UNIVERSIDADE ESTADUAL DA PARAÍBA

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA EM SAÚDE

PAULO CÉSAR OLIVEIRA BRITO

PLANO DE DESENVOLVIMENTO DE UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E ANÁLISES DE DADOS EM ESTUDOS DE CASOS DE OBESIDADE

CAMPINA GRANDE - PB

2020

PAULO CÉSAR OLIVEIRA BRITO

PLANO DE DESENVOLVIMENTO DE UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E ANÁLISES DE DADOS EM ESTUDOS DE CASOS DE OBESIDADE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como requisito para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Orientador: José Augusto de Oliveira Neto.

CAMPINA GRANDE - PB

2020

B862p Brito, Paulo César Oliveira.
Plano de desenvolvimento de uma ferramenta de aprendizado de máquina para previsão e análises de dados em estudos de casos de obesidade [manuscrito] / Paulo César Oliveira Brito. - 2020.
79 p.
Digitado.
Dissertação (Mestrado em Profissional em Ciência e Tecnologia em Saúde) - Universidade Estadual da Paraíba, Pró-Reitoria de Pós-Graduação e Pesquisa , 2020.
"Orientação : Prof. Dr. José Augusto Oliveira Neto , Departamento de Computação - CCT."
1. Aprendizado de máquina. 2. Inteligência artificial. 3. Obesidade. I. Título
21. ed. CDD 600

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

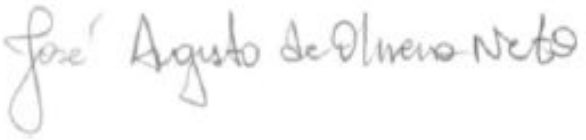
PAULO CÉSAR OLIVEIRA BRITO

PLANO DE DESENVOLVIMENTO DE UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E ANÁLISES DE DADOS EM ESTUDOS DE CASOS DE OBESIDADE

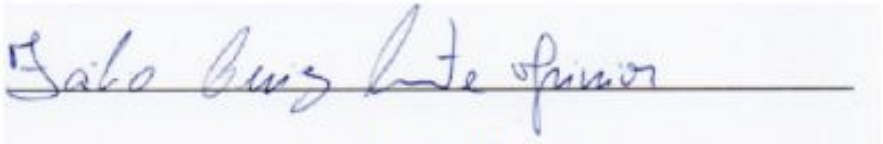
Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como requisito para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Dissertação aprovada em: 29/07/2020

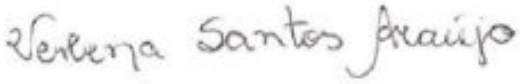
BANCA EXAMINADORA:



Prof. Dr. José Augusto de Oliveira Neto
Universidade Estadual da Paraíba (UEPB)



Prof. Me. Fábio Luiz Leite
Universidade Estadual da Paraíba (UEPB)



Profa. Dra. Verbena Santos Araújo
Universidade Federal do Rio Grande do Norte (UFRN)

RESUMO

Esse trabalho propôs-se a desenvolver uma ferramenta para classificação e análises de dados em estudos de casos de obesidade. A metodologia consiste em etapas de análise dos algoritmos encontrados no estado da arte em aprendizado de máquina até a apresentação de uma proposta de criação de uma ferramenta para o aprendizado de máquina em estudos de casos em obesidade. Através deste experimento e análises, buscou-se observar quais os benefícios da utilização de técnicas de aprendizado de máquina para a descoberta de conhecimento em estudos de casos de obesidade em adolescentes de escolas públicas.

Palavras chave: Aprendizado de Máquina, Inteligência Artificial, Obesidade, Classificação.

ABSTRACT

This work proposed to develop a tool for classification and analysis of data in obesity case studies. The methodology consists of stages of analysis of the algorithms found in the state of the art in machine learning until the presentation of a proposal to create a tool for machine learning in case studies on obesity. Through this experiment and analysis, we sought to observe the benefits of using machine learning techniques for the discovery of knowledge in case studies of obesity in adolescents from public schools.

Keywords: Machine Learning, Artificial Intelligence, Obesity, Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA

IEC - INTERNATIONAL ELECTROTECHNICAL COMMISSION

IMC - ÍNDICE DE MASSA CORPORAL

ISO - INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

KNN - ALGORITMO K-VIZINHOS MAIS PRÓXIMOS

NEPE - NÚCLEO DE ESTUDOS E PESQUISAS EPIDEMIOLÓGICAS

NUTES - NÚCLEO DE TECNOLOGIAS ESTRATÉGICAS EM SAÚDE

OMS - ORGANIZAÇÃO MUNDIAL DE SAÚDE

PIB - PRODUTO INTERNO BRUTO

RF - RANDOM FOREST

SUS - SISTEMA ÚNICO DE SAÚDE

SVM - MÁQUINA DE VETOR DE SUPORTE

UEPB - UNIVERSIDADE ESTADUAL DA PARAÍBA

SUMÁRIO

1 INTRODUÇÃO

Com os avanços em Inteligência Artificial e Aprendizado de Máquina observados ao longo dos últimos anos, verifica-se a importância da utilização destas ferramentas tecnológicas para apoiar o diagnóstico de problemas de saúde. Algumas áreas de saúde já utilizam técnicas de aprendizado de máquina para classificação ou criação de modelos de predição(NGUYEN et. al, 2017), a exemplo do estudo de Basso et. al(2014) foi apresentado um protótipo de aplicação que utiliza Redes Neurais Artificiais com dados do estado patológico do paciente, para apoiar o diagnóstico de Diabetes tipo 2.

Estudos relacionados com a temática diabetes tipo 2 foram publicados nos últimos dois anos, com resultados parciais, onde o foco está na criação de modelos preditivos. Olivera(2016) compara, em seu trabalho, modelos preditivos para detecção de diabetes não diagnosticado utilizando diferentes algoritmos de aprendizado de máquina. Neste estudo, foi desenvolvido um protótipo de plataforma para análise de risco de diabetes, utilizando dados como: histórico médico, medidas antropométricas, escolaridade e hábitos comportamentais. Ribeiro(2009) o qual relata a criação de um sistema de auxílio de diagnóstico de Diabetes tipo 2, baseado nas máquinas de vetor de suporte para uma classe e na codificação eficiente através da análise de componentes independentes, usados para classificar uma base de dados de pacientes em diabéticos e não-diabéticos.

A obesidade está muito relacionada a diversas doenças crônicas. O maior risco é para diabetes mellitus. No Brasil, a prevalência de diabetes em adultos com peso normal/baixo peso é de 5,4%, e na população com obesidade é mais que o dobro (14,0%)(FERREIRA et. al, 2013).

Observa-se, portanto, o interesse em trabalhos relacionados com a Diabetes Mellitus e outras temáticas de relevância significativa na área da saúde. Nesse contexto, vale destacar, por exemplo, preocupação atual das organizações mundiais de saúde em combater problemas de obesidade em todo o mundo. A Organização Mundial de Saúde (OMS) considera a obesidade como uma epidemia mundial condicionada principalmente pelo perfil alimentar e de atividade física (DIAS et. al, 2017).

A obesidade ganhou destaque na agenda pública internacional nas últimas três décadas, caracterizando-se como um evento de proporções globais e de prevalência crescente (DIAS et. al., 2017).

Observando a preocupação mundial em combater a obesidade e, ao mesmo tempo, analisando os avanços expressivos de resultados da aplicação de aprendizado de máquina e inteligência artificial em estudos de casos em saúde, surge a seguinte questão: técnicas de machine learning podem contribuir para o auxílio ao diagnóstico em estudos de casos de obesidade?

O presente estudo está delimitado na criação de um plano de desenvolvimento de uma ferramenta de aprendizado de máquina para atuar no suporte à decisão operacional ao profissional de saúde em estudos de casos de obesidade. Através deste estudo, busca-se observar quais as contribuições para o estado da arte em pesquisas de obesidade utilizando inteligência artificial e aprendizado de máquina, de tal forma que os resultados encontrados possam apoiar decisões médicas e contribuir para o estado da arte em aprendizado de máquina aplicado a estudos de obesidade.

2 HIPÓTESES

Ao longo da execução deste projeto de pesquisa, foi avaliado o estado da arte em aprendizado de máquina aplicado a estudos de obesidade, como também, foi proposto o desenvolvimento de uma ferramenta para apoio à decisão operacional ao profissional de saúde em estudos de casos de obesidade.

A partir desta descrição, foram formuladas as seguintes hipóteses:

H1 - A aplicação de algoritmos de aprendizado de máquina em estudos de casos de obesidade contribui para a decisão operacional do profissional de saúde.

H2 - O suporte visual gráfico contribui para o processo de tomada de decisão em estudos de casos de obesidade.

3 OBJETIVOS

3.1 Objetivo Geral

Desenvolver um plano de desenvolvimento de uma ferramenta de aprendizado de máquina para o apoio a decisões de profissionais de saúde em estudos de obesidade.

3.2 Objetivos Específicos

1. Aplicar técnicas de aprendizado de máquina no contexto de obesidade.
2. Identificar quais as melhores técnicas podem ser aplicadas no contexto de obesidade.
3. Avaliar junto a profissionais de saúde sobre os ganhos encontrados com a tomada de decisão disponibilizada através da ferramenta proposta.

4 REVISÃO DA LITERATURA

4.1 OBESIDADE

Atualmente, a obesidade tornou-se um problema de saúde pública, destacando-se no cenário epidemiológico mundial. Tanto países desenvolvidos como em desenvolvimento, entre eles o Brasil, possuem altos índices de pessoas acima do peso em suas populações(SCHMIDT et. al, 2011).

Sobrepeso e obesidade têm se tornado uma epidemia mundial nos últimos anos, além de se configurar um sério problema de saúde pública. Segundo dados da Organização Mundial de Saúde (OMS), a incidência mundial de obesidade dobrou desde 1980 . Em 2014, 39% dos adultos com mais de 18 anos de idade estavam com sobrepeso e 13% eram obesos(CORREA et. al, 2017).

Existem diferentes maneiras de mensuração da obesidade, sendo o índice de massa corporal (IMC) o principal indicador na avaliação do estado nutricional em adultos. O indicador é obtido por meio da razão entre o peso e o quadrado da altura do indivíduo e, segundo a classificação da Organização Mundial de Saúde (OMS), proposta em 1995, valores

maiores ou iguais a 25 kg/m2 indicam excesso de peso e valores maiores ou iguais a 30,0 kg/m2 caracterizam obesidade(FERREIRA et. al, 2013).

Como pode ser observado, a obesidade é um problema sério que afeta a maioria de toda a população mundial. Nos Estados Unidos, a obesidade já afeta 39,6% da população adulta, de acordo com relatórios governamentais publicados no ano passado (DA COSTA et al, 2015). De 2014 para 2015, o número de adolescentes com obesidade na faixa de 12 a 16 anos de idade, também subiu de 19,1% em 2014-15 para 19,8% (ABESO, 2016).

O Brasil segue essa mesma tendência de aumento da obesidade. Resultados do relatório da Organização das Nações Unidas para Alimentação e Agricultura e da Organização Pan-americana de Saúde (2017) mostram que, em 2010, 17,8% da população brasileira era obesa; passando para 20% em 2014 (DIAS et. al, 2017).

Os custos econômicos com obesidade têm se tornado preocupantes nos últimos anos. O custo de uma doença pode ser medido pelo impacto financeiro no sistema de saúde (custos diretos) e pela perda da produtividade e qualidade de vida (custos indiretos) da sociedade e do indivíduo(BAHIA, 2014). Os problemas de saúde causados pela falta de exercícios físicos diários custaram ao mundo cerca de US$ 67,5 bilhões em 2013 – mais do que o PIB de muitos países (BAILLOT et. al, 2013). No Brasil, o custo total para o SUS, estimado para um ano com todas as doenças relacionadas ao sobrepeso e à obesidade – câncer, diabetes e cardiológicas - é de US$ 20.152.102.171. As hospitalizações custam US$ 1.472.742.952, e os procedimentos de ambulatório, US$ 679.353.348 (ABESO, 2016).

A obesidade está muito relacionada a diversas doenças crônicas. O maior risco é para diabetes mellitus. No Brasil, a prevalência de diabetes em adultos com peso normal/baixo peso é de 5,4%, e na população com obesidade é mais que o dobro (14,0%). Vários estudos mostram que a obesidade também aumenta o risco de hipertensão arterial. Diversos tipos de câncer, como o colorretal, também apresentam forte associação com a obesidade(FERREIRA et. al,2013).

O aumento da prevalência de sobrepeso e obesidade em idades cada vez mais precoces tem despertado a preocupação de pesquisadores e profissionais da área de saúde, em razão dos danos e agravos à saúde provocados pelo excesso de peso, tais como hipertensão arterial, cardiopatias, diabetes, hiperlipidemias, dentre outras (MEDEIROS, 2014).

Uma alimentação não saudável e exercício físico insuficiente são os principais fatores

de risco para a obesidade. Indicadores que medem a frequência de atividade física, tanto no lazer como no trabalho, e o sedentarismo (horas assistidas de televisão por dia) são importantes para avaliar o estilo de vida das pessoas. Vários estudos nacionais e internacionais têm evidenciado uma associação entre horas de televisão assistidas e o excesso de peso e a obesidade na população em geral. O aumento da prevalência de obesidade em diversos países também pode ser explicado por um maior consumo de alimentos não saudáveis, constituindo uma categoria de alimentação chamada fast-food(FERREIRA et. al, 2013).

