERÊNCIA DO OBJETO DE PESQUISA AO PPGTIC

Programa de Pós-Graduação em Tecnologias da Informação e Comunicação (PPGTIC) busca auxiliar e solucionar problemas de natureza interdisciplinar. O mesmo está segmentado em três diferentes linhas de pesquisa, são elas: Tecnologia Computacional, Tecnologia Educacional e Tecnologia, Gestão e Inovação.

Esta dissertação foi elaborada no escopo da linha de pesquisa Tecnologia Computacional e está diretamente ligada com o objeto de formação do PPGTIC devido ao fato de utilizar mineração de dados e outras técnicas computacionais em dados da área da saúde buscando assim auxiliar na resolução de problemas de natureza da área da saúde. Em relação aos trabalhos já defendidos no PPGTIC, muitos são os trabalhos que utilizam mineração de dados e aprendizado de máquina, porém não há dissertações com a temática relacionada a predição das forças de reação do solo durante a caminhada e a corrida na água. Sendo assim, esta dissertação busca preencher essa lacuna encontrada.

2 REFERENCIAL TEÓRICO

Este capítulo expõe a sistematização de conceitos realizada considerando a temática principal de pesquisa desta dissertação. Sendo assim, são demonstrados os conceitos sobre saúde, locomoção humana, marcha na água, mineração de dados, mineração de dados na saúde, e ainda aprendizagem de máquina.

2.1 MINERAÇÃO DE DADOS

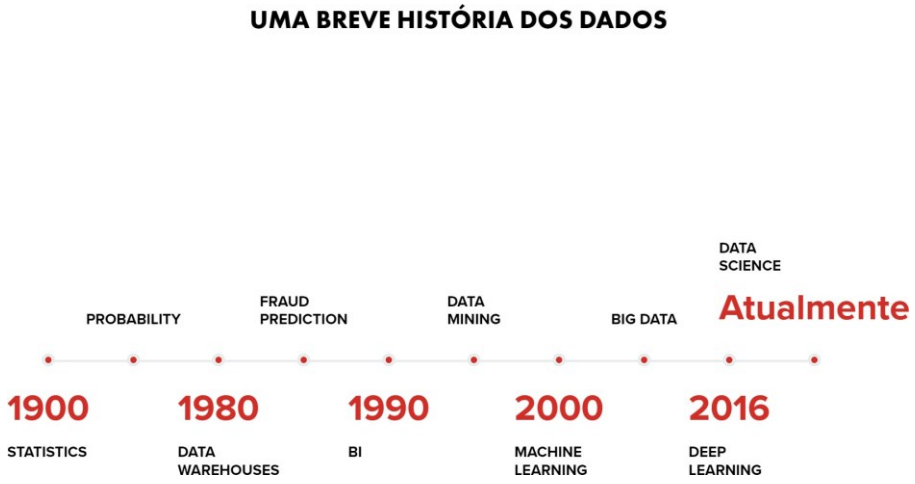
O rápido progresso das Tecnologias da Informação impulsionou a geração de dados digitais e tecnologias de armazenamentos dos mesmos, levando a um enorme e crescente volume de dados armazenados em bancos de dados, repositórios e na World Wide Web.

Segundo Zhou (2003), embora informações valiosas possam estar escondidas "por trás dos dados", o grande volume de dados torna complexa, se não impossível, para os seres humanos extraí-los de maneira manual. Sendo assim, existia uma riqueza de dados, porém pouca informação sendo gerada com os mesmos.

Para Zhou (2003), devido a essa necessidade, durante o final dos anos 1980, surgiu um novo processo chamado de mineração de dados, que se dedica a extrair conhecimento de grandes volumes de dados, com a ajuda da computação moderna, ou seja, do computador e suas ferramentas.

De acordo com Han, Kamber e Pei (2011), a mineração de dados como conhecemos hoje trata-se de uma área interdisciplinar, pois a mesma baseia-se em diversas disciplinas e áreas. Na Figura 1 conseguimos identificar a evolução das tecnologias até chegarmos no atual processo de mineração de dados.

Figura 1 – Evolução da Mineração de Dados.



Fonte: Elaborado pelo autor.

Para Zhou (2003) as áreas e tecnologias que contribuíram para mineração de dados são: recuperação de informações, bancos de dados, visualização de dados, estatística, aprendizado de máquina, computação paralela e até mesmo computação distribuída. Porém, Zhou (2003) destaca que por mais que diversas disciplinas e tecnologias tenham influenciado a mineração de dados, as principais contribuições vieram de fato das áreas de: banco de dados, aprendizado de máquina e estatística.

Existem inúmeras óticas sobre o conceito de mineração de dados, desta maneira destaca-se nesta dissertação três visões sobre o conceito da mesma:

1-Para Hand et al. (2001), baseando-se no panorama da estatística: Mineração de Dados é a análise de um conjuntos de dados, com o objetivo de descobrir relacionamentos inesperados e de abreviar os dados de uma forma que eles sejam compreensíveis e úteis.

2-Em Cabena et al. (1998), em relação à visão de banco de dados: Mineração de Dados é uma área interdisciplinar que agrega técnicas de máquinas de conhecimentos, estatísticas, reconhecimento de padrões, banco de dados e visualização, para que seja possível coletar informações de grandes bases de dados.

3-Para Fayyad et al. (1996), a mineração de dados tem uma perspectiva de aprendizado de máquina: Mineração de Dados é um dos diversos passos no processo de descoberta de conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob algumas limitações computacionais produzem um conjunto de padrões de certos dados.

Destaca-se ainda que segundo Han, Kamber e Pei (2011) muitos autores e pesquisadores da área veem a mineração de dados como uma fase crucial do processo de descoberta do conhecimento (do inglês Knowledge Discovery in Databases - KDD), processo este ilustrado pela Figura 2.

Baseado nesse contexto nota-se que a geração do conhecimento se dá por meio de uma sequência de fases que se realizadas de forma precisa geram informação e conhecimento útil. Para Fayyad et al. (1996), estas fases são resumidas a algumas etapas fundamentais, são elas:

- Compreensão do problema e definição do objetivo do processo de KDD;
- Seleção dos dados a serem utilizados;
- Pré-processamento;
- Transformação de dados;
- Mineração de dados;
- Análise dos resultados obtidos.

Figura 2 – O ciclo do processo de KDD.

Fonte: Adaptado de FAYYAD et al. (1996).

A primeira fase do processo de KDD é o entendimento do problema e a elucidação da meta que deseja-se atingir, com o processo de descoberta de conhecimento. Já na segunda fase a etapa a ser executada chama-se seleção. A mesma consiste em um processo ao qual é selecionado um conjunto de dados ou subconjunto. Este conjunto de dados oriundo da etapa de seleção, são os dados que serão utilizados nas demais fases do KDD para gerar informações.

O pré-processamento de dados é a terceira fase do processo de descoberta de conhecimento, fase ao qual tem os seguintes objetivos:

- Remover ruídos;

- Apurando as informações fundamentais para modelar ou explicar o ruído;
- Seleção de atributos relevantes;
- Formatação dos dados;
- Tratamento de campos ausentes.

Baseado nesses objetivos nota-se que o pré processamento tem a meta de realizar uma limpeza de modo geral. A transformação é a quarta fase, a mesma tem o propósito de redução da dimensionalidade dos dados ou em alguns casos complementar os dados. Sendo assim, observa-se que o número total de atributos (variáveis), usados nas demais fases, podem ser reduzidos ou permanecerem os mesmos.

Na quinta fase, o processo realizado chama-se mineração de dados. Processo ao qual é responsável pela utilização de algoritmos de computadores específicos de aprendizagem, tais como: algoritmos de associação, classificação, clusterização e etc. Destaca-se que a indicação da maioria dos autores é de que esse processo só seja executado de fato após as fases de seleção, pré-processamento e transformação já estarem concluídas.

Por meio do uso desses algoritmos busca-se encontrar padrões os quais posteriormente serão analisados e interpretados pelos especialistas. Para Alves, Cechinel e Queiroga (2018), por meio do uso desses algoritmos busca-se encontrar padrões os quais posteriormente serão analisados e interpretados pelos especialistas.

Na última etapa do KDD, chamada de interpretação e avaliação, deve-se interpretar os padrões encontrados e avaliar os mesmos com ajuda dos especialistas da área ao qual os dados pertencem. Destaca-se ainda que todo o processo de descoberta do conhecimento realizado deve ser documentado e todo o feedback entregue para os responsáveis.

As convicções e concepções sobre o processo de descoberta de conhecimento e apresentados neste capítulo são em sua maioria, fundamentados na obra de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996).

2.2 MINERAÇÃO DE DADOS NA SAÚDE

De acordo com Carvalho, Moser, Silva e Dallagassa (2012), a fisioterapia e a área de saúde em geral, vem passando por uma fase de grande crescimento na geração e armazenamento de dados. Dessa forma, aumenta-se, também, a probabilidade de obtenção de informações relevantes, para assim apoiar o processo decisório dos profissionais de saúde. No entanto,

Carvalho, Moser, Silva e Dallagassa (2012) destacam que diversas vezes, o volume de dados gerados é tão grande que gera uma dificuldade no processo de busca pela informação, necessitando então de processos mais sofisticados para a manipulação e utilização de tais dados.

Com base no que foi dito anteriormente, nota-se que essa dificuldade atrapalha os profissionais de saúde em promover a gestão e qualidade de cuidados tendo como métrica as bases de dados da saúde.

A dificuldade apresentada por profissionais e gestores da saúde, possivelmente acontece devido a frenética geração de dados, o que produz uma inaptidão natural no ser humano de explorar, extrair e interpretar os mesmos para obtenção de informação e conhecimento (STEINER; SOMA; SHIMIZU; NIEVOLA; STEINER NETO, 2006).

Neste contexto, a técnica de Mineração de Dados apresenta-se como uma alternativa para a área da saúde e da fisioterapia, para melhor aproveitar o potencial dos dados no que tange ao reconhecimento de informações e padrões de difícil identificação (GOMES et al., 2014).

Na área da saúde, especialmente na saúde pública, o uso da mineração de dados e do processo de KDD está sendo permitido como uma forma de acelerar a busca de conhecimento (GALVÃO; MARIN, 2009).

Sendo assim, o uso da MD nos sistemas de informação de saúde pública e em grandes bases de dados hospitalares ajuda de maneira significativa para encontrar relacionamentos, características dos pacientes (que influenciam em determinadas doenças), identificação de tratamentos médicos de êxito para diferentes doenças e identificação de diversos tipos de padrões ocultos até então (GALVÃO; MARIN, 2009).

Dessa forma, para Galvão e Marin (2009), quando a mineração de dados encontra padrões importantes, pode ser realizada uma predição de tendências futuras baseada no passado, seja no histórico de um paciente ou de doenças específicas.

Conforme Galvão e Marin (2009), nota-se que nos dias de hoje, a informação e o conhecimento são benefícios lícitos, técnicos e fundamentais para à procura de conhecimento em empresas de saúde, controle social, tornando-se importante para o processo de tomada de decisão com prazos cada vez menores e mais otimizados.

2.2.1 MÉTODOS PARA MINERAÇÃO DE DADOS NA SAÚDE

Segundo Domingues (2003), muitas técnicas são utilizadas para caracterizar os tipos de padrões que podem ser descobertos nas tarefas de mineração de dados. Essas técnicas se

referem a atividades que podem ser aplicadas individualmente ou em grupo para a descoberta de padrões.

Para Galvão e Marin (2009), os métodos utilizados para mineração de dados na saúde em sua maioria seguem os métodos utilizados para qualquer outra área ao qual é aplicado a mineração de dados. Assim, nos parágrafos a seguir está descrita uma curta introdução de alguns dos itens mais relevantes da área.

Com base na visão acima descrita podemos dizer que a seguinte taxonomia pode ser efetiva para diversos casos de mineração de dados na saúde:

- Predição
  - o Classificação
  - o Regressão
  - o Estimação de Densidade
- Agrupamento
- Mineração de relações
  - o Mineração de Regras de associação
  - o Mineração de Correlações
  - o Mineração de Padrões Sequenciais
  - o Mineração de Causas
- Destilação de dados para facilitar decisões humanas
- Descobertas com modelos

Para Baker, Isotani e Carvalho (2011), em relação a predição, o intuito é desenvolver modelos que deduzem tópicos particulares dos dados, tecnicamente chamados de variáveis preditivas (predicted variables), por meio da análise e fusão dos inúmeros aspectos identificados nos dados, conhecidos como variáveis preditoras (predictor variables). De acordo com Baker, Isotani e Carvalho (2011), a predição precisa que uma quantidade pertinente dos dados seja manualmente codificada para permitir uma identificação precisa de uma ou muitas variáveis preditoras primariamente conhecidas.

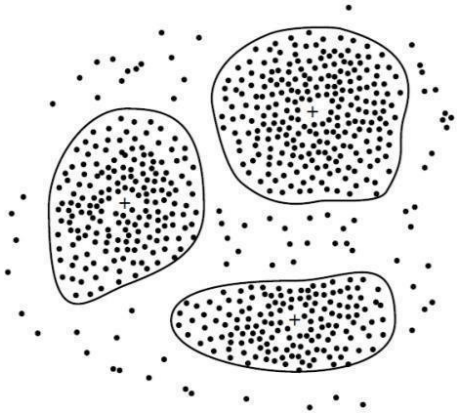
Assim como descrito na taxonomia, existem três tipos de predição: classificação, regressão, e estimação de densidade. Tratando-se de classificação, a variável preditora é binária ou categórica (BAKER; ISOTANI; CARVALHO, 2011).

Segundo Pessoa, Lima, Silva, Stephany, Strauss, Caetano e Ferreira (2012), na estimação de densidade utiliza-se uma função gaussiana, que possibilita formar um campo de densidade de ocorrências, que é suave e marca de forma evidente a região de ação convectiva a partir das descargas, as quais são bastante esparsas no espaço e no tempo.

Em ocasiões onde o campo preditor é um número, os algoritmos de regressão mais comuns utilizam: regressão linear, redes neurais, e máquinas de suporte vetorial. Métodos diferentes têm mais ou menos eficácia, conforme os aspectos das variáveis preditoras usadas (BAKER; ISOTANI; CARVALHO, 2011).

Para Camilo e Silva (2009), a tarefa de agrupamento (ou cluster) busca reconhecer e agrupar as instâncias semelhantes. Segundo Camilo e Silva (2009), um agrupamento é um conjunto de registros semelhantes mutuamente, todavia dessemelhantes das demais instâncias que estão em outros agrupamentos. Esta tarefa difere da classificação pois a mesma não tem a intenção de predizer, estimar ou classificar o valor de uma variável, ela meramente reconhece os grupos de dados semelhantes assim como exemplifica a Figura 3.

Figura 3 – Registros agrupados em três clusters.



Fonte: Adaptado de Han, Kamber e Pei (2011).

Para Baker, Isotani e Carvalho (2011), a **mineração de relações** tem como principal meta encontrar prováveis relações entre atributos que estão presentes nos bancos de dados. Segundo Alves, Cechinel e Queiroga (2018), uma das tarefas é reconhecer quais variáveis são mais fortemente relacionadas com um atributo particular. Observa-se que também é válido encontrar as relações entre quaisquer variáveis presentes nos dados. Existem quatro formas para reconhecer essas relações, são elas: regras de associação,sequências, correlações e causas (BAKER; ISOTANI; CARVALHO, 2011).

Para Baker, Isotani e Carvalho (2011), a **mineração de regras de associação** tem como objetivo fundamental identificar e gerar regras do tipo se-então (if-then).

Ou seja, podemos dizer que a representação usada é uma regra apontando o quanto o aparecimento de um conjunto de itens está ligado com a frequência que um outro conjunto de itens distintos aparece nos mesmos registros (PIVATO, 2006). Por exemplo, ao analisar um banco de dados seria viável reconhecer uma regra que indica a associação entre a variável "objetivo do paciente" a qual poderia ser uma variável binária contendo os valores não alcançado ou alcançado, e uma segunda variável binária chamada de "tratamento de fisioterapia" que pode ter os valores sim ou não.

Assim, caso um determinado paciente de um ortopedista tenha o objetivo de voltar a praticar determinado esporte após uma lesão, o mesmo deve solicitar ajuda dos profissionais de fisioterapia, gerando então a regra "tratamento de fisioterapia".

Tratando-se de **mineração de correlações**, o objetivo é encontrar correlações lineares (positivas ou negativas) entre variáveis (BAKER; ISOTANI; CARVALHO, 2011. Samiullah, Ahmed, Nishi, Fariha, Abdullah e Islam (2013) destacam que a mineração de correlação é reconhecida como uma das tarefas de mineração de dados mais importantes por sua capacidade de identificar dependências subjacentes entre objetos. Para Samiullah, Ahmed, Nishi, Fariha, Abdullah e Islam (2013), eventualmente, para extrair algum conhecimento muito útil de grandes bases de dados, técnicas de correlação são usadas.

A meta central da **mineração de sequências**, trata-se de encontrar associação temporal entre eventos e o efeito destes eventos no valor de uma variável (BAKER; ISOTANI; CARVALHO, 2011). Assim sendo, observa-se que é possível identificar qual trajeto de ações conseguem levar o paciente a alcançar uma recuperação efetiva.

Conhecendo essas informações faz-se práticável criar/testar diversas atividades e métodos de recuperação de lesão que poderiam melhorar a qualidade do tratamento (recuperação de lesões), ajudando assim o paciente a alcançar a sua meta final que é a recuperação total da lesão.

Na **mineração de causas**, os algoritmos e técnicas empregados têm a função de verificar se um evento causa outro evento por meio da análise dos padrões de covariância (BAKER; ISOTANI; CARVALHO, 2011).

Por meio dessa fundamentação seria viável compreender, por exemplo, quais as razões de um paciente ter tido um desempenho ruim no processo de recuperação de lesão. Existiria a possibilidade de caso finalizado o prazo dado pelo fisioterapeuta para a recuperação de lesão,

saber quais eventos fizeram com que o paciente tivesse um resultado negativo (a não recuperação da lesão no período estabelecido).

No campo de **destilação de dados para facilitar decisões humanas**, o objetivo deste segmento é mostrar dados que são complexos de um modo simples, que facilite o entendimento dos mesmos e apresente de forma objetiva suas características mais relevantes. Sendo assim, conforme Costa et al. (2008) o principal método dessa área da mineração de dados é o de visualização da informação.

Um dos recursos mais importantes para a destilação de dados para facilitar decisões humanas é a metodologia de repetição de texto. Essa metodologia consiste em demonstrar breves partes da base de dados em formato de texto, depois do processo de receberem rótulos por agentes humanos.

Sendo assim, destaca-se que os métodos dessa subárea facilitam e descomplicam a visualização da informação incluída nos dados oriundos da saúde e coletados por softwares ligados a mesma.

Por meio dos métodos demonstrados, torna-se viável obter informações que ajudem a aperfeiçoar o setor da saúde e fisioterapia. assim, é possível compreender de forma mais clara e adequada os pacientes, como eles reagem, o papel do contexto na qual a recuperação ocorre, além de fatores diversos que influenciam no resultado final do tratamento.

Um exemplo do uso das técnicas e métodos de mineração de dados voltada para saúde seria as clínicas de fisioterapia utilizarem a classificação, para uma vasta análise das características dos pacientes lesionados, ou fazer uso da estimativa para prever a probabilidade de uma variedade de resultados, como o sucesso do tratamento.

2.3ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Neste tópico será abordado a concepção de aprendizagem de máquina (AM) e o conceito dos algoritmos utilizados nesta dissertação. Algoritmos de Aprendizagem de Máquina são os responsáveis por viabilizar que os computadores aprendam, ou seja, os mesmos permitem aos computadores tomarem decisões baseados em tentativas de resoluções de problemas anteriores bem-sucedidas.

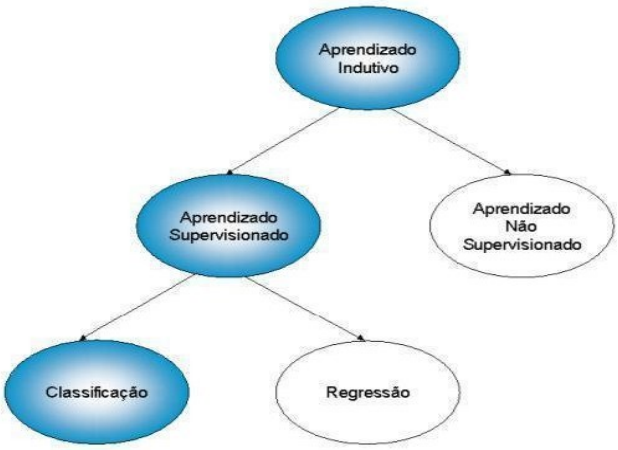
Sendo assim pode-se dizer que os algoritmos formam um padrão o qual o computador irá se basear para resolver os problemas de mesma espécie, a nomenclatura da aplicação desse processo denomina-se aprendizado de máquina (subárea da inteligência artificial).

O Aprendizado de Máquinas estuda métodos computacionais para obter novas informações, novas habilidades e novos meios de estruturar o conhecimento já existente (MITCHELL, 1997).

Segundo Souto et al. (2018), as técnicas de AM podem ser fragmentadas, de modo geral, em aprendizado supervisionado e aprendizado não supervisionado como demonstra a Figura 4. Se antes do processo de aprendizado o indutor recebe um grupo de exemplos, cada exemplo é formado por um conjunto de atributos de entrada e um conjunto de atributos de saída, então esse modo de aprendizado pode ser classificado como **aprendizado supervisionado**.

Para Souto et al. (2018), em contrapartida, **aprendizado não supervisionado** é efetuado quando, para cada exemplo, apenas os atributos de entrada estão disponíveis. Essas técnicas de aprendizado são usadas quando a meta for encontrar em um conjunto de dados padrões ou tendências (aglomerados) que ajudam na compreensão desses dados.

Figura 4 – A hierarquia do aprendizado.



Fonte: Monard et al. (2018).

Destaca-se que os algoritmos usados nesta dissertação, tratam-se de algoritmos de aprendizagem supervisionados.

2.3.1 ALGORITMOS DE REGRESSÃO

Para Willcox (1938), muitos estudiosos apontam o ano de 1663 como o ano de surgimento da estatística. Baseado em Graunt (1977), pode-se dizer que a regressão e a classificação datam da mesma época que a estatística.

Porém o uso de computadores e algoritmos para realizar regressão só vieram anos depois, Davidson et al. (1999) e Giustolisi & Savic (2006) encontram-se entre os primeiros a usar ferramentas de computação evolucionária para solucionar problemas de regressão.

As regressões tem o objetivo principal de efetuar previsões por meio dos dados obtidos previamente (OKAMURA, 2019).

Segundo Novaes (2015), a regressão conta com uma estimação numérica de uma função da variável de resposta e as variáveis de controle. O tipo de variável de resposta para regressão trata-se de variável do tipo real. Ou seja, quando as variáveis de representação do fenômeno possuem valores contínuos, refere-se a um problema de regressão (OKAMURA, 2019).

Conforme Novaes (2015), as técnicas de regressão podem ser utilizadas em diferentes tipos de base de dados e conjuntos de dados. Ou seja, destaca-se que a mesma pode ser aplicada em bases de dados da saúde e da fisioterapia.

Na regressão, podemos dizer que uma das metas a serem atingidas se aplica da seguinte forma: deseja-se regredir uma variável de resposta em função do menor número possível de variáveis independentes, de preferência a partir de um modelo que tenha a estrutura mais simples possível (NOVAES, 2015).

2.3.1.1 REGRESSÃO LINEAR

Para Sell (2018), a **regressão linear simples** estabelece uma equação matemática linear que expõe o relacionamento entre duas variáveis, uma dependente e outra independente, com o objetivo de estimar valores para uma variável, com base em valores conhecidos da outra.

Os algoritmos de regressão linear simples são uma das ações mais básicas do aprendizado supervisionado. Independente de sua simpleza os mesmos são muitíssimos efetivos e solucionam de forma satisfatória diversos problemas (OKAMURA, 2019). O termo linear sugere que a relação entre o parâmetro dependente e o independente pode ser retratada por meio de uma reta, ou algo muito similar à fórmula demonstrada na Fórmula 1.

Fórmula 1 – Fórmula regressão linear simples.

$$\boldsymbol{y = ax + B}$$

$$\boldsymbol{a = \frac{\sum xy - N\overline{x}\overline{y}}{\sum x^2 - N(\overline{x})^2}}$$

$$\boldsymbol{y = \overline{y} - a\overline{x}}$$

Fonte: Adaptado Okamura (2019).

Com base na fórmula matemática da regressão linear, nota-se que é possível identificar as variáveis y, a, x e b. Onde o coeficiente ’y’ retrata o termo independente que deseja-se classificar ou prever, ’x’ a variável dependente, ’a’ a inclinação da reta no plano cartesiano e b a constante que determina o valor de ’y’ quando ’x’ é 0 (OKAMURA, 2019).

Segundo Sell (2018), já na regressão linear múltipla, três ou mais variáveis são utilizadas, onde uma variável é dependente e duas ou mais variáveis independentes, com o objetivo de melhorar a capacidade de predição em comparação a regressão linear simples.

Com o intuito de atingir um desempenho superior a regressão linear simples, é utilizada a técnica de Backwards Elimination para geração do modelo de regressão linear multivariável (OKAMURA, 2019). Para entender essa técnica é indispensável compreender o conceito de p-value, que é uma medida estatística que auxilia o cientista a avaliar a exatidão de suas respostas (OKAMURA, 2019).

O p-value é usado para determinar se os resultados dos experimentos efetuados estão entre um mesmo alcance de valores para o evento observado (OKAMURA, 2019). O cientista especifica um p-value ideal antes de efetuar seus experimentos, e geralmente ele é determinado em 0,05. Caso o valor p-value observado esteja abaixo de 0,05, então a hipótese nula, na maioria dos casos, é descartada (OKAMURA, 2019).

2.3.1.2 ADABOOST

Para Duarte (2009), o AdaBoost ou Boosting Adaptativo, é um meta-algoritmo de aprendizado de máquina baseado em Boosting que pode ser usado "individualmente" ou em conjunto com praticamente qualquer outro algoritmo de aprendizado de máquina para melhorar o seu desempenho em um determinado conjunto de dados.

Segundo Duarte (2009), os diversos classificadores, que fazem parte do seu comitê, são criados sequencialmente. Cada classificador adicional é construído favorecendo os exemplos do conjunto de treinamento incorretamente classificados pelos classificadores anteriores.

O AdaBoost chama um algoritmo-base em várias iterações t, onde t ∈ [1..T]. Em cada iteração t, a distribuição de pesos do conjunto de treinamento é atualizada para uso do algoritmo-base (DUARTE, 2009). A atualização é efetuada de forma a, relativamente, aumentar os pesos dos exemplos erroneamente classificados em confronto com os pesos dos exemplos corretamente classificados (DUARTE, 2009).

2.3.1.3 BAYESIAN RIDGE REGRESSION

A abordagem bayesian ridge regression ou regressão linear bayesiana consiste na integração do teorema de Bayes no processo de aprendizado de máquina, de modo que capture as suposições dos coeficientes do modelo (w), como uma distribuição prévia de probabilidade p(w) e o efeito dos dados observados (D), dado pela probabilidade condicional p(D | w). A regra de Bayes pode ser descrita conforme a equação presente na Fórmula 2 (BISHOP, 2006).

Fórmula 2 – Regra de Bayes.

$$\boldsymbol{p(w|D) = \frac{p(D|w)p(w)}{p(D)}}$$

Fonte: Elaborado pelo autor.

Segundo Neal (2012), a função de verossimilhança, dada por p(D | w), concede apanhar o impacto das observações dos dados, como função de w, ou seja, mensura a probabilidade dos dados para a diversidade de configurações do vetor de parâmetros do modelo.

O denominador da equação apresentada na Fórmula 2 retrata uma constante de normalização e não depende de w, e é geralmente ignorada, tendo em vista que é irrelevante para o primeiro nível de inferência (MACKAY, 1992).

Dessa maneira para Rozin (2018), a apresentação do teorema de Bayes é dada em relação a priori verossimilhança, em que a priori é dada pela suposição prévia dos parâmetros. Sendo assim, Rozin (2018) considera que é interessante que os parâmetros sejam ajustados de maneira a maximizar a probabilidade do conjunto de dados, que pode ser conquistado por um estimador de máxima verossimilhança.

2.3.1.4 RIDGE

Este modelo soluciona um modelo de regressão em que a função de perda é a função de mínimos quadrados lineares e a regularização é dada pela norma 12 (SCIKIT-LEARN, 2021). Também nomeada como Regressão de cume ou regularização de Tikhonov. Este estimador tem suporte embutido para regressão multivariada (SCIKIT-LEARN, 2021). Ou seja, quando y é uma matriz 2d de forma (n_samples, n_targets).

2.3.1.5 RIDGECV

Regressão de Ridge, trata-se de uma regressão de Ridge com validação cruzada incorporada (SCIKIT-LEARN, 2021). Por padrão, este algoritmo executa a validação cruzada generalizada, que é uma forma de validação cruzada Leave-One-Out eficiente (SCIKIT-LEARN, 2021).

2.3.1.6 GRADIENT BOOSTING REGRESSOR

O mesmo é um Gradient Boosting (GB) para regressão. O GB constrói um modelo aditivo de forma avançada no estágio e também permite a otimização de funções de perda diferenciáveis arbitrárias (SCIKIT-LEARN, 2021). Em cada estágio, uma árvore de regressão é ajustada no gradiente negativo da função de perda fornecida Scikit-Learn (2021).

2.3.1.7 RANDOM FOREST REGRESSOR

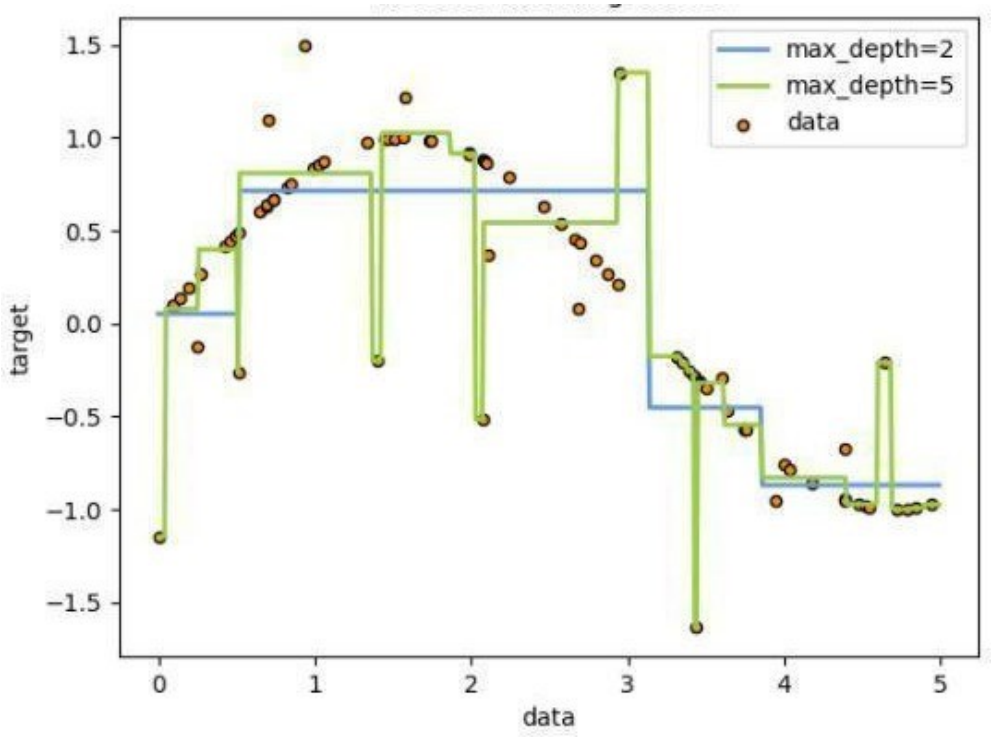
Também chamado de regressor de floresta aleatória é um meta-estimador que ajusta uma série de árvores de decisão de classificação em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e o sobreajuste de controle (SCIKIT-LEARN, 2021).

2.3.1.8 DECISION TREE REGRESSOR

O mesmo trata-se de um algoritmo de árvores de decisões, um método de aprendizado supervisionado não paramétrico usado para regressão . O objetivo é criar um modelo que preveja o valor de uma variável de destino, aprendendo regras de decisão simples inferidas dos recursos de dados. Uma árvore pode ser vista como uma aproximação constante por partes (SCIKIT-LEARN, 2021).

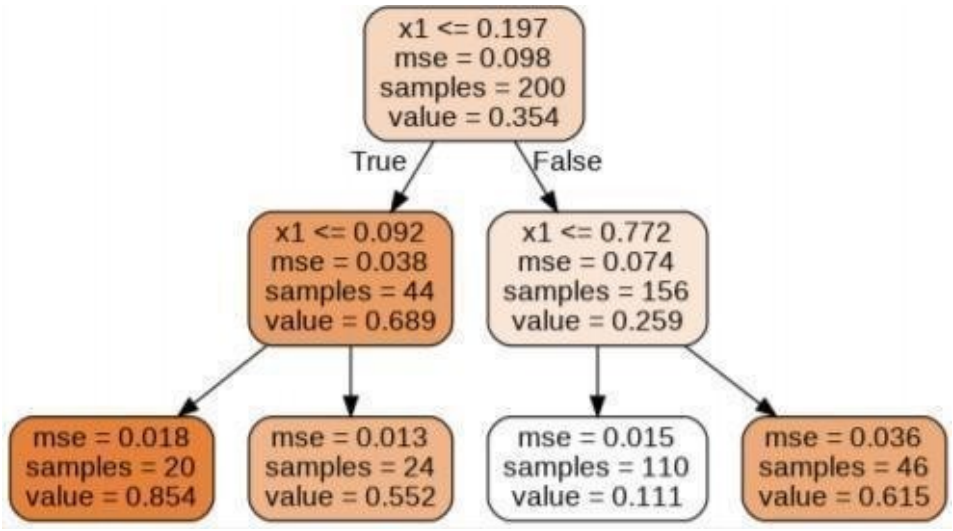
Por exemplo, a Figura 5 ilustra um exemplo onde as árvores de decisão aprendem com os dados a se aproximar de uma curva senoidal com um conjunto de regras de decisão if-then-else. Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais adequado é o modelo (SCIKIT-LEARN, 2021). Já a Figura 6 demonstra como podemos visualizar de maneira gráfica (em formato de árvore) um problema de árvore de regressão para uma quadrática com ruído.