A adolescência é uma fase de constantes transformações biopsicossociais, tendo a nutrição um papel importante e de grande complexidade. Evidências científicas apontam que os níveis de prevalência de sobrepeso e obesidade nos adolescentes são significativos. Estatísticas recentes do IBGE apontam que um em cada cinco jovens entre 10 e 19 anos apresentam excesso de peso(BREVIDELLI et. al, 2015). Entre os fatores de risco para a obesidade infantil e na adolescência estão o fato de ter pais obesos, a influência dos meios de comunicação, sedentarismo, alimentação inadequada, fatores genéticos, nível socioeconômico, entre outros (OLIVEIRA et. al, 2010).

Através do nosso trabalho, pretendeu-se atuar na busca de contribuições para o auxílio no diagnóstico à obesidade, oferecendo uma proposta de ferramenta que facilitasse o diagnóstico realizado pelo profissional de saúde, considerando todos os dados necessários que classificam um paciente como obeso ou não-obeso.

4.2 APRENDIZADO DE MÁQUINA APLICADO A ESTUDOS DE OBESIDADE

Aprendizado de Máquina é um subcampo da ciência da computação que evoluiu dos estudos sobre reconhecimento de padrões em Inteligência Artificial. Esta definição do termo "aprendizado de máquina" foi realizada por Arthur Samuel em 1959, o qual afirmava que aprendizado de máquina é um campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (SIMON, 2013).

Aprendizado de Máquina corresponde ao conjunto de técnicas que empregam o princípio de indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de exemplos (LORENA, 2007). Esta forma de aprendizado pode ser dividido em aprendizado supervisionado e não-supervisionado.

No aprendizado supervisionado tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por um conjunto de exemplos compostos por uma entrada e sua respectiva saída desejada. Neste caso, o algoritmo de aprendizado de máquina extrai a representação do conhecimento a partir desses exemplos (HAYKIN et al, 1999).

No aprendizado não-supervisionado não existe a presença de um professor e, neste caso, não existem rótulos de saída desejada. O algoritmo de aprendizado de máquina aprende a representar as entradas submetidas seguindo uma medida de qualidade (SOUTO et. al, 2003)

Um algoritmo nada mais é do que uma sequência de passos, capaz de resolver um problema e, consequentemente, gerar algum resultado. Os algoritmos de aprendizado de máquina operam na construção de um modelo a partir de entradas amostrais a fim de fazer previsões ou decisões guiadas pelos dados ao invés de simplesmente seguir instruções programadas (SIMON, 2013).

A partir dos anos 90, o aprendizado de máquina começou a se organizar como um campo separado da Inteligência Artificial, mudando seu foco de abordagens simbólicas herdadas da inteligência artificial para métodos de estatística e teoria da probabilidade(LANGLEY, 2011).

As aplicações para o aprendizado de máquina são inúmeras, as quais se destacam: Bioinformática, jogos de estratégia, detecção de fraude virtual, diagnóstico médico, mecanismos de busca, publicidade online e análises de mercado de ações. Albuquerque(2014) apresenta um estudo utilizando máquinas de vetor de suporte vetorial em finanças com resultados satisfatórios no mercado financeiro de ações, aplicado no contexto nacional. No contexto da saúde, várias pesquisas já aplicaram aprendizado de máquina para classificação e previsões em diversas áreas, tais como: fisioterapia, reconhecimento de padrões em estudos de combate ao câncer e análises de imagens de exames, estudos epidemiológicos, estudos integrados no contexto de Internet das Coisas, entre outros. Por exemplo, Santana et. al(2014) propoem um modelo de raciocínio probabilístico capaz de auxiliar a triagem de pacientes de risco no desenvolvimento de câncer cervical.

Ao longo desta revisão, foram estudados os seguintes algoritmos e suas características:

1. Floresta Aleatória

Florestas aleatórias são um método de aprendizado conjunto para classificação, regressão e outras tarefas que operam na construção de uma multiplicidade de árvores de decisão na fase de treinamento e gerando a classe que é o modo das classes (classificação) ou predição de média das árvores individuais. Em Floresta Aleatória, para classificar um novo objeto com base em atributos, cada árvore dá uma classificação e dizemos a árvore "vote" para essa classe. A floresta escolhe a classificação que tenha a maioria dos votos (FRIZZARINI e LAURETTO, 2013).

Cada árvore é plantada e cultivada como se segue:

- Se o número de casos no conjunto de treinamento é N, então amostra de casos de n é tomado ao acaso, mas com a substituição. Esta amostra será o conjunto de treinamento para o cultivo da árvore.
- Se existem M variáveis de entrada, um número m << M é especificado de modo a que em cada nó, as variáveis m sejam selecionadas aleatoriamente para fora do M e a melhor divisão sobre estes m é usado para dividir o nó. O valor de m seja mantido constante durante o crescimento florestal.
- Cada árvore é cultivada na maior extensão possível. Não há poda.

2. Redes Neurais Artificiais

São modelos computacionais inspirados pelo sistema nervoso central de um animal que são capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões. Redes neurais artificiais geralmente são apresentadas como sistemas de neurônios interconectados, que podem computar valores de entradas, simulando o comportamento de redes neurais biológicas(BASTOS, 2007).

3. Algoritmo k-means

Em mineração de dados, agrupamento k-means é um método de clustering que objetiva particionar n observações dentre k grupos onde cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de

Voronoi. A clusterização k-means tende a encontrar clusters de extensão espacial comparáveis enquanto o mecanismo de maximização da expectativa permite ter diferentes formas (VARGAS, 2012).

4. Algoritmo KNN

O KNN foi proposto por Fukunaga e Narendra em 1975. É um dos classificadores mais simples de ser implementado, de fácil compreensão e ainda hoje pode obter bons resultados dependendo de sua aplicação. A ideia principal do KNN é determinar o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento (MORIM, 2009).

5. Árvore de Decisão

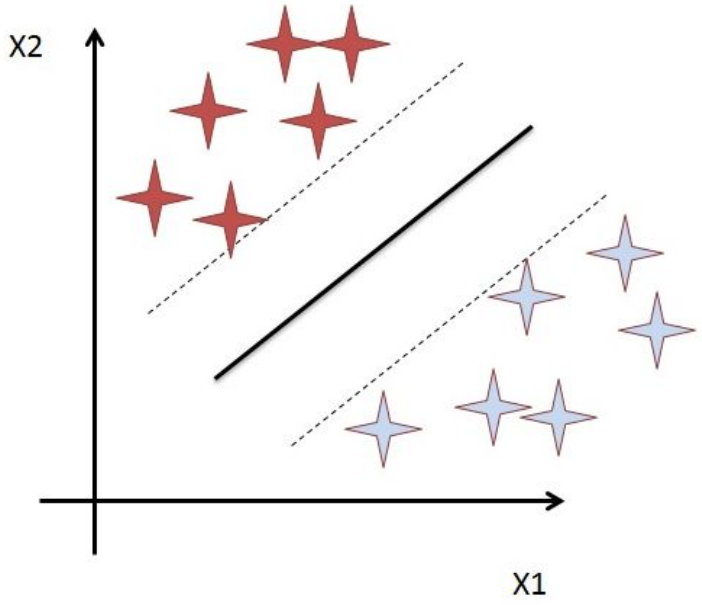
É uma abordagem comportamental que usa diagramas para mapear as várias alternativas e resultados de decisões de investimento, assim como as probabilidades de ocorrerem. Baseia-se em estimativas e probabilidades associadas aos resultados de cursos de ação que competem entre si. Essencialmente, árvores de decisões são diagramas que permitem representar e avaliar problemas que envolvem decisões sequenciais, colocando em destaque os riscos e os resultados financeiros identificados nos diversos cursos de ação (CAMILO e SILVA, 2009).

6. Máquina de Vetor de Suporte(SVM)

Uma máquina de vetores de suporte é um conjunto de métodos do aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados(MONARD e BARANAUSKAS, 2003).

Uma Máquina de Vetor de Suporte, ou SVM, encontra uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes. Essa distância entre o hiperplano e o primeiro ponto de cada classe costuma ser chamada de margem. A Máquina de Vetor de Suporte coloca em primeiro lugar a classificação das classes, definindo assim cada ponto pertencente a cada uma das classes, e em seguida maximiza a margem. Ou seja ela primeiro classifica as classes corretamente e depois em função dessa restrição define a distância entre as margens (NGUYEN, 2017).

Figura 01 - funcionamento da máquina de vetor de suporte



Fonte: https://pt.wikipedia.org/wiki/M%C3%A1quina_de_vetores_de_suporte

7. Árvore de modelo logístico

É uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias (ABREU et. al, 2009).

8. Rede Neural múltiplas camadas

É uma rede neural semelhante à perceptron, mas com mais de uma camada de neurônios em alimentação direta. Tal tipo de rede é composta por camadas de neurônios ligadas entre si por sinapses com pesos. O aprendizado nesse tipo de rede é geralmente feito através do algoritmo de retropropagação do erro (FERNEDA, 2006).

9. Classificador de Bayes

Tipo de classificador estatístico que classifica um determinado elemento em uma determinada classe baseando-se na probabilidade do elemento pertencer a essa classe. É um tipo de aprendizado supervisionado que se baseia no Teorema de Bayes (SANTOS e ALVES, 2008).

10. Hoeffding Tree(VFDT)

Uma árvore Hoeffding (VFDT) é um algoritmo de indução de árvores de decisão incremental, a qualquer momento, capaz de aprender a partir de fluxos de dados massivos, assumindo que os exemplos de geração de distribuição não mudam com o tempo. As árvores de Hoeffding exploram o fato de que uma amostra pequena geralmente pode ser suficiente para escolher um atributo de divisão ideal. Uma característica teoricamente atraente de Hoeffding Trees não compartilhada por outros aprendizes de árvore de decisão incremental é que ela tem garantias sólidas de desempenho (MERCADO et. al, 2017).

11. Rede Bayesiana

Uma rede bayesiana é um modelo gráfico probabilístico que representa um conjunto de variáveis e suas dependências condicionais por meio de um gráfico acíclico direcionado (DAG). Por exemplo, uma rede bayesiana poderia representar as relações probabilísticas entre doenças e sintomas. Dados os sintomas, por exemplo, a rede pode ser usada para calcular as probabilidades da presença de várias doenças (MERCADO et. al, 2017).

4.2.1 AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA

A avaliação de algoritmos de aprendizado de máquina consiste na geração de experimentos para classificação e/ou predição de dados, utilizando técnicas de aprendizado

de máquina, tais como, algoritmos de árvore de decisão, redes neurais, floresta aleatória e k-vizinhos mais próximos(ESTEBAN et. al, 2017). Atualmente, as pesquisas em obesidade que aplicam técnicas de aprendizado de máquina, utilizam modelos de dados bastante distintos, o que torna o estudo do estado da arte complexo, por conta da falta de resultados consistentes ou que tragam evolução para modelos de dados similares, analisados ao longo dos anos(DE GREGORY et. al, 2017).

Os modelos de dados, encontrados através da revisão realizada, tratam sobre o estudo de variantes genéticas, monitoramento da atividade física, dados antropométricos e estudos de microbiana intestinal. Fazil et. al. (2017) tratam sobre uma avaliação da relação entre a glicose e o colesterol nos indivíduos pré-diabéticos quanto à causa do diabetes e doenças cardiovasculares provenientes do diagnóstico da obesidade. Neste estudo, através da aplicação de uma árvore de decisão, foi observado que os dados referentes a idade, níveis elevados de pressão arterial, colesterol e IMC tem maior probabilidade de serem detectados como pré-diabéticos. Porém, pode ser considerado como um estudo de resultados parciais, onde mais análises precisam ser conduzidas para prevalecer resultados de relações de diabetes e doença cardiovascular entre envelhecimento.

Fergus et. al (2016) realizaram uma avaliação do algoritmo de uma rede neural artificial, com o objetivo de realizar a detecção e predição de tipos de atividade física em crianças e sua intensidade associada. Através deste estudo, foi observado que as atividades e conjuntos de recursos produzem melhores resultados quando a previsão é realizada através de uma rede neural artificial denominada Multilayer Perceptron (MLP). DOBBINS et. al (2015) tratam sobre a detecção de atividade física através do processamento de dados no contexto de obesidade. Esta detecção é realizada por algoritmos de aprendizado de máquina que, por sua vez, são avaliados sob o seu desempenho para esta tarefa. Este estudo apresenta melhorias para o estado da arte, em que a classificação dos dados alcança uma precisão de 99%.

Montanez et. al (2017) realizaram uma análise de variantes genéticas para previsão da obesidade, através da execução de algoritmos de aprendizado de máquina, tais como, rede neural, random forest, k vizinhos mais próximos, entre outros, que foram comparativamente avaliados em termos de sua capacidade de identificar as variáveis mais importantes do conjunto de dados genéticos disponíveis, para classificar o sujeito em relação a parâmetros de obesidade. Para o conjunto de dados analisado, a máquina de vetor de suporte gerou um

melhor resultado de classificação, por conseguir utilizar todos os recursos associados a traços da doença durante o treinamento.

Huang et. al (2014) realizaram um estudo com o objetivo de explorar a relação entre fatores de microbiana intestinal e a propensão do sujeito para a obesidade. Com estes dados referentes a bactérias intestinais, foi empregado o algoritmo Random Forest(RF) com eliminação de recurso recursivo para identificar espécies bacterianas que coletivamente geram a melhor capacidade preditiva para distinguir amostras de pessoas obesas e não-obesas. Este é um estudo inicial que não permitiu gerar resultados consistentes, mas que, pode servir como um campo de pesquisa viável em aprendizado de máquina aplicado a questões de obesidade. Por exemplo, na análise realizada, pode-se implementar um modelo híbrido, combinando outras fontes de dados e/ou técnicas a fim de gerar resultados consistentes.