Figura 5 – Exemplo Decision Tree Regressor.



Fonte: Scikit-Learn (2021).

Figura 6 – Exemplo de visualização Decision Tree Regressor.



Fonte: Pohlenz (2020).

Para Scikit-Learn (2021), existe algumas vantagens no uso de algoritmos de árvores de decisão, são elas:

- Simples de entender e interpretar pois as árvores podem ser visualizadas de maneira gráfica.
- Requer pouca preparação de dados em comparação a outras técnicas.
- O custo do uso da árvore (ou seja, dados de previsão) é logarítmico no número de pontos de dados usados para treinar a árvore.
- Capaz de lidar com dados numéricos e categóricos.
- Capaz de lidar com problemas de múltiplas saídas.
- Usa um modelo de caixa branca. Ou seja, se uma determinada situação é observável em um modelo, a explicação para a condição é facilmente explicada pela lógica booleana.
- Possível validar um modelo por meio de testes estatísticos. Assim sendo, torna-se possível contabilizar a confiabilidade do modelo.
- Apresenta um bom desempenho, mesmo que suas suposições sejam violadas de alguma forma pelo modelo verdadeiro a partir do qual os dados foram gerados.

2.3.1.9 NEAREST NEIGHBORS REGRESSION

A regressão baseada em vizinhos (KNN) pode ser usada nas situações em que os rótulos de dados são contínuos em vez de variáveis discretas (SCIKIT-LEARN, 2021). O rótulo atribuído a um ponto de consulta é calculado com base na média dos rótulos de seus vizinhos mais próximos (SCIKIT-LEARN, 2021).

A regressão básica de vizinhos mais próximos usa pesos uniformes, assim sendo, cada ponto na vizinhança local contribui uniformemente para a classificação de um ponto de consulta. Em muitos casos, pode ser proveitoso ponderar os pontos de forma que os pontos próximos contribuam mais para a regressão do que os pontos distantes (SCIKIT-LEARN, 2021).

2.3.1.10 LASSO

O algoritmo de Lasso é um modelo linear que estima coeficientes esparsos. É vantajoso em diversos contextos devido à tendência de preferir soluções com menos coeficientes diferentes de zero, reduzindo efetivamente o número de recursos dos quais a solução dada é dependente (SCIKIT-LEARN, 2021). Devido a esses motivos, o Lasso e suas variantes são fundamentais para o campo da detecção por compressão (SCIKIT-LEARN, 2021).

2.3.1.11 LARS

A regressão de ângulo mínimo (LARS) é um algoritmo de regressão para dados de alta dimensão. O LARS é muito parecido com a regressão progressiva passo a passo (SCIKIT-LEARN, 2021). Para cada tarefa, o LARS encontra o recurso mais relacionado ao destino. Quando há vários recursos com correlação igual, em vez de continuar ao longo do mesmo recurso, ele avança em uma direção equiangular entre os recursos (SCIKIT-LEARN, 2021).

2.3.1.12 SUPPORT VECTOR REGRESSION

Para Okamura (2019), o Support Vector Machines (SVM) são máquinas de aprendizado que executam o princípio indutivo de minimização do risco estrutural para atingir boas generalizações em um limitado número de padrões de aprendizado.

O algoritmo SVM obtém padrões sutis em bancos de dados complexos, performando um aprendizado de classificação discriminativa por exemplos para assim predizer dados que nunca foram apresentados ao algoritmo, por isso é classificado como um algoritmo de aprendizado supervisionado (OKAMURA, 2019).

Uma SVM pode ser utilizada tanto para problemas de regressão quanto para problemas de classificação. Quando a SVM é utilizada para problemas de regressão denomina-se a mesma de Support Vector Regression (SVR).

O modelo produzido pelo SVM depende apenas de um subconjunto dos dados de treinamento, porque a função de custo para construir o modelo não se preocupa com os pontos de treinamento que estão além da margem (SCIKIT-LEARN, 2021). Analogamente, o modelo produzido pelo Support Vector Regression depende apenas de um subconjunto dos dados de treinamento, pois a função de custo ignora as amostras cuja previsão está próxima de seu destino (SCIKIT-LEARN, 2021).

Drucker (1997) destaca que resultados positivos foram alcançados por meio do uso da SVR em predições de regressão e time series.

2.3.1.13 GENERALIZED LINEAR REGRESSION

Os Modelos Lineares Generalizados amplificam os modelos lineares de duas maneiras. No primeiro modo os valores previstos $\hat{y}$ estão ligados a uma combinação linear das variáveis de entrada X através de uma função de link inverso h como demonstra a Fórmula 3:

Fórmula 3 – Fórmula do primeiro modo do modelo linear generalizado.

$$\widehat{\boldsymbol{y}}(\boldsymbol{w}, \boldsymbol{X}) = \boldsymbol{h}(\boldsymbol{X}\boldsymbol{w})$$

Fonte: Scikit-Learn (2021).

Em segundo lugar, a função de perda quadrada é substituída pelo desvio de unidade d de uma distribuição na família exponencial, conforme demonstra a Fórmula 4. Onde α é a penalidade de regularização L2. Quando os pesos da amostra são fornecidos, a média se torna uma média ponderada.

Fórmula 4 – Fórmula do primeiro modo do modelo linear generalizado.

$$\min_{w} \frac{1}{2n_{samples}} \sum_{i} d(yi, \hat{y}i) + \frac{\alpha}{2} ||w||2,$$

Fonte: Scikit-Learn (2021).

2.4 FERRAMENTAS PARA MINERAÇÃO DE DADOS

Como demonstrado nos capítulos anteriores, sabe-se que nos dias atuais a mineração de dados está cada dia mais importante para diversos segmentos, inclusive para o setor da saúde e fisioterapia. Nesse sentido, pode-se dizer que a mineração de dados é fundamental para ofertar subsídios para as tomadas de decisões, buscando uma "vantagem competitiva" ou a própria evolução do setor em prol da ciência. Sendo assim, esse tópico tem o objetivo de apresentar as principais ferramentas de mineração de dados e descrever as escolhidas para utilização dessa pesquisa.

Na atualidade são muitas as ferramentas de mineração de dados ofertadas na Web. Segundo Queiroga (2017), a maioria dos fornecedores oferta um período de demonstração ou em diversos casos oferecem a ferramenta de forma ilimitada e gratuita, assim pode-se descobrir qual ferramenta será melhor para a sua pesquisa e para seus dados. Para Queiroga (2017), algumas das ferramentas mais usadas de mineração de dados são: NLTK, RapidMiner, Oracle Data Mining, Knime, Cognos, RProgramming, Weka, Orange e o SAS.

Além dessas ferramentas, também existe a possibilidade de utilizar linguagens de programação com bibliotecas específicas para a mineração de dados e ciência de dados. Uma dessas linguagens de programação chama-se Python e de acordo com Ribeiro e Oliveira (2018) o python é uma das linguagens mais utilizadas para técnicas de mineração de dados.

A mesma oferece aos usuários uma ótima experiência de uso, tendo em vista que conta com diversas bibliotecas, favorecendo a execução de técnicas de mineração de dados (RIBEIRO; OLIVEIRA, 2018).

Associado ao python muitos utilizam a plataforma Anaconda, que é uma das plataformas de Python mais usadas para ciências de dados e mineração de dados, pois conta com várias bibliotecas já instaladas na plataforma (RIBEIRO; OLIVEIRA, 2018).

Considerando o uso da plataforma anaconda, normalmente utiliza-se a ferramenta Jupyter para criação dos scripts, fazer treinamento e teste, pois a mesma conta com vários recursos que facilitam e simplificam o desenvolvimento (RIBEIRO; OLIVEIRA, 2018).

2.4.1 PYTHON

Para diversas pessoas, a linguagem de programação python possui uma grande influência no mundo da programação. A partir de sua criação em 1991, o python virou uma das linguagens de programação mais conhecidas (MCKINNEY, 2019).

Segundo McKinney (2019), o Python tornou-se popular por volta do ano de 2005, nessa época o mesmo era muito usado para construção de sites, devido aos seus diversos frameworks web como Django. Com certa regularidade o mesmo, é denominado de linguagem de scripting, pois muitas vezes é usado para criar pequenos programas ou scripts para automatizar várias tarefas. Deixa-se claro que o python também é usado para o desenvolvimento de softwares complexos e sérios.

Entre as muitas linguagens interpretadas, o Python foi uma das linguagens que mais desenvolveu sua comunidade, devido a muitas causas históricas e culturais (MCKINNEY, 2019). Sendo assim o python conta com uma comunidade ativa de processamento científico e análise de dados (MCKINNEY, 2019).

Nos últimos dez anos, o python deixou de ser apenas uma linguagem de computação científica inovadora, para tornar-se uma das linguagens mais relevantes para ciência de dados, aprendizado de máquina, mineração de dados e desenvolvimento de softwares em geral (MCKINNEY, 2019).

De acordo com McKinney (2019), naturalmente o Python, quando trata-se de análise de dados, processamento interativo e visualização de dados, gera comparações com outras linguagens de programação e ferramentas usadas na área de ciência de dados, como: MATLAB, R,sTATA, SAS, e outras.

Atualmente o python conta com muitas bibliotecas como pandas e scikit-learn. Devido a essas bibliotecas o mesmo transformou-se em uma das opções mais populares para as atividades de ciência de dados e mineração de dados McKinney (2019). Destaca-se ainda que o Python é uma ótima opção para a construção de softwares de aplicações de dados.

Os conceitos de Python, bem como os relatos sobre o mesmo foram baseados na obra de (MCKINNEY, 2019).

2.4.2 PANDAS

O pandas começou a ser desenvolvido em 2018 na AQR Capital Management. No final de 2009, o mesmo tornou-se código aberto e é ativamente apoiado até hoje por uma comunidade de indivíduos com ideias semelhantes em todo o mundo, que contribuem com seu tempo e trabalho para manter o pandas como uma biblioteca de código aberto (PANDAS, 2021).

O pandas oferta estruturas de dados de alto nível e funcionalidades, projetadas para tornar simples e rápido utilizar dados estruturados ou tabulares. O pandas, desde seu desenvolvimento, vem auxiliando a viabilizar o Python como uma das ferramentas mais poderosas para análise de dados (MCKINNEY, 2019).

Os objetos mais conhecidos do pandas são o Data Frame (DF) que refere-se a uma estrutura de dados tabular, orientada a colunas, e as séries que nada mais são do que array unidimensional, com rótulo (MCKINNEY, 2019).

O pandas conta com recursos de processamento de alto desempenho similares aos arrays da NumPy com os recursos flexíveis de manipulação de dados das planilhas e dos bancos de dados relacionais similares aos de banco de dados que usam SQL. O mesmo oferece uma função aprimorada de indexação para ajudar em tarefas como seleção, manipulação e limpeza dos dados (MCKINNEY, 2019).

Para Pandas (2021), os principais destaques do pandas são:

- Um objeto DataFrame rápido e eficiente para manipulação de dados com indexação integrada;
- Ferramentas para ler e gravar dados entre estruturas de dados na memória e diferentes formatos: arquivos CSV e de texto, Microsoft Excel, bancos de dados SQL e o formato rápido HDF5;
- Alinhamento inteligente de dados e manuseio integrado de dados perdidos: obtenha alinhamento automático baseado em rótulo em cálculos e manipule facilmente dados confusos em uma forma ordenada;
- Remodelagem e rotação flexível de conjuntos de dados;
- Inteligente, baseada em rótulo de corte, a indexação de fantasia, e subconjuntos de grandes conjuntos de dados;

- As colunas podem ser inseridas e excluídas de estruturas de dados para mutabilidade de tamanho;
- Agregar ou transformar dados com um poderoso grupo por mecanismo, permitindo operações dividir-aplicar-combinar em conjuntos de dados;
- Mesclagem e junção de conjuntos de dados de alto desempenho;
- A indexação de eixo hierárquica fornece uma maneira intuitiva de trabalhar com dados de alta dimensão em uma estrutura de dados de dimensão inferior;
- Funcionalidade da série temporal: geração de intervalo de datas e conversão de frequência, estatísticas de janela móvel, mudança de data e atraso. Crie até mesmo compensações de tempo específicas de domínio e junte séries temporais sem perder dados;
- Altamente otimizado para desempenho, com caminhos de código críticos escritos em Cython ou C.
- Python com pandas está em uso em uma ampla variedade de domínios acadêmicos e comerciais, incluindo finanças, neurociência, economia, estatística, publicidade, web analytics e muito mais.

2.4.3 JUPYTER E IPYTHON

O projeto IPython começou no ano de 2001 como um projeto secundário de Fernando Pérez para produzir um interpretador Python. Nos anos seguintes o IPython, tornou-se uma das ferramentas associadas ao python mais importantes.

Mesmo que o ipython sozinho não oferta nenhuma técnica de processamento ou de análise de dados, o mesmo foi criado desde o princípio para aumentar a sua produtividade, tanto em um processamento interativo, quanto no desenvolvimento de software.

Segundo (MCKINNEY, 2019), algumas das vantagens do IPython são:

- Conta com o fluxo de trabalho de execução-exploração;
- Acesso fácil ao shell e ao sistema de arquivos;
- Executa diversas tarefas de forma mais rápida.

No ano de 2014, a equipe do IPython e Fernando divulgaram uma ferramenta que foi denominada de Jupyter. Dessa forma o notebook web do IPython tornou-se o então conhecido

atualmente como notebook Jupyter, com suporte nos dias atuais para muitas linguagens de programação (MCKINNEY, 2019).

Conforme Jupyter (2021), O Projeto Jupyter é caracterizado como: um projeto de código aberto sem fins lucrativos, à medida que evoluiu para oferecer suporte à ciência de dados interativa e computação científica em muitas linguagens de programação.

2.4.4 SCIKIT-LEARN

A biblioteca Scikit-Learn implementa muitos algoritmos de aprendizado de máquina (WEIAND; WEIAND, 2018).A partir da elaboração do projeto, no ano de 2010, o scikit-learn tornou-se um kit de ferramentas fundamentais para aprendizado de máquina na linguagem Python. Junto com o pandas, o scikit-learn permite que o Python seja uma linguagem de programação eficiente para a ciência de dados (MCKINNEY, 2019).

Conforme McKinney (2019), scikit-learn conta com submódulos para módulos como:

- Classificação: SVM, nearest neighbors, floresta aleatória, random forest, regressão logística e etc.
- Regressäo: regressão linear, regressão de Lasso, regressão de ridge e etc.
- Clustering: k-means, clustering espectral e etc.
- Redução de dimensionalidade: PCA, seleção de atributos e etc.
- Seleção de modelos: grid search, validação cruzada, métricas.
- Pré-processamento: extração de atributos, normalização.

2.5 MEDIDA ESTATÍSTICAS

2.5.1 R-quadrado (R^2)

Segundo Maia (2017), o R-quadrado (R^2), também chamado de coeficiente de determinação, é uma medida estatística que retrata a proporção da variância de uma variável dependente que é explicada por uma variável ou variáveis independentes em um modelo de regressão.

Conforme Maia (2017), o R^2 varia entre 0 e 1, muitas vezes o mesmo é apresentado em termos percentuais. Nesse caso, representa a quantidade da variância dos dados que é explicada pelo modelo linear.

2.5.2 Erro Médio Absoluto (MAE)

Segundo Maia (2017), o erro médio absoluto, MAE (da sigla em inglês Mean Absolute Error), é calculado a partir da média dos erros absolutos. Sendo assim, utiliza-se do módulo de cada erro para fugir da subestimação, portanto, o valor é menos afetado por pontos especialmente extremos (outliers).

2.5.3 Erro Quadrático Médio (MSE)

Para Maia (2017), o erro quadrático médio, MSE (da sigla em inglês Mean Squared Error), é muito utilizado para averiguar a acurácia de modelos. O mesmo estabelce um maior peso aos maiores erros, pois cada erro é elevado ao quadrado individualmente e, após isso, a média desses erros quadráticos é calculada.

2.5.4 Raiz Quadrada do Erro Médio (RMSE)

A sigla RMSE é oriunda do inglês Root Mean Squared Error. Segundo Maia (2017), a raiz quadrada do erro médio (RMSE) trata-se da raiz quadrada do MSE, em que o erro retorna à unidade de medida do modelo.

2.6 LOCOMOÇÃO HUMANA

Apesar de a pesquisa em questão estar focada em exercícios dentro da água, é indispensável que os conceitos de locomoção humana fora da água, especialmente os conceitos de FRS, sejam igualmente explicados já que servirão de critério para identificar as mudanças na caminhada e corrida subaquática.

Segundo Dicharry (2010), os estudos apontam que existem três maneiras básicas de locomoção humana, são elas: andar, correr e saltar. Para Haupenthal (2013), com o objetivo de contribuir para o estudo e análise destes movimentos foram desenvolvidas nomenclaturas e

variáveis para estudar estes fenômenos. Em grande parte das pesquisas considera-se que a locomoção conta com um padrão cíclico que se repete indefinidamente a cada passo, as descrições geralmente investigam o que acontece num ciclo, presumindo que os ciclos sucessivos serão iguais (HAUPENTHAL, 2013).

Apesar de que essa suposição não seja verdadeira, é uma aproximação coerente, pois encontram-se diversos eventos observáveis que são praticamente iguais para qualquer ser humano, caminhando ou correndo (ROSE; GAMBLE, 1998).

Todavia Perry (1992) destaca que em diversos casos ocorre uma variação entre diferentes indivíduos ou para o mesmo indivíduo como efeitos de variação de velocidade, tipo de calçado, idade, tipo de atividade física praticada, entre outros.

Destaca-se ainda três fatos sobre a locomoção humana, o primeiro é que a marcha é o movimento humano mais habitual, mesmo sendo um dos mais complexos (WINTER, 1991). O segundo fato é que, para Zatsiorsky (2004), diversos movimentos humanos são definidos pela repetição contínua de um padrão fundamental. Para finalizar esses fatos destaca-se que as pessoas, têm uma capacidade fantástica de modificar propositadamente a velocidade, distância e cadência a fim de atingir as demandas do ambiente (ZATSIORSKY, 2004).

Um ciclo de marcha ou corrida pode ser determinado como o intervalo de tempo no período em que uma sequência de acontecimentos sucessivos e regulares se completa (HAUPENTHAL, 2013). Estes acontecimentos sequenciais começam no momento em que um pé toca o solo e termina quando o mesmo pé atinge o solo novamente (HAUPENTHAL, 2013). O ciclo citado pode ser separado em duas etapas principais: apoio e oscilação/balanço (ROSE; GAMBLE, 1998).

Torna-se significativo retratar uma diferenciação entre caminhada e corrida. A marcha humana acompanha um ritmo contínuo onde acontece a transição da caminhada para a corrida e da mesma para o tiro ou "sprint" com um "grande" acréscimo de velocidade de deslocamento.

Torna-se significativo retratar uma diferenciação entre caminhada e corrida (HAUPENTHAL, 2013). A marcha humana acompanha um ritmo contínuo onde acontece a transição da caminhada para a corrida e da mesma para o tiro ou "Sprint" com um "grande" acréscimo de velocidade de deslocamento (NOVACHECK,1998) .

A definição entre a caminha e a corrida ocorre quando o período de duplo apoio durante a fase de contato com o solo não ocorre mais (HAUPENTHAL, 2013). Se ampliarmos a velocidade da caminhada, em um certo momento, o movimento da corrida forma-se, no qual não existe uma fase com os dois pés no solo. Na corrida nota-se uma sucessão de pequenos

saltos onde existe uma pequena fase de voo, na qual nenhum dos pés está em contato com o solo (HAUPENTHAL, 2013).

Para Zatsiorsky (2004), a mudança do padrão de caminhada para corrida ocorre acontece por volta da velocidade de 2,0 m.s-1. Já o tiro ou "Sprint" transcorre quando o toque do pé com o solo não é mais realizado pelo calcanhar, mas sim com o antepé (NOVACHECK, 1998).

É no decorrer da fase de apoio da marcha que acontece a coleta da FRS e a descrição quantitativa dos aspectos dinamométricos da marcha, por meio das plataformas de força (HAUPENTHAL, 2013).

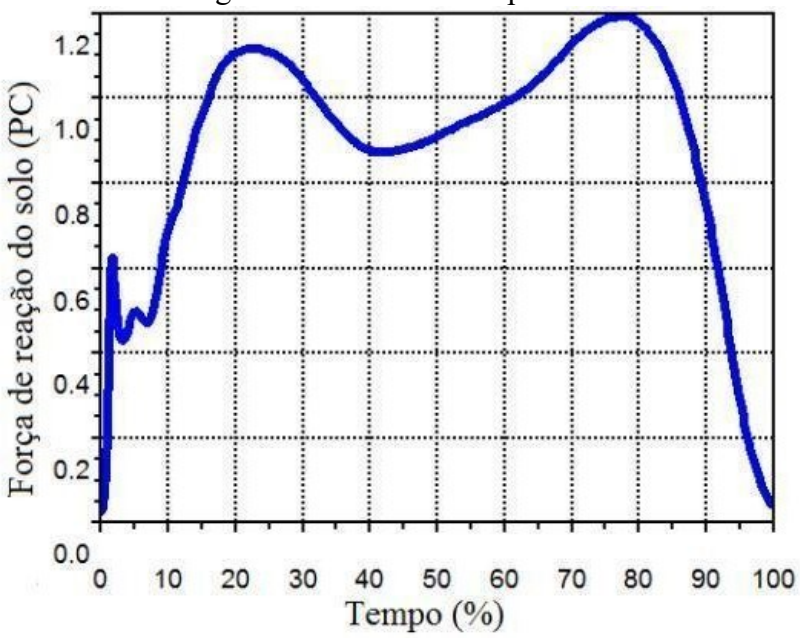
Existe uma grande variação e baixa relevância da componente médio-lateral para a força resultante (HAUPENTHAL, 2013). Por este motivo, esta pesquisa optou por predizer apenas dois componentes da FRS são eles: o componente vertical e a componente ântero-posterior.

2.6.1 COMPONENTE VERTICAL DA FRS

Para Rose e Gamble (1998) o componente da FRS mais dominante é a vertical, devido a sua grandeza de força em relação às outras. A atividade da componente vertical se associa com a ação da gravidade e para a marcha fora da água necessita dispor de no mínimo à força do peso corporal.

A componente vertical é o principal componente da FRS devido a sua magnitude, sendo a componente que melhor define a carga sobre o aparelho locomotor. (ZATSIORSKY, 2004). Na caminhada fora da água a componente vertical dispõe dois picos de força, os mesmos podem ser analisados na Figura 7.

Figura 7 – Curva da componente vertical.



Fonte: Haupenthal (2013).

Segundo Perry (1992) o primeiro pico de força na componente vertical da FRS possui seu princípio por meio do toque do pé no solo em resposta à carga corporal do sujeito.

Para Novacheck (1998), o segundo pico de força refere-se a etapa ativa do movimento na ocasião em que os músculos do membro inferior em apoio estão efetuando a etapa de propulsão da marcha.

Quando trata-se de corrida fora da água considera-se que a componente vertical da FRS tem um pico de impacto e um pico ativo. Em um primeiro momento o pico de impacto é influenciado pelo toque do pé no solo (ZATSIORSKY, 2004).No contexto de corredores que tocam o solo com o médio pé e ante pé, normalmente nas análises não se encontra o pico de impacto (ZATSIORSKY, 2004). O segundo pico é influenciado pela atuação muscular no decorrer do apoio (ZATSIORSKY, 2004).

A relevância dos picos de força altera-se conforme: a velocidade, idade das pessoas, calçado utilizado, atividade física executada, sexo e etc (PERRY, 1992). Haupenthal (2013) destaca que entre todas as variáveis influenciadoras a alteração da velocidade é o fator que mais interfere.

2.6.2 COMPONENTE DA ÂNTERO-POSTERIOR DA FRS

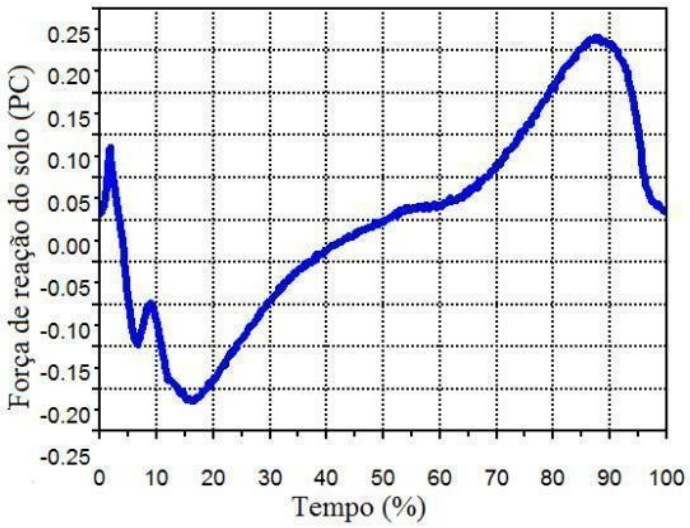
Conforme Haupenthal (2013), o elemento ântero-posterior está ligado com as acelerações de progressão. Segundo Rose e Gamble (1998), esse elemento é essencial para começar ou terminar momentos de locomoção e para alterar a velocidade da marcha (HAUPENTHAL, 2013).

O componente ântero-posterior demonstra dois picos de força onde um é negativo e o outro é positivo (HAUPENTHAL, 2013). Ao longo da etapa negativa o pé está empurrando o solo anteriormente, e no solo causa uma reação na direção posterior para desacelerar o movimento (HAUPENTHAL, 2013). Para Perry (1992), já na etapa positiva o pé empurra o solo na direção posterior, e no solo gera uma força de reação na direção anterior no sentido de acelerar o movimento.

Conforme Haupenthal (2013), o elemento ântero-posterior conta com dois picos de força, na marcha fora da água a componente ântero-posterior conta dois picos de força muito explícitos, os mesmos podem ser vistos na Figura 8.

No primeiro momento acontece o pico negativo, após o pico negativo ocorre um aumento na força de reação ao solo gerando então o pico positivo (PERRY, 1992). Levando em consideração os dois picos de força, que possuem intensidades de polaridades distintas porém semelhantes em magnitude, a curva torna-se similar a imagem de um dente de serra.

Figura 8 – Curva da componente ântero-posterior.



Fonte: Haupenthal (2013).

Para a caminhada os valores encontram-se de 120 a 150 % do peso corporal (PC). Já na corrida os valores estão entre 160 a 500 % em relação a componente vertical (HAUPENTHAL, 2013; apud HAMILL e KNUTZEN, 1999; NIGG e HERZOG, 1994; NOVACHECK, 1998; VIEL, 2001; ZATSIORSKY, 2004).

A componente ântero-posterior é fundamentalmente ligada com as acelerações de progressão, assim a mesma é imprescindível para alterar a velocidade da marcha. Destaca-se que para a caminhada, seus valores são de 20 % do peso corporal, enquanto para a corrida são de 40 % a 50 % do peso corporal (HAUPENTHAL, 2013; apud NIGG e HERZOG, 1994; PERRY, 1992; ROSE E GAMBLE, 1998; WINTER, 1991).

Em relação aos atributos das curvas, para a componente vertical encontra-se uma fase originária que é chamada de pico de impacto que acontece até (em torno) de 50 ms. Na próxima fase da caminhada existem dois picos de força enquanto para a corrida apenas um pico. Já em relação a componente ântero-posterior, a mesma conta com dois picos onde o primeiro é negativo e está relacionado à desaceleração do movimento, enquanto o segundo pico (pico positivo) é ligado à aceleração do movimento.

Por meio das referências citadas nesse tópico pode-se dizer que a marcha humana é um movimento que possui estudos profundos e com características definidas quanto a FRS.

2.7 MARCHA NA ÁGUA

Sabe-se que o ser humano relaciona-se com o meio aquático desde o início da humanidade, com objetivos de: exploração, condicionamento e recreação.

Segundo Haupenthal (2013), em relação ao uso da água para reabilitação, encontra-se registro da utilização da água com objetivo terapêutico a partir de 2400 antes de Cristo. Estes documentos indicam o emprego de fontes minerais no Egito antigo, na Assíria e na cultura Mohammed durante o século XVIII.

São muitas as situações em que os profissionais de saúde podem se favorecer com a diminuição da força resultante, o que acaba gerando uma menor sustentação do peso. Alguns exemplos de benefícios que pode-se citar são: Os processos de recuperação funcional das fraturas de fêmur e tíbia, problemas de coluna, pós-operatórios de próteses, recuperação entre dois eventos competitivos, capacidade aeróbia, reconstrução de ligamentos e tendões, melhora ou manutenção de força, etc (BATES; HANSON, 1998).

No tocante às propriedades físicas da água, sabe-se que quanto maior o nível de imersão, menores serão as cargas nas estruturas musculoesqueléticas (HAUPENTHAL, 2013).

Quando utiliza-se a água o paciente que sofreu uma lesão ou passou por algum tipo de cirurgia, dentro de pouco tempo pode ser submetido a começar seu processo de recuperação por meio do treinamento da marcha ou corrida (na água), para que assim consiga se recuperar o mais rápido possível sem ter prejuízos nas estruturas que precisam de reabilitação. Lembrando que após o processo de recuperação na água o paciente na maioria dos casos pode voltar a executar atividades de marcha ou corrida fora da água (SKINNER; THOMSON, 1985).

Sendo assim, podemos dizer que a marcha e a corrida na água aceleram e impulsionam o processo de reabilitação, evitando danos aos procedimentos cirúrgicos (na etapa de pós-operatória). Porém, para que de fato todos esses benefícios ocorram na prática, o profissional de saúde deve saber estimar a carga, profundidade de imersão, e as circunstâncias em que carece realizar a marcha ou corrida subaquática. Todos os fatores citados acima podem diminuir o tempo e os custos com tratamento.

Apesar desse tipo de tratamento e exercício aquático ser bem reconhecido, o avanço dessa área no tocante às pesquisas que estudam a força de reação do solo nos exercícios aquáticos não acompanhou a popularidade da aplicação dos mesmos (exercícios subaquáticos) (HAUPENTHAL, 2013).

2.8 TRABALHOS RELACIONADOS

Na atualidade o cenário de mineração de dados na saúde tem evoluído, sendo assim esta importante área de pesquisa vem permitindo a análise de um imenso volume de dados, auxiliando na solução de problemas voltados à saúde.

Haupenthal (2013) desenvolveu uma pesquisa com o objetivo de contribuir na prescrição dos exercícios de caminhada e corrida na água, no que refere-se ao conhecimento e controle das forças de reação do solo (FRS) durante o contato com o fundo da piscina. A pesquisa de Haupenthal (2013) objetivou elaborar, avaliar e validar um modelo de regressão para o pico das componentes vertical (Fy), ântero-posterior (Fx) e da resultante (FR) da FRS durante a caminhada e corrida na água.

Participaram do trabalho de Haupenthal (2013) 143 pessoas fragmentadas em dois grupos: grupo de regressão (n=119) e grupo de validação (n=24). Conforme Haupenthal (2013),

os sujeitos efetuaram a caminhada e a corrida na água em um dos cinco níveis de imersão (0,75, 0,90, 1,05, 1,20 e 1,35 m) com alternância da velocidade (lenta, normal e rápida).

As variáveis independentes usadas foram: imersão, velocidade e as características antropométricas (estatura, massa, índice de conicidade e as circunferências de abdômen, quadril, coxa e perna). Haupenthal (2013) notou que as variáveis estatura e imersão tinham forte relação entre si e para esquivar-se da multicolinearidade foram associadas por meio da criação da variável: razão_EI.

Haupenthal (2013), utilizou esses dados no software SPSS (aplicando regressão linear), para assim gerar modelos para a predição de FR, Fy e Fx nos exercícios de caminhada e corrida. Destaca-se ainda que Haupenthal (2013) também apresentou um modelo para ambos os exercícios (corrida e caminhada), onde o mesmo juntou os dados de caminhada e corrida para gerar o modelo de predição para FR, Fy e Fx.

No exercício de caminhada, Haupenthal (2013) alcançou um R^2 de 0,88 para FR, 0,87 para Fy e 0,90 para Fx. Já quando trata-se do exercício de corrida os valores de R^2 obtidos para FR, Fy e Fx foram respectivamente, R^2= 0,67, R^2=0,62 e R^2=0,84.

No experimento em que Haupenthal (2013), juntou os dois exercícios para realizar a predição de FR, Fy e Fx, o R^2 de FR foi 0,70 enquanto o Fy alcançou apenas 0,66 e Fx 0,87.

Mello (2010) em seu experimento realizou um estudo epidemiológico sobre a prevalência de lesões em atletas do gênero feminino, na modalidade futebol, identificando os mecanismos e os aspectos dessas lesões. Mello (2010) utilizou a mineração de dados para realizar uma análise de sessenta e nove fichas de avaliação das atletas de futebol feminino da Associação Desportiva do Guarujá. Dessa maneira, 84 lesões foram identificadas, das quais 57,1% transcorreram por contato direto e 47,6% das lesões foram contusões. Outro padrão identificado foi que o local mais acometido foi o joelho (26,1%), e em sua maioria a lesão de joelho ocorre durante períodos de competições.

A pesquisa de Chester, Khondoker, Shepstone, Lewis e Jerosch-Herold (2019), utilizou técnicas de mineração de dados como a regressão e a classificação para predizer pessoas que têm tendência a sofrer com dores nos ombros. Identificando preditores desse tipo de dor e permitindo que os médicos e fisioterapeutas agrupasse as pessoas em grupos de risco para dor persistente no ombro e incapacidade ou em grupos que são menos propensos a sofrer com esse tipo de dor. Foram utilizados dados de 810 pessoas, coletados durante 6 meses.

Outro estudo onde a mineração de dados foi aplicada para auxiliar na saúde e na fisioterapia é o caso do estudo de Carvalho, Moser, Silva e Dallagassa (2012).

A pesquisa de Carvalho, Moser, Silva e Dallagassa (2012) tinha o objetivo de debater o potencial de utilização do processo KDD sobre um conjunto de dados de acompanhamento fisioterapêutico de pacientes, bem como sua utilidade na tomada de decisões terapêuticas ou profiláticas. Segundo Carvalho, Moser, Silva e Dallagassa (2012), selecionou-se um subconjunto de dados, relacionados a prontuários em uma clínica de fisioterapia, do qual foram coletados três grandes grupos-alvo de tarefas de Mineração de Dados: associação, classificação e agrupamento.

No trabalho de Carvalho, Moser, Silva e Dallagassa (2012), foram coletados padrões por meio dos dados, de modo que fosse possível ao leitor entender passo a passo do processo utilizado para obter os mesmos, ampliando seu entendimento dos resultados alcançados. Muitos padrões foram descobertos, os quais evidenciaram as possíveis relações entre as variáveis utilizadas.

3 METODOLOGIA E EXPERIMENTOS

Esta seção apresenta o contexto em que os dados foram utilizados, assim como os processos usados na metodologia, do desenvolvimento dessa pesquisa que foram: Seleção e Pré-processamento dos dados, Transformação dos dados, Seleção dos dados, Geração e avaliação dos modelos de predição. Além disso, essa seção exibe também a descrição dos experimentos realizados.

3.1 CONTEXTO

Para a elaboração deste trabalho foram usados e analisados os dados contidos em estudos anteriores do Laboratório de Pesquisas em Biomecânica Aquática da UDESC os quais também foram usados na pesquisa de Haupenthal (2013), sendo os mesmos fornecidos via e-mail por Haupenthal (2013).

O banco de dados conta com dados de 143 sujeitos suas características antropométricas como a massa corporal, estatura, circunferência da coxa, cintura, abdômen, e diversos outros fenômenos que interferem na predição da FRS. Essa base de dados conta com 6 tipos de exercícios, os quais este estudo tem a intenção de predizer, são eles:

- Caminhada lenta

- Caminhada normal
- Caminhada rápida
- Corrida lenta
- Corrida normal
- Corrida rápida

O arquivo fornecido por Haupenthal (2013), contém todos os dados agrupados em um único arquivo. Salienta-se ainda que os dados foram disponibilizados em formato “.CSV".

Os dados foram coletados durante a realização do estudo de (HAUPENTHAL, 2013). Os mesmos foram anteriormente aprovados por um Comitê de Ética em Pesquisa em Seres Humanos – da Universidade do Estado de Santa Catarina, durante o estudo de Haupenthal (2013). Aprovação esta que está documentada sobre o número de referencia 52/2008.

A coleta de dados foi efetuada na piscina e no Laboratório de Pesquisas em Biomecânica Aquática do CEFID/UDESC, usando uma plataforma de força subaquática posicionada no centro de uma passarela de 8 m e conectada ao Sistema de Aquisição de Dados ADS2002-IP.

Os dados contam com as instâncias de todos os sujeitos que fizeram parte dos experimentos finais de Haupenthal (2013), ou seja, a base contém todos os indivíduos que foram testados nos seguintes exercícios subaquáticos: caminhada lenta, caminhada normal, caminhada rápida, corrida lenta, corrida normal e corrida rápida. Sendo assim a base de dados consta com 858 instâncias. Restringindo a base de dados, para cada um dos exercícios subaquáticos teríamos então 143 instâncias para cada um dos tipos de exercícios conforme demonstra a Tabela 1. Salienta-se que os mesmos sujeitos foram submetidos a todos os experimentos.

Fundamentado no contexto da base de dados citada anteriormente e nos resultados do estudo de Haupenthal (2013), acredita-se que a mesma tem potencial para a realização de predições dos picos de FRS em diversos tipos de exercícios subaquáticos.

Assim sendo esta pesquisa propõem modelos de predição dos picos de FRS em diversos tipos de exercícios subaquáticos, pois acredita-se que dessa forma os profissionais de saúde e especialistas da área (saúde) possam melhorar seus trabalhos o que consequentemente gera benefícios (ligados a recuperação de lesões) a população.

Tabela 1 – Quantitativos dos totais de instâncias.

Tipo Registro	Quantidade de registros
Caminhada lenta	143
Caminhada normal	143
Caminhada rápida	143
Corrida lenta	143
Corrida normal	143
Corrida rápida	143
Total de Registros	**858**

Fonte: Elaborado pelo autor.

Como dito anteriormente, todos os sujeitos foram submetidos a todos os experimentos, porém cada sujeito pertence a um grupo específico de imersão (profundidade da água). A Tabela 2 demonstra os níveis de imersão existentes no estudo, sua profundidade em metros e quantos sujeitos fazem parte de cada grupo.

Tabela 2 – Níveis de imersão.

Nível	Valor do nível em metros	Quantidade de sujeitos no grupo
Nível 1	0,75 m	35
Nível 2	0,90 m	22
Nível 3	1,05 m	29
Nível 4	1,20 m	24
Nível 5	1,35 m	33

Fonte: Elaborado pelo autor.

Sendo assim, os sujeitos (e seus respectivos grupos de imersão) foram submetidos à predição dos picos de FRS em 6 experimentos distintos, onde cada uma das atividades subaquáticas anteriormente citadas é um experimento. A Figura 9 demonstra de forma mais clara a caracterização de cada experimento.

No estudo de Haupenthal (2013), o mesmo deixa claro que as velocidades foram classificadas de acordo com as médias obtidas por meio da velocidade dos sujeitos. Baseado

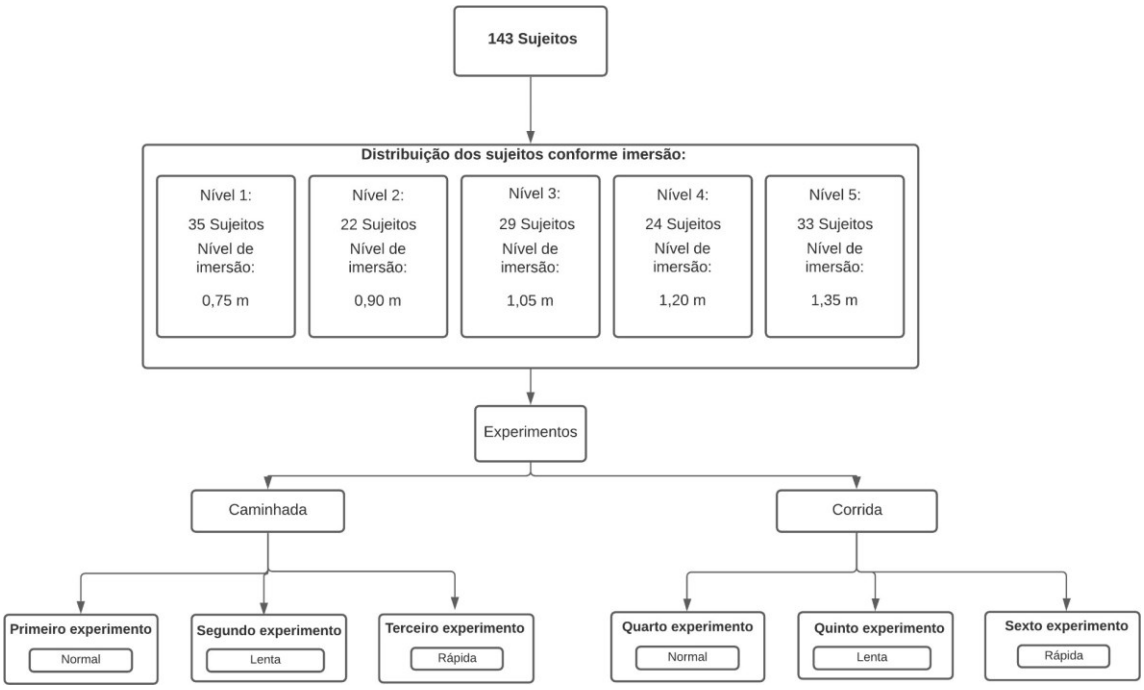
nessa média, para ser caracterizado como velocidade normal deveria estar dentro de uma variação de 10% da média. Para ser classificado como velocidade lenta, o participante teria que realizar a velocidade de deslocamento inferior à média menos 20 %. Já para ser considerado velocidade rápida o valor deveria ser o valor da média mais 20%. A Tabela 3 exemplifica de forma simples essas situações.

Tabela 3 - Valores de velocidade.

Exercício	Velocidade lenta	Velocidade normal	Velocidade rápida
Caminhada	≤ 20% da média	Média registrada ± 10%	≥ 20% da média
Corrida	≤ 20 % da média	Média registrada ± 10%	≥ 20% da média

Fonte Adaptado de Haupenthal (2013).

Figura 9 – Taxonomia dos experimentos.



Fonte: Elaborado pelo autor.

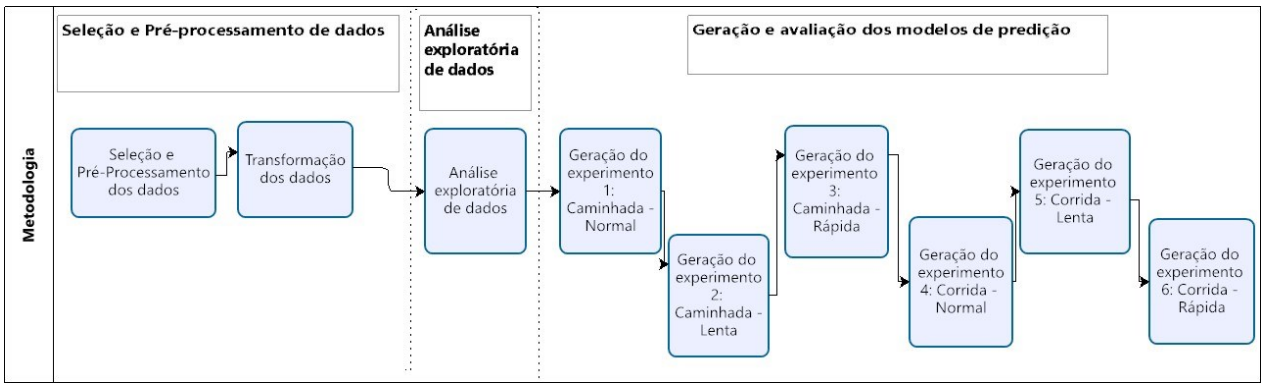
3.2 METODOLOGIA

A metodologia empregue para a elaboração desta pesquisa é composta das seguintes fases: Seleção e Pré-processamento dos dados, Análise exploratória de dados, geração e avaliação dos modelos de predição.

Destaca-se que o processo de pré-processamento de dados e transformação de dados foi aplicado de maneira igualitária para todos os experimentos. Após a transformação de dados, uma análise exploratória dos dados foi realizada e em seguida todos os experimentos foram gerados e avaliados.

A Figura 10 expressa de forma mais clara os processos e subprocessos usados na metodologia. Já o Quadro 1 esclarece a ordem e a base de dados usadas em cada experimento. Ao longo deste capítulo será explicado de forma detalhada cada etapa da metodologia e processos adotados.

Figura 10 – Processos utilizados na metodologia.



Fonte: Elaborado pelo autor.

Quadro 1 – Dados usados em cada experimento.

Experimento	**Caminhada/Corrida**	**Ritmo**
Primeiro experimento	Caminhada	Normal
Segundo experimento	Caminhada	Lenta
Terceiro experimento	Caminhada	Rápida
Quarto experimento	Corrida	Normal
Quinto experimento	Corrida	Lenta
Sexto experimento	Corrida	Rápida

Fonte: Elaborado pelo autor.

3.3 SELEÇÃO E PRÉ-PROCESSAMENTO DE DADOS

Após o download dos exercícios subaquáticos recebidos via e-mail, notou-se que que os dados não se encontravam divididos por tipo de exercício. Os mesmos encontravam-se juntos todos

no mesmo arquivo .CSV, ao qual a variável "Exercicio" continha as categorias do tipo de exercício. Notou-se também que o arquivo contava com 42 variáveis, as quais muitas não são necessárias e relevantes para o estudo em questão como demonstra o Apêndice C.

Sendo assim, no primeiro momento julgou-se necessário separar essa base em 6 partes de acordo com o tipo de atividade subaquática. Essa segmentação foi feita por meio do leitor e editor de arquivos ".CSV" (microsoft office excel versão 2010) e seus recursos de filtros, pois como a quantidade de dados não é grande não foi necessária a criação de scripts por meio de linguagens de programação.

Em um segundo momento considerou-se importante a criação de uma base de dados, por meio de um banco de dados e de uma Linguagem de definição de dados (DDL), para depositar todas as informações relevantes presentes no arquivo original. Devido a esse contexto foi preciso fazer um corte (desconsiderar) em grande parte das variáveis, as quais de acordo com o estudo de Haupenthal (2013) não possuem importância para esse tipo de estudo. A Figura 11 ilustra de maneira mais clara as tabelas criadas no banco de dados.

Assim, as variáveis utilizadas nesse estudo são as variáveis a serem preditas: FR, Fx e Fy e as variáveis de entrada (independentes): Velocity, Sex, Age, Mass_kg, Waist_cm, Thigh_cm, IH_ratio, descritas com maiores detalhes no dicionário de dados presente no anexo A.

Sendo assim, foi desenvolvido um script para preencher o banco de dados, local onde os dados usados neste trabalho ficaram salvos. Script esse que é mostrado no Apêndice D. A ferramenta escolhida para efetuar essa tarefa foi o anaconda[1] e o jupyter[2] notebook utilizando linguagem python.

Constatou-se que para o preenchimento do banco de dados era necessário realizar a etapa de pré-processamento, etapa essa que teve o objetivo de transformar as vírgulas que separavam as casas decimais em pontos. Essa etapa foi realizada em todas as variáveis selecionadas para compor o banco de dados, como demonstra o Apêndice D.

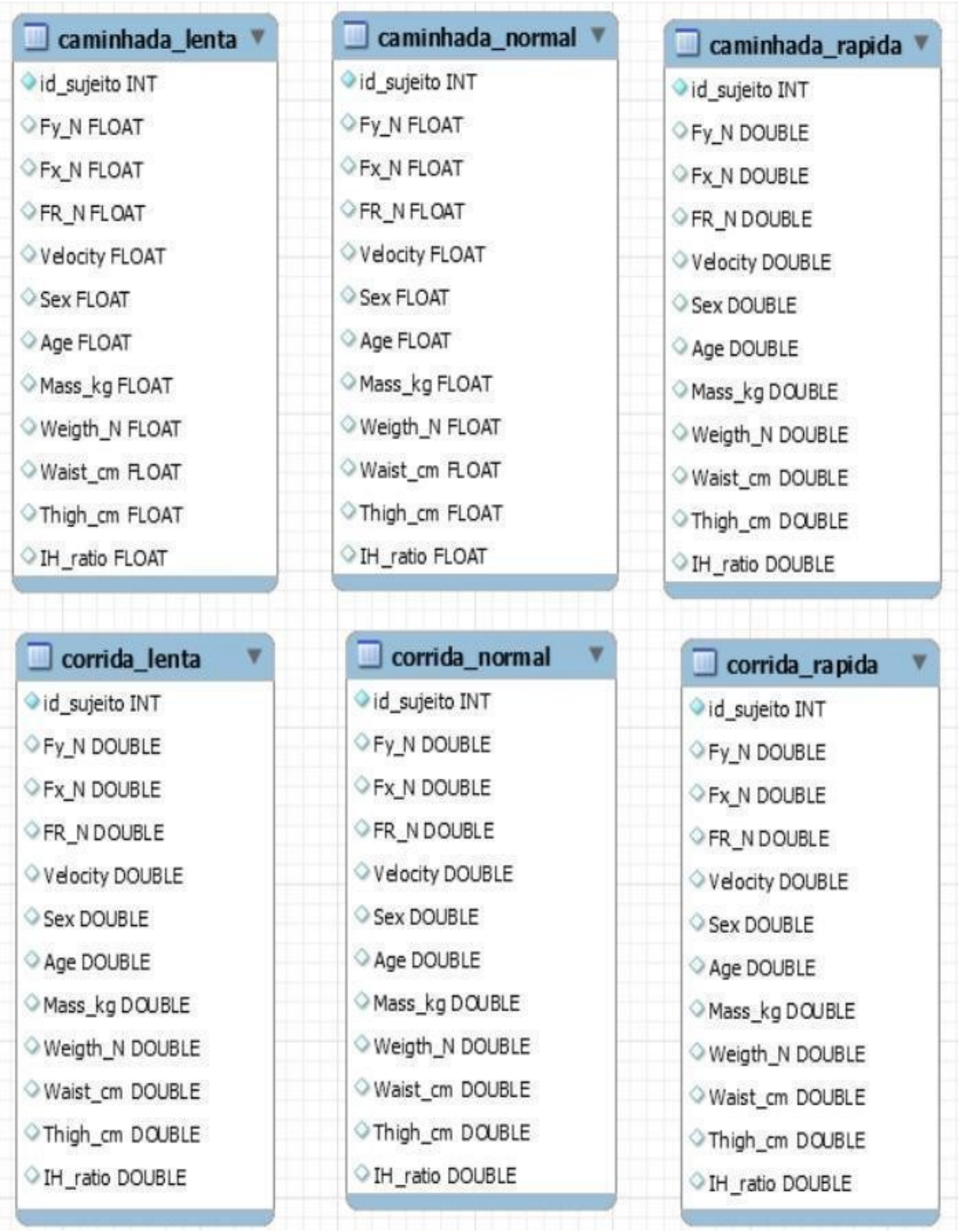
Sendo assim, nota-se que os dados originais foram carregados na linguagem python por meio de uma biblioteca chamada pandas. Após o carregamento, foram submetidos ao pré-processamento e depois enviados ao banco de dados por meio de uma biblioteca de comunicação com o banco de dados Mysql.

---

[1] https://www.anaconda.com/ - Versão utilizada: Anaconda Navigator 1.10.0

[2] https://jupyter.org/ - Versão utilizada: Jupter 1.0.0

Figura 11 – Banco de dados criado.



Fonte: Elaborado pelo autor.

3.4 TRANSFORMAÇÃO DOS DADOS

Com a meta de utilizar os dados, existentes na base, como input para os algoritmos de mineração de dados, é necessário efetuar a tarefa de transformação dos dados, para que os mesmos tornem-se apropriados para esses algoritmos e para teoricamente aumentar o desempenho dos mesmos.

Sendo assim, depois de completar as tarefas de pré-processamento dos dados, o próximo passo a ser efetuado neste trabalho foi a transformação dos dados. Uma das etapas mais relevantes efetuadas nesta fase de transformações dos dados foi a normalização, pois por meio da mesma torna-se viável normalizar as variáveis fazendo com que as grandezas de todas as variáveis fiquem com a mesma ordem de grandeza (base, unidade). Dessa maneira a diferença do valor de cada variável em relação a sua própria média é o que vai ser ressaltado, podendo assim fazer com que os algoritmos de mineração de dados obtenham resultados mais efetivos.

As variáveis que foram submetidas a esse processo de normalização tratam-se das variáveis independentes. A Tabela 4 demonstra os campos submetidos ao processo citado,suas médias e a fórmula matemática usada para a normalização.

Tabela 4 - Variáveis submetidas a normalização.

Nome da variável	**Média**	**Fórmula matemática**
Velocity	0.603471830985915	(Velocity/Média Velocity)
Sex	1.5774647887323943	(Sex/Média Sex)
Age	24.739436619718308	(Age/Média Age)
Mass_kg	69.51126760563378	(Mass_kg/Média Mass_kg)
Waist_cm	81.76760563380282	(Waist_cm/Média Waist_cm)
Thigh_cm	51.54225352112676	(Thigh_cm/Média Thigh_cm)
IH_ratio	0.6074744366197182	(IH_ratio/Média IH_ratio)

Fonte: Elaborado pelo autor.

Para que a normalização fosse de fato concretizada, foi realizado um script em python que efetuava a importação dos dados presentes no banco de dados, por meio de uma Query SQL (Linguagem de Consulta Estruturada), importando a base de dados no pandas para que em seguida fosse feito os cálculos de médias de cada uma das variáveis de entrada (independentes) como apresenta o Apêndice E.

Após a descoberta das médias de cada variável foi criado uma função em python para normalizar os dados de todas as variáveis. Essa função fazia a leitura de cada uma das linhas presentes no banco de dados, e dividia o valor de cada variável por sua respectiva média, e em seguida por meio de uma conexão com o banco de dados fazia a atualização do valor de todas as variáveis através dos comandos SQL de UPDATE. O Apêndice E expõe de forma mais detalhada o processo utilizado.

3.5 ANÁLISE EXPLORATÓRIA DE DADOS

Em relação à análise exploratória de dados, o primeiro procedimento adotado foi a realização de gráficos por meio da biblioteca Seaborn, para poder verificar a influência das variáveis independentes sobre as variáveis a serem preditas (Fx, Fy e FR). O Apêndice F mostra o script utilizado para a geração dos gráficos.

Em um segundo momento foi verificada a distribuição normal dos dados por meio da biblioteca scipy.stats (utilizando uma função chamada de normaltest) onde a significância usada foi de 0,05. Outras informações averiguadas foram moda, mediana, média, desvio padrão, desvio absoluto, amplitude, variância, simetria, covariância e correlação. Destaca-se que essa análise foi feita após a etapa de transformação dos dados, onde os mesmos já tinham sido submetidos ao processo de normalização. O Apêndice F demonstra com clareza o algoritmo utilizado para efetuar as tarefas citadas. Salienta-se que esse script foi aplicado para todos os 6 grupos de dados separadamente.

3.6 GERAÇÃO E AVALIAÇÃO DOS MODELOS DE PREDIÇÃO

Essa etapa refere-se ao processo de mineração de dados, que é o centro desse trabalho, a tarefa de procura da informação aplicada nessa pesquisa, foi efetuada por meio das técnicas de regressão.

Os experimentos efetuados nesta dissertação foram feitos, focados no objetivo inicial. Dessa forma, as variáveis selecionadas foram testadas para verificar se os resultados seriam positivos ou negativos em relação a sua eficácia perante aos algoritmos.

Os experimentos foram realizados usando a linguagem de programação python por meio do jupyter notebook (ferramenta do software anaconda) e diversas bibliotecas como Pandas, NumPy e Scikit-learn (utilizada para gerar os modelos de predição). Destaca-se que nessa etapa do trabalho a biblioteca utilizada para a geração de gráficos foi a Plotly. A seguir será elucidado como os experimentos foram executados bem como esclarecido suas particularidades em relação aos dados usados em cada um dos experimentos.

Foi escolhida a linguagem de programação python e suas bibliotecas para aplicar os algoritmos de mineração de dados, pois o mesmo vem se destacando no cenário de mineração de dados e ciências de dados, devido a sua facilidade e seus resultados robustos e positivos.

Ao longo da formação dessa pesquisa, inúmeros experimentos foram feitos precedentemente a delimitação total da metodologia usada no mesmo. Sendo assim notou-se que 6 experimentações ganharam destaque em relação às demais e por esse motivo tornaram-se as experimentações finais deste trabalho.

O **primeiro experimento**, consistia em gerar modelos preditivos por meio doconjunto de dados de caminhada normal. Baseado nisso, por meio da linguagem de programação python e suas bibliotecas (scikit-learn, pandas e etc), as variáveis independentes foram utilizadas para compor o input dos algoritmos usados, onde as variáveis a serem preditas foram FR, Fx e Fy.

Desta forma é notório que os mesmos scripts foram aplicados para as 3 variáveis que desejava-se predizer (FR, Fx e Fy), apenas intercalando no script o atributo a ser predito.

Destaca-se que diversas combinações de variáveis foram testadas, porém o conjunto que representava todos os atributos independentes foi o que obteve melhores resultados, sendo assim as seguintes variáveis foram usadas como input:

- Velocity
- Sex
- Age
- Mass_kg
- Waist_cm
- Thigh_cm
- IH_ratio

Os algoritmos usados para gerar os modelos de predição, foram selecionados com base nos algoritmos de regressão da biblioteca scikit-learn. Dessa maneira os algoritmos aplicados foram :

- Linear Regression
- SVR
- AdaBoost Regressor
- Gradient Boosting Regressor
- Tweedie Regressor
- Lasso Lars
- Lasso
- Bayesian Ridge
- Ridge Cross-Validation
- Ridge
- Random Forest Regressor
- K-Neighbors Regressor
- Decision Tree Regressor
- Multi-layer Perceptron regressor

No tocante ao método utilizado para treinamento e teste, foi empregue a técnica de Cross-Validation K-Fold, onde o K foi adotado como K=10. Salienta-se ainda que testes foram realizados com a técnica Cross-Validation Leave-one-out, porém com a técnica de K-Fold encontra-se menos sobreposição nas amostras de validação cruzada, do que com Leave-one-out, o que consiste em menos correlação entre as estimativas de validação cruzada.

O uso de recurso computacional é menor na técnica de K-Fold (comparada com a técnica de Leave-one-out), o que a torna mais fácil de ser replicada. Outro motivo para adoção dessa técnica de K-Fold onde k=10 é que aparentemente esse é o método mais utilizado nos trabalhos e pesquisas encontrados, que envolvem mineração de dados na saúde. Porém salienta-se que os resultados obtidos com Leave-one-out estão descritos no anexo B.

Em relação aos parâmetros usados nos algoritmos, adotou-se os seus respectivos padrões da biblioteca Scikit-Learning (estão retratados no Apêndice G). No entanto, vale ressaltar que testes de hiper-parametrização foram realizados por meio do método

GridSearchCV e também RandomizedSearchCV, porém os resultados mostraram-se negativos e foram desprezados.

Os **demais experimentos** seguiram exatamente a mesma metodologia do primeiro experimento, com a diferença de que os dados utilizados fazem parte de outro grupo de exercício, como demonstra o Quadro 1.

4 RESULTADOS E DISCUSSÕES

Depois de realizar todos os experimentos, iniciou-se a análise e discussão dos resultados obtidos, onde primeiramente realizou-se uma análise exploratória dos dados perante sua distribuição. Já em um segundo momento procurou-se entender e apresentar os resultados de cada experimento individualmente, por meio dos resultados obtidos com a aplicação dos algoritmos, fazendo um comparativo entre os mesmos. Em um terceiro momento comparou-se os resultados dos experimentos entre os mesmos. Já em um quarto momento a comparação ocorreu por meio dos resultados das pesquisas e dos trabalhos relacionados.

4.1 ANÁLISE GERAL DOS DADOS

Essa seção apresenta gráficos que foram gerados durante a exploração de dados, para que fosse possível ter ideia de quais variáveis poderiam ser importantes para os algoritmos de regressão. Testes de distribuição normal dos dados também são apresentados onde a significância usada foi de 0.05. Outras informações apresentadas nesta seção são: moda, mediana, média, desvio padrão, desvio absoluto, amplitude, variância, simetria, covariância e correlação.

Destaca-se que é apresentado nesse tópico os dados referentes aos exercícios de caminhada normal e de corrida normal, os dados referentes aos demais grupos encontram-se no Apêndice A.

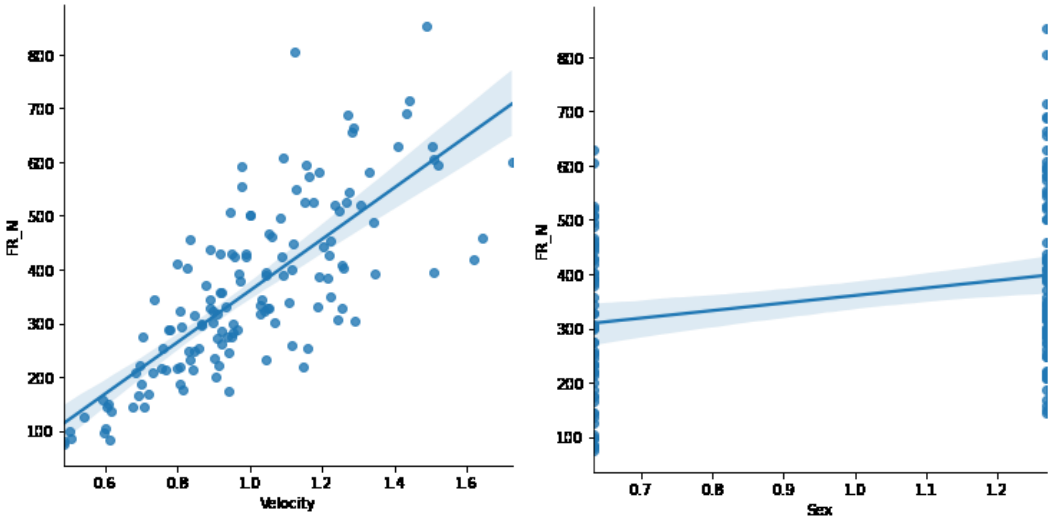
4.1.1 INFLUÊNCIA DAS VARIÁVEIS NOS DADOS DE CAMINHADA NORMAL

Neste tópico será apresentado a influência de cada variável de input sobre os atributos a serem preditos FR, Fy E Fx. A Figura 12 demonstra a influência das variáveis Velocity e Sex

em relação a FR. A Figura 13 e Figura 14 mostram a importância dos atributos Thigh_cm e IH_ratio Age, Mass_kg e Waist_cm também em relação ao FR.

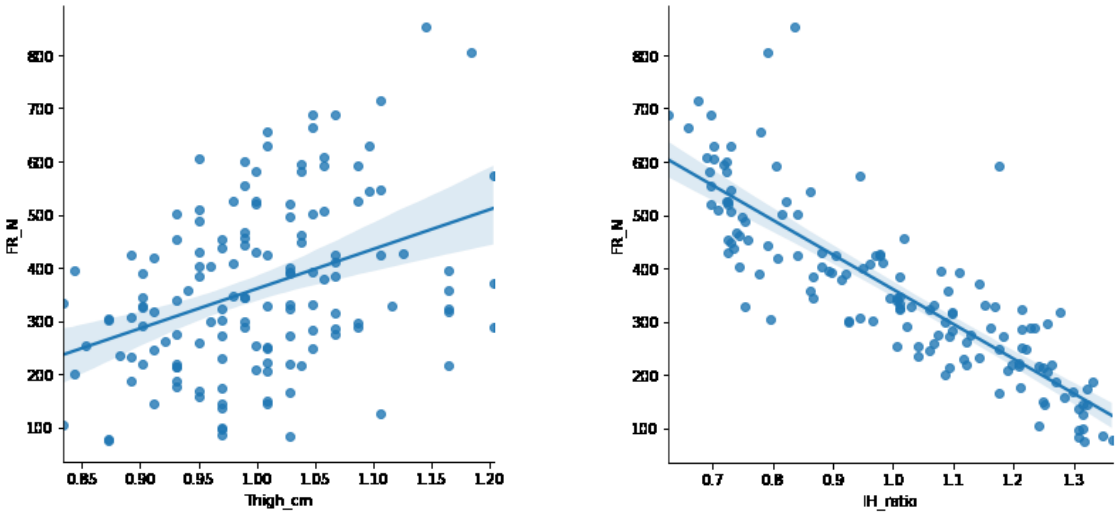
Analisando essas figuras podemos concluir que as variáveis Velocity e IH_ratio são as que mais influenciam em FR quando trata-se do exercício de caminhada normal, ou seja são as variáveis mais importantes para a predição da FR usando algoritmos (nesse tipo de exercício).

Figura 12 – Influência das variáveis Velocity e Sex em relação a FR no exercício de caminhada normal.



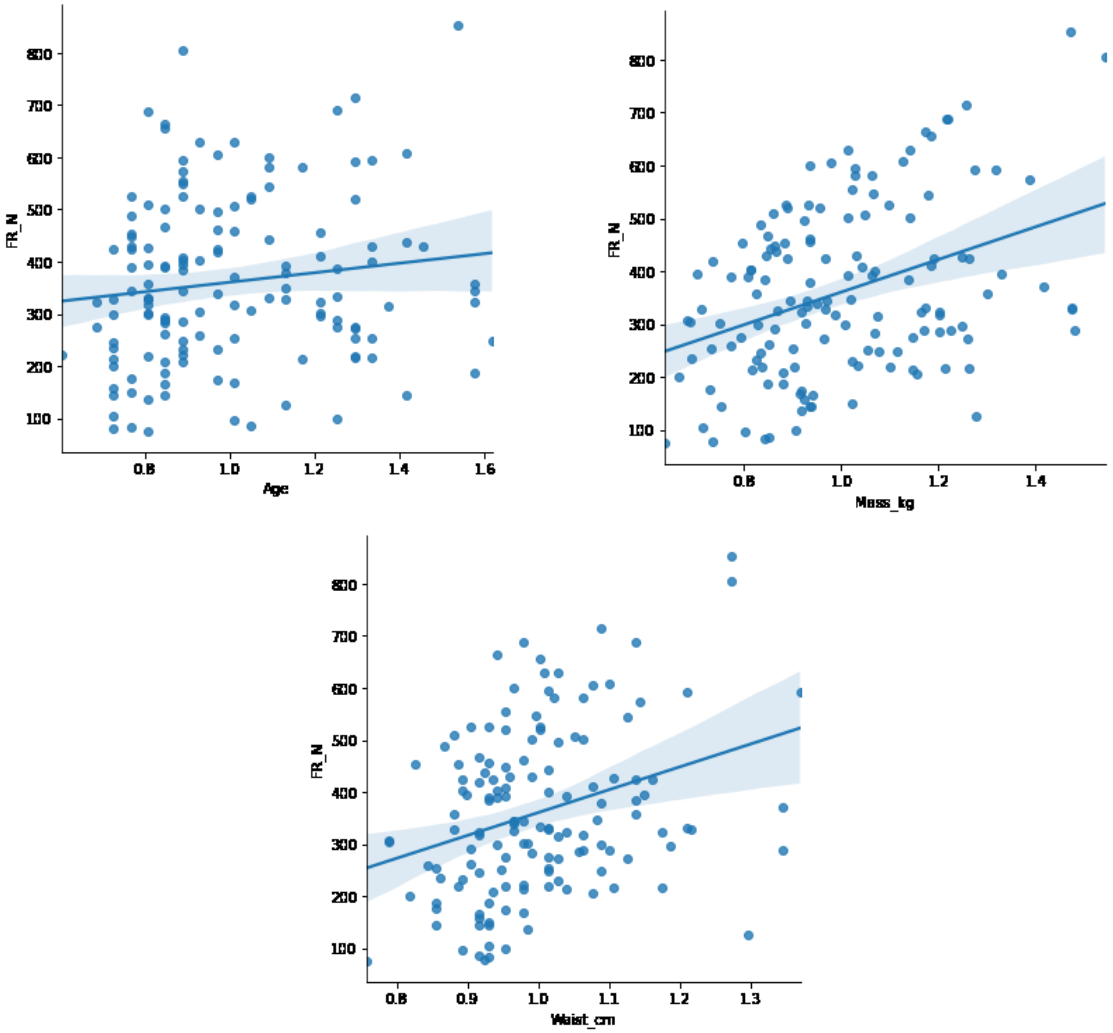
Fonte: Elaborado pelo autor.