Através da leitura dos artigos mencionados, pode-se verificar a falta de uma padronização na busca e utilização de algoritmos mais eficientes para estudos de caso relacionados à obesidade. Em nosso estudo, pretendeu-se também realizar um estudo de verificação e testagem de algoritmos para a verificação de quais os algoritmos que melhor se comportam na geração de resultados em estudos de caso de obesidade.

4.2.2 DESENVOLVIMENTO DE FERRAMENTAS

Além da avaliação de algoritmos de aprendizado de máquina, alguns trabalhos apresentam o desenvolvimento de ferramentas para aplicação no diagnóstico e intervenções em estudos de obesidade. Estas ferramentas estão cada vez mais integradas com outras tecnologias, tais como, sensores de internet das coisas e análise de redes sociais, com o objetivo de fornecer resultados cada vez mais precisos. Em outras palavras, as ferramentas de aprendizado de máquina utilizam uma base de dados histórica para “aprender” sobre esses dados e, quando uma funcionalidade é executada após esta fase de aprendizado, ao receber novos dados, históricos ou coletados em tempo real, o algoritmo de aprendizado de máquina consegue responder uma informação sobre aquele novo dado analisado, com base no aprendizado realizado anteriormente.

Harous et. al (2017) trata sobre um modelo de monitoramento de obesidade híbrido que se baseia em dados coletados a partir de sensores e dados participativos coletados de uma

comunidade de redes sociais estabelecida para fornecer conscientização sobre obesidade, seu monitoramento e prevenção. Com este modelo, foram realizadas predições sobre os dados através da utilização do algoritmo KNN e, com isso, os resultados da predição geram recomendações que desencadeiam intervenções sempre que necessário.

Nguyen et. al (2017) desenvolveram uma ferramenta para contagem e detecção do número de deglutição durante a ingestão de alimentos. A ferramenta desenvolvida combina sensores internet das coisas com aprendizado de máquina, através da utilização de algoritmos de redes neurais e random forest para classificação de padrões alimentares. As redes neurais são modelos computacionais inspirados pelo sistema nervoso central de um animal que são capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões (BASTOS, 2007). Random Forest é um algoritmo de aprendizado conjunto para classificação, regressão e outras tarefas que operam na construção de uma multiplicidade de árvores de decisão (FRIZZARINI e LAURETTO, 2013). Com esta proposta, a pesquisa realizada tenta reduzir o peso dos participantes, através da verificação de problemas na ingestão incorreta de alimentos.

Julian et. al (2017) apresentam uma ferramenta de triagem automática para o diagnóstico de obesidade em crianças mexicanas. Este diagnóstico possui o objetivo de ajudar na tomada de decisão para comunidades-alvo para programas sociais de prevenção da saúde. Além do desenvolvimento da ferramenta, o trabalho apresenta uma avaliação dos algoritmos de aprendizado de máquina e, demonstra que a árvore de modelo logístico e o algoritmo Simple Logistic foram os melhores classificadores observados ao longo de suas execuções.

4.3 APRENDIZADO DE MÁQUINA APLICADO A TEMAS RELACIONADOS À OBESIDADE - DIABETES MELLITUS

Estudos de avaliação de experimentos utilizando técnicas de aprendizado de máquina são realizados também para o diagnóstico de diabetes mellitus tipo 2, que é uma doença relacionada com a detecção de obesidade em pacientes. Assim, torna-se necessária a verificação do estado da arte para este tema relacionado, visto que existem trabalhos publicados nesta área e que possuem resultados relevantes. Desta forma, como nos artigos

apresentados nas seções anteriores, as pesquisas que envolvem obesidade e diagnóstico de diabetes mellitus possuem modelos de dados distintos, mas convergem no emprego de técnicas de aprendizado de máquina para classificação e/ou predição de dados.

Um modelo híbrido também foi implementado por Dalakleid et. al (2013), utilizando uma combinação de um algoritmo genético com um classificador de vizinhos mais próximos para a previsão do risco de complicações cardiovasculares em pacientes com diabetes tipo 2. Para este estudo, foi verificado que dados como, idade em que o paciente foi diagnosticado, pressão arterial média, glicose, colesterol, sexo, presença de diabetes mellitus familiar e o uso de antagonistas de cálcio, representam o subconjunto que fornece maior eficiência para que o algoritmo classificador forneça um melhor resultado de precisão e desempenho.

Kumar e Pranavi (2017) realizaram uma análise com o objetivo de aprender coisas do conjunto de dados de um Big Data para, prever o tipo de diabetes e verificar o modelo de aprendizado de máquina adequado para fornecer previsões precisas. Desta forma, foi realizado um estudo comparativo das métricas coletadas durante a execução dos algoritmos, tais como, precisão, kappa, sensibilidade e especificidade, e foi visto que o algoritmo Random Forest possui maior precisão para classificar os dados de forma correta. Mercaldo et. al (2017) tratam sobre um método capaz de classificar os pacientes afetados pelo diabetes, utilizando um conjunto de características selecionadas de acordo com os critérios da Organização Mundial de Saúde. A classificação é realizada e comparada em algoritmos de aprendizado de máquina, tais como, Random Forest, Multilayer Perceptron, Hoeffding Tree e BayesNet. Neste estudo, foi observado que o algoritmo Hoeffding Tree obteve uma maior precisão na classificação, correspondente a 85%.

Zhu et. al (2016) realizaram uma avaliação do risco de readmissão de pacientes de diabetes em hospitais. Nesta pesquisa, foram utilizados algoritmos de aprendizado de máquina, tais como, regressão e floresta aleatória, com o objetivo de identificar e priorizar os recursos de risco para a causa da diabetes. Para este estudo, o algoritmo que teve melhor comportamento para classificação dos dados foi o algoritmo floresta aleatória manipulada(Random Forest). Porém, o conjunto de dados utilizado não consegue diferenciar dados para os dois tipos de diabetes e, para nosso foco de estudo, é priorizada a diabetes mellitus por ser uma doença relacionada a obesidade.

Wang et. al (2013) realizaram uma pesquisa de avaliação de uma abordagem de classificação, utilizando uma rede neural artificial, sem a utilização de parâmetros

bioquímicos para identificar o alto risco de diabetes mellitus tipo 2 em adultos. Com este estudo, foram identificados alguns fatores associados ao risco de diabetes mellitus tipo 2, tais como, a idade, sexo, histórico familiar, atividades físicas, taxa de pulso e pressão de pulso.

Nirmala et. al (2013) utilizaram um modelo de algoritmo k-vizinhos mais próximos para classificar bancos de dados de pessoas diabéticas. Neste estudo, o algoritmo utilizado consegue uma precisão de 97.4% na classificação dos dados. Além disso, é feita uma comparação com resultados encontrados na literatura e afirma-se que a precisão obtida com este estudo é um dos melhores resultados já encontrados no estado da arte.

Vaischali et. al (2017) trata sobre uma metodologia para melhorar a previsão do diagnóstico de diabetes tipo 2, através da utilização de algoritmos de aprendizado de máquina. Para isso, foi utilizado um algoritmo de seleção de características genéticas e um modelo de algoritmo Fuzzy Evolutivo. Desta forma, o estudo aponta uma taxa de classificação dos dados com precisão igual a 83.4%.

Lee et. al (2014) em seu trabalho, tratam sobre um estudo preditivo de informações que são utilizadas para o diagnóstico de diabetes tipo 2 através de uma combinação de medidas antropométricas. Os algoritmos de aprendizado de máquina utilizados(Classificador de Bayes e Regressão Logística) obtiveram resultados relevantes e mostraram que a precisão destes resultados foi superior a medidas individuais em homens e mulheres.

5 MATERIAL E MÉTODO

A princípio será apresentado o local de estudo da pesquisa, suas denominações e características. Em seguida, será apresentado o delineamento da pesquisa, explicando o tipo e natureza da pesquisa, bem como, a coleta de dados da pesquisa. Por fim, serão apresentadas as ferramentas adotadas para a análise e o desenvolvimento da pesquisa proposta.

5.1 Local do Estudo

A pesquisa foi desenvolvida no Núcleo de Tecnologias Estratégicas em Saúde (NUTES), situado no CAMPUS I da Universidade Estadual da Paraíba(UEPB). Por se tratar de um estudo de análise de algoritmos e criação de um plano de desenvolvimento de ferramentas tecnológicas, o local de pesquisa se limitou ao ambiente do laboratório NUTES,

pois não foi necessária nenhuma intervenção externa para implementar as etapas de execução do projeto.

5.2 Delineamento da Pesquisa

Para que os objetivos da pesquisa fossem atendidos, este trabalho fundamentou-se em análise exploratória da literatura, através de fontes secundárias e terciárias, tais como artigos entre outras publicações, sobre os aspectos relacionados ao desenvolvimento de ferramentas, técnicas e algoritmos de aprendizado de máquina para utilização no contexto de obesidade, e desenvolvimento de plano de uma ferramenta tecnológica.

O estudo dessas técnicas traz em sua principal característica tentar explicar e entender os benefícios a serem implantados para os pacientes com a utilização desta tecnologia que poderá possibilitar avanços no diagnóstico realizado por profissionais de saúde e na qualidade de vida do usuário.

5.2.1 - Etapas da metodologia

A metodologia da pesquisa realizada foi composta pelas seguintes fases e etapas:

Fase 1 - Etapa para verificação do estado da arte em ferramentas e técnicas de aprendizado de máquina utilizadas em estudos de obesidade e temas relacionados.

- Etapa 1.1 - Análise exploratória da literatura através da utilização de um protocolo para revisão sistematizada.
- Etapa 1.2 - Especificação de critérios para aceitação de algoritmos de aprendizado de máquina.
- Etapa 1.3 - Seleção de algoritmos de aprendizado de máquina de acordo com os critérios estabelecidos.

Fase 2 - Testagem das técnicas de aprendizado de máquina selecionadas no contexto da amostra de dados em análise.

- Etapa 2.1 - Processamento de dados e adequação de acordo com o contexto da utilização de técnicas de aprendizado de máquina.

- Etapa 2.2 - Criação de modelos para utilização em técnicas de aprendizado de máquina selecionadas na fase anterior.
- Etapa 2.3 - Criação de testes para execução das técnicas de aprendizado de máquina.
- Etapa 2.4 - Execução e testagem de técnicas de aprendizado de máquina selecionadas na fase anterior.
- Etapa 2.5 - Verificação de resultados da execução das técnicas através de atributos coletados em cada algoritmo.
- Etapa 2.6 - Execução e testagem de algoritmos não previstos na fase anterior.
- Etapa 2.7 - Verificação de resultados da execução das técnicas implementadas na etapa 2.6.
- Etapa 2.8 - Apresentação de resultados encontrados nesta fase da pesquisa.

Fase 3 - Criação de um plano de desenvolvimento de uma ferramenta para previsão e análise de dados que possa ser utilizado de forma eficiente no contexto de estudos em obesidade a partir de algoritmos de aprendizado de máquina analisados na fase anterior.

- Etapa 3.1 - Criação do plano de desenvolvimento da ferramenta.
- Etapa 3.2 - Aplicação de questionário sobre proposta de ferramenta junto a profissionais de saúde.
- Etapa 3.3 - Análise de dados obtidos com a aplicação de questionário.
- Etapa 3.4 - Produção de documento de resultados da pesquisa.
- Etapa 3.5 - Apresentação dos resultados.

5.3 Protocolo de revisão sistematizada

Para efetivação do processo de revisão do estado da arte dos assuntos abordados nesta pesquisa, foi organizado um protocolo de revisão sistematizada, com o objetivo de permitir uma busca eficiente de trabalhos relacionados com a pesquisa. Os subtópicos a seguir apresentam os processos realizados.

5.3.1 Escopo da Pesquisa

Para realização das buscas, foram utilizadas as bibliotecas digitais IEExplore, Scielo e ScienceDirect, através dos seus mecanismos de busca avançada, utilizando palavras-chave e strings de busca que serão apresentados nos próximos tópicos.

5.3.2 Critérios Adotados para Seleção das Fontes

Por se tratar de um estudo interdisciplinar, envolvendo áreas como Tecnologia da Informação e Saúde, foram selecionadas as bibliotecas digitais que tivessem trabalhos relacionados nestas áreas, em suas bases de dados. Inicialmente, essa verificação foi realizada de forma manual, através da pesquisa por palavras-chave, tais como: "algorithms", "obesity" e "machine learning". Além disso, foi utilizado o critério de seleção de bibliotecas digitais que possuíssem um sistema de busca avançada para expressão de busca personalizada.

5.3.3 Restrições

Esta revisão sistematizada está restrita a análise das publicações, obtidas pelos mecanismos de busca avançada das bibliotecas digitais, escolhidas pelo autor desta revisão sistemática. Para realização das buscas, com o objetivo de verificar o estado da arte, foram utilizados apenas os trabalhos com data de publicação menor que cinco(5) anos.

5.3.4 Idiomas

Os idiomas das publicações aceitos para pesquisa foram inglês e português. O idioma inglês foi considerado pelo fato de ser utilizado como idioma padrão na maioria das bibliotecas digitais que possuem um bom QUALIS. Já o idioma português foi considerado para critério de análise do ambiente brasileiro de pesquisa e suas contribuições para o estado da arte.

5.3.5 Métodos de Busca de Publicações

5.3.5.1 Expressão Geral de Busca

A seguinte expressão será considerada como expressão geral de busca: *"machine learning algorithms" AND classification AND obesity*.

5.3.5.2 Busca Manual

Para a realização de busca manual, foi utilizada a pesquisa nos títulos e resumos(abstracts) dos artigos pelas palavras-chave que compõem a expressão de busca desta revisão sistemática.

5.3.6 Procedimentos de Seleção e Critérios

5.3.6.1 Procedimentos de Seleção

i) Seleção e Catalogação Preliminar das Publicações

Foram selecionados artigos que apresentem informações no título e no abstract/resumo relacionados à questão de pesquisa. Cada publicação foi catalogada em um banco de dados criado especificamente para este fim e armazenada em um repositório para análises posteriores.

ii) 1 Filtro - Seleção das publicações relevantes

Critérios de Inclusão(CI) do Primeiro Filtro(1F). Foi incluída toda publicação que:

- CI1F-01: Define e/ou apresenta instrumentos teóricos e/ou práticos voltados para a análise de algoritmos de aprendizado de máquina aplicados em estudos de caso em obesidade.

- CI1F-02: Investiga, compara e/ou avalia instrumentos teóricos e/ou práticos voltados para a análise de algoritmos de aprendizado de máquina aplicados em estudos de caso em obesidade.
- CI1F-03: Apresenta a aplicação de instrumentos teóricos e/ou práticos voltados para a análise de algoritmos de aprendizado de máquina aplicados em estudos de obesidade.

Critérios de Exclusão(CE) do Primeiro Filtro(1F). Será excluída toda publicação que:

- CE1F-01: Não faz nenhum tipo de avaliação ou análise.
- CE1F-02: Refere-se ao objeto de estudo desta Revisão Sistemática apenas como proposta de trabalho futuro.
- CE1F-03: Não está relacionado com aprendizado de máquina e/ou inteligência artificial.
- CE1F-04: Não está relacionado com questões relativas a estudos de casos em obesidade.

iii) 2 Filtro - Seleção das publicações a partir de critérios de qualidade

Critérios de Exclusão(CE) do Segundo Filtro(2F). Foi excluída toda publicação que:

- Ao executar a busca pela aplicação Start, o Score da publicação não for suficientemente relevante, ou seja, a publicação possui pouca ou nenhuma relação com o artigo, de acordo com as palavras-chave analisadas.

iv) 3 Filtro - Seleção dos dados relevantes

Critérios de Seleção(CS) do Terceiro Filtro(3F). Foi incluída toda publicação que contiver:

- ETP(Elemento Teórico e/ou Prático): informações relacionadas à qual elemento teórico e/ou prático a publicação aborda.
- ANA(Análise): possui foco ou informações sobre a fase de análise de algoritmos de aprendizado de máquina.

- AVA(Avaliação): possui foco ou informações sobre a fase de avaliação de algoritmos de aprendizado de máquina.
- CLA(Classificação): possui foco ou informações sobre classificação utilizando algoritmos de aprendizado de máquina.
- PRE(Predição): aborda questões referentes a predição de resultados utilizando algoritmos de aprendizado de máquina.
- OBE(Obesidade): aborda questões de pesquisas referentes a obesidade.

Caso a publicação contenha todos os critérios supracitados, receberia a seguinte sigla:
CS3F+ ETP+ ANA + AVA + CLA+ PRE + OBE

5.3.7 Procedimentos para extração de dados

5.3.7.1 Seleção e Catalogação Preliminar dos Dados Coletados

Os dados coletados foram catalogados em arquivo contendo uma tabela composta pelos informações a serem extraídas das publicações.

5.3.7.2 Na Seleção dos Dados Relevantes

A seleção dos dados foi realizada através da utilização da ferramenta Start que, por sua vez, permitiu associar cada publicação aos critérios de inclusão, exclusão e qualidade definidos. Desta forma, ao finalizar o processo(3º filtro), finalizou-se a lista de publicações selecionadas, que foram catalogadas utilizando a ferramenta Mendeley.

5.3.7.3 Extração dos Dados

Foram extraídos os seguintes dados das publicações:

- Título: título da publicação analisada;
- Ano: ano em que o trabalho foi publicado;
- Autores: nomes dos autores;
- Veículo: nome do periódico;

- QUALIS: avaliação QUALIS/CAPES do periódico ao qual o trabalho foi publicado;
- Abstract/Resumo: resumo em inglês/português da publicação;
- Objetivo do artigo: qual objetivo do artigo que está sendo analisado;
- Base de Busca: biblioteca digital que a publicação encontra-se armazenada;
- URL: DOI da publicação analisada;
- Técnicas/Abordagens/Métodos: Quais técnicas/abordagens/métodos utilizados no artigo;
- Qual tipo de pesquisa?: Tipo de pesquisa descrita no artigo(teórica/empírica);
- Qual técnica/algoritmo de machine learning empregado na pesquisa do artigo;
- Resultados: quais resultados os autores obtiveram com a pesquisa;
- Limitações: quais limitações foram encontradas pelos autores ao longo do desenvolvimento da pesquisa;

5.3.7.4 Sumarização dos Resultados

Os resultados desta revisão sistemática serão sumarizados na forma de tabelas e gráficos, com o objetivo de gerar um documento de delimitação do escopo do estudo em algoritmos de aprendizado de máquina utilizados em estudos de caso em obesidade.

5.3.8. Procedimentos para Análise

5.3.8.1 Análise Quantitativa

Foram extraídas as seguintes informações quantitativas do montante das publicações:

- Quantidade de publicações selecionadas nas pesquisas manuais;
- Quantidade de publicações selecionadas por cada engenho de busca das bibliotecas digitais;
- Quantidade de publicações duplicadas;
- Quantidade de publicações aceitas e rejeitadas após a execução de cada uma das etapas da revisão sistemática;

- Quantidade de algoritmos de aprendizado de máquina identificados na publicação.

5.3.8.2 Análise Qualitativa

A análise qualitativa desta revisão sistemática utilizou os dados quantitativos e realizar considerações pertinentes ao estudo, com o objetivo de avaliar os achados com relação a questão de pesquisa descrita no início deste protocolo.

5.4 Execução da Revisão Sistematizada

5.4.1 Definição das Palavras Chave e Calibração da Expressão de Busca

Com a definição da questão de pesquisa, foram selecionadas as seguintes palavras-chave que comporam a expressão de busca: obesity, classification e "machine learning algorithms".

A máquina de busca utilizada para o processo de calibração foi a ScienceDirect, disponível em: (https://www.sciencedirect.com/search/advanced).

O processo de calibração da expressão de busca envolveu as seguintes etapas: definição da base de dados de busca para testes de protocolo, identificação de publicações para o grupo de controle, criação da expressão de busca inicial, testes utilizando a expressão de busca e análises dos resultados encontrados com a expressão de busca inicial.

5.4.1.1 Primeira Rodada

Os testes para a primeira rodada foram iniciados com a expressão de busca "machine learning algorithms"AND classification AND obesity. Os resultados obtidos foram agrupados em dois grupos: Grupo 1 - grupo de controle e Grupo 2 - artigos com indícios de aceitação.

Nesta primeira consulta foram retornados 188 resultados, dos quais 8 foram classificados como dentro do grupo de controle(Grupo 1) e 20 outros artigos apresentaram indícios de que estariam no escopo da busca(Grupo 2).

Nesta primeira rodada, foram identificadas mais algumas palavras chave que foram adicionadas na expressão de busca, resultado em: ("machine learning algorithms") AND (classification OR classify) AND (obesity OR obese). Esta nova busca retornou 200 publicações, sendo 9 consideradas no grupo 1 e 22 no grupo 2.

5.4.1.2 Segunda Rodada

Para a segunda rodada, foram adicionadas novas palavras-chave relacionadas a palavra-chave "machine learning algorithms". Desta forma, foi gerada a seguinte expressão de busca: ("machine learning algorithms" OR "intelligent algorithms") AND (classification OR classify) AND (obesity OR obese). Esta nova pesquisa retornou 204 publicações, a qual retornou 10 publicações pertencentes ao grupo de controle - Grupo 1.

5.4.1.3 Terceira Rodada

A terceira rodada deste processo de calibração, foram adicionadas mais algumas palavras-chaves, onde a expressão de busca foi definida como: ("machine learning algorithms" OR "machine learning algorithm" OR "intelligent algorithms") AND (classification OR classifications OR classify OR classifying) AND (obesity OR obese OR obesity disease). Esta nova pesquisa retornou 208 publicações, a qual retornou 10 publicações pertencentes ao grupo de controle - Grupo 1.

5.4.2 Definição das Máquinas de Busca

A pesquisa foi realizada utilizando as bibliotecas IEExplore, ScienceDirect e Scielo. Foram utilizadas as seguintes expressões de buscas para suas respectivas bibliotecas digitais:

Expressão de Busca na Biblioteca Digital da IEExplore

(("machine learning algorithms" OR "machine learning algorithm" OR "intelligent algorithms") AND (classification OR classifications OR classify OR classifying) AND (obesity OR obese OR obesity disease))

Expressão de Busca na Biblioteca Digital da ScienceDirect

("machine learning algorithms" OR "machine learning algorithm" OR "intelligent algorithms") AND (classification OR classifications OR classify OR classifying) AND (obesity OR obese OR obesity disease)

Expressão de Busca na Biblioteca Digital da Scielo

("machine learning algorithms" or "algoritmos de aprendizado de máquina" or "algoritmos inteligentes" or "machine learning algorithm" or "intelligent algorithms") [All indexes] and (classification or classificação or classifications or classify or classifying) [All indexes] and (obesity or obesidade or obese or obesity disease) [All indexes]

5.4.3 Instrumento para Consulta Manual

Para a realização da consulta manual, devem ser pesquisadas as palavras-chave, listadas abaixo, nos títulos e resumos/abstracts dos artigos e registrar os dados na forma de tabela, contendo as seguintes informações: artigo, autores, ano, idioma e palavras-chave correspondentes.

Palavras-chave da string de busca a serem utilizadas para esta consulta manual:

- Machine Learning Algorithms, Intelligent Algorithms, Algoritmos de Aprendizado de Máquina;
- Classification, Classify, Classifying, Classificação;
- Obesity, Obesidade;

5.4.4 Identificação do Período de Busca

O período de busca considerado para esta revisão sistematizada foi considerado como os últimos cinco(5) anos, ou seja, serão consideradas publicações realizadas no intervalo 2013-2018.

5.5 Coleta de Dados

Serão utilizados dados reais para a realização de experimentos com as técnicas de machine learning, para ser aplicada em estudos de casos de obesidade. Estes dados são fruto de trabalho do Núcleo de Estudos e Pesquisas Epidemiológicas(NEPE) da Universidade Estadual da Paraíba(UEPB), financiado pelo CNPQ, vinculado ao Mestrado em Saúde Pública, disponibilizados pela Professora Dra. Carla Campos Muniz Medeiros e a Professora Dra. Danielle Franklin Carvalho.

Assim, os dados utilizados foram secundários oriundos de um estudo transversal realizado em escolas públicas estaduais em Campina Grande, PB, Brasil, com 579 adolescentes de 15 a 19 anos, investigando aspectos socioeconômicos; demográfico; estilo de vida; e variáveis clínicas, coletados no ano de 2013. Os dados foram coletados por meio de um formulário validado, abrangendo dados de antropometria; pressão sanguínea; medições; e testes laboratoriais. Em sua maioria, os estudantes que responderam a pesquisa foram considerados não-obesos. Assim, pode-se refletir inicialmente que os algoritmos podem traçar um perfil melhor para estudantes não-obesos e isso, em algum momento, poderia interferir também na classificação de novos dados de estudantes e/ou pacientes a serem analisados durante a utilização da ferramenta proposta. Porém, o processo de testagem de diferentes algoritmos permitiu verificar as melhores técnicas que conseguiram classificar de forma mais eficiente os dados de pacientes obesos e não-obesos, considerando as possíveis limitações dos dados.

Os dados coletados foram analisados de acordo com padrões de classificação de dados já fornecidos pelo estado da arte em pesquisas de avaliação nutricional e estudos de obesidade. Vale salientar que, para classificação é definido um Escore-z como instrumento de medida que avalia se o paciente possui baixo peso, peso ideal, sobrepeso ou obesidade. Escore-z (FEIJÓ et. al, 1997) é um termo estatístico que quantifica a distância do valor observado em relação à mediana dessa medida ou ao valor que é considerado normal na população. Além deste instrumento de medida, a literatura também apresenta a definição de um Percentil para classificação e diagnóstico do estado nutricional. Percentil (BAUM et. al, 2017) é um termo estatístico e refere-se à posição ocupada por determinada observação no interior de uma distribuição.

A tabela abaixo, apresentada por Feijó et. al(1997), apresenta as equivalências entre esses dois instrumentos de medidas:

Escore-z	Percentil	Interpretação
-3	0,1	Espera-se que em uma população saudável sejam encontradas 0,1% das crianças abaixo desse valor.
-2	2,3	Espera-se que em uma população saudável sejam encontradas 2,3% das crianças abaixo desse valor. Convenciona-se que o equivalente ao escore-z -2 é o percentil 3.
-1	0,1	Espera-se que em uma população saudável sejam encontradas 15,9% das crianças abaixo desse valor.
0	50,0	É o valor que corresponde à média da população, isto é, em uma população saudável, espera-se encontrar 50% da população acima e 50% da população abaixo desse valor.
+1	84,1	Espera-se que em uma população saudável sejam encontradas 84,1% das crianças abaixo desse valor, ou seja, apenas 15,9% estariam acima desse valor. Convenciona-se que o equivalente ao escore-z +1 é o percentil 85.
+2	97,7	Espera-se que em uma população saudável sejam encontradas 97,7% das crianças abaixo desse valor, ou seja, apenas 2,3% estariam acima desse valor. Convenciona-se que o equivalente ao escore-z +2 é o percentil 97.
+3	99,9	Espera-se que em uma população saudável sejam encontradas 99,9% das crianças abaixo desse valor, ou seja, apenas 0,1% estariam acima desse valor.