Figura 13 - Influência das variáveis Thigh_cm e IH_ratio em relação a FR no exercício de caminhada normal.



Fonte: Elaborado pelo autor.

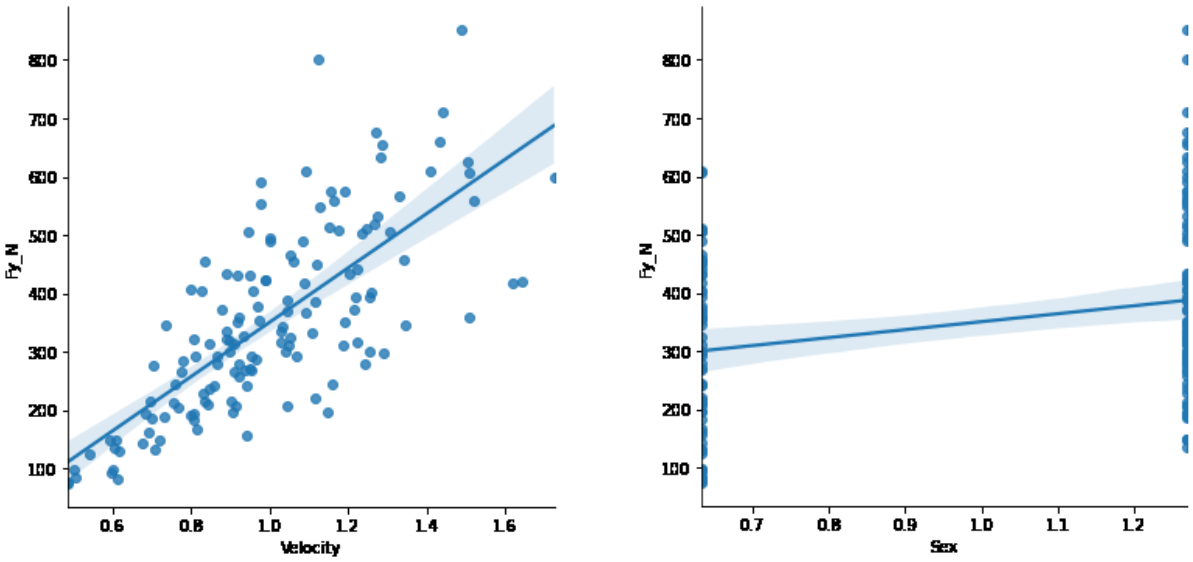
Figura 14 – Influência das variáveis Age, Mass_kg e Waist_cm em relação a FR no exercício de caminhada normal.



Fonte: Elaborado pelo autor.

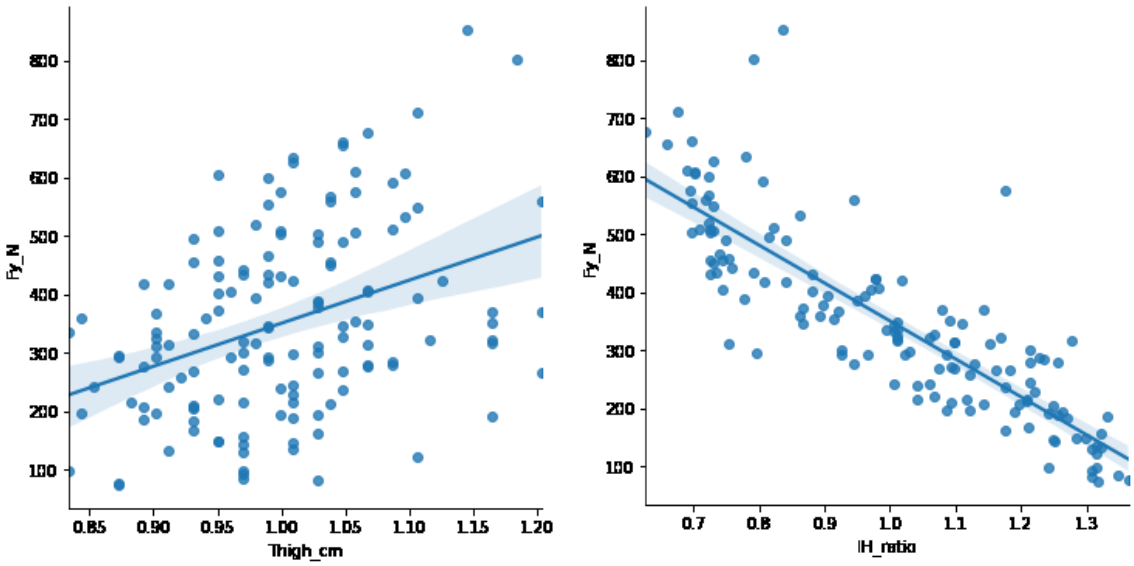
No tocante à influência dos atributos em relação a variável FY, a Figura 15 demonstra essa relação de Fy com Velocity e Sex. A Figura 16 e Figura 17 demonstram a relevância dos atributos Age, Mass_kg, Waist_cm, Thigh_cm e IH_ratio do mesmo modo em referência ao Fy. Assim como no caso do FR, as variáveis que mais demonstraram influência em Fy foram Velocity e IH_ratio.

Figura 15 – Influência das variáveis Velocity e Sex em relação a FY no exercício de caminhada normal.



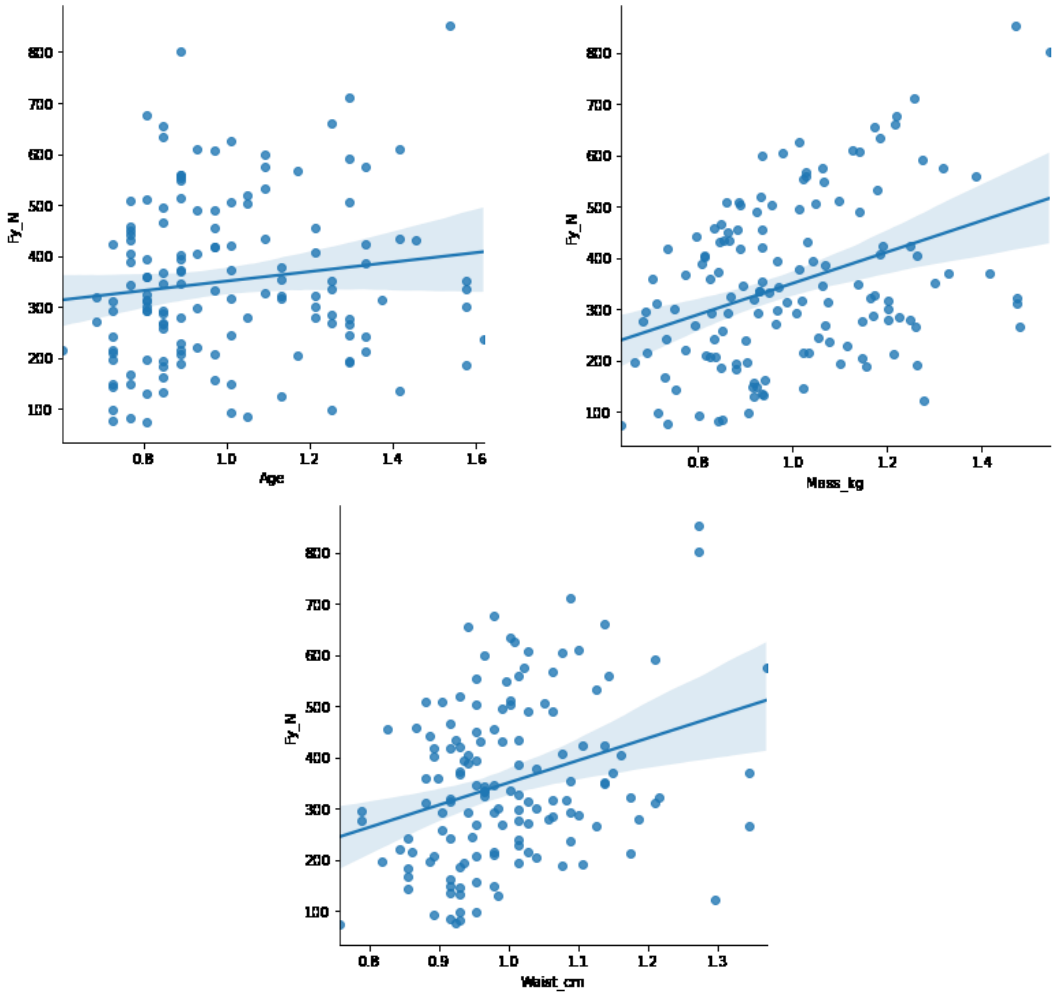
Fonte: Elaborado pelo autor.

Figura 16 - Influência das variáveis Thigh_cm e IH_ratio em relação a FY no exercício de caminhada normal.



Fonte: Elaborado pelo autor.

Figura 17 – Influência das variáveis Age, Mass_kg e Waist_cm em relação a FY no exercício de caminhada normal.

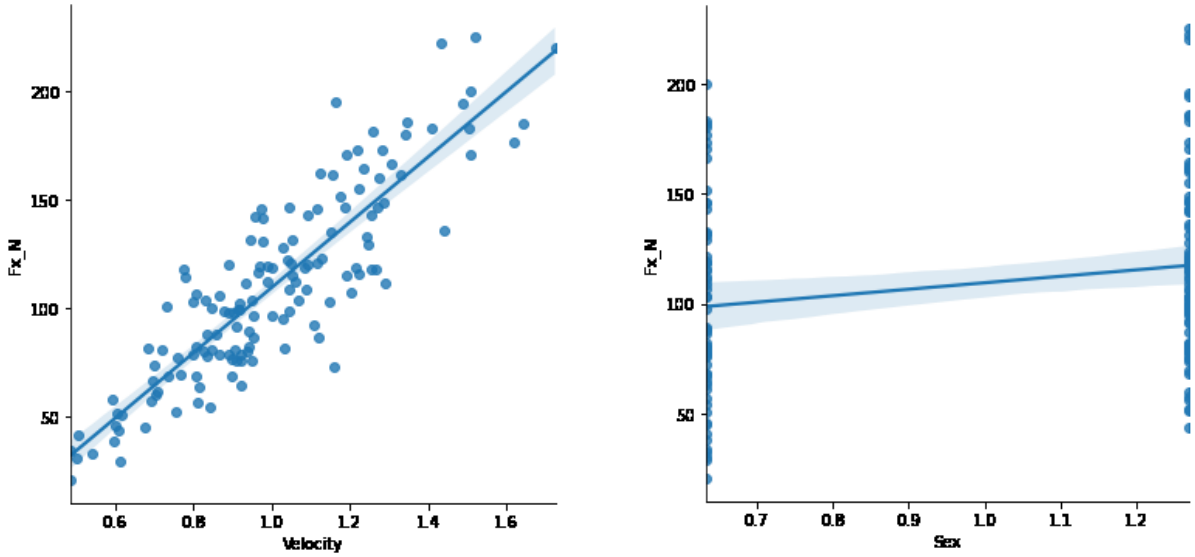


Fonte: Elaborado pelo autor.

Para a predição de Fx nota-se uma redução da influência dos atributos Velocity e principalmente de IH_ratio se comparado com a predição de FR e Fy. Mesmo com essa redução o atributo Velocity continua sendo o atributo mais importante para a predição de Fx.

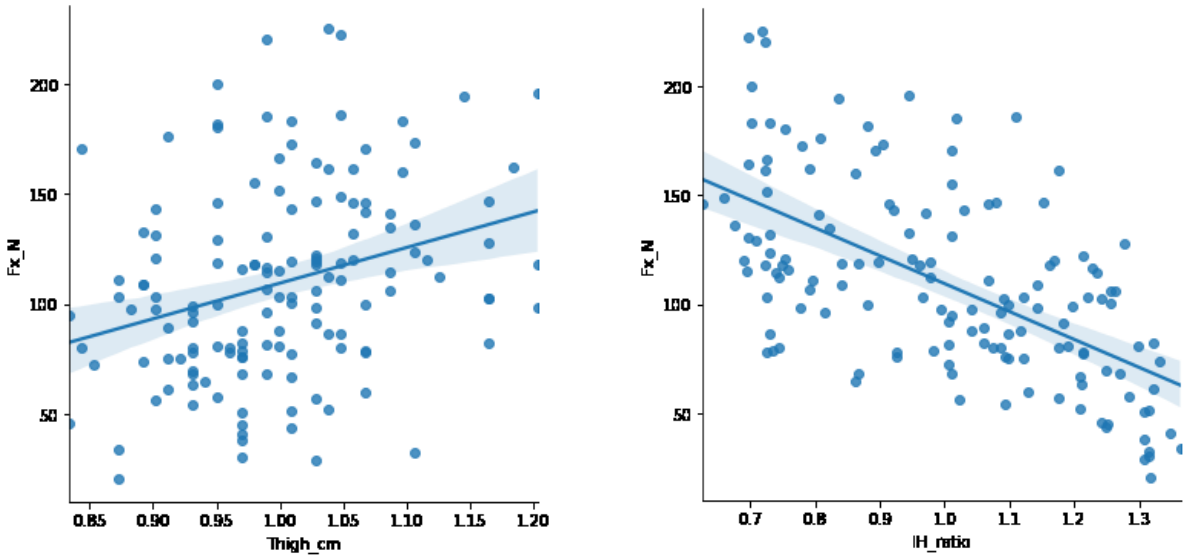
O atributo IH ratio já não se trata de uma variável tão relevante para a predição de Fx como é nos casos de FR e Fy, passando a ser uma variável de influência similar a todas as demais. A Figura 18, Figura 19 e Figura 20 atestam o cenário citado.

Figura 18 – Influência das variáveis Velocity e Sex em relação a Fx no exercício decaminhada normal.



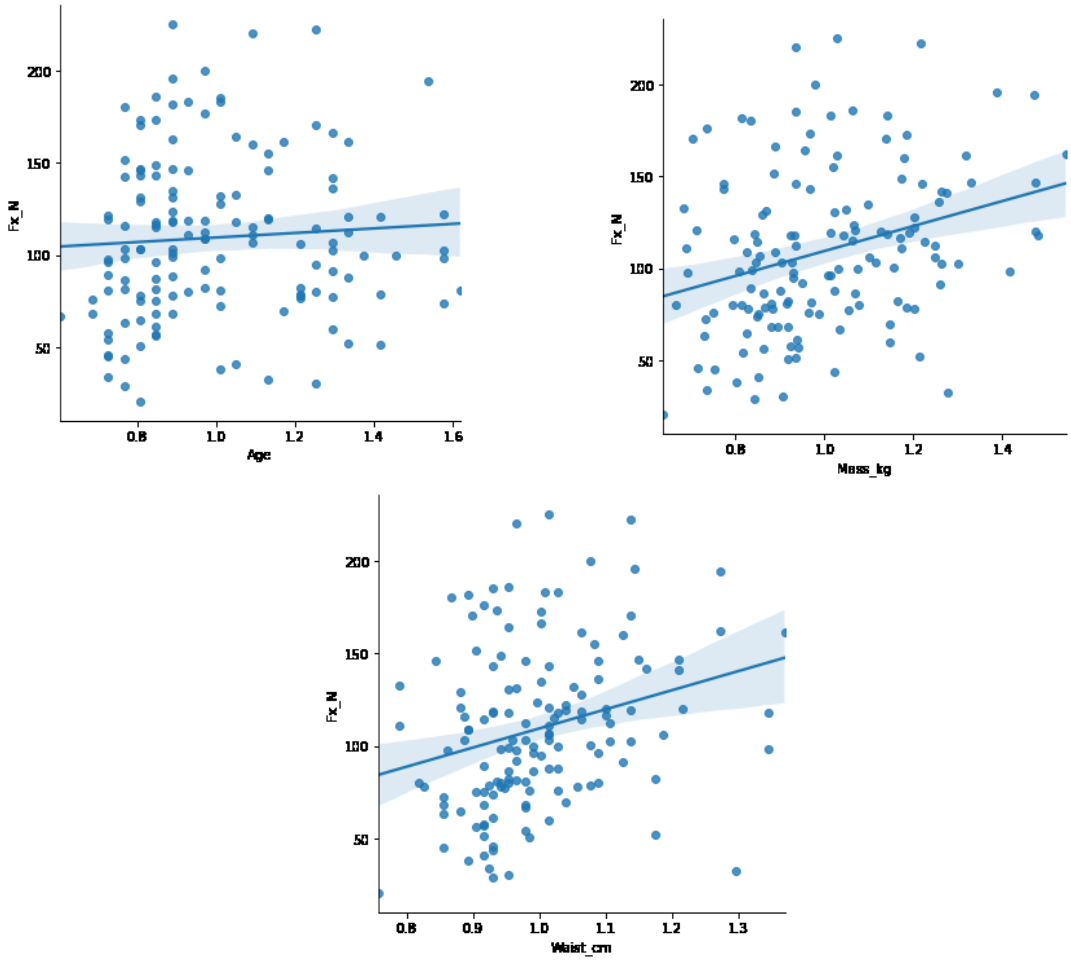
Fonte: Elaborado pelo autor.

Figura 19 - Influência das variáveis Thigh_cm e IH_ratio em relação a Fx no exercício de caminhada normal.



Fonte: Elaborado pelo autor.

Figura 20 - Influência das variáveis Age, Mass_kg e Waist_cm em relação a Fx noexercício de caminhada normal.



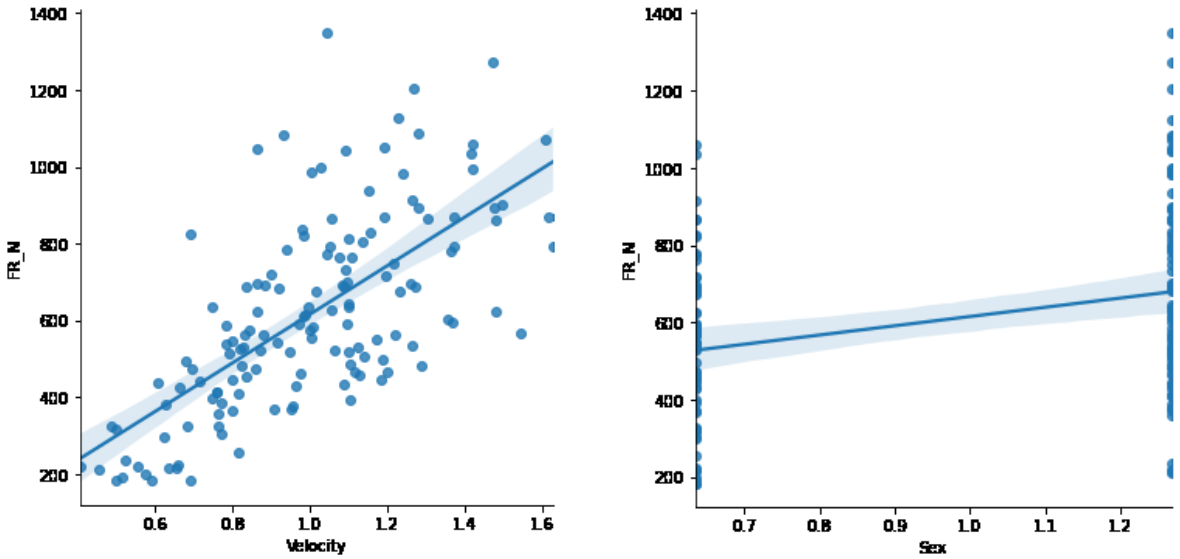
Fonte: Elaborado pelo autor.

4.1.2 INFLUÊNCIA DAS VARIÁVEIS NOS DADOS DE CORRIDA NORMAL

Neste tópico será apresentado a influência de cada variável de input sobre os atributos a serem preditos FR, FY E FX, quando trata-se do exercício de corrida normal. Sendo assim por meio de gráficos será feita uma análise. A Figura 21 demonstra a influência das variáveis Velocity e Sex em relação a FR. A Figura 22 e Figura 23 mostram a importância dos atributos Thigh_cm e IH_ratio, Age, Mass_kg e Waist_cm também em relação ao FR.

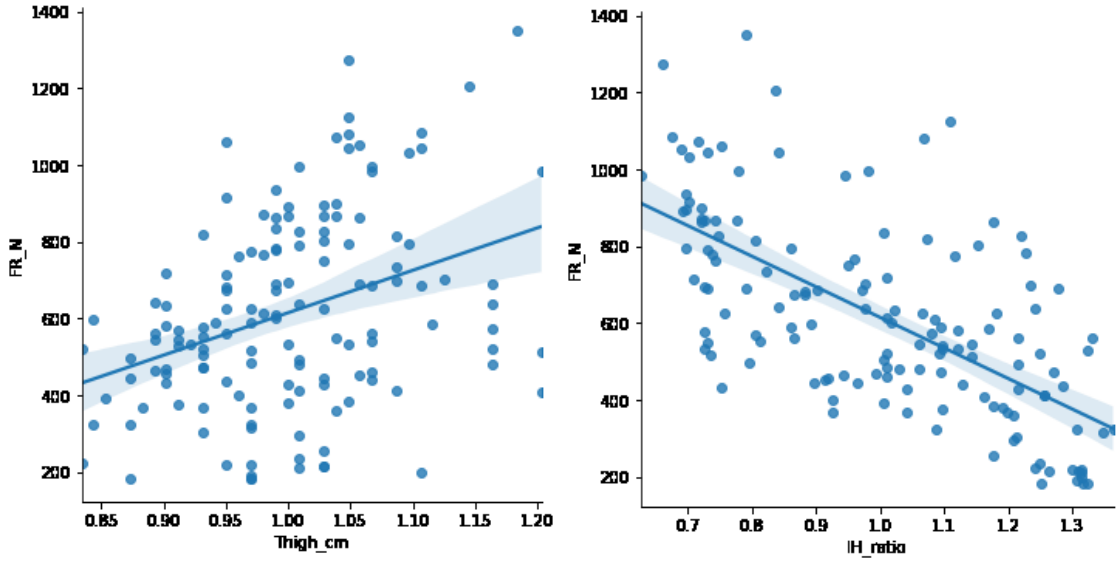
Analisando essas figuras podemos concluir que as variáveis Velocity e IH_ratio são as que mais influenciam em FR quando trata-se do exercício de corrida normal, ou seja são as variáveis mais importantes para a predição da FR usando algoritmos (nesse tipo de exercício).

Figura 21 – Influência das variáveis Velocity e Sex em relação a FR no exercício de corrida normal.



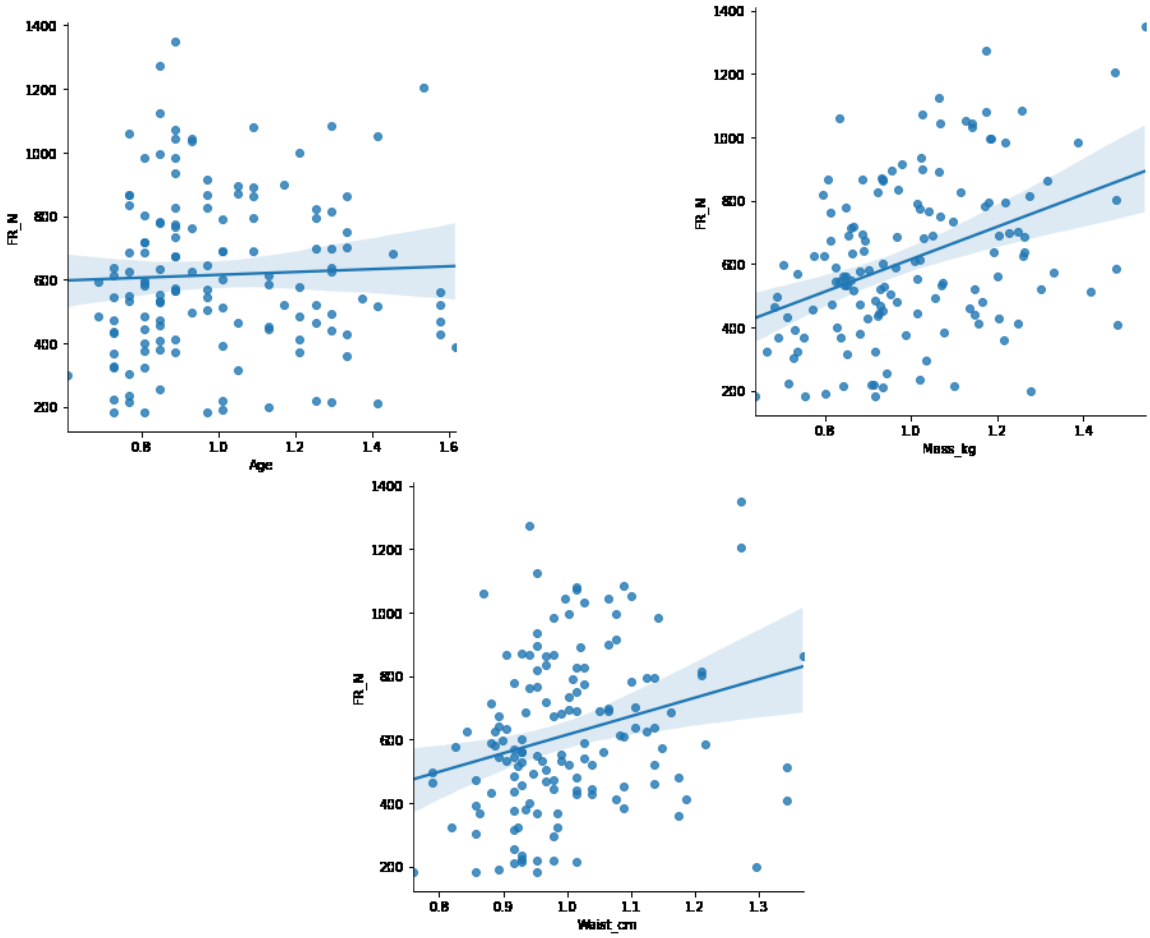
Fonte: Elaborado pelo autor.

Figura 22 - Influência das variáveis Thigh_cm e IH_ratio em relação a FR no exercício de corrida normal.



Fonte: Elaborado pelo autor.

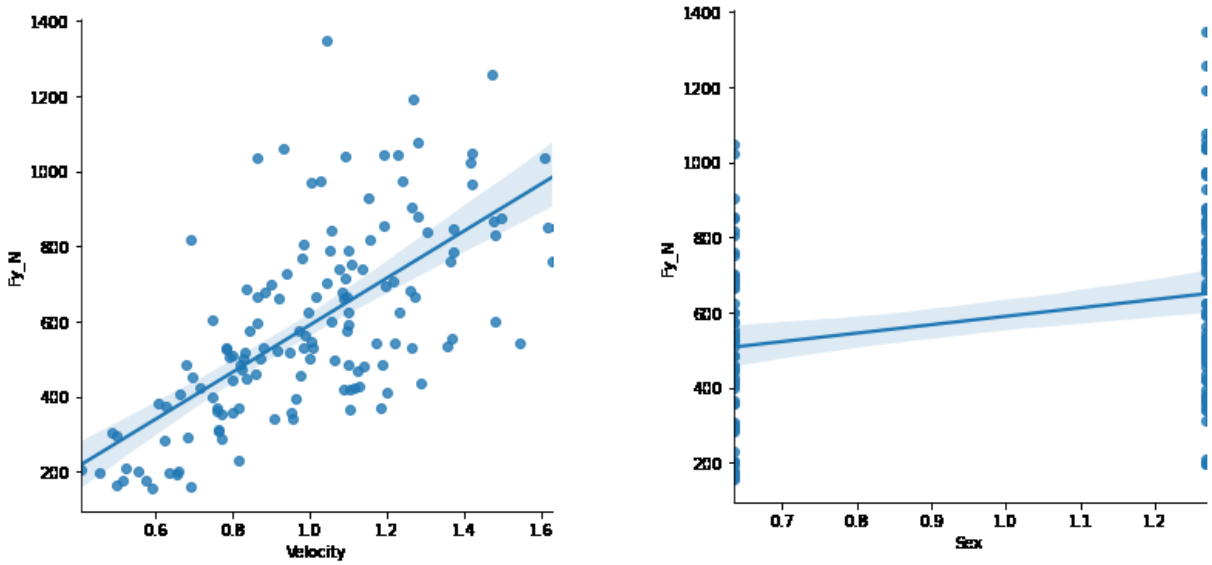
Figura 23 – Influência das variáveis Age, Mass_kg e Waist_cm em relação a FR no exercício de corrida normal.



Fonte: Elaborado pelo autor.

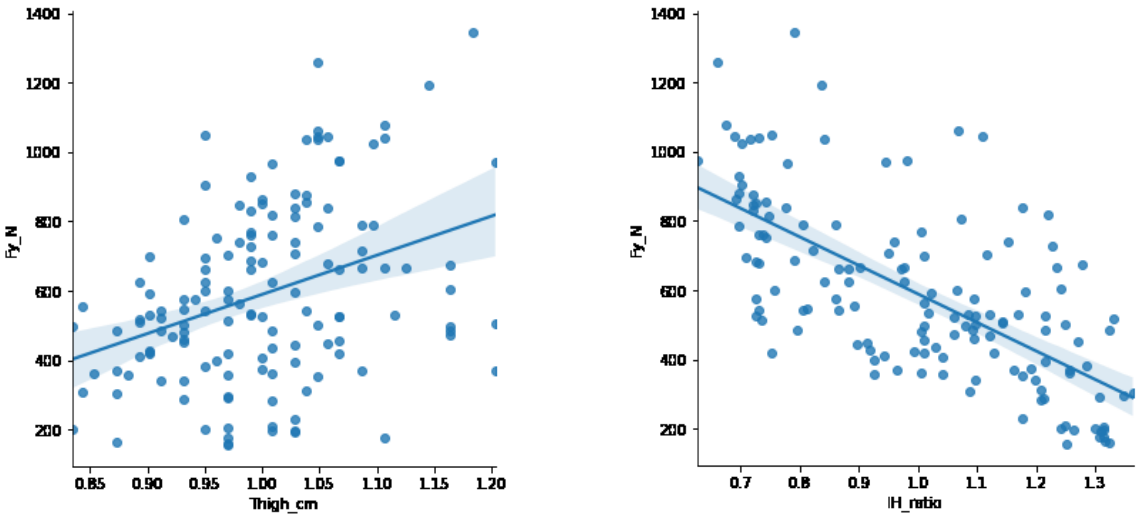
Em relação à influência dos atributos sobre Fy, a Figura 24, Figura 25 e Figura 26 demonstra essa relação. Como no exercício de caminhada normal, já apresentados anteriormente, as variáveis que mais influenciam em Fy quando o exercício é corrida normal são Velocity e IH_ratio.

Figura 24 – Influência das variáveis Velocity e Sex em relação a Fy no exercício de corrida normal.