Tabela 01 - Equivalências entre Escore-z e Percentil

Através das definições das equivalências de valores para classificação, utilizando Escore-z e Percentil, as seguintes definições são apresentadas em Da Costa et. al(2017) para o diagnóstico do estado nutricional:

VALORES CRÍTICOS		DIAGNÓSTICO NUTRICIONAL
< Percentil 0,1	< Escore-z -3	Magreza acentuada
≥ Percentil 0,1 e < Percentil 3	≥ Escore-z -3 e < Escore-z -2	Magreza
≥ Percentil 3 e ≤ Percentil 85	> Escore-z -2 e ≤ Escore-z +1	Eutrofia
> Percentil 85 e ≤ Percentil 97	> Escore-z +1 e ≤ Escore-z +2	Sobrepeso
> Percentil 97 e ≤ Percentil 99,9	> Escore-z +2 e ≤ Escore-z +3	Obesidade
> Percentil 99,9	> Escore-z +3	Obesidade grave

Tabela 02 - Pontos de corte de IMC-para-idade para crianças dos 5 aos 10 anos.

O banco de dados a ser utilizado apresentou algumas variáveis que foram consideradas relevantes para a fase de experimento dos dados na ferramenta, de tal forma que esses dados possam gerar predições de obesidade e condições associadas ao estado nutricional. Podemos citar, por exemplo, as seguintes variáveis: totafis - total de minutos em atividades físicas durante o dia; hrsono - total de horas de sono; IMC; pnascg - peso ao nascer; idade; sexo; hrtv - total de minutos assistindo TV por dia; hrgames - total de minutos jogando videogame por dia; hrpc - total de minutos utilizando computador por dia; hrsedent - acumulado de horas sedentárias (hrtv + hrgames + hrpc).

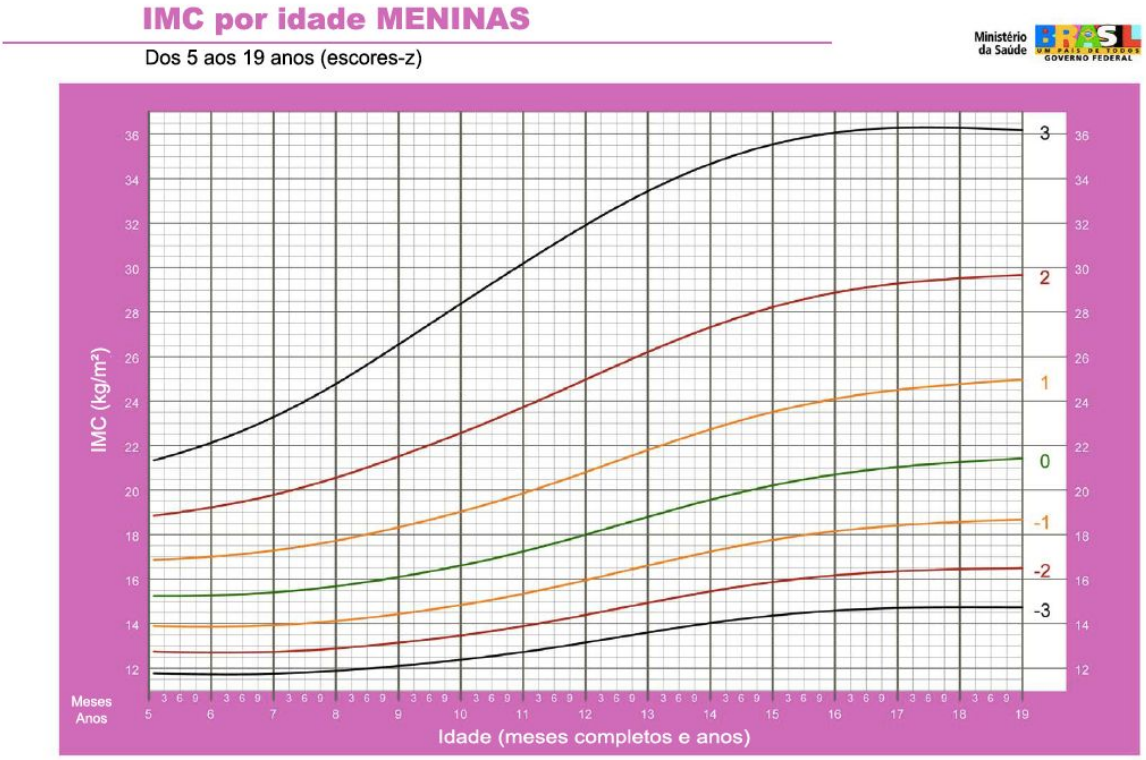
O estado da arte em avaliação nutricional e estudos de casos de obesidade, apresentou também, algumas classificações relevantes que serviram para o aprendizado de máquina da ferramenta a ser desenvolvida (DIAS et. al, 2017). Dentre as classificações, podemos citar a classificação do paciente quanto às atividades físicas praticadas, classificação do sedentarismo e a classificação do sono (BAUM et al, 2017).

1. Classificação do paciente quanto às atividades físicas praticadas:
    a. Inativo = 0 minutos
    b. Insuficientemente ativo I = 1 a 149 minutos
    c. Insuficientemente ativo II = 150 a 229 minutos
    d. Ativo = 300 minutos ou mais
2. Classificação do sedentarismo:
    a. Sedentário: 2 horas ou mais

b. não sedentário: <2 horas

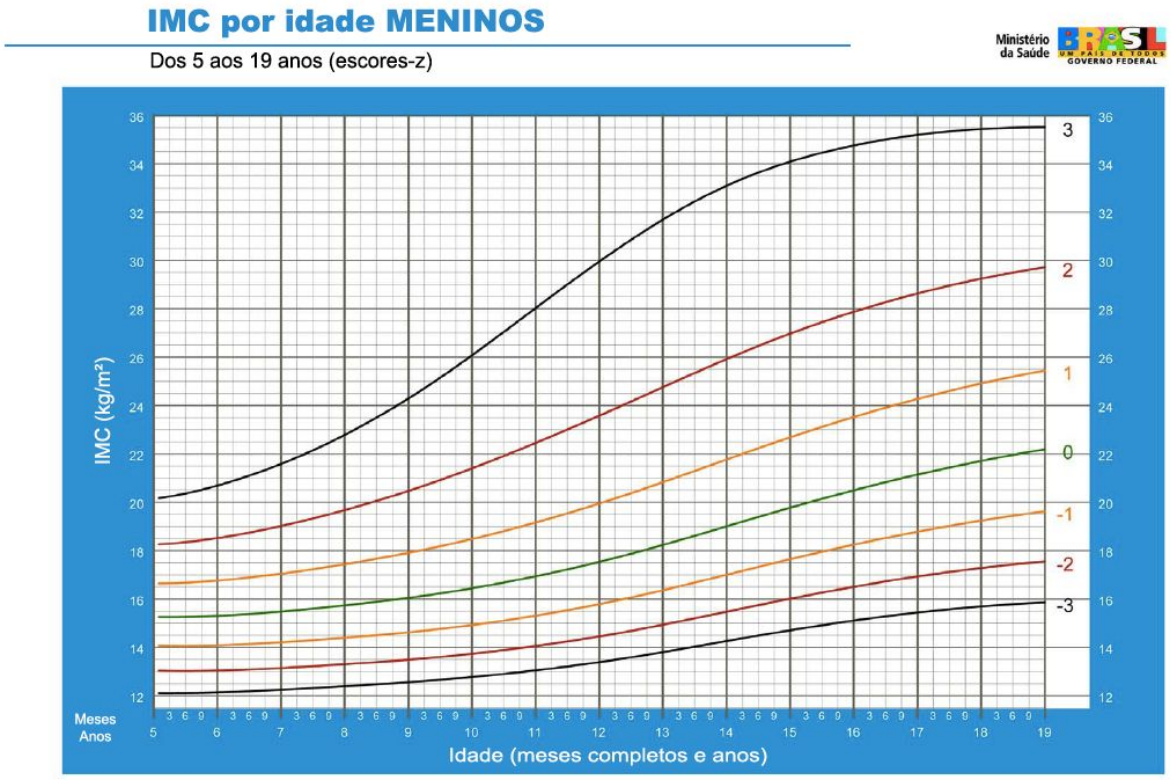
3. Classificação do sono:
    a. Curto: < 8 horas
    b. Média: 8-10 horas
    c. Longo: >= 10 horas

Os gráficos abaixo apresentam o comportamento do Escore-z, apresentado nas Tabelas 01 e 02, relacionado à medida do IMC e idade de meninos e meninas, dos 5 aos 19 anos.



Fonte: WHO Growht reference data for 5-19 years, 2007 (http://www.who.int/growthref/en/)

Figura 02 - IMC por idade meninas



Fonte: WHO Growht reference data for 5-19 years, 2007 (http://www.who.int/growthref/en/)

Figura 03 - IMC por idade meninos

5.6 Tecnologias para análises de dados e execução das técnicas de aprendizado de máquina

Para realização das análises dos algoritmos de aprendizado de máquina e validação das técnicas com a utilização da base de dados escolhida, as seguintes tecnologias foram utilizadas:

5.6.1 R Studio

R Studio é um software livre de ambiente de desenvolvimento integrado para R, uma linguagem de programação para gráficos e cálculos estatísticas e análises de dados.

5.6.2 Plataforma Collaboratory

Plataforma Google para codificação de programas em Python. Mais informações sobre esta plataforma no ANEXO III deste documento.

5.6.3 Library Caret

O pacote Caret é um conjunto de funções que tentam simplificar o processo de criação de modelos preditivos. Este pacote contém ferramentas para pré-processamento, divisão de dados, estimativa de importância de uma variável, bem como outras funcionalidades.

6 RESULTADOS

6.1 Critérios para Seleção de Algoritmos

Ao longo do processo de revisão sistematizada, foram vários os algoritmos analisados sob o contexto de pesquisas em obesidade. Os critérios definidos inicialmente e de prioridade alta para o processo de seleção foram os seguintes:

Critério 1 - Utilização no contexto de obesidade;
Critério 2 - Atualidade da pesquisa em que o algoritmo foi aplicado - foram consideradas pesquisas com até oito(8) anos da data de sua publicação;

A segunda fase de definição de critérios para seleção de algoritmos corresponde a uma separação entre os tipos de algoritmos que realizam tarefas de classificação e predição. Nesta etapa, foram descartados os algoritmos que realizam apenas a tarefa de classificação de dados, visto que o estudo realizado apresenta um plano de desenvolvimento de uma ferramenta que possui funcionalidades de previsão de informações. Assim, foi definido o seguinte critério e organizado na Tabela 03 as informações de cada algoritmo e sua respectiva classificação de acordo com o critério estabelecido:

Critério 3 - Algoritmo deve realizar tarefas de predição de dados;

Tabela 03 - Classificação dos algoritmos em relação ao critério 3

Algoritmo	Classificação do Critério 3
Floresta Aleatória	Classificação e Predição

Redes Neurais Artificiais	Predição
Algoritmo k-means	Predição
Algoritmo KNN	Classificação
Árvore de Decisão	Predição
Máquina de Vetor de Suporte(SVM)	Classificação e Predição
Árvore de modelo logístico	Predição
Rede Neural múltiplas camadas	Predição
Classificador de Bayes	Classificação
Hoeffding Tree	Predição
Rede Bayesiana	Predição

O último critério definido consistiu na verificação da precisão do resultado da execução de cada algoritmo no contexto em que foi aplicado. Sendo assim, foi definido o critério na seguinte forma:

Critério 4 - Algoritmo deve ter precisão nos resultados das pesquisas com uma taxa maior que 90%;

Após a verificação deste último critério, os seguintes algoritmos foram filtrados:

- Floresta Aleatória com precisão igual a 92%;
- Máquina de Vetor de Suporte com precisão igual a 95%;

6.2 Algoritmos Selecionados

Após a definição dos critérios e validação com as pesquisas encontradas através da revisão sistematizada, foi definida a utilização dos Algoritmos Floresta Aleatória e Máquina de Vetor de Suporte para o contexto da pesquisa.

6.3 Critérios para Seleção de Variáveis para o Aprendizado

Várias variáveis foram coletadas de pacientes ao longo da pesquisa realizada e, precisaram ser selecionadas para obtermos um resultado mais eficiente no processo de tomada de decisão através da utilização da proposta de ferramenta que foi desenvolvida. Desta forma, foram definidos os seguintes critérios para seleção de variáveis:

Critério 1 - Variável a ser utilizada apresentou resultados expressivos em trabalhos publicados pelo grupo de pesquisa;
Critério 2 - Variável a ser utilizada apresentou resultados expressivos em trabalhos publicados no estado da arte em pesquisas de aprendizado de máquina aplicado a obesidade.