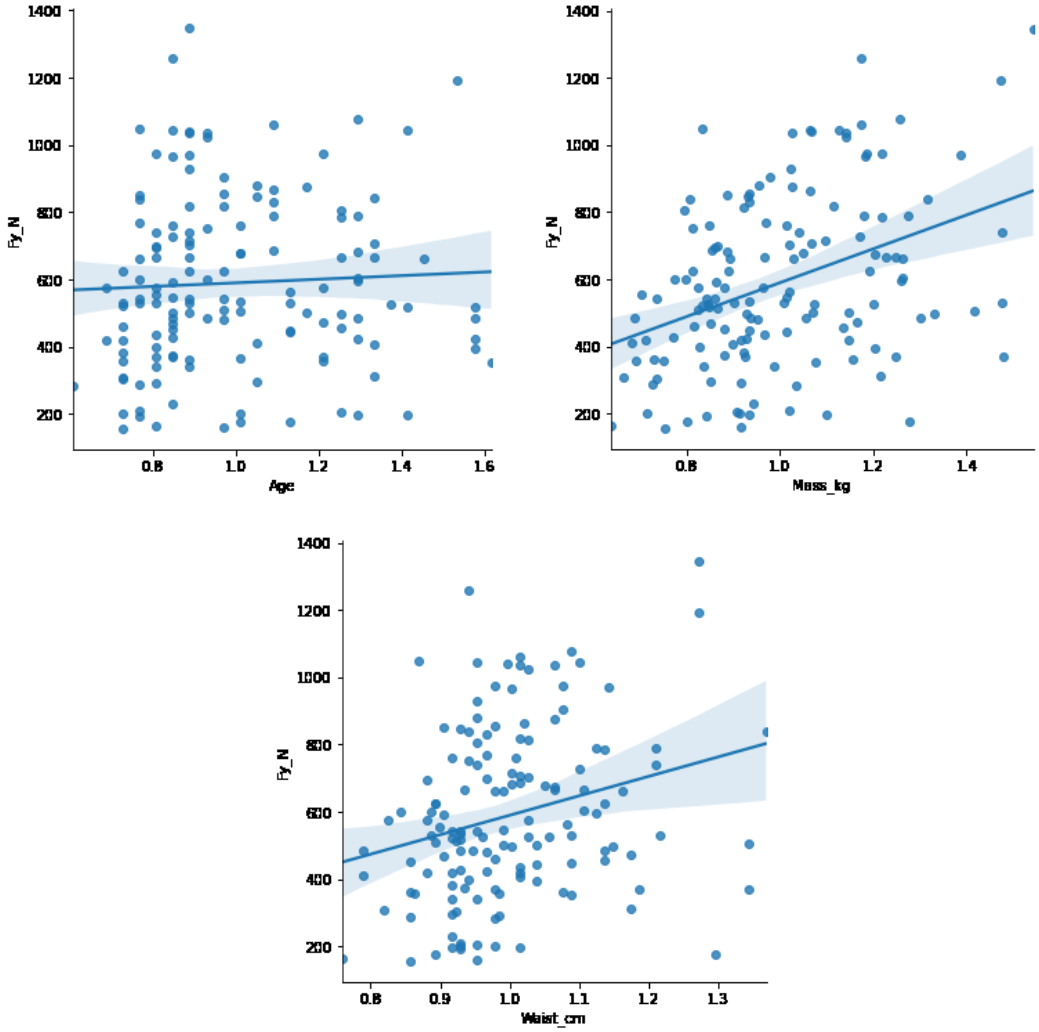
Fonte: Elaborado pelo autor.

Figura 25 – Influência das variáveis Thigh_cm e IH_ratio em relação a Fy no exercício de corrida normal.



Fonte: Elaborado pelo autor.

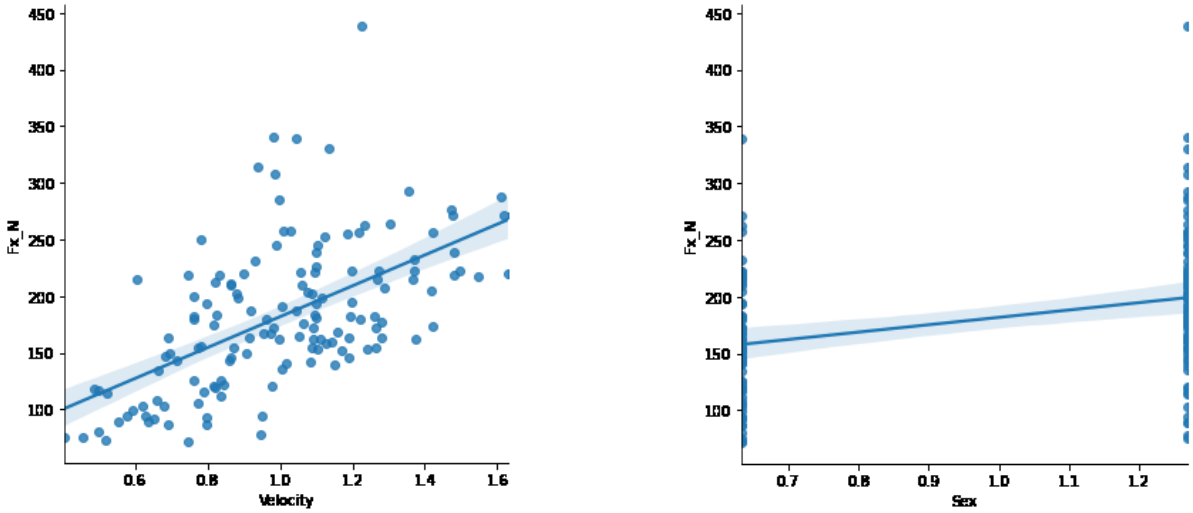
Figura 26 – Influência das variáveis Age, Mass_kg e Waist_cm em relação a Fy no exercício de corrida normal.



Fonte: Elaborado pelo autor.

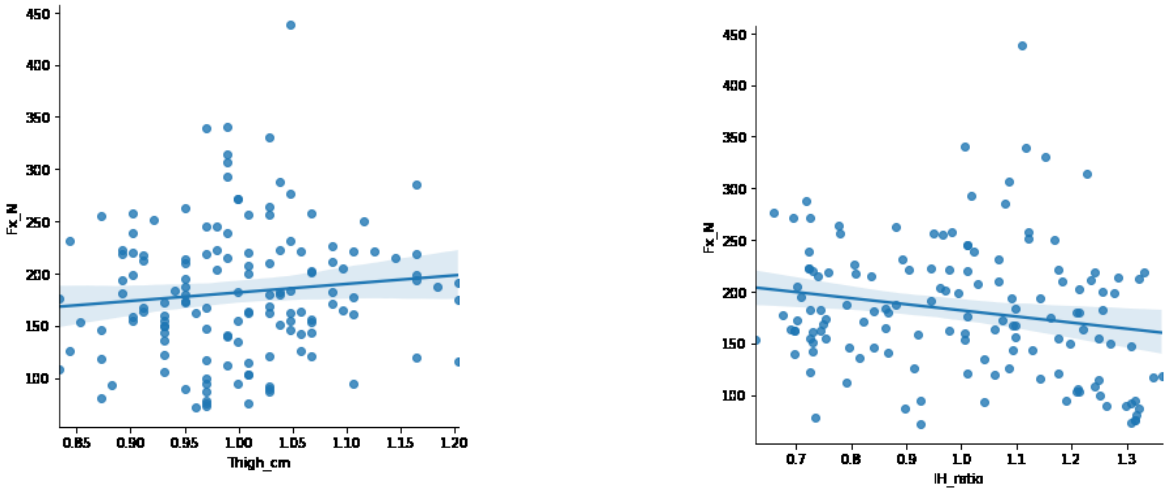
Assim como no exercício de caminhada normal, para o exercício de corrida normal também ocorreu uma redução da influência dos atributos Velocity e especialmente de IH_ratio se confrontado com a predição de FR e Fy. Mesmo com essa diminuição de influência a variável Velocity continua sendo o atributo mais relevante para a predição de Fx. A Figura 27, Figura 28 e Figura 29 mostram de forma clara o cenário citado.

Figura 27 – Influência das variáveis Velocity e Sex em relação a Fx no exercício de corrida normal.



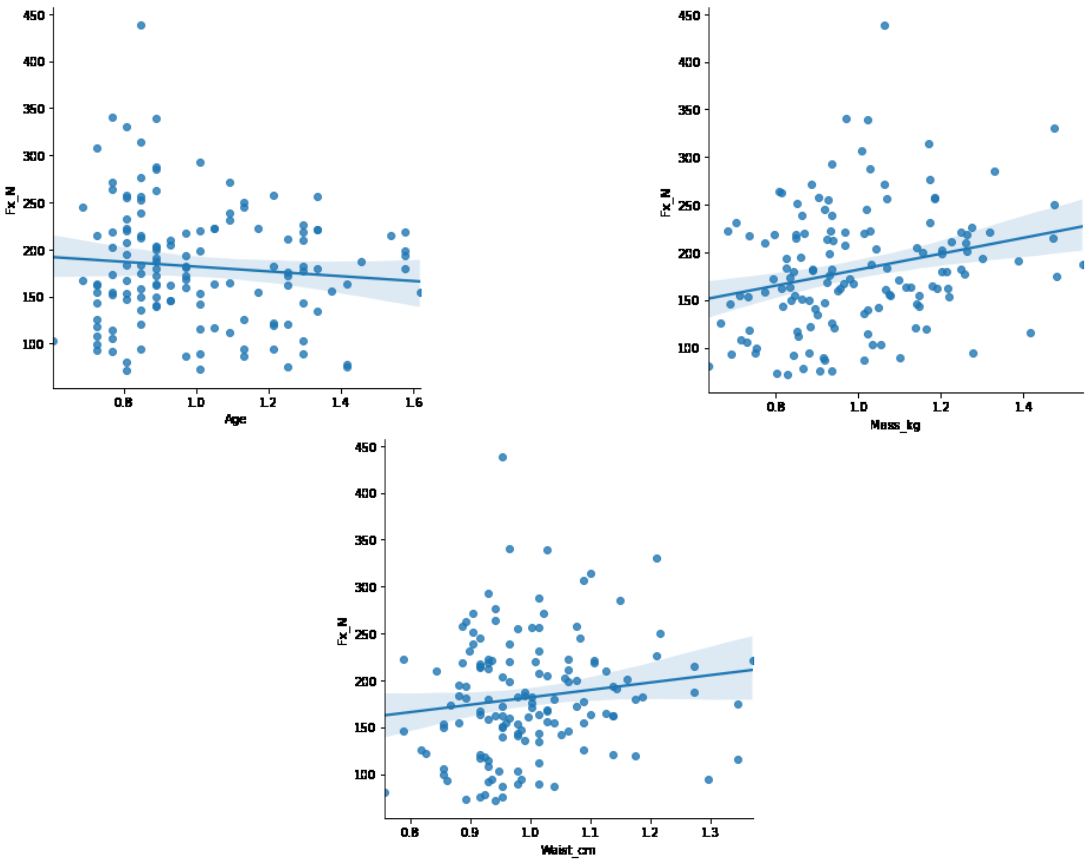
Fonte: Elaborado pelo autor.

Figura 28 – Influência das variáveis Thigh_cm e IH_ratio em relação a Fx no exercício de corrida normal.



Fonte: Elaborado pelo autor.

Figura 29 – Influência das variáveis Age, Mass_kg e Waist_cm em relação a Fx no exercício de corrida normal.



Fonte: Elaborado pelo autor.

4.1.3 DISTRIBUIÇÃO NORMAL

Nesta seção será apresentado informações sobre um teste de distribuição normal dos dados, onde o objetivo era explorar e ter clareza sobre os dados antes de gerar os modelos de predição. A significância foi definida com o valor de 0,05, sendo assim o valor P (P Value) deve ser >0,05 para que a amostra seja considerada normal. Dessa forma será apresentado informações (relacionado a cada uma das variáveis) como P Value, statistic, moda, mediana, média, desvio padrão, desvio absoluto, amplitude, variância, simetria, covariância e correlação. Destaca-se que serão apresentados os dados apenas dos exercícios de caminhada normal e corrida normal, as informações dos demais exercícios estarão no Apêndice B.

4.1.3.1 DISTRIBUIÇÃO NORMAL DOS DADOS DE CAMINHADA NORMAL E CORRIDA NORMAL

Como o Quadro 2 demonstra, podemos observar que as variáveis: Velocity, Sex, Waist_cm, Thigh_cm, IH_ratio e FR são normais para ambos os exercícios. As variáveis Age, Mass_kg não são normais para nenhum dos dois conjuntos de dados citados. Já o atributo FX é normal para o exercício de caminhada normal, mas é considerado não normal para o exercício de corrida normal, enquanto que FY é o oposto de FX quanto à normalidade. Sendo assim podemos observar que a maioria dos atributos possuem uma distribuição normal.

A Tabela 5 apresenta os valores de P Value para cada uma das variáveis no exercício de caminhada normal e a Tabela 6 exibi o P Value para os atributos do exercício de corrida normal. No tocante a simetria dos dados o Quadro 3 expõe a situação completa.

Na Tabela 5 também é possível verificar dados como a mediana, média, desvio padrão, amplitude dos dados, statistic, variância dos dados e desvio absoluto para o exercício de caminhada normal. Os mesmos dados estatísticos podem ser consultados para o exercício de corrida normal por meio da Tabela 6.

A covariância dos dados de caminhada normal, estão contidas na Tabela 7. A correlação dos dados do mesmo exercício estão presentes na Tabela 8. Já em relação aos dados de corrida normal a Tabela 9 apresenta a covariância e a Tabela 10 a correlação. Os dados gráficos de cada variável se encontram no Apêndice b.

Quadro 2 - Distribuição dos dados de caminhada normal e corrida normal.

Variável	**Distribuição caminhada normal**	**Distribuição corrida normal**
Velocity	normal	normal
Sex	normal	normal
Age	Não é normal	Não é normal
Mass_kg	Não é normal	Não é normal
Waist_cm	normal	normal
Thigh_cm	normal	normal
IH_ratio	normal	normal

FX	normal	Não é normal
FY	Não é normal	normal
FR	normal	normal

Fonte: Elaborado pelo autor.

Quadro 3 - Simetria dos dados de caminhada normal e corrida normal.

Variável	**Simetria caminhada normal**	**Simetria corrida normal**
Velocity	Assimétrica a esquerda	Assimétrica a direita
Sex	Assimétrica a esquerda	Assimétrica a esquerda
Age	Assimétrica a direita	Assimétrica a direita
Mass_kg	Assimétrica a direita	Assimétrica a direita
Waist_cm	Assimétrica a direita	Assimétrica a direita
Thigh_cm	Assimétrica a direita	Assimétrica a direita
IH_ratio	Assimétrica a esquerda	Assimétrica a esquerda
FX	Assimétrica a direita	Assimétrica a direita
FY	Assimétrica a direita	Assimétrica a direita
FR	Assimétrica a direita	Assimétrica a direita

Fonte: Elaborado pelo autor.

Tabela 5 – Estatísticas dos dados de caminhada normal.

Variável	P Value	Significancia comparativa	Statistic	Moda	Mediana	Média	Desvio padrão	Desvio absoluto	Amplitude	Variância
Velocity	0,26	0,05	2,66	1,04	0,96	1	0,25	0,22	1,23	0,06
Sex	1,21	0,05	998,92	1,26	1,26	0,99	0,31	0,3	0,63	0,09
Age	0	0,05	13,26	0,88	0,92	1	0,23	0,19	1,01	0,05
Mass_kg	0,03	0,05	6,98	0,93	0,96	0,99	0,19	0,16	0,9	0,03
Waist_cm	3,08	0,05	20,77	0,91	0,97	1	0,11	0,08	0,61	0,01
Thigh_cm	0,26	0,05	2,66	0,97	0,99	1	0,08	0,06	0,36	0
IH_ratio	4,13	0,05	89,26	1,01	1,01	0,99	0,18	0,18	0,73	0,04
FX	0,11	0,05	4,28	20,51	104,74	109,49	43,87	34,96	204,56	1924,89
FY	0,03	0,05	6,68	73,1	326,27	350,24	159,66	127,98	778,82	25494
FR	0.053	0,05	5,84	74,65	332,12	360,54	159,68	128,52	778,07	25499,3

Fonte: Elaborado pelo autor.

Tabela 6 – Estatísticas dos dados de corrida normal.

Variável	P Value	Significancia comparativa	Statistic	Moda	Mediana	Média	Desvio padrão	Desvio absoluto	Amplitude	Variância
Velocity	0,35	0,05	2,04	0,69	1	0,99	0,27	1,22	1,22	0,07
Sex	1,21	0,05	998,92	1,26	0,92	0,99	0,31	0,63	0,63	0,09
Age	0	0,05	13,26	0,88	0,92	1	0,23	1,01	1,01	0,05
Mass_kg	0,03	0,05	6,98	0,93	0,96	0,99	0,19	0,9	0,9	0,03
Waist_cm	3,08	0,05	20,77	0,91	0,97	1	0,11	0,61	0,61	0,01
Thigh_cm	0,26	0,05	2,66	0,97	0,99	1	0,08	0,36	0,36	0
IH_ratio	4,13	0,05	89,26	1,01	1,01	0,99	0,2	0,73	0,73	0,04
FX	0	0,05	14,37	71,46	176,64	181,78	63,88	366,92	366,92	4081,84
FY	0,058	0,05	5,67	177,59	541,27	589,68	254,78	1191,76	1191,76	64915,1
FR	0,09	0,05	4,72	181,27	578,27	614,79	252,98	1168,92	1168,92	64002,6

Fonte: Elaborado pelo autor.

Tabela 7 - Covariância dos dados de caminhada normal.

Variáveis	FY	FX	FR	Velocity	Sex	Age	Mass_kg	Waist_cm	Thigh_cm	IH_ratio
FY	1.000.000	0.747567	0.997875	0.745457	0.269314	0.137156	0.373443	0.310703	0.377263	-0.852187
FX	0.747567	1.000.000	0.778539	0.878931	0.210380	0.066035	0.301909	0.268066	0.301651	-0.606490
FR	0.997875	0.778539	1.000.000	0.767483	0.273096	0.134284	0.377553	0.313125	0.381483	-0.848699
Velocity	0.745457	0.878931	0.767483	1.000.000	0.105943	-0.006789	0.077895	0.047715	0.078428	-0.707952
Sex	0.269314	0.210380	0.273096	0.105943	1.000.000	0.348489	0.643061	0.480911	0.435141	0.037810
Age	0.137156	0.066035	0.134284	-0.006789	0.348489	1.000.000	0.337843	0.365338	0.283625	0.039489
Mass_kg	0.373443	0.301909	0.377553	0.077895	0.643061	0.337843	1.000.000	0.874805	0.818892	0.045240
Weigth_N	0.373443	0.301909	0.377553	0.077895	0.643061	0.337843	1.000.000	0.874805	0.818892	0.045240
Waist_cm	0.310703	0.268066	0.313125	0.047715	0.480911	0.365338	0.874805	1.000.000	0.745528	0.055880
Thigh_cm	0.377263	0.301651	0.381483	0.078428	0.435141	0.283625	0.818892	0.745528	1.000.000	-0.070929
IH_ratio	-0.852187	-0.606490	-0.848699	-0.707952	0.037810	0.039489	0.045240	0.055880	-0.070929	1.000.000

Fonte: Elaborado pelo autor.

Tabela 8 - Correlação dos dados de caminhada normal.

Variáveis	FY	FX	FR	Velocity	Sex	Age	Mass_kg	Waist_cm	Thigh_cm	IH_ratio
FY	25.494.012.570	5.236.874.370	25.442.467.754	30.417.530	13.512.833	5.137.510	11.652.753	5.650.584	4.919.852	-28.334.116
FX	5.236.874.370	1.924.891.390	5.454.404.277	9.854.623	2.900.525	0.679671	2.588.592	1.339.594	1.080.925	-5.540.933
FR	25.442.467.754	5.454.404.277	25.499.268.669	31.319.531	13.704.017	5.030.438	11.782.207	5.695.211	4.975.393	-28.221.072
Velocity	30.417.530	9.854.623	31.319.531	0.065308	0.008508	-0.000407	0.003890	0.001389	0.001637	-0.037674
Sex	13.512.833	2.900.525	13.704.017	0.008508	0.098750	0.025691	0.039492	0.017213	0.011168	0.002474
Age	5.137.510	0.679671	5.030.438	-0.000407	0.025691	0.055035	0.015489	0.009762	0.005434	0.001929
Mass_kg	11.652.753	2.588.592	11.782.207	0.003890	0.039492	0.015489	0.038192	0.019473	0.013071	0.001841
Weigth_N	11.652.753	2.588.592	11.782.207	0.003890	0.039492	0.015489	0.038192	0.019473	0.013071	0.001841
Waist_cm	5.650.584	1.339.594	5.695.211	0.001389	0.017213	0.009762	0.019473	0.012973	0.006936	0.001325
Thigh_cm	4.919.852	1.080.925	4.975.393	0.001637	0.011168	0.005434	0.013071	0.006936	0.006671	-0.001206
IH_ratio	-28.334.116	-5.540.933	-28.221.072	-0.037674	0.002474	0.001929	0.001841	0.001325	-0.001206	0.043362

Fonte: Elaborado pelo autor.

Tabela 9 – Covariância dos dados de corrida normal.

Variáveis	FY	FX	FR	Velocity	Sex	Age	Mass_kg	Waist_cm	Thigh_cm	IH_ratio
FY	1.000.000	0.514839	0.997468	0.666642	0.276436	0.049383	0.387035	0.257816	0.362049	-0.673473
FX	0.514839	1.000.000	0.570264	0.577833	0.319217	-0.093808	0.255758	0.141068	0.104717	-0.192633
FR	0.997468	0.570264	1.000.000	0.677928	0.296613	0.041418	0.395852	0.261884	0.357095	-0.655120
Velocity	0.666642	0.577833	0.677928	1.000.000	0.091392	-0.071702	0.019596	-0.019269	0.038247	-0.799638
Sex	0.276436	0.319217	0.296613	0.091392	1.000.000	0.348489	0.643061	0.480911	0.435141	0.037810
Age	0.049383	-0.093808	0.041418	-0.071702	0.348489	1.000.000	0.337843	0.365338	0.283625	0.039489
Mass_kg	0.387035	0.255758	0.395852	0.019596	0.643061	0.337843	1.000.000	0.874805	0.818892	0.045240
Weigth_N	0.387035	0.255758	0.395852	0.019596	0.643061	0.337843	1.000.000	0.874805	0.818892	0.045240
Waist_cm	0.257816	0.141068	0.261884	-0.019269	0.480911	0.365338	0.874805	1.000.000	0.745528	0.055880
Thigh_cm	0.362049	0.104717	0.357095	0.038247	0.435141	0.283625	0.818892	0.745528	1.000.000	-0.070929
IH_ratio	-0.673473	-0.192633	-0.655120	-0.799638	0.037810	0.039489	0.045240	0.055880	-0.070929	1.000.000

Fonte: Elaborado pelo autor.

Tabela 10 – Correlação dos dados de corrida normal.

Variáveis	FY	FX	FR	Velocity	Sex	Age	Mass_kg	Waist_cm	Thigh_cm	IH_ratio
FY	64.915.122.421	8.380.556.374	64.294.053.809	46.054.494	22.132.802	2.951.676	19.271.153	7.481.891	7.534.051	-35.731.315
FX	8.380.556.374	4.081.848.385	9.217.285.599	10.010.066	6.408.897	-1.406.006	3.193.319	1.026.565	0.546431	-2.562.801
FR	64.294.053.809	9.217.285.599	64.002.627.135	46.503.846	23.580.762	2.458.134	19.571.151	7.546.339	7.378.539	-34.512.419
Velocity	46.054.494	10.010.066	46.503.846	0.073521	0.007787	-0.004561	0.001038	-0.000595	0.000847	-0.045150
Sex	22.132.802	6.408.897	23.580.762	0.007787	0.098750	0.025691	0.039492	0.017213	0.011168	0.002474
Age	2.951.676	-1.406.006	2.458.134	-0.004561	0.025691	0.055035	0.015489	0.009762	0.005434	0.001929
Mass_kg	19.271.153	3.193.319	19.571.151	0.001038	0.039492	0.015489	0.038192	0.019473	0.013071	0.001841
Weigth_N	19.271.153	3.193.319	19.571.151	0.001038	0.039492	0.015489	0.038192	0.019473	0.013071	0.001841
Waist_cm	7.481.891	1.026.565	7.546.339	-0.000595	0.017213	0.009762	0.019473	0.012973	0.006936	0.001325
Thigh_cm	7.534.051	0.546431	7.378.539	0.000847	0.011168	0.005434	0.013071	0.006936	0.006671	-0.001206
IH_ratio	-35.731.315	-2.562.801	-34.512.419	-0.045150	0.002474	0.001929	0.001841	0.001325	-0.001206	0.043362

Fonte: Elaborado pelo autor.

4.2 RESULTADOS DO PRIMEIRO EXPERIMENTO (CAMINHADA NORMAL)

O primeiro experimento, que como citado previamente tem o objetivo de gerar um modelo preditivo para cada uma das variáveis FR, Fy e Fx (no exercício de caminhada normal), obteve-se um R^2 de 0,932 para FR, 0,920 para Fy, ou seja, uma explicação de 93,2%, 92,0% respectivamente da variabilidade dos dados. Esses resultados foram obtidos por meio do algoritmo SVR, tendo em vista que, como demonstra a Tabela 11 e Tabela 12, foi o algoritmo que teve melhor desempenho para a predição da FR e Fy. Já para Fx o R^2 foi de 0,804, ou seja, 80,4% de explicação da variabilidade dos dados (por meio do algoritmo Lasso Lars Regression) como apresenta a Tabela 13.

Os quadros também demonstram os valores de erros para cada um dos algoritmos testados no experimento em questão. As medidas de erros demonstradas nos quadros são: R^2 (Coeficiente de determinação), Mean absolute error (MAE), Erro quadrático médio (MSE) e Raiz quadrada do erro-médio (RMSE).

A Figura 30 demonstra que para o algoritmo SVR a variável mais importante na predição de FR é a variável IH_ratio, seguida por Mass_kg e velocity respectivamente. O mesmo ocorre para Fy como apresenta a Figura 31. Já no caso da variável Fx os atributos que mais influenciam no algoritmo Lasso Lars Regression são a Velocity e Mass_kg, como mostra a Figura 32.

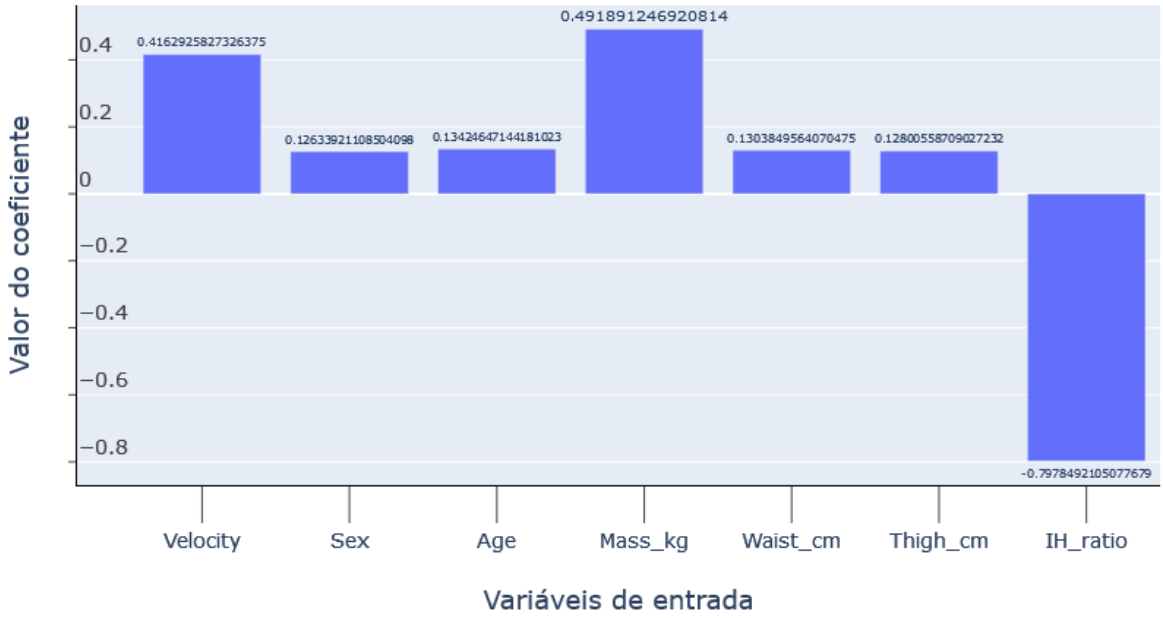
Tabela 11 - Resultados da predição de FR para caminhada normal.

Posição	**Algoritmo**	**R^2**	**MAE**	**MSE**	**RMSE**
1° Lugar	SVR	0,93	28,30	1541,61	37,22
2° Lugar	Gradient Boosting Regressor	0,92	30,09	1819,35	39,94
3° Lugar	Tweedie Regressor	0,91	32,61	1947,28	42,09
4° Lugar	Lasso	0,90	31,26	2227,33	45,11
5° Lugar	Ridge Cross-Validation	0,90	32,01	2290,73	46,12
6° Lugar	Random Forest Regressor	0,90	33,89	2316,38	45,92
7° Lugar	Bayesian Ridge Regression	0,90	32,09	2319,07	46,30
8° Lugar	ridge regression	0,90	32,89	2326,28	46,74

9° Lugar	Lasso Lars Regression	0,90	31,62	2352,25	46,20
10° Lugar	Linear Regression	0,89	32,45	2345,11	46,53
11° Lugar	AdaBoost regressor	0,85	42,11	3551,84	57,26
12° Lugar	K-Neighbors Regressor	0,82	42,73	4340,80	61,48
13° Lugar	Decision Tree Regression	0,80	48,96	4774,81	66,52
14° Lugar	Multi-layer Perceptron regressor	-5,58	353,41	150069	385,17

Fonte: Elaborado pelo autor.

Figura 30 – Importância das variáveis para o melhor modelo de FR para caminhada normal.



Fonte: Elaborado pelo autor.

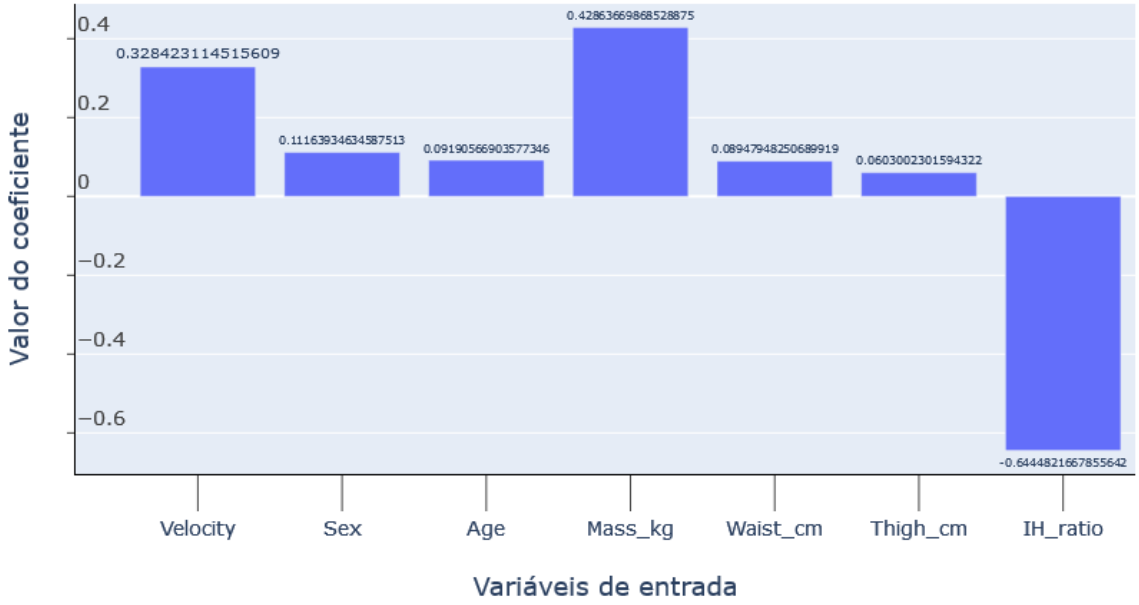
Tabela 12 - Resultados da predição de FY para caminhada normal.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	SVR	0,92	30,28	1827,70	40,90
2° Lugar	Gradient Boosting Regressor	0,91	31,68	2052,63	42,41
3° Lugar	Tweedie Regressor	0,91	33,17	1976,33	42,82
4° Lugar	Lasso	0,89	33,27	2445,65	47,40
5° Lugar	Random Forest Regressor	0,89	35,12	2443,02	47,00
6° Lugar	Ridge Cross-Validation	0,89	34,09	2494,39	48,28

7° Lugar	Lasso Lars Regression	0,89	33,90	2524,86	48,27
8° Lugar	Bayesian Ridge Regression	0,89	34,06	2518,11	48,43
9° Lugar	Linear Regression	0,89	34,42	2541,32	48,61
10° Lugar	ridge regression	0,88	35,34	2559,82	49,11
11° Lugar	AdaBoost regressor	0,84	42,76	3647,67	58,01
12° Lugar	K-Neighbors Regressor	0,79	45,96	4864,43	65,95
13° Lugar	Decision Tree Regression	0,75	52,08	5653,50	71,76
14° Lugar	Multi-layer Perceptron regressor	-5,28	343,12	142958	375,65

Fonte: Elaborado pelo autor.

Figura 31 – Importância das variáveis para o melhor modelo de Fy para caminhada normal.



Fonte: Elaborado pelo autor.

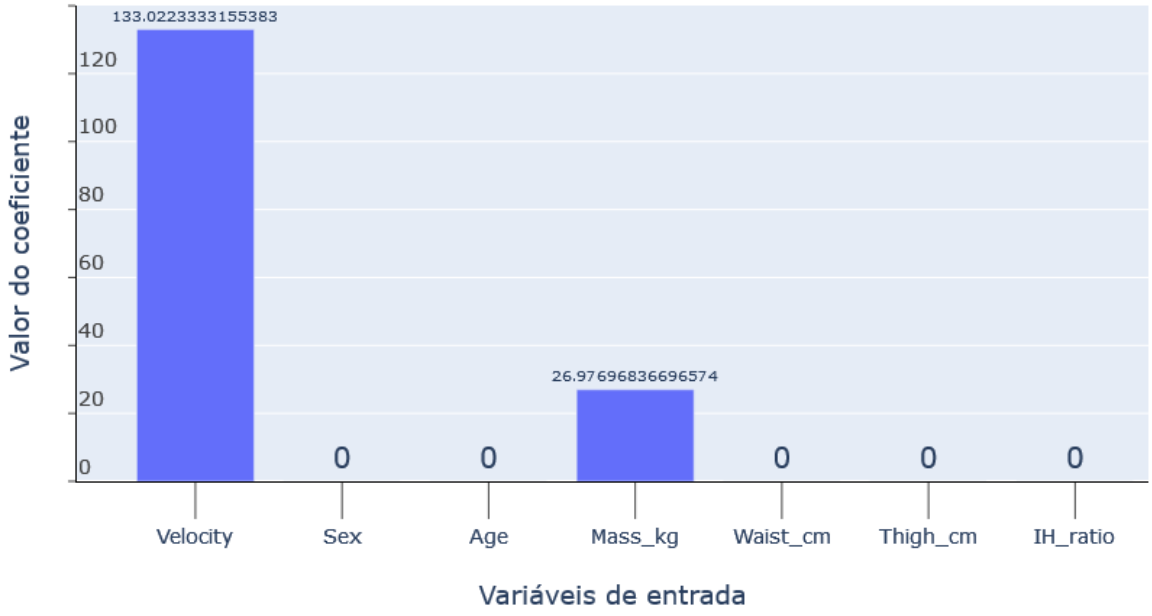
Tabela 13 – Resultados da predição de FX para caminhada normal.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	Lasso Lars Regression	0,80	14,64	339,34	18,05
2° Lugar	Ridge Cross-Validation	0,80	14,74	344,69	18,17
3° Lugar	Bayesian Ridge Regression	0,80	14,76	345,39	18,19
4° Lugar	ridge regression	0,79	14,85	348,94	18,35

5° Lugar	Linear Regression	0,79	14,85	357,94	18,49
6° Lugar	Lasso	0,78	15,48	378,95	19,17
7° Lugar	SVR	0,77	15,71	398,79	19,57
8° Lugar	Tweedie Regressor	0,76	16,10	404,42	19,65
9° Lugar	AdaBoost regressor	0,73	16,64	465,32	21,08
10° Lugar	Random Forest Regressor	0,73	16,37	447,40	20,85
11° Lugar	Gradient Boosting Regressor	0,73	17,32	468,28	21,22
12° Lugar	K-Neighbors Regressor	0,58	21,21	731,92	26,70
13° Lugar	Decision Tree Regression	0,53	21,88	799,15	27,83
14° Lugar	Multi-layer Perceptron regressor	-6,37	103,15	12544,20	111,50

Fonte: Elaborado pelo autor.

Figura 32 – Importância das variáveis para o melhor modelo de Fx para caminhada normal.



Fonte: Elaborado pelo autor.

4.3 RESULTADOS DO SEGUNDO EXPERIMENTO (CAMINHADA LENTA)

No segundo experimento, ensaio ao qual foi gerado um modelo preditivo para cada uma das forças de FRS no exercício de caminhada lenta, atingiu-se um R^2 de 0,906 para a variável FR por meio de um algoritmo denominado de Tweedie Regressor. Já para a Fy obteve-se um R^2 de 0,912 (utilizando o algoritmo Tweedie Regressor), enquanto para Fx foi de 0,643 com o algoritmo SVR. Ou seja, alcançou-se um explicação de 90,6% para FR, 91,2% para Fy e 64,3% para Fx. A Tabela 14, Tabela 15 e Tabela 16 demonstram os valores obtidos por cada um dos algoritmos testados.

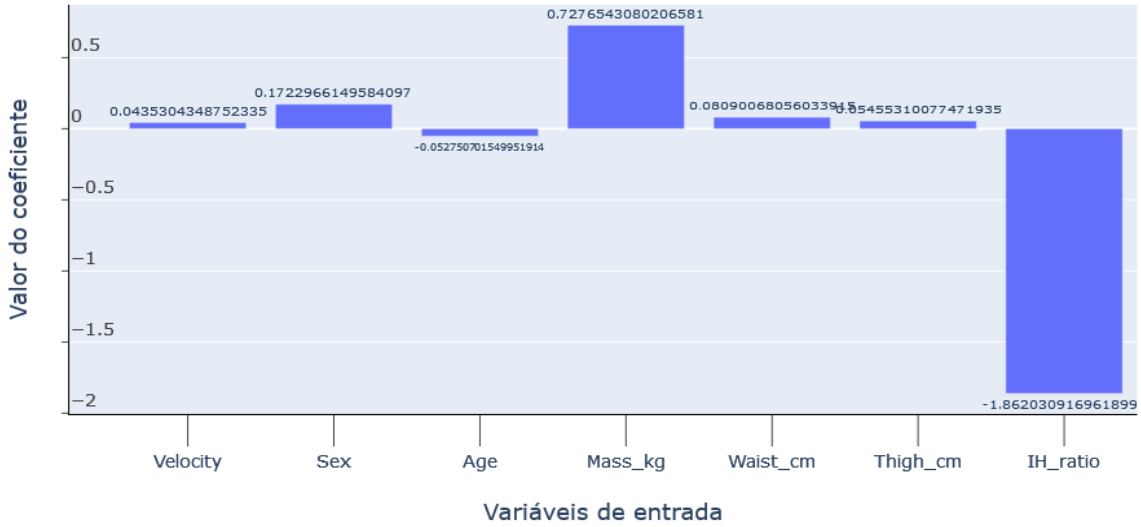
A Figura 33 demonstra que para o algoritmo Tweedie Regressor as variáveis mais importantes na predição de FR são respectivamente: IH_ratio, Mass_kg e Sex. O mesmo ocorre para Fy como apresenta a Figura 34. Já no caso da variável Fx os atributos que mais influenciam no algoritmo SVR é a Velocity e também IH_ratio como mostra a Figura 35.

Tabela 14 – Resultados da predição de FR para caminhada lenta.

Posição	**Algoritmo**	**R^2**	**MAE**	**MSE**	**RMSE**
1° Lugar	Tweedie Regressor	0,90	30,34	1966,32	41,35
2° Lugar	Random Forest Regressor	0,90	30,43	2139,61	43,92
3° Lugar	Gradient Boosting Regressor	0,90	29,52	2112,63	43,69
4° Lugar	Lasso	0,89	31,94	2171,94	44,94
5° Lugar	Ridge Cross-Validation	0,89	32,43	2274,50	45,74
6° Lugar	Bayesian Ridge	0,89	32,54	2294,26	45,69
7° Lugar	Linear Regression	0,89	32,73	2307,27	45,74
8° Lugar	Ridge	0,88	33,42	2414,96	47,82
9° Lugar	Lasso Lars	0,87	35,02	2644,27	48,81
10° Lugar	AdaBoost	0,85	39,50	3095,90	53,66
11° Lugar	KNN	0,77	49,82	4682,92	65,97
12° Lugar	Decison Tree	0,71	48,04	5504,06	71,81
13° Lugar	SVM	0,27	40,35	13254,30	78,57
14° Lugar	MLPRegressor	-5,39	329,12	131366	360,09

Fonte: Elaborado pelo autor.

Figura 33 – Importância das variáveis para o melhor modelo de FR para caminhada lenta.



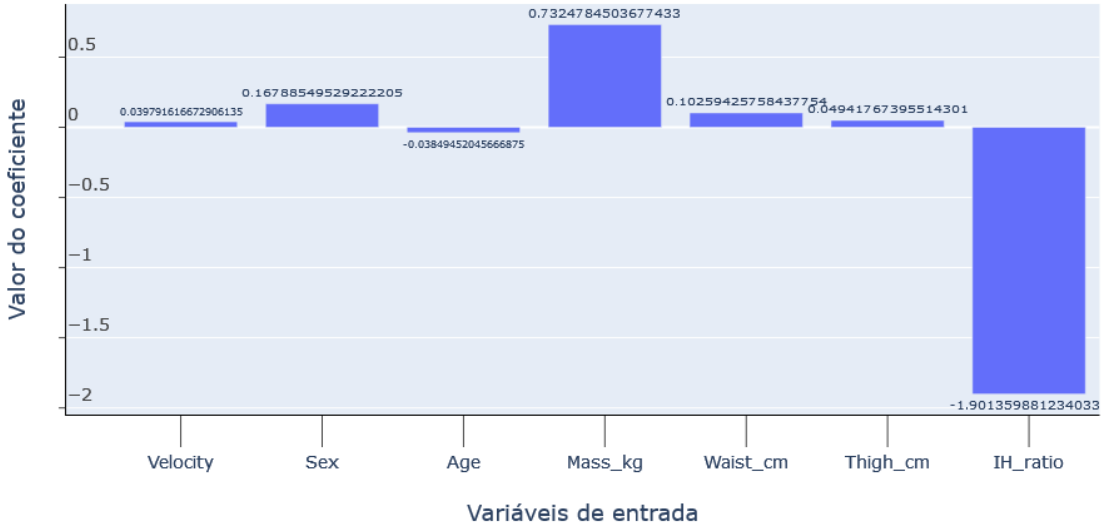
Fonte: Elaborado pelo autor.

Tabela 15 – Resultados da predição de Fy para caminhada lenta.

Posição	Algoritmo	R²	MAE	MSE	RMSE
1° Lugar	Tweedie Regressor	0,91	29,82	1843,10	40,20
2° Lugar	Lasso	0,90	31,18	2081,57	44,10
3° Lugar	Lasso Lars Regression	0,90	31,52	2105,20	44,00
4° Lugar	Random Forest Regressor	0,89	31,39	2187,07	44,60
5° Lugar	Ridge Cross-Validation	0,89	31,67	2172,64	44,84
6° Lugar	Bayesian Ridge Regression	0,89	32,03	2191,23	44,78
7° Lugar	Gradient Boosting Regressor	0,89	30,99	2216,58	45,12
8° Lugar	Linear Regression	0,89	32,31	2204,63	44,84
9° Lugar	ridge regression	0,88	32,37	2317,02	46,998
10° Lugar	AdaBoost regressor	0,86	39,82	2990,87	52,77
11° Lugar	K-Neighbors Regressor	0,77	51,23	4906,97	67,46
12° Lugar	Decision Tree Regression	0,70	50,39	5515,07	72,18
13° Lugar	SVR	0,39	39,38	10959,30	73,61
14° Lugar	Multi-layer Perceptron regressor	-5.21987	325.686	129349	357.273

Fonte: Elaborado pelo autor.

Figura 34 – Importância das variáveis para o melhor modelo de Fy para caminhada lenta.



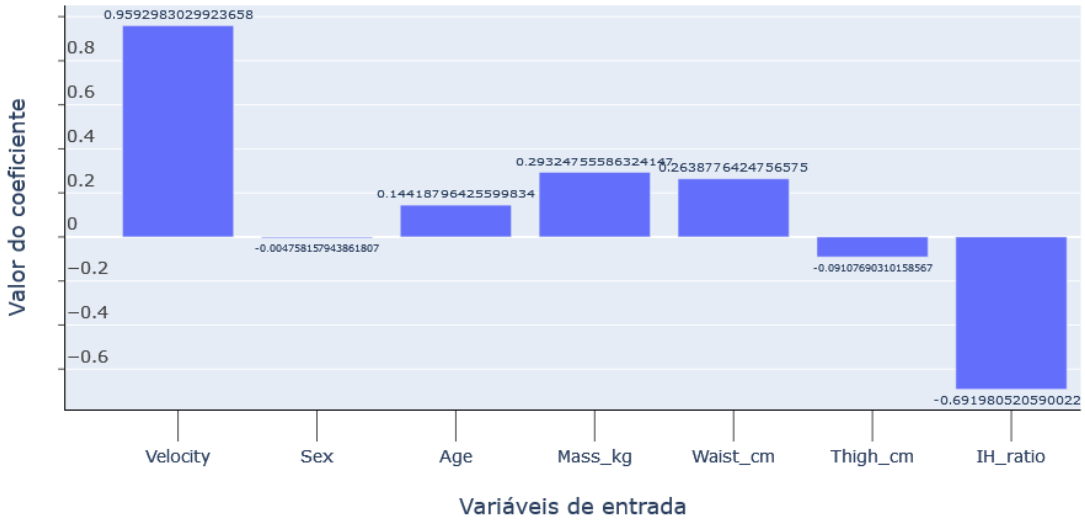
Fonte: Elaborado pelo autor.

Tabela 16 – Resultados da predição de Fx para caminhada lenta.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	SVR	0,64	12,68	260,97	15.79
2° Lugar	Random Forest Regressor	0,59	12,87	269,14	16.09
3° Lugar	AdaBoost regressor	0,59	12,93	279,56	16.36
4° Lugar	Gradient Boosting Regressor	0,58	13,58	287,49	16.57
5° Lugar	K-Neighbors Regressor	0,45	14,92	351,04	18.62
6° Lugar	Decision Tree Regression	0,44	15,38	375,98	19.10
7° Lugar	Lasso Lars Regression	0,30	17,05	511,28	21.91
8° Lugar	Linear Regression	0,28	16,93	534,58	22.20
9° Lugar	ridge regression	0,27	16,88	517,27	22.01
10° Lugar	Bayesian Ridge Regression	0,27	16,88	519,18	22.04
11° Lugar	Tweedie Regressor	0,24	17,73	568,23	22.84
12° Lugar	Ridge Cross-Validation	0,19	17,62	546,60	22.64
13° Lugar	Lasso	0,17	18,81	579,03	23.55
14° Lugar	Multi-layer Perceptron regressor	-6,38	65,11	5072,15	70.85

Fonte: Elaborado pelo autor.

Figura 35 – Importância das variáveis para o melhor modelo de Fx para caminhada lenta.



Fonte: Elaborado pelo autor.

4.4 RESULTADOS DO TERCEIRO EXPERIMENTO (CAMINHADA RÁPIDA)

Em relação aos resultados do terceiro experimento, o algoritmo que mais teve sucesso para a predição da FR e Fx no exercício de caminhada rápida foi o SVR. O mesmo obteve um R^2 de 0.8241 para FR e 0.6938 para Fx, enquanto que o R^2 alcançado para Fy foi de 0.8312 por meio do algoritmo Tweedie Regressor.

Dessa maneira, a precisão obtida para FR foi de 82,4% e 83,1% para Fy sendo que para Fx foi de apenas 69,3%. A Tabela 17, Tabela 18 e Tabela 19 demonstram a aplicação e resultados de todos os algoritmos testados.

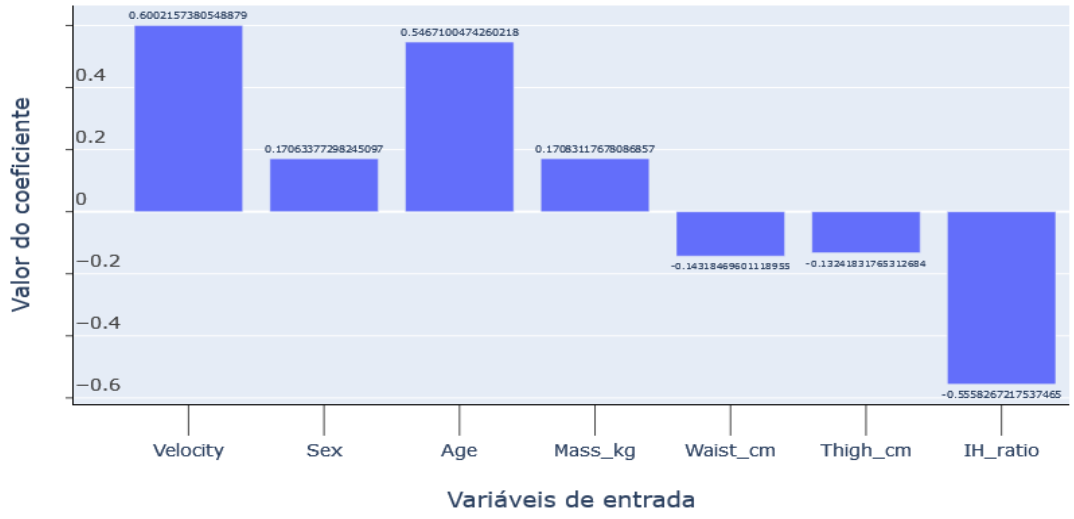
A Figura 36 demonstra que para o algoritmo SVR as variáveis mais importantes na predição de FR são Velocity, IH_ratio e Age . Já para Fy, os atributos que mais influenciam no uso do algoritmo Tweedie Regressor são IH_ratio, Thigh_cm e Velocity como apresenta a Figura 37. Para Fx os atributos que mais afetam o algoritmo SVR são Age e Thigh_cm como mostra a Figura 38.

Tabela 17 – Resultados da predição de FR para caminhada rápida.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	SVR	0,82	55,10	5115,25	70,16
2° Lugar	Tweedie Regressor	0,81	55,35	5256,04	70,48
3° Lugar	Lasso	0,81	57,32	5667,30	73,46
4° Lugar	ridge regression	0,81	58,31	5660,17	73,62
5° Lugar	Ridge Cross-Validation	0,81	59,05	5617,98	73,42
6° Lugar	Bayesian Ridge Regression	0,81	59,19	5628,06	73,49
7° Lugar	Lasso Lars Regression	0,81	59,52	5722,07	74,17
8° Lugar	Random Forest Regressor	0,80	56,47	5869,18	73,99
9° Lugar	Linear Regression	0,80	60,63	5844,54	75,01
10° Lugar	AdaBoost regressor	0,79	60,42	6139,72	76,39
11° Lugar	Gradient Boosting Regressor	0,77	62,72	6466,83	78,89
12° Lugar	K-Neighbors Regressor	0,73	67,03	7484,26	85,33
13° Lugar	Decision Tree Regression	0,55	84,22	12812,60	110,62
14° Lugar	Multi-layer Perceptron regressor	-6,62	433,21	222444	468,87

Fonte: Elaborado pelo autor.

Figura 36 – Importância das variáveis para o melhor modelo de FR para caminhada rápida.



Fonte: Elaborado pelo autor.

Tabela 18 – Resultados da predição de Fy para caminhada rápida.

Posição	Algoritmo	R²	MAE	MSE	RMSE
1° Lugar	Tweedie Regressor	0,83	51,94	4499,41	65,74
2° Lugar	SVR	0,82	54.94	4845,17	68,72
3° Lugar	Random Forest Regressor	0,81	54,00	5168,24	69,68
4° Lugar	Lasso Lars Regression	0,81	56,66	5089,63	70,35
5° Lugar	Gradient Boosting Regressor	0,81	56,37	5334,38	71,56
6° Lugar	Ridge Cross-Validation	0,80	57,73	5290,05	71,93
7° Lugar	Lasso	0,80	55,97	5336,85	72,08
8° Lugar	Bayesian Ridge Regression	0,80	57,85	5300,93	72,00
9° Lugar	ridge regression	0,80	57,00	5343,74	72,22
10° Lugar	Linear Regression	0,79	59,20	5506,78	73,43
11° Lugar	AdaBoost regressor	0,79	59,02	5800,82	73,89
12° Lugar	K-Neighbors Regressor	0,74	64,54	6719,98	81,32
13° Lugar	Decision Tree Regression	0,59	75,02	10733,60	101,57
14° Lugar	Multi-layer Perceptron regressor	-5,90	399,28	191897	435,06

Fonte: Elaborado pelo autor.

Figura 37 – Importância das variáveis para o melhor modelo de Fy para caminhada rápida.

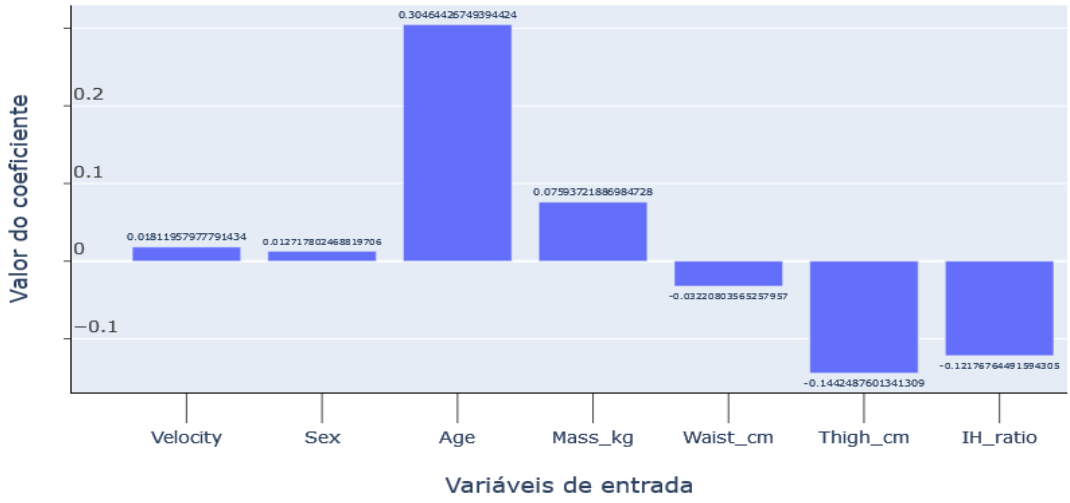
Fonte: Elaborado pelo autor.

Tabela 19 – Resultados da predição de Fx para caminhada rápida.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	SVR	0,69	25,99	1363,83	35,66
2° Lugar	Lasso Lars Regression	0,68	26,81	1391,40	36,06
3° Lugar	Lasso	0,68	26,96	1435,85	36,56
4° Lugar	ridge regression	0,68	26,97	1415,72	36,37
5° Lugar	Bayesian Ridge Regression	0,68	27,07	1414,77	36,37
6° Lugar	Ridge Cross-Validation	0,68	27,12	1413,19	36,36
7° Lugar	Linear Regression	0,67	27,67	1447,45	36,82
8° Lugar	Random Forest Regressor	0,63	28,92	1570,71	38,75
9° Lugar	AdaBoost regressor	0,63	29,53	1639	39,79
10° Lugar	Tweedie Regressor	0,61	28,60	1606,97	38,80
11° Lugar	Gradient Boosting Regressor	0,56	31,26	1881,34	42,21
12° Lugar	K-Neighbors Regressor	0,47	33,88	2265,49	46,59
13° Lugar	Decision Tree Regression	0,14	43,77	3475,81	58,43
14° Lugar	Multi-layer Perceptron regressor	-7,69	179,19	37213,70	192,16

Fonte: Elaborado pelo autor

Figura 38 – Importância das variáveis para o melhor modelo de Fx para caminhada rápida.



Fonte: Elaborado pelo autor.

4.5 RESULTADOS DO QUARTO EXPERIMENTO (CORRIDA NORMAL)

O quarto experimento, que como citado anteriormente tem a meta de gerar um modelo preditivo para cada uma das variáveis FR, Fy e Fx (no exercício de corrida normal), obteve-se um R^2 de 0.609 para FR, 0.608 para FY e 0.526 para Fx, ou seja, uma explicação de 60,9%, 60,8% e 52,6% respectivamente da variabilidade dos dados. Esses resultados foram obtidos por meio do algoritmo Ridge Cross-Validation para FR e por meio do Tweedie Regressor para Fy. Já o resultado obtido para Fx foi atingido por meio do algoritmo SVR.

A Tabela 20, Tabela 21 e Tabela 22 apresentam os valores de erros para cada um dos algoritmos testados no experimento em questão. As medidas de erros demonstradas nos quadros são: R^2 (Coeficiente de determinação), Mean absolute error (MAE), Erro quadrático médio (MSE) e Raiz quadrada do erro-médio (RMSE).

A Figura 39 demonstra que para o algoritmo Ridge Cross-Validation as variáveis mais importantes na predição de FR é são respectivamente: Velocity, Mass_kg e Age. Já para o algoritmo Tweedie Regressor na predição de Fy os atributos que mais influenciam são Mass_kg, IH_ratio e Waist_cm como apresenta a Figura 40. Já no caso da variável Fx os atributos que mais influenciam no algoritmo SVR é são Velocity e a Mass_kg , como mostra a Figura 41.

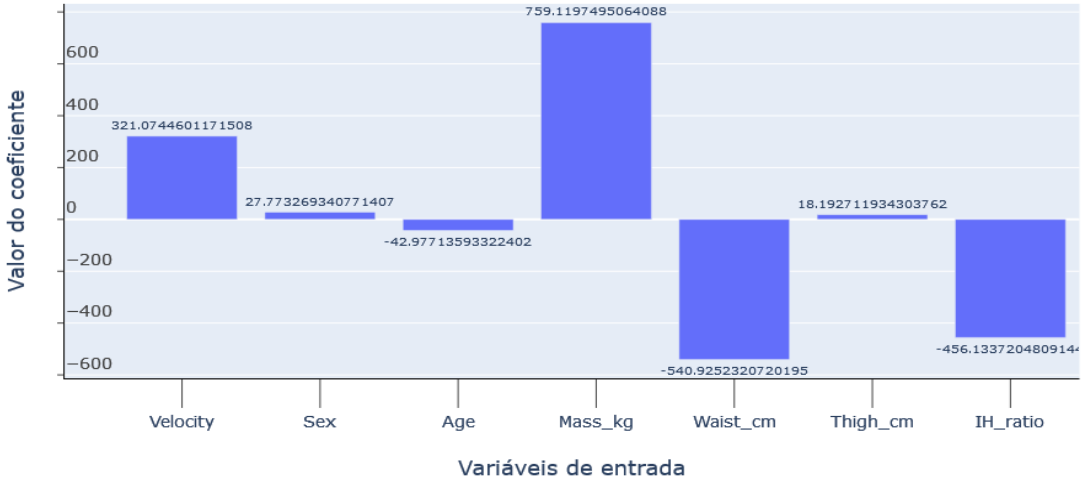
Tabela 20 – Resultados da predição de FR para corrida normal.

Posição	**Algoritmo**	**R^2**	**MAE**	**MSE**	**RMSE**
1° Lugar	Ridge Cross-Validation	0,60	119,84	22731,20	148,96
2° Lugar	Lasso Lars Regression	0,60	119,95	22771,30	149,09
3° Lugar	ridge regression	0,60	119,85	228520	149,25
4° Lugar	Bayesian Ridge Regression	0,60	120,11	22858,50	149,38
5° Lugar	Tweedie Regressor	0,60	119,18	22835,20	149,39
6° Lugar	Lasso	0,60	120,85	23300,10	150,63
7° Lugar	Linear Regression	0,59	121,84	23185,40	150,64
8° Lugar	SVR	0,59	120,03	23308,90	150,35
9° Lugar	Random Forest Regressor	0,57	125,00	25419,70	155,42

10° Lugar	Gradient Boosting Regressor	0,55	128,80	26609,60	159,59
11° Lugar	AdaBoost regressor	0,54	132,67	27322,60	162,01
12° Lugar	K-Neighbors Regressor	0,43	146,12	33170,70	179,84
13° Lugar	Decision Tree Regression	0,25	159,60	43295,10	205,79
14° Lugar	Multi-layer Perceptronregressor	-6,82	608,54	433871	656,10

Fonte: Elaborado pelo autor.

Figura 39 – Importância das variáveis para o melhor modelo de FR para corrida normal.



Fonte: Elaborado pelo autor.

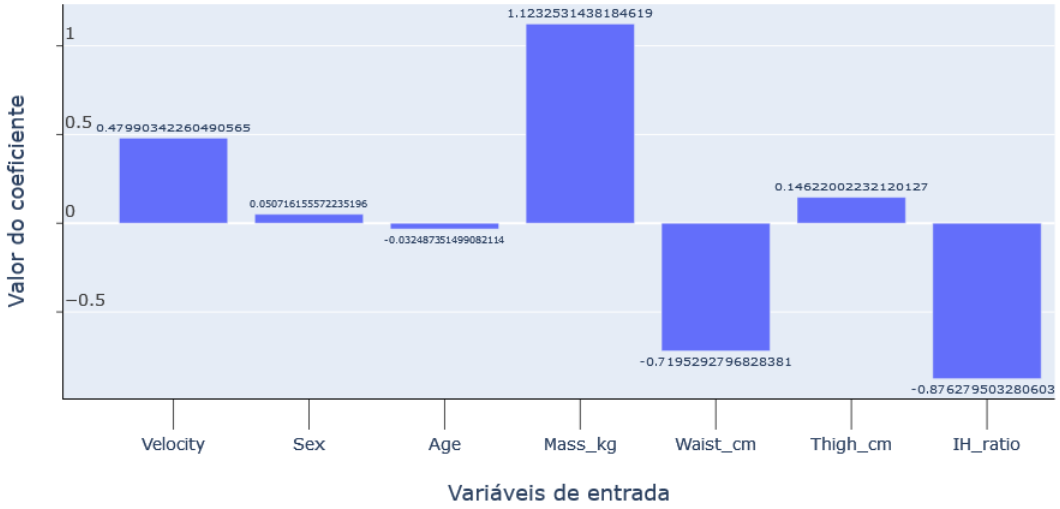
Tabela 21 – Resultados da predição de Fy para corrida normal.

Posição	**Algoritmo**	**R²**	**MAE**	**MSE**	**RMSE**
1° Lugar	Tweedie Regressor	0,60	118,34	22776,20	149,04
2° Lugar	Lasso Lars Regression	0,60	120,00	23119,30	150,26
3° Lugar	Ridge Cross-Validation	0,60	120,15	23115,50	150,27
4° Lugar	ridge regression	0,60	120,61	23266,90	150,71
5° Lugar	Bayesian Ridge Regression	0,60	120,59	23250,50	150,73
6° Lugar	Random Forest Regressor	0,60	122,34	23849,40	151,09
7° Lugar	Lasso	0,60	121,56	23605,90	151,72
8° Lugar	Linear Regression	0,59	122,04	23592,10	152,0
9° Lugar	SVR	0,58	123,09	24275,80	153,35

10° Lugar	AdaBoost regressor	0,54	132,57	27867,80	163,39
11° Lugar	Gradient Boosting Regressor	0,53	132,66	27394,80	162,55
12° Lugar	K-Neighbors Regressor	0,43	146,45	33349,40	180,39
13° Lugar	Decision Tree Regression	0,25	165,47	44480,70	207,73
14° Lugar	Multi-layer Perceptron regressor	-6,11	582,65	403973	632,77

Fonte: Elaborado pelo autor.

Figura 40 – Importância das variáveis para o melhor modelo de Fy para corrida normal.



Fonte: Elaborado pelo autor.

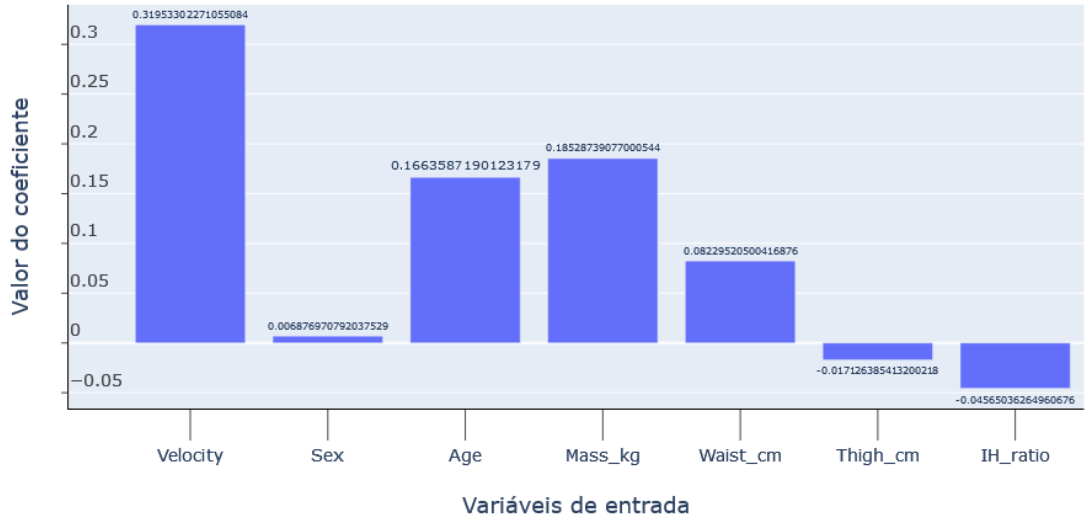
Tabela 22 – Resultados da predição de Fx para corrida normal.

Posição	**Algoritmo**	**R²**	**MAE**	**MSE**	**RMSE**
1° Lugar	SVR	0,52	30,40	1694,41	40,36
2° Lugar	Lasso Lars Regression	0,50	31,99	1838,90	41,99
3° Lugar	Ridge Cross-Validation	0,48	32,25	1831,59	42,11
4° Lugar	Bayesian Ridge Regression	0,48	32,29	1835,48	42,15
5° Lugar	ridge regression	0,47	32,99	1891,50	42,74
6° Lugar	Tweedie Regressor	0,46	32,90	1905,36	42,98
7° Lugar	Linear Regression	0,46	32,48	1872,05	42,62
8° Lugar	Lasso	0,43	33,56	2066,69	44,42
9° Lugar	AdaBoost regressor	0,34	38,26	2300,42	47,08
10° Lugar	Random Forest Regressor	0,30	38,22	2390,41	48,12

11° Lugar	Gradient Boosting Regressor	0,24	40,20	2500,12	49,58
12° Lugar	K-Neighbors Regressor	-0,05	44,32	3415,35	57,23
13° Lugar	Decision Tree Regression	-0,41	49,22	4431,12	66,04
14° Lugar	Multi-layer Perceptron regressor	-9,49	175,75	34957,70	185,99

Fonte: Elaborado pelo autor.

Figura 41 – Importância das variáveis para o melhor modelo de Fx para corrida normal.



Fonte: Elaborado pelo autor.

4.6 RESULTADOS DO QUINTO EXPERIMENTO (CORRIDA LENTA)

No quinto experimento, foi gerado um modelo preditivo para cada uma das forças de FRS no exercício de corrida lenta, obteve-se um R^2 de 0,595 para a variável FR por meio de um algoritmo denominado de Tweedie Regressor. Já para a FY atingiu-se um R^2 de 0,607 (utilizando o algoritmo Random Forest Regressor), enquanto para Fx foi de 0.4552 usando o algoritmo Bayesian Ridge Regression. Ou seja, alcançou-se uma explicação de 59,5% para FR, 60,7% para Fy e 45,5% para Fx.

A Tabela 23, Tabela 24 e Tabela 25 demonstram os valores obtidos por cada um dos algoritmos testados. A Figura 42 demonstra que para o algoritmo Tweedie Regressor as variáveis mais importantes na predição de FR são os atributos IH_ratio, Velocity e Mass_kg. Já para o algoritmo Random Forest Regressor na predição de Fy a ordem dos atributos que mais influenciam é respectivamente: IH_ratio, Mass_kg e Velocity como apresenta a Figura 43. Já

no caso da variável Fx os atributos que mais influenciam no algoritmo Bayesian Ridge Regression é a Velocity e IH_ratio, como mostra a Figura 44.