6.4 Variáveis selecionadas

Considerando os critérios definidos no tópico anterior, as seguintes variáveis do banco de dados foram selecionadas para a realização de experimentos com a proposta de ferramenta que foi desenvolvida:

- sexo - sexo do paciente;
- idade - idade do paciente na data da coleta de dados;
- escore z - estado nutricional do paciente;
- imc - índice de massa corporal;
- pnascg - peso do paciente ao nascer;
- hrsono - total de horas diárias de sono;
- altura - altura do paciente;
- peso - peso do paciente;
- hrtv - total de minutos assistindo TV por dia;

- hrgames - total de minutos jogando vídeo game por dia;
- hrpc - total de minutos na frente do computador por dia;
- hrsedent - acumulado de horas sedentárias.
- pressão arterial sistólica;
- pressão arterial diastólica;
- glicemia em jejum;
- hemoglobina glicada;
- proteína c reativa ultrassensível;
- espessamento da carótida;
- histórico de obesidade familiar;
- histórico de diabetes familiar;
- histórico de doença cardiovascular familiar;
- hdl - taxa de colesterol hdl;
- ldl - taxa de colesterol ldl;
- totafis - total de horas de atividades físicas por semana;
- triglic - taxa de triglicerídeos do paciente;
- colesterol total;
- ca - circunferência abdominal;

6.5 Etapas de verificação da técnicas de aprendizado de máquina

A metodologia de verificação das técnicas de aprendizado de máquina obedeceu à sequência das seguintes etapas apresentadas na Figura 05, as quais são: Processar dados, gerar modelos, gerar testes, implementar algoritmos, executar algoritmos e apresentar os resultados obtidos. Seguindo esta sequência, o processo de execução da pesquisa foi finalizado.

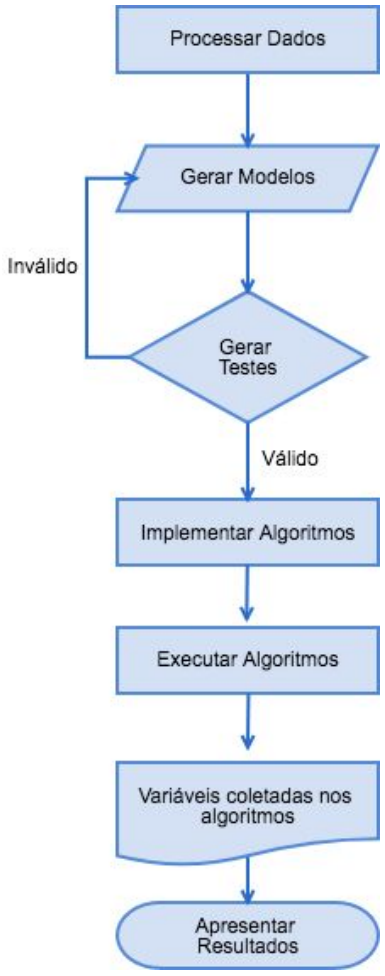


Figura 05 - Etapas de verificação

6.5.1 Processamento, limpeza e leitura dos dados

Em posse da base de dados que seria analisada, alguns procedimentos foram realizados para prosseguir na implementação das técnicas de machine learning. As seguintes etapas deste processo foram executadas:

- Inicialmente, foram retiradas todas as variáveis que não foram consideradas interessantes para análise de dados. Apenas as variáveis selecionadas de acordo com critérios estabelecidos e analisados por especialista dos dados(Tópico referenciar) permaneceram na base de dados.
- Ao executar uma proposta inicial de modelo para os dados, algumas variáveis geraram multi classificações e, com isso, afetariam o resultado da execução das técnicas. Sendo assim, as seguintes variáveis foram excluídas do processo: "hrsedcal", "hrsedent", "escorez", "escorez1".

- A variável "escorez" apresentava um diagnóstico sobre o estado nutricional de cada paciente. Essa classificação previamente realizada por especialista dos dados permitiu agilidade durante o processo de classificação pelo especialista de dados. Como os resultados geraram multi classificações, foi criada uma variável denominada "classifica" que apresenta a classificação nutricional para "obeso" ou "não-obeso" de acordo com os critérios estabelecidos para a análise do Escore-Z(Função peso) de cada paciente.

6.5.2 Preparação da plataforma para processamento de dados

Inicialmente, na plataforma a ser utilizada para a análise dos dados, R Studio, as bibliotecas a serem utilizadas são importadas no documento e os dados deverão ser carregados. A figura abaixo apresenta a sequência de códigos que foram executados para a fase de preparação da plataforma e carregamento inicial dos dados.

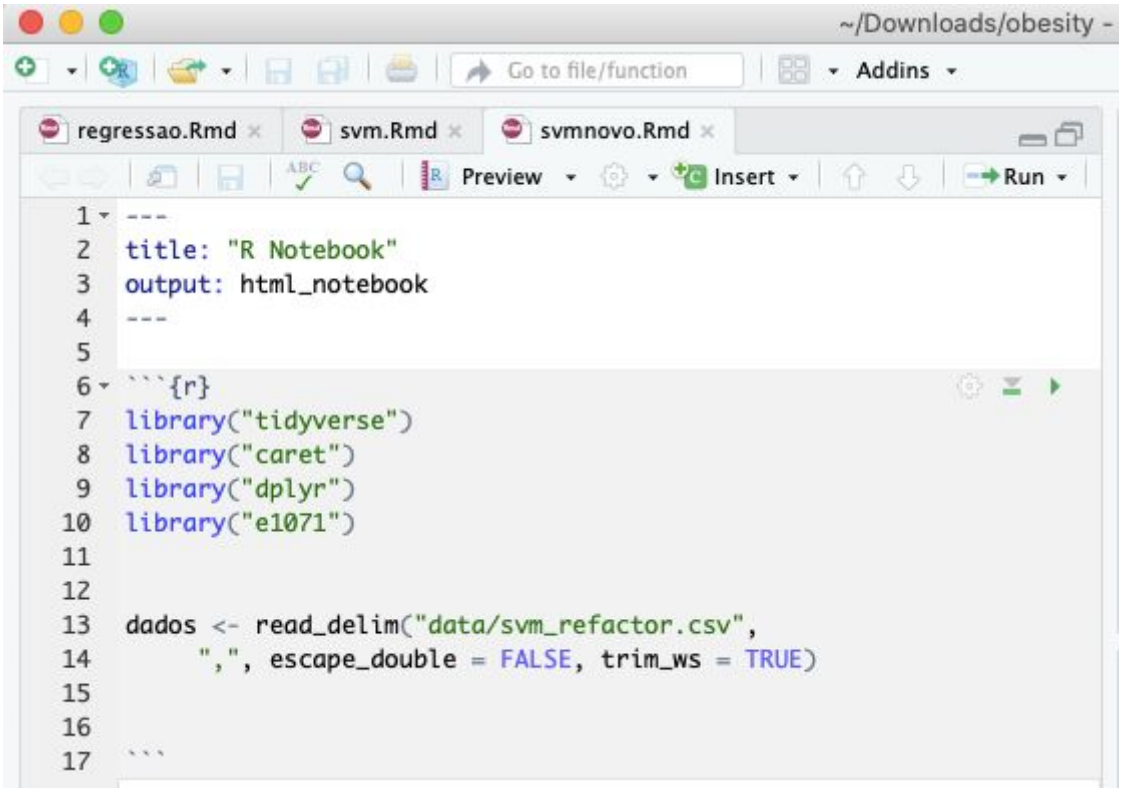


Figura 05 - Carregamento inicial dos dados

6.5.3 Gerar modelo de dados e testes de classificação

Após a execução da etapa inicial de processamento, limpeza e leitura dos dados, uma etapa importante no processo de análise de dados é a criação de um modelo de dados que possa ser interpretado por algoritmos de aprendizado de máquina. Para realização desta etapa, foi utilizada a biblioteca "Caret" disponível na ferramenta R Studio. Inicialmente, os dados foram divididos em conjuntos para treinamento e teste, usando 70% para treinamento. O código abaixo nos mostra como o modelo e os conjuntos para testes foram implementados:

```
18
19 ```{r}
20 treino_idx <- createDataPartition(y=dados$classifica, p=.7,
   list=FALSE)
21 treino <- dados[treino_idx,]
22 teste <- dados[-treino_idx,]
23 names(treino) <- names(dados)
24 names(teste) <- names(dados)
25 ```
```

Figura 06 - Divisão dos dados para testes

Após a criação do modelo de dados e conjuntos para testes, foi criada uma função que facilitasse o processo de análise de dados, em que, ao inserir o nome do algoritmo classificador como entrada da função, a mesma retornaria algumas informações importantes sobre a execução do algoritmo no conjunto de testes implementado. O código abaixo nos mostra como a função foi implementada:

```r
treinaEAvalia <- function(metodo) {
  set.seed(23)
  modelo <- train(classifica~., data=treino, method=metodo,
trControl=trainControl(method="cv", number=10), prox=TRUE,
allowParallel=TRUE)
  #print(modelo$finalModel)
  predicoes <- predict(modelo, newdata=teste)
  print(confusionMatrix(table(data=predicoes, teste$classifica)))
  precisao <- posPredValue(predicoes, teste$classifica)
  recuperacao <- sensitivity(predicoes, teste$classifica)
  f1 <- (2 * precisao * recuperacao) / (precisao + recuperacao)
  print(precisao)
  print(recuperacao)
  print(f1)
  return(f1)
}
```

Figura 07 - Implementação de função

Para utilizar a função implementada e apresentada na figura anterior, o seguinte código foi executado:

```r
f1_rf <- treinaEAvalia("rf")
```

Figura 08 - Execução da função implementada

Como pode ser observado, a função treinaEAvalia recebe como parâmetro o nome ou abreviação do método de aprendizado de máquina que será utilizado. No caso da figura acima foi utilizada a método referente ao algoritmo Random Forest. Ao executar à função implementada, à seguinte saída foi apresentada:

```r
```{r}
f1_rf <- treinaEAvalia("rf")
```
```

```
Confusion Matrix and Statistics


data        nao_obeso obeso
  nao_obeso       123     5
  obeso             6    27

               Accuracy : 0.9317
                 95% CI : (0.881, 0.9654)
    No Information Rate : 0.8012
    P-Value [Acc > NIR] : 3.257e-06

                  Kappa : 0.788

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9535
            Specificity : 0.8438
         Pos Pred Value : 0.9609
         Neg Pred Value : 0.8182
             Prevalence : 0.8012
         Detection Rate : 0.7640
   Detection Prevalence : 0.7950
      Balanced Accuracy : 0.8986

       'Positive' Class : nao_obeso
```

Figura 09 - Execução do Random Forest

O resultado da função é um conjunto de informações cruciais à respeito da execução do algoritmo Random Forest sobre a massa de dados em estudo. Inicialmente, foi apresentada uma matriz de confusão como informações relativas a classificação do dados em obeso e não-obeso. Em sequência, foram calculados atributos importantes para a análise dos algoritmos, como a sua acurácia, sensitividade e especificidade.

Para uma melhor visualização de informações sobre os métodos de aprendizado de máquina utilizados neste trabalho, o seguinte link pode ser acessado que trata sobre à biblioteca utilizada: http://topepo.github.io/caret/available-models.html

6.5.4 Comparação entre algoritmos

Esta fase do processo de análise de dados utilizando técnicas de aprendizado de máquina, corresponde a execução dos algoritmos no modelo de dados e testes criados a fim de verificar quais algoritmos poderão ser utilizados no contexto de estudos de caso em obesidade. O processo de comparação entre os algoritmos obedeceu a metodologia e planejamento do experimento, utilizando os algoritmos Random Forest e Máquina de Vetor de Suporte(SVM). Após a realização desta análise comparativa, também foram executados outros algoritmos que não estavam previstos no escopo da pesquisa, a fim de que pudéssemos verificar a possibilidade de inserção de outras técnicas de aprendizado de máquina no contexto de estudos de caso em obesidade. Esses algoritmos executados foram considerados nessa segunda etapa de acordo com a visualização de sua aplicabilidade em trabalhos relacionados e citados ao longo deste trabalho.

A seguinte execução, apresentada nas figuras abaixo, foi realizada para todos os algoritmos selecionados nesta pesquisa. Através dos resultados apresentados com as variáveis coletadas, foi possível realizar comparações consistentes entre os algoritmos analisados.

```r
f1_logreg <- treinaEAvalia("LogitBoost")
```

```
Confusion Matrix and Statistics


data         nao_obeso obeso
  nao_obeso        120     5
  obeso              9    27

               Accuracy : 0.913
                 95% CI : (0.8584, 0.9516)
    No Information Rate : 0.8012
    P-Value [Acc > NIR] : 8.508e-05

                  Kappa : 0.7392

 Mcnemar's Test P-Value : 0.4227

            Sensitivity : 0.9302
            Specificity : 0.8438
         Pos Pred Value : 0.9600
         Neg Pred Value : 0.7500
             Prevalence : 0.8012
         Detection Rate : 0.7453
   Detection Prevalence : 0.7764
      Balanced Accuracy : 0.8870

       'Positive' Class : nao_obeso
```

Figura 10 - Execução do algoritmo LogicBoost

```r
```{r}
f1_knn <- treinaEAvalia("kknn")
```
```

```
Confusion Matrix and Statistics


data        nao_obeso obeso
  nao_obeso       118    25
  obeso            11     7

               Accuracy : 0.7764
                 95% CI : (0.7041, 0.8382)
    No Information Rate : 0.8012
    P-Value [Acc > NIR] : 0.81411

                  Kappa : 0.1598

 Mcnemar's Test P-Value : 0.03026

            Sensitivity : 0.9147
            Specificity : 0.2188
         Pos Pred Value : 0.8252
         Neg Pred Value : 0.3889
             Prevalence : 0.8012
         Detection Rate : 0.7329
   Detection Prevalence : 0.8882
      Balanced Accuracy : 0.5667

       'Positive' Class : nao_obeso
```

Figura 11 - Execução do algoritmo KNN

```
```{r}
f1_svm <- treinaEAvalia("svmLinearWeights")
```
```

```
Confusion Matrix and Statistics


data        nao_obeso obeso
  nao_obeso       123    17
  obeso             6    15

               Accuracy : 0.8571
                 95% CI : (0.7934, 0.9072)
    No Information Rate : 0.8012
    P-Value [Acc > NIR] : 0.04261

                  Kappa : 0.4849

 Mcnemar's Test P-Value : 0.03706

            Sensitivity : 0.9535
            Specificity : 0.4688
         Pos Pred Value : 0.8786
         Neg Pred Value : 0.7143
             Prevalence : 0.8012
         Detection Rate : 0.7640
   Detection Prevalence : 0.8696
      Balanced Accuracy : 0.7111

       'Positive' Class : nao_obeso
```

Figura 12 - Execução do algoritmo SVM

6.5.4.1 Algoritmos selecionados para o escopo da pesquisa

Como informado no item 6.2, os algoritmos selecionados para o escopo da pesquisa, com base em critérios estabelecidos, foram os algoritmos Random Forest e Máquina de Vetor de Suporte(SVM). Ao serem executados com o modelo de dados implementado e apresentado nos tópicos anteriores, os seguintes apresentados nas Tabelas 4 e 5 foram obtidos:

Random Forest	
Acurácia	93%
Sensibilidade	95%
Especificidade	84%

Tabela 04 - Resultados Random Forest

Máquina de Vetor de Suporte(SVM)	
Acurácia	85%
Sensibilidade	95%
Especificidade	46%

Tabela 05 - Resultados SVM

Como pode ser observado, Random Forest possui uma acurácia maior se comparado com à execução do algoritmo SVM, isso já nos mostra indícios de que à estratégica do RF é bem mais eficiente.