Tabela 23 - Resultados da predição de FR para corrida lenta.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	Tweedie Regressor	0,59	116,31	22636,60	147,489
2° Lugar	Ridge Cross-Validation	0,59	119,11	22996,30	149,55
3° Lugar	Random Forest Regressor	0,58	118,05	23383,30	150,259
4° Lugar	Bayesian Ridge Regression	0,58	119,80	23228,40	150,287
5° Lugar	Linear Regression	0,58	118,84	23111,10	149,857
6° Lugar	ridge regression	0,58	121,63	23530	151,129
7° Lugar	Gradient Boosting Regressor	0,57	117,53	23679,20	149,998
8° Lugar	Lasso	0,57	122,08	23930,10	152,475
9° Lugar	SVR	0,56	116,12	24172,80	152,449
10° Lugar	Lasso Lars Regression	0,56	122,22	24110	153,339
11° Lugar	AdaBoost regressor	0,54	130,58	26322,40	159,912
12° Lugar	K-Neighbors Regressor	0,50	134,52	29123,10	167,164
13° Lugar	Decision Tree Regression	0,41	137,27	33365,60	175,002
14° Lugar	Multi-layer Perceptron regressor	-5,24	544,67	359608	596,486

Fonte: Elaborado pelo autor.

Figura 42 – Importância das variáveis para o melhor modelo de FR para corrida lenta.

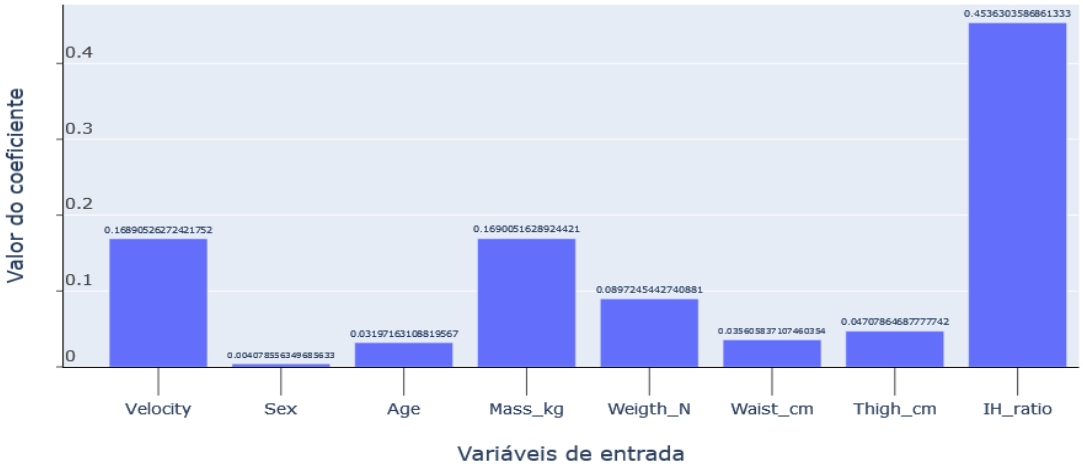
Fonte: Elaborado pelo autor.

Tabela 24 – Resultados da predição de Fy para corrida lenta.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	Random Forest Regressor	0,60	117,24	22982,90	149,11
2° Lugar	Gradient Boosting Regressor	0,60	114,16	22845,70	148,31
3° Lugar	Tweedie Regressor	0,59	117,04	23180,80	149,09
4° Lugar	Lasso Lars Regression	0,59	120,02	23512	151,32
5° Lugar	Ridge Cross-Validation	0,59	120,34	23654,80	151,62
6° Lugar	Bayesian Ridge Regression	0,58	120,93	23869,80	152,30
7° Lugar	Linear Regression	0,58	120,10	23795,40	151,99
8° Lugar	ridge regression	0,58	122,70	24175,30	153,17
9° Lugar	Lasso	0,57	122,96	24517,60	154,34
10° Lugar	SVR	0,57	115,74	24553,90	153,39
11° Lugar	AdaBoost regressor	0,54	130,58	26582,70	160,88
12° Lugar	K-Neighbors Regressor	0,50	136,52	29779,80	169,20
13° Lugar	Decision Tree Regression	0,36	143,26	36881,70	184,73
14° Lugar	Multi-layer Perceptron regressor	-4,90	533,78	349372	587,68

Fonte: Elaborado pelo autor.

Figura 43 – Importância das variáveis para o melhor modelo de Fy para corrida lenta.



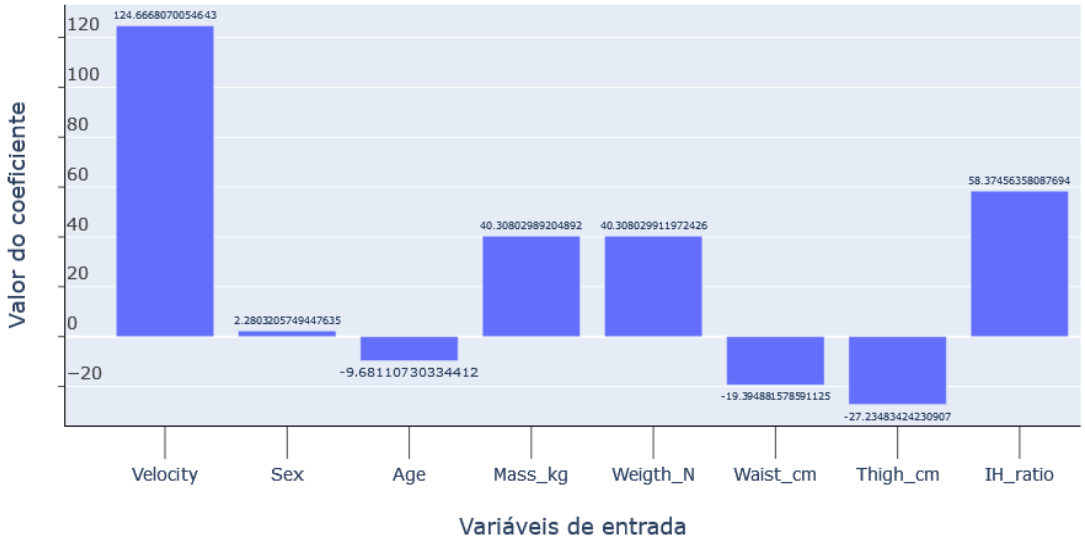
Fonte: Elaborado pelo autor.

Tabela 25 – Resultados da predição de Fx para corrida lenta.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	Bayesian Ridge Regression	0,45	22,28	854,67	28,11
2° Lugar	Ridge Cross-Validation	0,45	22,25	851,48	28,06
3° Lugar	ridge regression	0,45	22,38	866,23	28,30
4° Lugar	SVR	0,44	22,10	863,49	28,31
5° Lugar	Lasso Lars Regression	0,44	22,69	889,43	28,62
6° Lugar	Linear Regression	0,43	22,44	866,02	28,34
7° Lugar	Tweedie Regressor	0,43	22,61	870,33	28,63
8° Lugar	Lasso	0,36	23,55	1016,40	30,67
9° Lugar	Gradient Boosting Regressor	0,34	25,14	1056,68	31,26
10° Lugar	Random Forest Regressor	0,34	24,61	1082,46	31,51
11° Lugar	AdaBoost regressor	0,21	26,48	1265,04	34,02
12° Lugar	K-Neighbors Regressor	0,03	28,50	1462,78	37,38
13° Lugar	Decision Tree Regression	-0,41	33,94	2168,02	44,09
14° Lugar	Multi-layer Perceptronregressor	-10,59	117,99	15720,10	124,91

Fonte: Elaborado pelo autor.

Figura 44 – Importância das variáveis para o melhor modelo de Fx para corrida lenta.



Fonte: Elaborado pelo autor.

`

4.7 RESULTADOS DO SEXTO EXPERIMENTO (CORRIDA RÁPIDA)

Em relação aos resultados do sexto experimento, o algoritmo que mais teve sucesso para a predição da FR e Fy no exercício de corrida rápida foi o Linear Regression, o mesmo obteve um R^2 de 0,581 para FR e um R^2 de 0,561 para Fy. O R^2 obtido em Fx foi de apenas 0,560 usando ridge regression.

Dessa maneira, a precisão obtida para FR foi de 58,1%, enquanto para Fy foi de 56,1%, sendo que para Fx foi de 56,0%. A Tabela 26, Tabela 27 e a Tabela 28 demonstram a aplicação e resultados de todos os algoritmos testados.

A Figura 46 demonstra que para o algoritmo Linear Regression as variáveis mais importantes na predição de FR são os atributos Mass_kg, Waist_cm e IH_ratio. Para Fy as variáveis mais significativas são respectivamente: IH_ratio, Waist_cm e Mass_kg como apresenta a Figura 47. Já no caso da variável Fx os atributos que mais influenciam no algoritmo ridge regression são Velocity e Mass_kg como mostra a Figura 48.

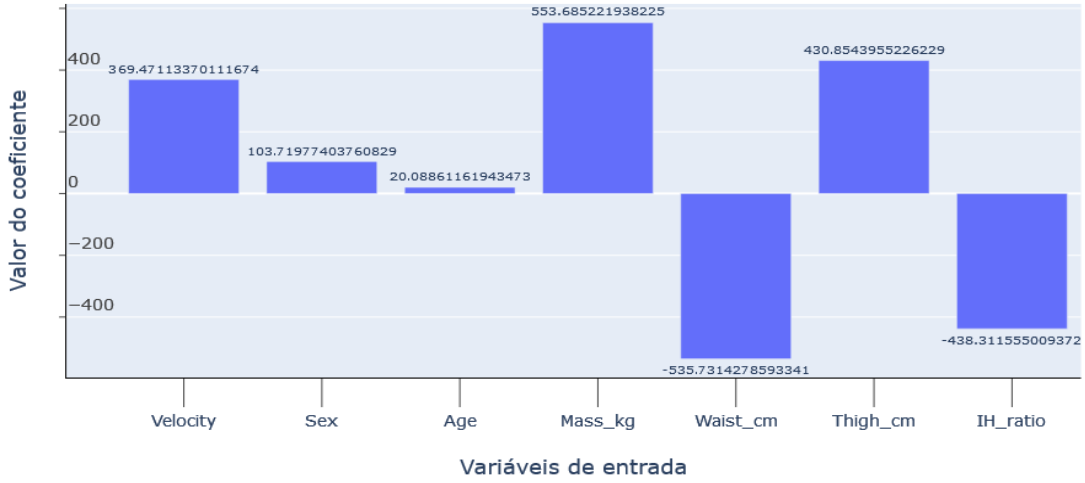
Tabela 26 – Resultados da predição de FR para corrida rápida.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	Linear Regression	0,58	129,20	29048,60	166,02
2° Lugar	ridge regression	0,56	132,36	30043,40	168,79
3° Lugar	Bayesian Ridge Regression	0,56	132,35	30240	169,22
4° Lugar	Lasso	0,56	132,40	30139,20	169,12
5° Lugar	Ridge Cross-Validation	0,56	133,58	30578,50	170,08
6° Lugar	Random Forest Regressor	0,55	129,70	30033,40	170,84
7° Lugar	Lasso Lars Regression	0,55	132,59	30840,10	170,51
8° Lugar	SVR	0,54	139,13	32386,20	174,58
9° Lugar	Tweedie Regressor	0,52	136,48	32694,60	175,49
10° Lugar	Gradient Boosting Regressor	0,50	137,04	34582,50	183,74
11° Lugar	AdaBoost regressor	0,48	137,20	35883,10	184,88
12° Lugar	K-Neighbors Regressor	0,44	156,95	38997,90	195,60

13° Lugar	Decision Tree Regression	0,20	180,04	55417,80	233,05
14° Lugar	Multi-layer Perceptron regressor	-9,25	793,91	708771	838,47

Fonte: Elaborado pelo autor.

Figura 45 – Importância das variáveis para o melhor modelo de FR para corrida rápida.



Fonte: Elaborado pelo autor.

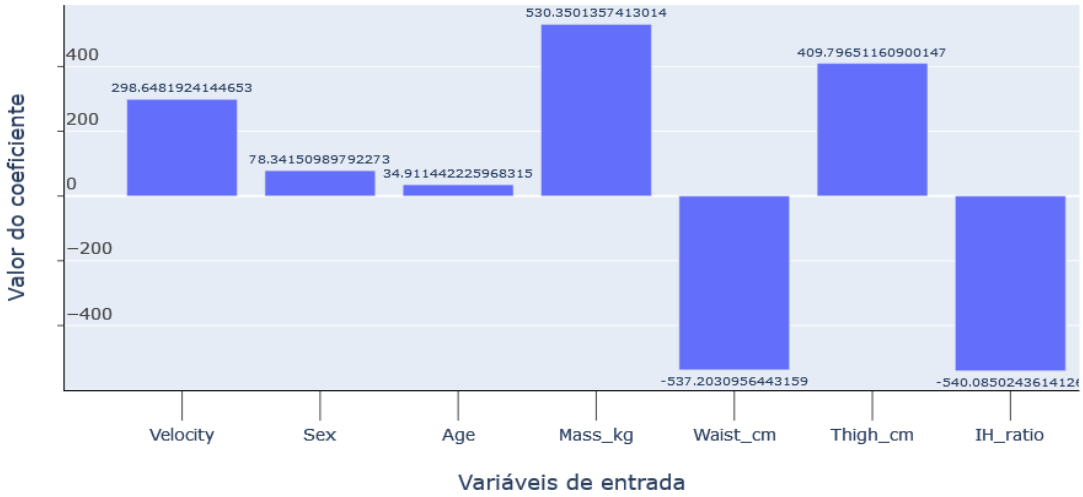
Tabela 27 – Resultados da predição de Fy para corrida rápida.

Posição	**Algoritmo**	**R^2**	**MAE**	**MSE**	**RMSE**
1° Lugar	Linear Regression	0,56	130,47	29447,90	167,31
2° Lugar	ridge regression	0,54	134,01	30257,80	169,59
3° Lugar	Random Forest Regressor	0,54	132,06	29998,30	170,71
4° Lugar	Bayesian Ridge Regression	0,54	134,17	30494,40	170,09
5° Lugar	Ridge Cross-Validation	0,54	135,05	30931,70	171,13
6° Lugar	Lasso	0,54	134,53	30577,10	170,58
7° Lugar	Lasso Lars Regression	0,53	134,45	31158,60	171,54
8° Lugar	Tweedie Regressor	0,51	137,30	32269,40	174,14
9° Lugar	SVR	0,50	140,68	34035,10	178,64
10° Lugar	AdaBoost regressor	0,46	139,43	35736,50	184,88
11° Lugar	Gradient Boosting Regressor	0,44	142,35	37131,20	190,08
12° Lugar	K-Neighbors Regressor	0,39	159,91	40682,80	199,88
13° Lugar	Decision Tree Regression	0,22	164,04	51048,10	222,72

14° Lugar	Multi-layer Perceptron regressor	-8,32	735,96	616794	781,70

Fonte: Elaborado pelo autor.

Figura 46 – Importância das variáveis para o melhor modelo de Fy para corrida rápida.



Fonte: Elaborado pelo autor.

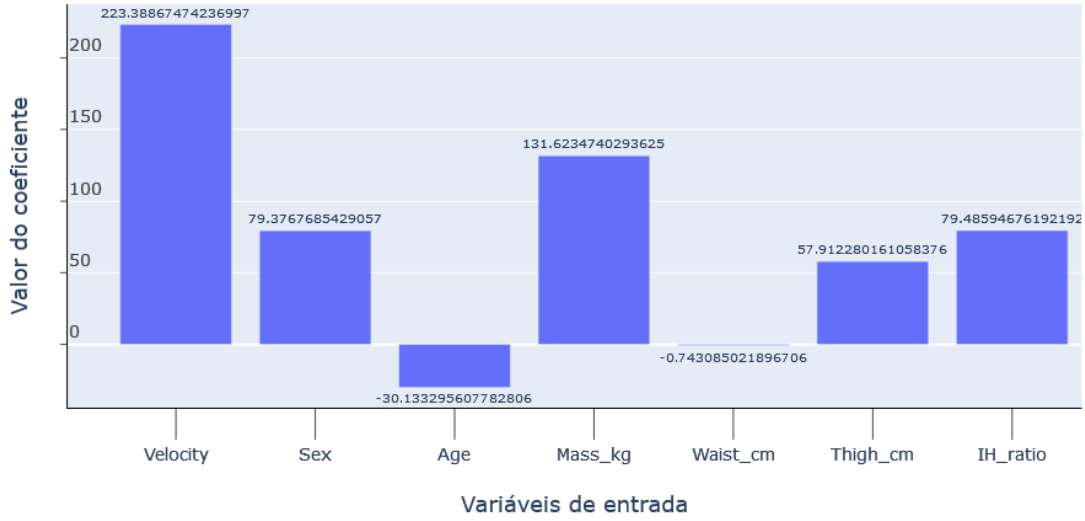
Tabela 28 – Resultados da predição de Fx para corrida rápida.

Posição	Algoritmo	R^2	MAE	MSE	RMSE
1° Lugar	ridge regression	0,56	47,91	3801,68	61,20
2° Lugar	Bayesian Ridge Regression	0,55	47,91	3790,40	61,16
3° Lugar	Ridge Cross-Validation	0,55	47,83	3784,99	61,13
4° Lugar	Lasso Lars Regression	0,55	47,75	3799,89	61,22
5° Lugar	Linear Regression	0,55	47,79	3817,06	61,33
6° Lugar	Lasso	0,54	49,28	3976,77	62,53
7° Lugar	Tweedie Regressor	0,51	50,62	4185,89	64,22
8° Lugar	SVR	0,47	50,89	4543,42	66,86
9° Lugar	Random Forest Regressor	0,45	52,69	4603,85	67,21
10° Lugar	K-Neighbors Regressor	0,43	55,51	4946,19	69
11° Lugar	AdaBoost regressor	0,42	54,80	4839,88	69,02
12° Lugar	Gradient Boosting Regressor	0,35	55,81	5351,71	71,99

13° Lugar	Decision Tree Regression	0,00	68	8297,13	88,20
14° Lugar	Multi-layer Perceptron regressor	-11,00	304,67	102401	319,01

Fonte: Elaborado pelo autor.

Figura 47 – Importância das variáveis para o melhor modelo de Fx para corrida rápida.



Fonte: Elaborado pelo autor.

4.8 DISCUSSÃO DOS RESULTADOS

Comparando todos os experimentos, nota-se que os algoritmos utilizados possuem uma predição melhor para os exercícios de caminhada do que de corrida. No caso dos exercícios de caminhada o algoritmo que obteve mais resultados positivos foi o SVM. Já para exercícios de corrida os algoritmos que melhor desempenharam foram sucesso o Tweedie Regressor e o Linear Regression. A predição mais efetiva foi alcançada para o exercício de caminhada normal (primeiro experimento). A Tabela 29 apresenta os melhores resultados obtidos em cada experimento.

Tabela 29 – Comparativo da predição dos 6 experimentos.

Experimento	**Algoritmo**	**R^2**	**explicação**
Primeiro Experimento - Caminhada Normal - FR	SVR	0,93	93,2%
Primeiro Experimento - Caminhada Normal - FY	SVR	0,92	92,0%

Primeiro Experimento - Caminhada Normal - FX	Lasso LarsRegression	0,80	80,4%
Segundo Experimento - Caminhada Lenta - FR	Tweedie Regressor	0,90	90,6%
Segundo Experimento - Caminhada Lenta - FY	Tweedie Regressor	0,91	91,2%
Segundo Experimento - Caminhada Lenta - FX	SVR	0,64	64,3%
Terceiro Experimento - Caminhada Rápida - FR	SVR	0,82	82,4%
Terceiro Experimento - Caminhada Rápida - FY	Tweedie Regressor	0,83	83,1%
Terceiro Experimento - Caminhada Rápida - FX	SVR	0,69	69,8%
Quarto Experimento - Corrida Normal - FR	Ridge Cross-Validation	0,60	60,9%
Quarto Experimento - Corrida Normal- FY	Tweedie Regressor	0,60	60,8%
Quarto Experimento - Corrida Normal - FX	SVR	0,52	52,6%
Quinto Experimento - Corrida Lenta - FR	Tweedie Regressor	0,59	59,5%
Quinto Experimento - Corrida Lenta - FY	Random ForestRegressor	0,60	60,7%
Quinto Experimento - Corrida Lenta - FX	Bayesian RidgeRegression	0,45	45,5%
Sexto Experimento - Corrida Rápida - FR	Linear Regression	0,58	58,1%
Sexto Experimento - Corrida Rápida - FY	Linear Regression	0,56	56,1%
Sexto Experimento - Corrida Rápida - FX	ridge regression	0,56	56,0%

Fonte: Elaborado pelo autor.

4.9 COMPARAÇÃO COM OS TRABALHOS RELACIONADOS

São poucos os estudos realizados até hoje que tem o objetivo de prever a FRS em exercícios realizados na água. Porém, como dito anteriormente, Haupenthal (2013) realizou um estudo que tinha como principal objetivo: elaborar, avaliar e validar um modelo de regressão para o pico das componentes vertical (Fy), ântero-posterior (Fx) e da resultante (FR) da FRS durante a caminhada e corrida na água. Esta dissertação tem grande relação com o trabalho de Haupenthal (2013), considerando que os mesmos têm objetivos bem semelhantes e utilizaram os mesmos dados.

No entanto, como pode-se notar no capítulo 2.7 (Trabalhos Relacionados), Haupenthal (2013) utilizou uma metodologia diferente da empregada nesse trabalho. O mesmo gerou um modelo de predição para cada exercício (caminhada e corrida) e também gerou um modelo para ambos (juntos), enquanto a metodologia adotada nesta dissertação foi a geração de modelos para predição da FRS também em exercícios de caminhada e corrida, porém fazendo a distinção de suas faixas de velocidade.

Sendo assim nesta dissertação foram gerados modelos para caminhada normal, caminhada lenta e caminhada rápida. Tratando-se da corrida, foram gerados modelos para corrida normal, corrida lenta e corrida rápida. Portanto, Haupenthal (2013) realizou 3 experimentos, enquanto este trabalho efetuou 6 experimentos.

Baseado neste cenário, nota-se que os resultados de Haupenthal (2013) são mais generalizados e os desse estudo mais específicos, o que acaba dificultando a comparação. Mas mesmo assim é possível comparar o R^2 dos experimentos considerando os exercícios.

Comparando os resultados de Haupenthal (2013) para caminhada, com o primeiro experimento (caminhada normal) e também com o segundo experimento (caminhada lenta), realizados nesta dissertação nota-se que o R^2 obtido neste estudo foi superior para FR e Fy, enquanto o resultado de Haupenthal (2013) foi melhor para Fx. No tocante ao terceiro experimento (caminhada rápida), o estudo de Haupenthal (2013) alcançou valores de R^2 superiores para FR, Fy e Fx.

Em relação aos demais experimentos: quarto experimento (corrida normal), quinto experimento (corrida lenta) e sexto experimento (corrida rápida), e comparando os resultados dos mesmos com a predição para a corrida de Haupenthal (2013), os resultados de Haupenthal (2013) são superiores para FR, Fy e Fx.

A Tabela 30 e a Tabela 31 deixam de forma mais clara essas comparações.

Tabela 30 – Resultados das predições de Haupenthal (2013).

Experimento	**R^2**
Primeiro Experimento - FR - Caminhada	0,88
Primeiro Experimento - Fy - Caminhada	0,87
Primeiro Experimento - Fx- Caminhada	0,90
Segundo Experimento - FR - Corrida	0,67
Segundo Experimento - Fy - Corrida	0,62
Segundo Experimento - Fx- Corrida	0,84

Fonte: Elaborado pelo autor.

Tabela 31 – Comparativo do R^2 com o trabalho de Haupenthal (2013).

Experimento desta dissertação	**R^2 desta dissertação**	**Experimento de Haupenthal (2013) usado na comparação**	**R^2 do experimento de Haupenthal (2013)**
Primeiro Experimento - Caminhada Normal - FR	0,93	1° Experimento - FR - Caminhada	0,88
Primeiro Experimento - Caminhada Normal - FY	0,92	1° Experimento - Fy - Caminhada	0,87
Primeiro Experimento - Caminhada Normal - FX	0,80	1° Experimento - Fx- Caminhada	0,90
Segundo Experimento - Caminhada Lenta - FR	0,90	1° Experimento - FR - Caminhada	0,88
Segundo Experimento - Caminhada Lenta - FY	0,91	1° Experimento - Fy - Caminhada	0,87
Segundo Experimento - Caminhada Lenta - Fx	0,64	1° Experimento - Fx- Caminhada	0,90
Terceiro Experimento - Caminhada Rápida - FR	0,82	1° Experimento - FR -Caminhada	0,88
Terceiro Experimento - Caminhada Rápida - FY	0,83	1° Experimento - Fy - Caminhada	0,87

Terceiro Experimento - Caminhada Rápida - FX	0,69	1° Experimento - Fx- Caminhada	0,90
Quarto Experimento - Corrida Normal - FR	0,60	2 ° Experimento - FR -Corrida	0,67
Quarto Experimento - Corrida Normal- FY	0,60	2 ° Experimento - Fy -Corrida	0,62
Quarto Experimento - Corrida Normal - FX	0,52	2 ° Experimento - Fx- Corrida	0,84
Quinto Experimento - Corrida Lenta - FR	0,59	2 ° Experimento - FR -Corrida	0,67
Quinto Experimento - Corrida Lenta - FY	0,60	2 ° Experimento - Fy -Corrida	0,62
Quinto Experimento - Corrida Lenta - FX	0,45	2 ° Experimento - Fx- Corrida	0,84
Sexto Experimento - Corrida Rápida - FR	0,58	2 ° Experimento - FR -Corrida	0,67
Sexto Experimento - Corrida Rápida - FY	0,56	2 ° Experimento - Fy -Corrida	0,62
Sexto Experimento - Corrida Rápida - FX	0,56	2 ° Experimento - Fx- Corrida	0,84

Fonte: Elaborado pelo autor.

Os demais estudos apresentados no capítulo 2.7 (Trabalhos Relacionados), não podem ter seus resultados comparados com os desta dissertação, pois apesar de tratarem-se de trabalhos que utilizaram a mineração de dados aplicada na saúde, possuem dados e objetivos diferentes.

5 CONCLUSÃO

Ainda que venha ocorrendo o crescimento na produção de pesquisas na área de mineração de dados ligada à saúde, é necessário que esta área seja mais estudada e pesquisada. Um fato que corrobora com a afirmativa anterior é que não se encontrou pesquisas com uma problemática bem próxima da desse trabalho, onde a mineração de dados e diferentes algoritmos foram utilizados para a predição da FRS em exercícios subaquáticos. Ainda que como exposto anteriormente, as forças de reação ao solo são de extrema importância para a prescrição de exercícios na água.

Por meio dos modelos de predição gerados nessa pesquisa conclui-se que é possível prever com um mínimo de exatidão a força de reação do solo durante a caminhada e corrida na água. Todavia salienta-se que é preciso trabalhos futuros voltados ao mesmo tema e meta dessa dissertação, para que seja possível melhorar significativamente os modelos de predição, principalmente para os exercícios de corrida.

Por meio dos gráficos gerados, as variáveis que demonstram ter mais impacto na FRS trata-se do IH_ratio, Velocity e Mass_kg no caso da predição de FR e Fy. Já para Fx a variável que mais destoou das demais foi Velocity. Sendo assim, é notório que estudiosos e especialistas da área da saúde e fisioterapia devem analisar com mais atenção as mesmas.

Notou-se que os algoritmos com melhor desempenho foram o SVR e o Tweedie Regressor, onde de maneira geral pode-se dizer que o SVR gera os melhores modelos para Fx, enquanto o Tweedie Regressor gera os melhores modelos para Fy, e ambos têm bons modelos para FR.

Destaca-se ainda que para estudos futuros seria importante a participação de mais sujeitos, gerando uma maior quantidade de dados e de casos, o que poderia trazer ainda mais padrões à tona por meio da aplicação da mineração de dados. Além disso sugere-se a coleta de mais variáveis antropométricas e também a adição de variáveis ligadas a histórico de lesões, pratica de atividades físicas, presença de comorbidades ou não, profissão e entre outras.

O objetivo fundamental desta dissertação foi efetuar um estudo e aplicar as técnicas e métodos de mineração de dados, para gerar modelos para predição da força de reação do solo durante a caminhada e corrida na água por meio de algoritmos de regressão como SVM, regressão linear e etc. Baseado nisso, essa pesquisa procurou auxiliar os profissionais de fisioterapia e da saúde a prescrever exercícios subaquáticos com maior eficiência, por meio de modelos de predição da FRS.

REFERÊNCIAS

ALVES, Rafael Damiani et al. **Predição do desempenho da redação do enem utilizando técnicas de mineração de dados**. 2018.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03, 2011.

BATES, A.; HANSON, N. **Exercícios aquáticos terapêuticos**. São Paulo: Manole, 1998.

BISHOP, Christopher M. Pattern recognition. **Machine learning**, v. 128, n. 9, 2006.

CABENA, Peter et al. **Discovering data mining: from concept to implementation**. Prentice-Hall, Inc., 1998.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), v. 29, 2009.

CARVALHO, Deborah Ribeiro et al. Mineração de dados aplicada à fisioterapia. **Fisioterapia em Movimento**, v. 25, p. 595-605, 2012.

CHESTER, Rachel et al. Autoeficácia e risco de dor persistente no ombro: resultados de uma análise de árvore de classificação e regressão (CART). **Jornal britânico de medicina esportiva** , v. 53, n. 13, pág. 825-834, 2019.

COSTA, Evandro et al. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações.
Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1-29, 2013.

DAVIDSON, James W.; SAVIC, D.; WALTERS, Godfrey A. Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. **Journal of Hydroinformatics**, v. 1, n. 2, p. 115-126, 1999.

DE SOUTO, M. C. P. et al. Técnicas de aprendizado de máquina para problemas de biologia molecular.
Sociedade Brasileira de Computação, v. 1, n. 2, 2003.

DICHARRY, J. Kinematics and kinetics of gait: from lab to clinic. **Clinical Sports Medicine**, v. 29, p. 347-364, 2010.

DOMINGUES, Miriam Lúcia Campos Serra. **Mineração de Dados Utilizando Aprendizado Não-Supervisionado**: um estudo de caso para bancos de dados da saúde. 2003.

DUARTE, Julio Cesar. **O Algoritmo Boosting at Start e suas Aplicações**. 2009. Tese de Doutorado. Tese (Doutorado)—PUC-RJ, 2009. 45.

DRUCKER, H. **Support vector regression machines**. Cambridge, p. 155–161, 1997.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da **literatura**. Acta Paulista de Enfermagem, v. 22, n. 5, p. 686-690, 2009.

GIUSTOLISI, Orazio; SAVIC, Dragan A. A symbolic data-driven technique based on evolutionary polynomial regression. **Journal of Hydroinformatics**, v. 8, n. 3, p. 207-222, 2006.

GOMES, Denilsen Carvalho et al. Mineração de Dados no Serviço de Atendimento de Urgências. **Journal of Health Informatics**, v. 6, n. 4, 2014.

GRAUNT, John. Natural and political observations mentioned in a following index, and made upon the bills of mortality. In: **Mathematical Demography**. Springer, Berlin, Heidelberg, 1977. p. 11-20.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining: Concepts and Techniques**. 3. ed. Amsterdam: Elsevier, 2011. 744 p.

HAND, David; MANNILA, H.; SMYTH, Padhraic. **Principles of data mining**. 5ed. London: MIT Press, 2001. 93-102p.

HAUPENTHAL, Alessandro et al. **Proposição de modelo preditivo da força de reação do solo durante a caminhada e corrida na água.** 2013.

JUPYTER. **Sobre nós**. 2021. Disponível em: https://jupyter.org/about. Acesso em: 29 jun. 2021.

KELLER, Tony S. et al. Relationship between vertical ground reaction force and speed during walking, slow jogging, and running. **Clinical biomechanics**, v. 11, n. 5, p. 253-259, 1996.

MACKAY, D. J. C. Bayesian interpolation. **Neural computation**, MIT Press, v. 4, n. 3, p. 415–447, 1992.

MAIA, Alexandre Gori. **Econometria**: conceitos e aplicações. Saint Paul Institute Of Finance, 2017.

MARQUES, Roberto Ligeiro; DUTRA, I. N. Ê. S. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. **Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil**, 2002.

MASUMOTO, K.; TAKASUGI, S.; HOTTA, NOBORU; FUJISHIMA, KAZUTAKA; IWAMOTO, Y.
Electromyographic analysis of walking in water in healthy humans. **Journal of Physiology and Anthropology Apply to Human Science**, v. 23, p. 119-27, 2004.

MCKINNEY, Wes**. Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython**. Novatec Editora, 2019.

MELLO, Beatriz Ribeiro de. **Epidemiologia de lesões no futebol feminino amador-Um estudo utilizando mineração de dados**. 2010.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina.
Sistemas inteligentes-Fundamentos e aplicações, v. 1, n. 1, p. 32, 2003.

NEAL, R. M. **Bayesian learning for neural networks**. [S.l.]: Springer Science & Business Media, 2012. v. 118 NOVACHECK, T. Biomechanics of running. Gait and Posture, v. 7, p. 77-95, 1998.

NOVACHECK, T. Biomechanics of running. Gait and Posture, v. 7, p. 77-95, 1998.

NOVAES, André Luiz Farias. **Programação Genética Econométrica: uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais**. 2015. Tese de Doutorado. PUC-Rio.

OKAMURA, Dalton Akio. **Análise de algoritmos de regressão aplicados a mercado financeiro**. 2019. PANDAS. Sobre pandas. Disponível em: https://pandas.pydata.org/about/. Acesso em: 29 jun. 2021.
PERRY, J. Gait analysis: normal and pathological function. New York: MacGraw-Hill, 1992.

PANDAS. **Sobre pandas**. Disponível em: https://pandas.pydata.org/about/. Acesso em: 29 jun. 2021.

PERRY, J. **Gait analysis: normal and pathological function**. New York: MacGraw-Hill, 1992.

PESSOA, Alex Sandro Aguiar et al. Mineração de dados meteorológicos para previsão de eventos severos.
Revista Brasileira de Meteorologia, v. 27, p. 61-74, 2012.

PIVATO, Marina Abichabki. **Mineração de regras de associação em dados georreferenciados**. 2012. Tese de Doutorado. Universidade de São Paulo.

POHLENZ, Vitor. **Decision Trees, Random Forests e Ensemble**. Florianópolis: Universidade Federal de Santa Catarina, 2020. 7 slides, color.

QUEIROGA, Emanuel Marques. **Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados**. 2017. Dissertação de Mestrado. Universidade Federal de Pelotas.

PIVATO, Marina Abichabki. **Mineração de regras de associação em dados georreferenciados**. 2012. Tese de Doutorado. Universidade de São Paulo.

RIBEIRO, João Luiz Oliveira et al. **Uso de técnicas de mineração de dados em Python para classificação de pássaros baseado nas medidas dos ossos**. 2017.

ROESLER, H.; HAUPENTHAL, A.; SCHÜTZ, G. R.; DE SOUZA, PATRÍCIA V. Dynamometric analysis of the maximum force applied in aquatic human gait at 1.3 m of immersion. **Gait and Posture**, v. 24, p. 412-17, 2006.

ROSE, J.; GAMBLE, J. G. **Marcha humana**. São Paulo: Premier, 1998.

ROZIN, Nicole Amanda. **PREVISÃO DO DESLOCAMENTO DE TEMPESTADES SEVERAS:
ABORDAGENS POR APRENDIZADO DE MÁQUINA**. 2018. 104 f. Monografia (Especialização) - Curso de Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2018.

SAMIULLAH, Md et al. Correlation mining in graph databases with a new measure. In: **Asia-Pacific Web Conference**. Springer, Berlin, Heidelberg, 2013. p. 88-95.

SCIKIT-LEARN. **Regressão Bayesiana**. 2021. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression. Acesso em: 28 jun. 2021.

SELL, Isair. Utilização da regressão linear como ferramenta de decisão na gestão de custos. In: **Anais do Congresso Brasileiro de Custos-ABC**. 2005.

SKINNER, A. T.; THOMSON, A. M. **Duffield: exercícios na água**. 3rd ed. São Paulo: Manole, 1985.

STEINER, Maria Teresinha Arns et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gestao & producao**, v. 13, n. 2, p. 325-337, 2006. MITCHELL, Tom M. et al. Machine learning. 1997.

TAKESHIMA, N.; ROGERS, M.; WATANABE, E. et al. Water-based exercise improves health- related aspects of fitness in older women. **Medicine and Science of Sports Exercise**, v. 34, n. 3, p. 544-51, 2002.

WEIAND, Augusto; WEIAND, Fernanda Rodrigues Ribeiro. **Análise de sentimentos do Twitter com Naïve Bayes e NLTK**. Scientiatec, [S.L.], v. 4, n. 3, p. 46-57, 24 abr. 2018. Instituto Federal de Educacao - Ciencia e Tecnologia do Rio Grande do Sul. http://dx.doi.org/10.35819/scientiatec.v4i3.2188.

WILLCOX, Walter F. The founder of statistics. **Revue de l'Institut International de Statistique**, p. 321-328, 1938.

WINTER, D. **The biomechanics and motor control of human gait: normal, elderly and pathological**. Canada: Waterloo Cover, 1991.

ZATSIORSKY, V. M. **Biomecânica no Esporte: Performance do Desempenho e Prevenção de Lesão**. São Paulo: Guanabara Koogan, 2004.

ZHANG, H. The optimality of naive Bayes. AA, [S.l.], v.1, n.2, p.3, 2004.

ZHOU, Zhi-Hua. **Three perspectives of data mining. Artificial Intelligence**, v. 143, n. 1, p. 139-146, 2003.

APÊNDICE A – INFLUÊNCIA DAS VARIÁVEIS DE ENTRADA (INDEPENDENTES)

As informações sobre a influência das variáveis de entrada (independentes) para FR, FY e FX referentes aos exercicios de caminhada lenta, caminhada rápida, corrida lenta e corrida rápidaseencontramnolink: https://drive.google.com/drive/folders/1rxDhHrbMVJ5my3hnRSZFxDfVX5zPci2Z?usp=sharing

Optou-se por disponibilizar os mesmos dessa maneira devido a grande quantidade de informações e por eles seguirem os mesmos padrões da caminhada normal e da corrida normal.

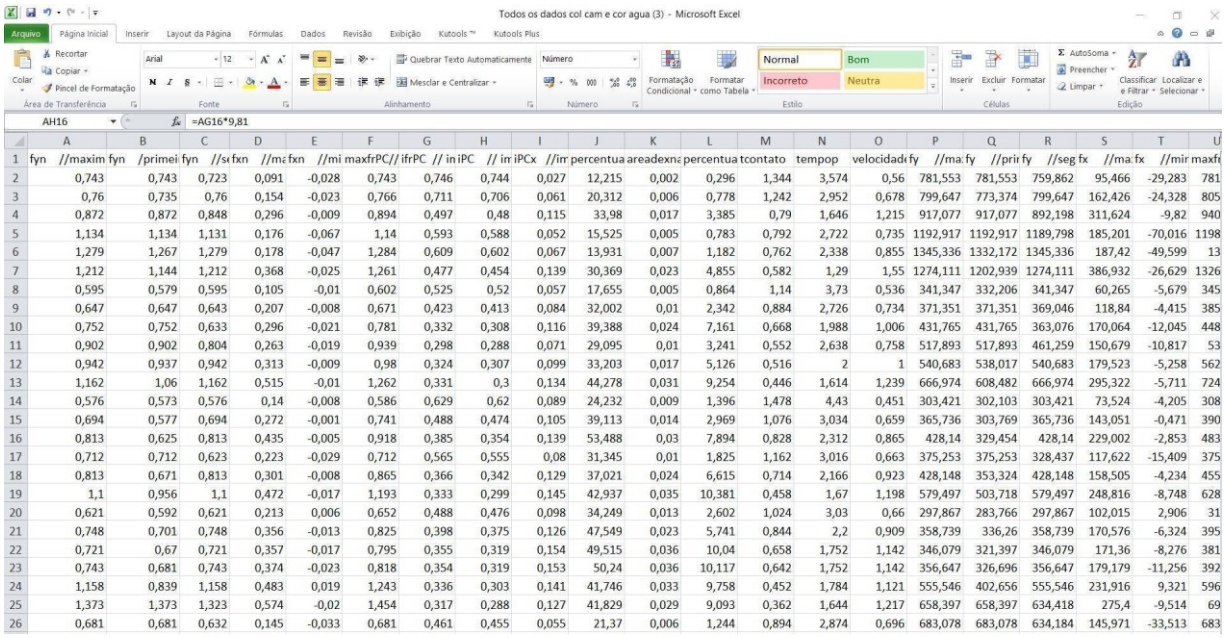
APÊNDICE B – INFLUÊNCIA DAS VARIÁVEIS DE ENTRADA (INDEPENDENTES)

As informações sobre a distribuição normal para os exercicios de caminhada lenta, caminhada rápida, corrida lenta e corrida rápida se encontram no link: https://drive.google.com/drive/folders/1NH8g0we-0E6pMggIASPyxgRt32eAHsKZ?usp=sharing

Optou-se por disponibilizar os mesmos dessa maneira devido a grande quantidade de informações e por eles seguirem os mesmos padrões da caminhada normal e da corrida normal.

APÊNDICE C – BASE DE DADOS ORIGINAL

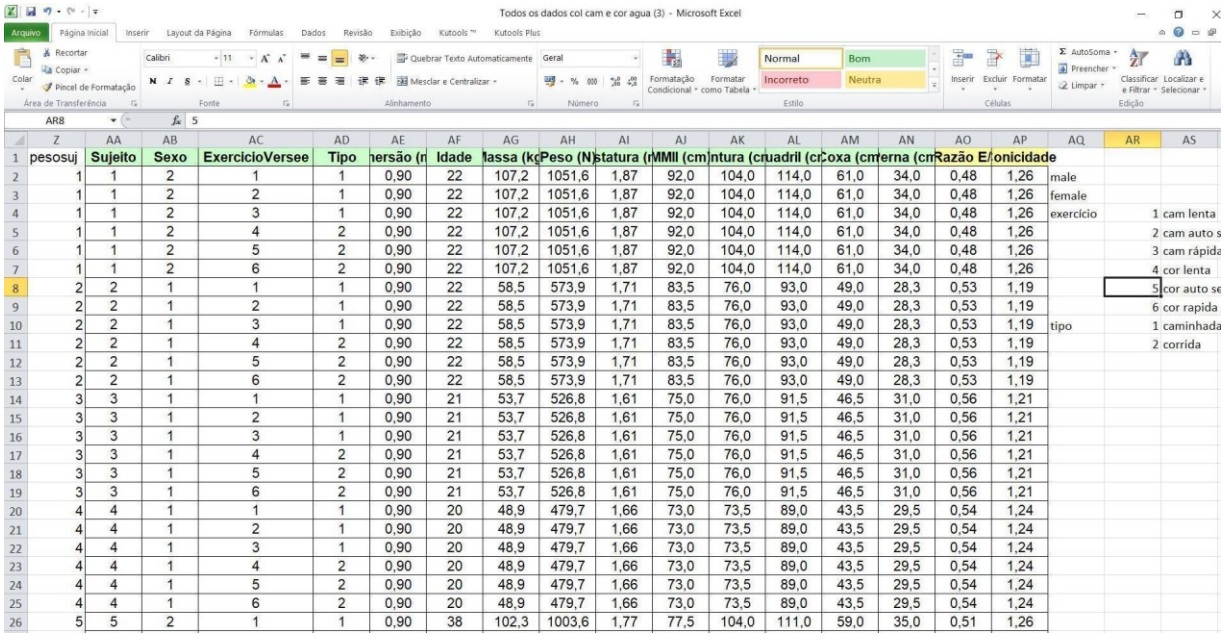
Figura 48 – Dados iniciais parte 1.



fyn	//maxim	fyn	/primei	fyn	//se	fxn	//ma	fxn	//mi	maxfrPC/	ifrPC	// in	iPC	// in	iPCx	//in	percentua	areadexna	percentua	tcontato	tempop	velocidad	fy	//ma	fy	//prir	fy	//seg	fx	//ma	fx	//mir	maxfr
	0,743		0,743		0,723		0,091		-0,028		0,743		0,746		0,744		0,027		12,215		0,002		0,296		1,344		3,574		0,56		781,553		781,553		759,862		95,466		-29,283		781
	0,76		0,735		0,76		0,154		-0,023		0,766		0,711		0,706		0,061		20,312		0,006		0,778		1,242		2,952		0,678		799,647		773,374		799,647		162,426		-24,328		805
	0,872		0,872		0,848		0,296		-0,009		0,894		0,497		0,48		0,115		33,98		0,017		3,385		0,79		1,646		1,215		917,077		917,077		892,198		311,624		-9,82		940
	1,134		1,134		1,131		0,176		-0,067		1,14		0,593		0,588		0,052		15,525		0,005		0,783		0,792		2,722		0,735		1192,917		1192,917		1189,798		185,201		-70,016		1198
	1,279		1,267		1,279		0,178		-0,047		1,284		0,609		0,602		0,067		13,931		0,007		1,182		0,762		2,338		0,855		1345,336		1332,172		1345,336		187,42		-49,599		13
	1,212		1,144		1,212		0,368		-0,025		1,261		0,477		0,454		0,139		30,369		0,023		4,855		0,582		1,29		1,55		1274,111		1202,939		1274,111		386,932		-26,629		1326
	0,595		0,579		0,595		0,105		-0,01		0,602		0,525		0,52		0,057		17,655		0,005		0,864		1,14		3,73		0,536		341,347		332,206		341,347		60,265		-5,679		345
	0,647		0,647		0,643		0,207		-0,008		0,671		0,423		0,413		0,084		32,002		0,01		2,342		0,884		2,726		0,734		371,351		371,351		369,046		118,84		-4,415		385
	0,752		0,752		0,633		0,296		-0,021		0,781		0,332		0,308		0,116		39,388		0,024		7,161		0,668		1,988		1,006		431,765		431,765		363,076		170,064		-12,045		448
	0,902		0,902		0,804		0,263		-0,019		0,939		0,298		0,288		0,071		29,095		0,01		3,241		0,552		2,638		0,758		517,893		517,893		461,259		150,679		-10,817		53
	0,942		0,937		0,942		0,313		-0,009		0,98		0,324		0,307		0,099		33,203		0,017		5,126		0,516		2		1		540,683		538,017		540,683		179,523		-5,258		562
	1,162		1,06		1,162		0,515		-0,01		1,262		0,331		0,3		0,134		44,278		0,031		9,254		0,446		1,614		1,239		666,974		608,482		666,974		295,322		-5,711		724
	0,576		0,573		0,576		0,14		-0,008		0,586		0,629		0,62		0,089		24,232		0,009		1,396		1,478		4,43		0,451		303,421		302,103		303,421		73,524		-4,205		308
	0,694		0,577		0,694		0,272		-0,001		0,741		0,488		0,474		0,105		39,113		0,014		2,969		1,076		3,034		0,659		365,736		303,769		365,736		143,051		-0,471		390
	0,813		0,625		0,813		0,435		-0,005		0,918		0,385		0,354		0,139		53,488		0,03		7,894		0,828		2,312		0,865		428,14		329,454		428,14		229,002		-2,853		483
	0,712		0,712		0,623		0,223		-0,029		0,712		0,565		0,555		0,08		31,345		0,01		1,825		1,162		3,016		0,663		375,253		375,253		328,437		117,622		-15,409		375
	0,813		0,671		0,813		0,301		-0,008		0,865		0,366		0,342		0,129		37,021		0,024		6,615		0,714		2,166		0,923		428,148		353,324		428,148		158,505		-4,234		455
	1,1		0,956		1,1		0,472		-0,017		1,193		0,333		0,299		0,145		42,937		0,035		10,381		0,458		1,67		1,198		579,497		503,718		579,497		248,816		-8,748		628
	0,621		0,592		0,621		0,213		0,006		0,652		0,488		0,476		0,098		34,249		0,013		2,602		1,024		3,03		0,66		297,867		283,766		297,867		102,015		2,906		31
	0,748		0,701		0,748		0,356		-0,013		0,825		0,398		0,375		0,126		47,549		0,023		5,741		0,844		2,2		0,909		358,739		336,26		358,739		170,576		-6,324		395
	0,721		0,67		0,721		0,357		-0,017		0,795		0,355		0,319		0,154		49,515		0,036		10,04		0,658		1,752		1,142		346,079		321,397		346,079		171,36		-8,276		381
	0,743		0,681		0,743		0,374		-0,023		0,818		0,354		0,319		0,153		50,24		0,036		10,117		0,642		1,752		1,142		356,647		326,696		356,647		179,179		-11,256		392
	1,158		0,839		1,158		0,483		0,019		1,243		0,336		0,303		0,141		41,746		0,033		9,758		0,452		1,784		1,121		555,546		402,656		555,546		231,916		9,321		596
	1,373		1,373		1,323		0,574		-0,02		1,454		0,317		0,288		0,127		41,829		0,029		9,093		0,362		1,644		1,217		658,397		658,397		634,418		275,4		-9,514		69
	0,681		0,681		0,632		0,145		-0,033		0,681		0,461		0,455		0,055		21,37		0,006		1,244		0,894		2,874		0,696		683,078		683,078		634,184		145,971		-33,513		683

Fonte: Elaborado pelo autor.

Figura 49 – Dados iniciais parte 2.



pesosuj	Sujeito	Sexo	ExercicioVersee	Tipo	hersão (r	Idade	lassa (kg	Peso (N)	statura (r	MMII (cm)	ntura (cr	uadril (cr	oxa (cm	erna (cm	Razão E	onicidade			
1	1	2	1	1	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26	male		
1	1	2	2	1	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26	female		
1	1	2	3	1	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26	exercício	1	cam lenta
1	1	2	4	2	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26		2	cam auto s
1	1	2	5	2	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26		3	cam rápida
1	1	2	6	2	0,90	22	107,2	1051,6	1,87	92,0	104,0	114,0	61,0	34,0	0,48	1,26		4	cor lenta
2	2	1	1	1	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19		5	cor auto se
2	2	1	2	1	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19		6	cor rapida
2	2	1	3	1	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19	tipo	1	caminhada
2	2	1	4	2	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19		2	corrida
2	2	1	5	2	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19			
2	2	1	6	2	0,90	22	58,5	573,9	1,71	83,5	76,0	93,0	49,0	28,3	0,53	1,19			
3	3	1	1	1	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
3	3	1	2	1	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
3	3	1	3	1	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
3	3	1	4	2	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
3	3	1	5	2	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
3	3	1	6	2	0,90	21	53,7	526,8	1,61	75,0	76,0	91,5	46,5	31,0	0,56	1,21			
4	4	1	1	1	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
4	4	1	2	1	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
4	4	1	3	1	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
4	4	1	4	2	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
4	4	1	5	2	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
4	4	1	6	2	0,90	20	48,9	479,7	1,66	73,0	73,5	89,0	43,5	29,5	0,54	1,24			
5	5	2	1	1	0,90	38	102,3	1003,6	1,77	77,5	104,0	111,0	59,0	35,0	0,51	1,26			

Fonte: Elaborado pelo autor.

APÊNDICE D – SCRIPTS DO PRÉ-PROCESSAMENTO E DO PREENCHIMENTO DO BANCO DE DADOS

Figura 50 – Pré-processamento.

```
In [20]: import mysql.connector # importando conector do mysql com o python
         import pandas as pd # importando a binlioteca do pandas
         import inflection # para mudar automaticamente nome das colunas
         import math # verificar se tem nan
         from datetime import datetime # importando biblioteca de modificacoes de datas
         pd.set_option('display.max_columns', None) # dizendo ao pandas que quero visualizar todas as colunas

         mydb = mysql.connector.connect( # informando o login e conectando com o banco de dados
             host="localhost",
             database="experimentoagua",
             user="root",
             passwd="231327ra")
         print(mydb)
         conn=mydb # declarando a conexao a variavel conn
         df= pd.read_csv('C:\Caminhada_Normal_Para_Escrita_Valores_Originais.Data', delimiter=';') #carregando .csv
                 #df = pd.read_csv(file_csv, na_values=' ')
         df['FR_N'] = [x.replace(",", ".") for x in df['FR_N']]
         df['FR_N'] = df['FR_N'].astype(float)
         df['Fy_N'] = [x.replace(",", ".") for x in df['Fy_N']]
         df['Fy_N'] = df['Fy_N'].astype(float)
         df['Fx_N'] = [x.replace(",", ".") for x in df['Fx_N']]
         df['Fx_N'] = df['Fx_N'].astype(float)
         df['Velocity'] = [x.replace(",", ".") for x in df['Velocity']]
         df['Velocity'] = df['Velocity'].astype(float)
                 #df['Immersion_m'] = [x.replace(",", ".") for x in df['Immersion_m']]
                 #df['Immersion_m'] = df['Immersion_m'].astype(float)
         df['Mass_kg'] = [x.replace(",", ".") for x in df['Mass_kg']]
         df['Mass_kg'] = df['Mass_kg'].astype(float)
         df['Weigth_N'] = [x.replace(",", ".") for x in df['Weigth_N']]
         df['Weigth_N'] = df['Weigth_N'].astype(float)
                 #publi['Stature_m'] = [x.replace(",", ".") for x in publi['Stature_m']]
                 #publi['Stature_m'] = publi['Stature_m'].astype(float)
         df['Waist_cm'] = [x.replace(",", ".") for x in df['Waist_cm']]
         df['Waist_cm'] = df['Waist_cm'].astype(float)
         df['Thigh_cm'] = [x.replace(",", ".") for x in df['Thigh_cm']]
         df['Thigh_cm'] = df['Thigh_cm'].astype(float)
         df['IH_ratio'] = [x.replace(",", ".") for x in df['IH_ratio']]
         df['IH_ratio'] = df['IH_ratio'].astype(float)
         #df['Sex'] = [x.replace(",", ".") for x in df['Sex']]
         #df['Sex'] = df['Sex'].astype(float)
         #df['Age'] = [x.replace(",", ".") for x in df['Age']]
         #df['Age'] = df['Age'].astype(float)
```

Fonte: Elaborado pelo autor.

Figura 51 – Script para preenchimento do banco de dados.

```
sql = "INSERT INTO experimentoagua.caminhada_normal (id_sujeito,Fy_N,Fx_N,FR_N,Velocity,Sex,Age,Mass_kg,Weigth_N,Waist_
cursor = conn.cursor() # atribuindo acao

for index, row in df.iterrows(): # for para percorrer todas as linhas

        # atribui valor da linha por coluna a uma variavel
        Id_sujeito=row['Id_sujeito']
        Fy_N= row['Fy_N']
        Fx_N= row['Fx_N']
        FR_N= row['FR_N']
        Velocity= row['Velocity']
        Sex= row['Sex']
       # Immersion_m= row['Immersion_m']
        Age= row['Age']
        Mass_kg= row['Mass_kg']
        Weigth_N= row['Weigth_N']
       # Stature_m= row['Stature_m']
        Waist_cm= row['Waist_cm']
        Thigh_cm= row['Thigh_cm']
        IH_ratio= row['IH_ratio']
        sql_data=(int(Id_sujeito),float(Fy_N),float(Fx_N),float(FR_N), float(Velocity),float(Sex),float(Age),float(Mass_
        print(sql_data)
      #  print('id_atleta:',id_atleta, 'nome_n:', nome_n,'tipo_de_atividade:',tipo_de_atividade, 'duracao:', duracao, '
        cursor.execute(sql,sql_data) # inseri no banco de dados com base no sql a ser executado e nas variaveis passada
        print("for ok")
conn.commit() # commita ou seja persiste de fato as alteracoes no banco de dados
print("feito")
```

Fonte: Elaborado pelo autor.

APÊNDICE E – SCRIPTS DE TRANSFORMAÇÃO DOS DADOS

Figura 52 – Descoberta da média das variáveis de entrada (independentes).

```
sql = "SELECT * FROM caminhada_normal"
df= pd.read_sql(sql, conn)
#media_Fy_N=0
#media_Fx_N=0
#media_FR_N=0
media_Velocity=0
media_Sex=0
media_Age=0
media_Mass_kg=0
media_Weigth_N=0
media_Waist_cm=0
media_Thigh_cm=0
media_IH_ratio=0
i=0

for index, row in df.iterrows(): # for para percorrer todas as linhas


        Fy_N= row['Fy_N']
        Fx_N= row['Fx_N']
        FR_N= row['FR_N']
        Velocity= row['Velocity']
        Sex= row['Sex']
        Age= row['Age']
        Mass_kg= row['Mass_kg']
        Weigth_N= row['Weigth_N']
        Waist_cm= row['Waist_cm']
        Thigh_cm= row['Thigh_cm']
        IH_ratio= row['IH_ratio']

       # media_Fy_N=media_Fy_N+Fy_N
       # media_Fx_N=media_Fx_N+Fx_N
       # media_FR_N=media_FR_N+FR_N
        media_Velocity=media_Velocity+Velocity
        media_Sex=media_Sex+Sex
        media_Age=media_Age+Age
        media_Mass_kg=media_Mass_kg+Mass_kg
        media_Weigth_N=media_Weigth_N+Weigth_N
        media_Waist_cm=media_Waist_cm+Waist_cm
        media_Thigh_cm=media_Thigh_cm+Thigh_cm
        media_IH_ratio=media_IH_ratio+IH_ratio

        i=i+1



#media_Fy_N=media_Fy_N/i
#media_Fx_N=media_Fx_N/i
#media_FR_N=media_FR_N/i
media_Velocity=media_Velocity/i
media_Sex=media_Sex/i
media_Age=media_Age/i
media_Mass_kg=media_Mass_kg/i
media_Weigth_N=media_Weigth_N/i
media_Waist_cm=media_Waist_cm/i
media_Thigh_cm=media_Thigh_cm/i
media_IH_ratio=media_IH_ratio/i
```

Fonte: Elaborado pelo autor.

Figura 53 – Normalização das variáveis de entrada (independentes).

```
sql = "UPDATE experimentoagua.caminhada_normal SET id_sujeito = %s, Velocity = %s, Sex = %s, Age = %s, Mass_kg = %s, We
cursor = conn.cursor() # atribuindo acao

for index, row in df.iterrows(): # for para percorrer todas as linhas

        # atribui valor da linha por coluna a uma variavel
        id_sujeito=row['id_sujeito']
        Fy_N= row['Fy_N']
        Fx_N= row['Fx_N']
        FR_N= row['FR_N']
        Velocity= row['Velocity']
        Sex= row['Sex']
       # Immersion_m= row['Immersion_m']
        Age= row['Age']
        Mass_kg= row['Mass_kg']
        Weigth_N= row['Weigth_N']
       # Stature_m= row['Stature_m']
        Waist_cm= row['Waist_cm']
        Thigh_cm= row['Thigh_cm']
        IH_ratio= row['IH_ratio']
        sql_data=(int(id_sujeito),float(Velocity/media_Velocity),float(Sex/media_Sex),float(Age/media_Age),float(Mass_k
        print(sql_data)
     #  print('id_atleta:',id_atleta, 'nome_n:', nome_n,'tipo_de_atividade:',tipo_de_atividade, 'duracao:', duracao, '
        cursor.execute(sql,sql_data) # inseri no banco de dados com base no sql a ser executado e nas variaveis passada
        print("for ok")
conn.commit() # commita ou seja persiste de fato as alteracoes no banco de dados
print("feito")
```

Fonte: Elaborado pelo autor.

APÊNDICE F – SCRIPTS DA ANÁLISE EXPLORATÓRIA DE DADOS

Figura 54 – Script de verificação da forma normal dos dados.

```
from scipy.stats import normaltest

significancia=0.05 #chance de erro
normais_n={}
moda_n={}
media_n={}
mediana_n={}
desviopadrao_n={}
desvioabsoluto_n={}
amplitude_n={}
variancia_n={}
simetria_n={}
stat_test_n={}
p_valor_n={}

print('significancia:', significancia)
for elemento in lista:
    nome=elemento
    analise=publi[nome]



    stat_test, p_valor = normaltest(analise)
    stat_test_n[nome]=stat_test
    p_valor_n[nome]=p_valor

    print(p_valor-significancia)
    verifica=p_valor <= significancia
    print(verifica)

    if (verifica==False):

        normais_n[nome] = 'É normal'
    else:

        normais_n[nome] = 'Não é normal'


    Moda = analise.mode()[0]
    moda_n[nome]=Moda

    Mediana =analise.median()
    mediana_n[nome]=Mediana

    Media = analise.mean()
    media_n[nome]=Media

    DesvioPadrao= analise.std()
    desviopadrao_n[nome]=DesvioPadrao

    DesvioAbsoluto = analise.mad()
    desvioabsoluto_n[nome]=DesvioAbsoluto

    amplitude= analise.max() - analise.min()
    amplitude_n[nome]=amplitude

    variancia= analise.var()
    variancia_n[nome]=variancia

    if((Media-Moda) > 0 ):

        simetria_n[nome]=(('Assimetrica a direita (Inclinado Positivamente)', (Media-Moda) ))



    elif((Media-Moda) < 0):

        simetria_n[nome]= (('Assimetrica a esquerda (Inclinado Negativamente) ', (Media-Moda)))


    elif((Media-Moda)==0):
        simetria_n[nome]=(('Simétrica', (Media-Moda)))


    else:

        print('erro')
```

Fonte: Elaborado pelo autor.

Figura 55 – Script para obtenção da moda, mediana, média, desvio padrão, desvio absoluto, amplitude, variância, simetria, covariância e correlação.

```
p_valor
for elemento in p_valor_n:
    print('Nome Variavel:', elemento, '         Valor:',  p_valor_n[elemento])
for elemento in stat_test_n:
    print('Nome Variavel:', elemento, '          Valor:',  stat_test_n[elemento])
for elemento in normais_n:
    print('Nome Variavel:', elemento, '          Valor:',  normais_n[elemento])
for elemento in moda_n:
    print('Nome Variavel:', elemento, '          Valor:',  moda_n[elemento])
for elemento in mediana_n:
    print('Nome Variavel:', elemento, '          Valor:',  mediana_n[elemento])
for elemento in media_n:
    print('Nome Variavel:', elemento, '          Valor:',  media_n[elemento])
for elemento in desviopadrao_n:
    print('Nome Variavel:', elemento, '          Valor:',  desviopadrao_n[elemento])
for elemento in desvioabsoluto_n:
    print('Nome Variavel:', elemento, '          Valor:',  desvioabsoluto_n[elemento])
for elemento in amplitude_n:
    print('Nome Variavel:', elemento, '          Valor:',  amplitude_n[elemento])
for elemento in variancia_n:
    print('Nome Variavel:', elemento, '          Valor:',  variancia_n[elemento])
for elemento in simetria_n:
    print('Nome Variavel:', elemento, '          Valor:',  simetria_n[elemento])
```

Fonte: Elaborado pelo autor.

APÊNDICE G – PARÂMETROS

Quadro 4 – Parâmetros dos algoritmos.

Algoritmo:	Parâmetros:
Linear Regression	Padrão
SVR	(kernel='poly')
AdaBoost Regressor	(random_state=0, n_estimators=100)
Gradient Boosting Regressor	(random_state=1)
Tweedie Regressor	(power=1, alpha=0.5, link='log')
Lasso Lars	(alpha=.1)
Lasso	Padrão
Bayesian Ridge	Padrão
Ridge Cross-Validation	Padrão
Ridge	(alpha=.5)
Random Forest Regressor	Padrão
K-Neighbors Regressor	(n_neighbors=2)
Decision Tree Regressor	Padrão
Multi-layer Perceptron regressor	Padrão

Fonte: Elaborado pelo autor.

ANEXO A – DICIONÁRIO DE DADOS

Variáveis dependentes: Variáveis a serem preditas (FR, Fy, Fx)

- Pico da resultante da FRS (FR): é o valor máximo da resultante de Fy e Fx durante a realização do contato com a plataforma
- Pico da componente vertical da FRS (Fy): é o valor máximo da componente vertical da FRS durante a realização do contato com a plataforma.
- Pico da componente ântero-posterior da FRS (Fx): é o valor máximo da componente ântero-posterior da FRS durante a realização do contato com a plataforma.

Variáveis independentes: Usadas como entrada nos algoritmos.

- Circunferência do abdômen (Waist_cm): medida da circunferência na altura da cicatriz umbilical do participante, expressa em centímetros.
- Circunferência da coxa (Thigh_cm): medida da circunferência na metade da coxa expressa em centímetros.
- Massa corporal (Mass_kg): é o valor da massa corporal total do sujeito, expressa em quilogramas (kg).
- Velocidade (Velocity): trata-se da velocidade de deslocamento do sujeito durante a corrida ou caminhada na água. É expressa em m/s.
- Sexo (Sex): o sexo dos participantes foi incluído como variável categórica nominal, sendo caracterizado no arquivo csv como valor zero (0) o homem e um (1) a mulher.
- IH_ratio: – Razão entre o nível de imersão e a estatura do indivíduo.
- Idade (Age): Idade dos sujeitos.

Observações:

- Imersão: A profundidade de imersão é a medida da distância entre a linha da água e a tampa da plataforma de força, em metros (m). Neste estudo foram analisadas cinco profundidades de imersão: 0,75, 0,90, 1,05, 1,20, e 1,35 m.
- Estatura: A estatura é a distância perpendicular entre o plano transverso do vértex e a porção mais inferior dos pés, em metros (m).

ANEXO B – RESULTADOS DOS EXPERIMENTOS USANDO A TÉCNICA DE LEAVE-ONE-OUT

Os melhores resultados para cada experimento usando a técnica de Leave-one-out estão apresentadas na Tabela 35. Os resultados específicos dos 6 experimentos podem ser consultados no link: https://drive.google.com/drive/folders/1EwlKWeMzrhFpf9NSp-ciwXDdZL7VXsYE?usp=sharing

Tabela 35 – Resultados dos experimentos usando a técnica de Leave-one-out.

Experimento	**Algoritmo**	**R^2**	**explicação**
Primeiro Experimento - Caminhada Normal - FR	Lasso	0.92	92,0%
Primeiro Experimento - Caminhada Normal - FY	Lasso	0.91	91,0%
Primeiro Experimento - Caminhada Normal - FX	AdaBoost R.	0.83	83,0%
Segundo Experimento - Caminhada Lenta - FR	Lasso	0.92	92,0%
Segundo Experimento - Caminhada Lenta - FY	Lasso	0.92	92,0%
Segundo Experimento - Caminhada Lenta - FX	Tweedie R.	0.49	49,0%
Terceiro Experimento - Caminhada Rápida - FR	SVR	0.82	82,0%
Terceiro Experimento - Caminhada Rápida - FY	Tweedie R.	0.83	83,0%
Terceiro Experimento - Caminhada Rápida - FX	SVR	0.69	69,0%
Quarto Experimento - Corrida Normal - FR	Ridge CV	0.60	60,0%
Quarto Experimento - Corrida Normal- FY	Tweedie R.	0.60	60,0%
Quarto Experimento - Corrida Normal - FX	SVR	0.52	52,0%
Quinto Experimento - Corrida Lenta - FR	Multi-layer P. R.	0.66	66,0%
Quinto Experimento - Corrida Lenta - FY	Multi-layer P. R.	0.66	66,0%
Quinto Experimento - Corrida Lenta - FX	Lasso	0.56	56,0%
Sexto Experimento - Corrida Rápida - FR	Ridge R.	0.67	67,0%
Sexto Experimento - Corrida Rápida - FY	Ridge R.	0.65	65,0%
Sexto Experimento - Corrida Rápida - FX	Random Forest R.	0.65	65,0%

Fonte: Elaborado pelo autor.