6.5.4.2 Algoritmos não-selecionados para o escopo da pesquisa

Como segue o protocolo metodológico apresentado, foram executados os algoritmos SVM e RF apresentados no tópico anterior e verificadas suas particularidades em relação ao teste no banco de dados verificado. Porém, como forma de complementação para esta pesquisa, os algoritmos Regressão Logística, Naive Bayes e KNN foram executados na base de dados, com o objetivo de verificar se algum dos algoritmos apresentou comportamento superior ao SVM ou RF.

Regressão Logística	
Acurácia	91%
Sensibilidade	93%
Especificidade	84%

Tabela 06 - Resultados Regressão Logística

Naive Bayes	
Acurácia	88%
Sensibilidade	95%
Especificidade	84%

Tabela 07 - Resultados Naive Bayes

KNN	
Acurácia	77%
Sensibilidade	91%
Especificidade	21%

Tabela 08 - Resultados KNN

Como está apresentado nas Tabelas 6, 7 e 8, pode ser observado que o algoritmo de Regressão Logística possui uma boa acurácia, igual à 91%, podendo também ser utilizado para futuros testes.

6.5.5 Ranking entre técnicas de aprendizado de máquina

Como forma de visualização comparativa entre as técnicas de aprendizado de máquina utilizadas, podemos elencar em formato de ranking o comportamento dos algoritmos na execução dos testes, assim obtivemos a seguinte classificação:

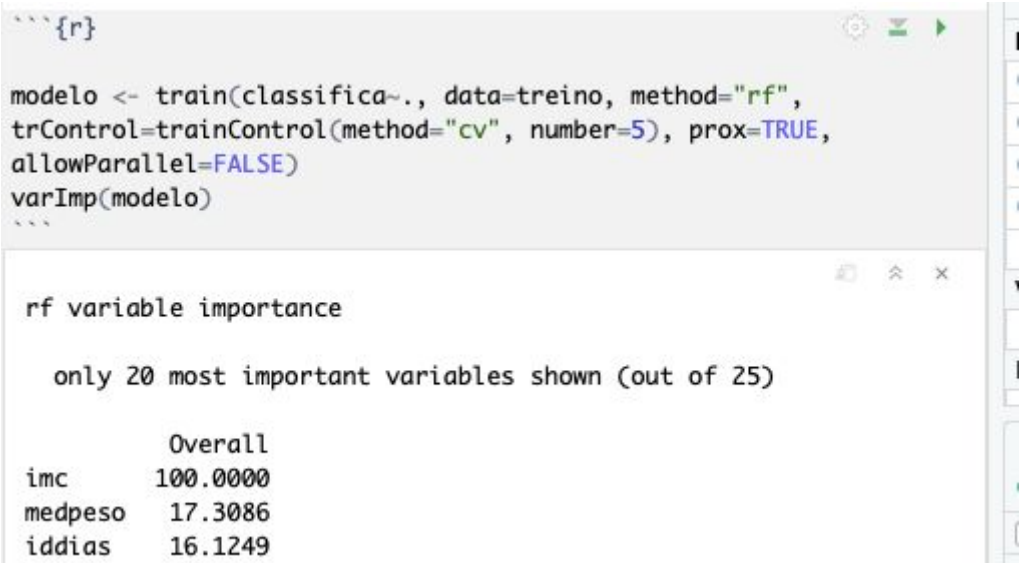
1 - Random Forest;
2 - Regressão Logística;
3 - Naive Bayes;
4 - Máquina de Vetor de Suporte(SVM);
5 - KNN;

Para efetivação do ranking acima, foram considerados as variáveis coletadas nos algoritmos, apresentadas nos tópicos anteriores, que são: acurácia, sensibilidade e especificidade.

6.5.6 Importância das variáveis no contexto de obesidade na utilização de técnicas de aprendizado de máquina

Após o processo de execução das técnicas de aprendizado de máquina e, observando que o algoritmo Random Forest apresentou a melhor proposta para utilização em estudos de

casos de obesidade, foi calculada a importância de cada variável do modelo de dados para a predição de dados utilizando o algoritmo Random Forest.



```{r}
modelo <- train(classifica~., data=treino, method="rf",
trControl=trainControl(method="cv", number=5), prox=TRUE,
allowParallel=FALSE)
varImp(modelo)
```

```
rf variable importance

  only 20 most important variables shown (out of 25)

          Overall
imc      100.0000
medpeso   17.3086
iddias    16.1249
```

Figura 14 - Execução de código sobre importância das variáveis

A tabela abaixo apresenta os valores de importância encontrados:

Nome da variável	**Percentual de importância**
imc	100
medpeso	17.30
iddias	16.12
idcri	7.78
medcabdo	5.87
ct	2.26
n_hdl	1.99
medcpesc	1.82
vldl	1.67
medestat	1.47
glicemia	1.35
triglic	1.22

ldl	0.97
medpad	0.94
medfc	0.90
medpas	0.88
hrtv	0.80
hrsono	0.75
pnascg	0.74
hba1c	0.72

Tabela 09 - Percentual de importância das variáveis

6.6 Resultados com coleta de entrevistas com profissionais de saúde

Após a realização das etapas de verificação da utilização de técnicas de aprendizado de máquina em base de dados e revisão sistematizada, buscou-se validar as hipóteses elencadas no tópico 2.

Para efeito de validação, foi realizada uma entrevista com profissionais de saúde que atuam em estudos de casos de obesidade, com o objetivo de responder as hipóteses desta pesquisa.

6.6.1 Critérios para seleção de participantes da entrevista

Para realização da pesquisa com os profissionais de saúde, os quais foram considerados médicos, nutricionistas e enfermeiros, foram definidos determinados critérios, com o objetivo de selecionar profissionais que possuíssem experiência comprovada na área de estudo deste trabalho. Assim, foram definidos os seguintes critérios:

CRITÉRIO 1 - Profissional possui formação na área da saúde;

CRITÉRIO 2 - Profissional possui experiência comprovada na área de estudo obesidade;

CRITÉRIO 3 - Profissional possui pós-graduação ou curso de atualização na área de estudo obesidade;

CRITÉRIO 4 - Profissional atua ou já atuou na área da saúde pública;

Através da definição acima, foi possível apresentar respostas para o questionário aplicado, com base nos conhecimentos adquiridos ao longo da carreira dos profissionais de saúde que participaram desta pesquisa.

6.6.2 Organização das entrevistas

Para realização desta coleta de dados, foi utilizada a plataforma Google Formulários que, por sua vez, é um site disponível gratuitamente e está integrado a todos os recursos do Google. Com esta plataforma, é possível criar um questionário, adicionar imagens, vídeos e links para arquivos externos e, após a criação do formulário, é gerado um link para divulgação e compartilhamento para que outras pessoas possam responder ao mesmo.

6.6.3 Aplicação de questionário

O questionário apresentado abaixo, foi aplicado num total de 10 (dez) profissionais de saúde que, por sua vez, atenderam a todos os critérios apresentados no tópico anterior. As respostas foram coletadas de forma não-presencial, a partir do envio do link do formulário via e-mail para os participantes da entrevista.

As entrevistas foram compostas de perguntas chave para o processo de validação, a saber:

P1 - Em estudos de aplicação de aprendizado de máquina, ou seja, aprendizado de máquina, a Idade, Altura, Peso e IMC foram considerados como fatores determinantes no processo de aprendizado do algoritmo que identifica se os dados de um novo paciente inserido no banco de dados é diagnosticado como obeso ou não-obeso. Tomando como base essa afirmação, estes resultados são significantes e/ou contribuem para o processo de diagnóstico do profissional de saúde em estudos de casos de obesidade? (Se possível, justificar sua resposta)

P2 - O seguinte gráfico apresenta a classificação de dados de pacientes como "obesos" e "não-obesos". Esta técnica de aprendizado de máquina, denominada "Máquina de Vetor de Suporte" aprende sobre uma base de dados de pacientes previamente diagnosticados e, ao receber um novo dado como entrada de um novo paciente, classifica este paciente e apresenta os resultados na forma deste gráfico. A imagem abaixo auxilia o profissional de saúde no processo de tomada de decisão para novos estudos de casos de pacientes? (Se possível, justificar sua resposta)

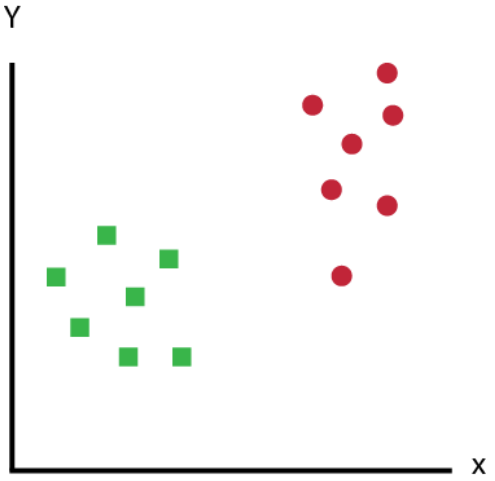


Figura 14 - Gráfico SVM. Legenda: pontos em verde(não-obeso) e pontos em vermelho(obeso), X: IMC, Y: idade.

6.6.4 Apresentação de resultados com a entrevista

Após à realização das entrevistas com os profissionais de saúde, os seguintes resultados foram encontrados e serão discutidos ao longo da apresentação dos dados:

PERGUNTA 1	
SIM	90%
NÃO	10%

Tabela 10 - Resultados pergunta 01

Para a PERGUNTA 1, algumas observações foram realizadas pelos profissionais de saúde, as quais, podemos citar:

- Outros dados devem ser explorados;
- São dados significativos e podem ser mais explorados;
- Para o diagnóstico da obesidade utiliza-se o IMC;
- Não são determinantes em sua totalidade, se faz necessário explorar outras variáveis no caso de uma avaliação médica;
- São significativos e contribuem, mas nem sempre são suficientes.

PERGUNTA 2	
SIM	80%
NÃO	20%

Tabela 11 - Resultados pergunta 02

Para a PERGUNTA 2, algumas observações foram realizadas pelos profissionais de saúde, as quais, podemos citar:

- Mostra que os dados estão agrupados e diferenciados;
- São necessários mais gráficos explicativos.

Após organização dos resultados e leitura das observações elencadas, reflete-se que o estudo possui resultados satisfatórios, podem auxiliar o processo de tomada de decisão médica, mas a avaliação médica de cada caso pode definir a importância de outras informações que são conclusivas para o processo de diagnóstico da obesidade. Os resultados apresentam indícios de que o desenvolvimento de uma ferramenta poderá auxiliar neste processo de tomada de decisão médica. Para isso, as próximas seções deste trabalho apresentam um esboço inicial, através de uma elicitação de requisitos para a construção futura de uma plataforma que integre todas as tecnologias apresentadas neste trabalho, para o auxílio na tomada de decisão médica.

6.7 PLANO DE DESENVOLVIMENTO DA FERRAMENTA

6.7.1 Plano de Requisitos da Ferramenta

O seguinte plano de requisitos da ferramenta proposta, apresenta a elicitação de requisitos funcionais e não-funcionais que devem ser implementados durante o processo de desenvolvimento da ferramenta. Esta elicitação de requisitos foi escrita de acordo com o padrão ISO/IEC 9126-1:2001 que descreve os atributos de qualidade de um software que devem compor os requisitos e sua arquitetura.

6.7.2 Requisitos Funcionais

(RF-01) O usuário deve ser capaz de realizar o cadastro com seus dados.
(RF-02) O usuário deve ser capaz de visualizar informações sobre seus pacientes.
(RF-03) O usuário deve ser capaz de selecionar quais informações do paciente serão utilizadas para gerar previsões.
(RF-04) O usuário deve ser capaz de visualizar lembretes, avisos ou indicativos sobre o estado futuro de saúde dos pacientes.
(RF-05) O usuário deve ser capaz de gerar relatórios individuais ou por grupos de pacientes.
(RF-06) O usuário deve ser capaz de inserir informações sobre seus pacientes.
(RF-07) O usuário deve ser capaz de modificar a base de dados utilizada pela máquina de aprendizado.
(RF-08) O usuário deve ser capaz de realizar filtragens de dados e consultas simples.
(RF-09) O usuário deve ser capaz de visualizar previsões sobre os dados inseridos no sistema.
(RF-10) O usuário deve ser capaz modificar as variáveis a serem utilizadas no processo de previsão e classificação dos dados.
(RF-11) O usuário deve ser capaz de visualizar informações dos algoritmos que foram utilizados no processo de previsão dos dados.
(RF-12) O usuário deve ser capaz de retornar feedback sobre as previsões geradas no sistema, como também, alertar sobre anomalias no processo de execução de determinados algoritmos de aprendizado de máquina.

6.7.3 Requisitos Não-Funcionais

(RNF-01) O sistema deve permitir o uso por diversas interfaces diferentes: tablet, navegador de internet, celular, notebook e computador convencional.
(RNF-02) O sistema deve permitir que o usuário realize login com seus dados básicos e senha.
(RNF-02) O sistema deve ser capaz de realizar as funções que foram especificadas nos requisitos funcionais.
(RNF-03) O sistema deve ser capaz de manter um nível de desempenho maior que 70% quando funcionar sob circunstâncias de baixa conexão e grandes quantidades de requisições.
(RNF-04) O sistema deve ser implementado de acordo com as normas de usabilidade de software.
(RNF-05) O sistema deve ser capaz de ser modificado em seu processo de evolução, ou seja, após o lançamento de suas primeiras versões.
(RNF-06) O sistema deve ser disponibilizado de forma segura para que seus dados sejam protegidos de ataques e invasões.
(RNF-07) O sistema deve disponibilizar diversas modalidades de gráficos que possam ser utilizadas pelo usuário.
(RNF-08) O sistema deve permitir a escolha por gráficos de acordo com a necessidade do usuário.
(RNF-09) O sistema deve fornecer todo o apoio necessário para o entendimento por parte do usuário de todas as tecnologias utilizadas e os resultados apresentados na tela.

6.7.4 Tecnologias para desenvolvimento da solução proposta

Os sub-tópicos seguintes apresentam as tecnologias elencadas para o desenvolvimento de uma solução proposta, baseado no estado da arte dessas tecnologias e sua efetiva utilização em plataformas web de gerenciamento de dados de saúde. Essas mesmas tecnologias já foram utilizadas, pelo autor deste trabalho, para o desenvolvimento de sistemas de gerenciamento de clínicas médicas.

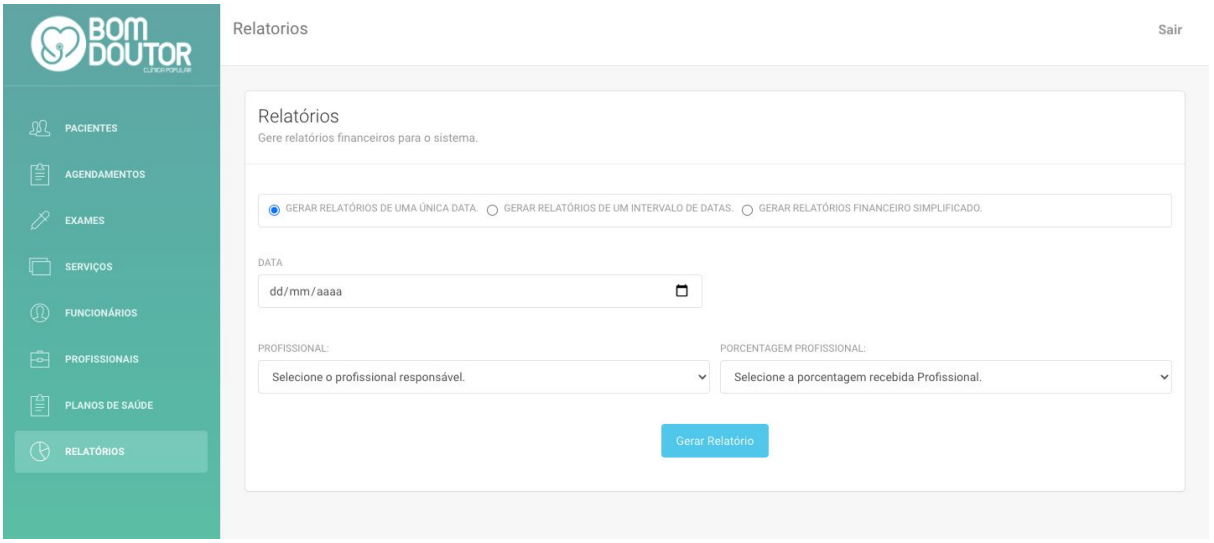


Figura 15 - Exemplo de sistema desenvolvido

A figura abaixo apresenta, de forma organizada, o relacionamento arquitetural entre as tecnologias propostas para o desenvolvimento da ferramenta.

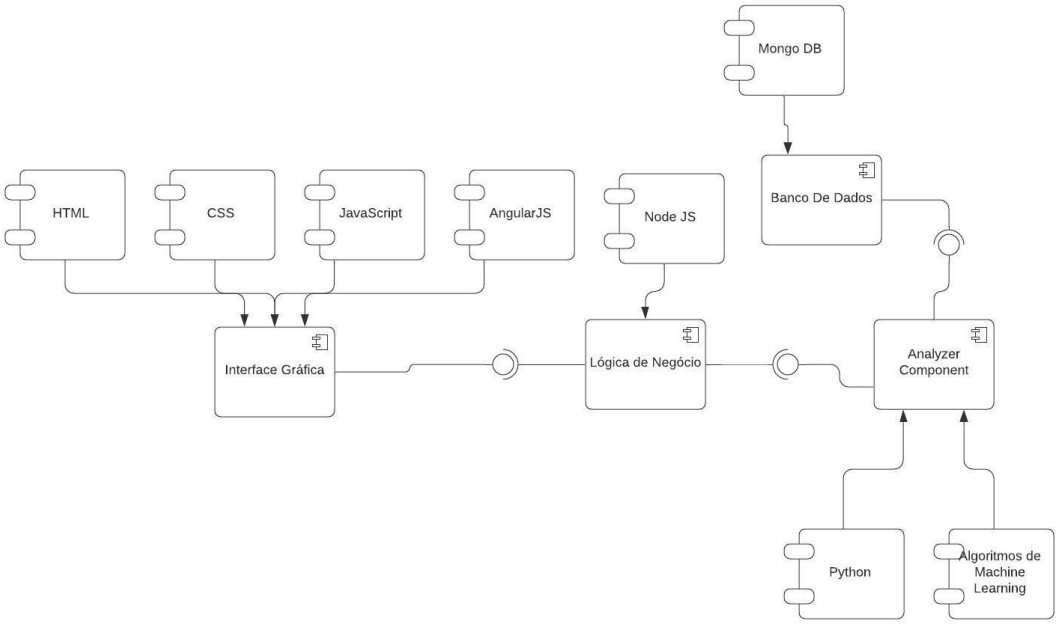


Figura 16 - Diagrama de arquitetura inicial do sistema

6.7.4.1 Node.JS

Node.js é um interpretador de código JavaScript com o código aberto, focado em migrar o Javascript do lado do cliente para servidores. Seu objetivo é ajudar programadores na criação de aplicações de alta escalabilidade, com códigos capazes de manipular dezenas de milhares de conexões simultâneas, numa única máquina física. O Node.js é baseado no interpretador V8 JavaScript Engine (interpretador de JavaScript open source implementado pelo Google em C++ e utilizado pelo Chrome). Foi criado por Ryan Dahl em 2009, e seu desenvolvimento é mantido pela fundação Node.js em parceria com a Linux Foundation.

6.7.4.2 Desenvolvimento Front-End

6.7.4.2.1 HTML

HTML é uma das linguagens que utiliza-se para desenvolver websites. O acrônimo HTML vem do inglês e significa Hypertext Markup Language ou em português Linguagem de Marcação de Hipertexto. O HTML é a linguagem base da internet. Foi criada para ser de fácil entendimento por seres humanos e também por máquinas, como por exemplo o Google ou outros sistemas que percorrem a internet capturando informação.

6.7.4.2.2 CSS

O Cascading Style Sheets (CSS) é uma linguagem utilizada para definir a apresentação de documentos que adotam para o seu desenvolvimento linguagens de marcação (como XML, HTML e XHTML). O CSS define como serão exibidos os elementos contidos no código de um documento e sua maior vantagem é efetuar a separação entre o formato e o conteúdo de um documento.

6.7.4.2.3 AngularJS

AngularJS é um framework JavaScript de código aberto, mantido pelo Google, que auxilia na execução de single-page applications. Seu objetivo é aumentar aplicativos que podem ser acessados por um navegador web, foi construído sob o padrão model-view-view-model (MVVM), em um esforço para facilitar tanto o desenvolvimento

quanto o teste dos aplicativos. A biblioteca lê o HTML que contém atributos especiais e então executa a diretiva na qual esta tag pertence, e faz a ligação entre a apresentação e seu modelo, representado por variáveis JavaScript comuns. O valor dessas variáveis JavaScript podem ser setadas manualmente, ou via um recurso JSON estático ou dinâmico.

6.7.4.2.4 - Banco de Dados Mongo DB

MongoDB é um software de banco de dados orientado a documentos livre, de código aberto e multiplataforma, escrito na linguagem C++. Classificado como um programa de banco de dados NoSQL, o MongoDB usa documentos semelhantes a JSON com esquemas. É desenvolvido pela MongoDB Inc. e publicado sob uma combinação da GNU Affero General Public License e Licença Apache.

Suas características permitem com que as aplicações estruturem informações de modo muito mais natural, pois os dados podem ser aninhados em hierarquias complexas e continuar a ser indexáveis e fáceis de buscar.

6.7.4.2.5 - Python

Python é uma linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991. Atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation. Apesar de várias partes da linguagem possuírem padrões e especificações formais, a linguagem como um todo não é formalmente especificada. O padrão *de facto* é a implementação CPython.

A linguagem foi projetada com a filosofia de enfatizar a importância do esforço do programador sobre o esforço computacional. Prioriza a legibilidade do código sobre a velocidade ou expressividade. Combina uma sintaxe concisa e clara com os recursos poderosos de sua biblioteca padrão e por módulos e *frameworks* desenvolvidos por terceiros.

7 CONCLUSÕES

Após o término de todas as etapas executadas, é possível verificar que, de acordo com as hipóteses apresentadas, técnicas de aprendizado de máquina que foram estudadas e selecionadas no escopo desta pesquisa, conseguem apresentar respostas relevantes para o processo de aprendizado em estudos de casos relacionados à obesidade.

Foi verificado também que a proposta de criação de uma ferramenta tecnológica, composta por técnicas de aprendizado de máquina avaliadas nesta pesquisa, é viável e pode contribuir no processo de tomada de decisão do profissional de saúde.

Trabalhos futuros poderão estar relacionados com a implementação de uma ferramenta composta pelos artefatos tecnológicos elencados ao longo desta pesquisa, como também, a possibilidade de integração da ferramenta a ser desenvolvida com outras bases de dados coletadas através de outros artefatos tecnológicos, mas que estejam relacionados com o campo de estudo deste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, M. N. S.; SIQUEIRA, A. L.; CAIFFA, W. T. *Regressão Logística Ordinal em Estudos Epidemiológicos.* Revista Scielo Saúde Pública, V. 43, p. 183-194, 2009.

ALBUQUERQUE, P. H. M. (2014). *Previsão de Séries Temporais Financeiras por Meio de Máquinas de Suporte Vetorial e Ondaletas.* Programa de Pós-Doutorado da Universidade de São Paulo. São Paulo.

BAUM, A. et al. *Targeting Weight loss interventions to reduce cardiovascular complication of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trid.* Lancet Diabetes Endrocrinal. 2017.

BASTOS, E. N. F. *Uma Rede Neural Auto-organizável construtiva para Aprendizado Perpétuo de Padrões Espaço-Temporais.* Repositório Digital da UFRGS. Universidade Federal do Rio Grande do Sul, 2007.

BAHIA, L. R.; ARAÚJO, D. V. *Impacto Econômico da Obesidade no Brasil.*, v. 13, n. 1, p. 13-17. Revista HUPE: Rio de Janeiro, 2017.

BAILLOT, A. et al. *Feasibility and Impacts of Supervised Exercise Training in Subjects with Obesity Awaiting Bariatric Surgery: a Pilot Study.* Volume 23, Issue 7, p. 882-891, Obesity Surgery, 2013.

BASSO, M. et al. *Sistema Inteligente para Apoio ao Diagnóstico de Diabetes Empregando Redes Neurais.* Anais do EATI - Encontro Anual de Tecnologia da Informação, Ano 4, n. 1, p. 56-63. Frederico Westphalen-RS, 2014.

BREVIDELLI, M. M. et al. *Prevalência e Fatores Associados ao Sobrepeso e Obesidade entre Adolescentes de uma Escola Pública.* Revista Brasileira em Promoção em Saúde. Fortaleza-CE, Vol. 28, n. 3, p. 379-386, 2015.

CAMILO, C. O.; SILVA, J. C. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.* Relatório Técnico(Mestrado em Ciência da Computação) - Universidade Federal de Goiás, 2009.

CHETTY, N.; VAISLA, K. S.; PATIL, N.. *An Improved Method for Disease Prediction using Fuzzy Approach.* Second International Conference on Advances in Computing Communication Engineering. p 568-572, 2015.

CORREA, E. N.; SCHMITZ, B. A. S.; VASCONCELOS, F. A. G. *Aspects of the built environment associated with obesity in children and adolescents: a narrative review.* Rev Nutr 2015. 28:237-40, 2015.

Custos de Doenças Ligadas à Obesidade para o SUS. Revista ABESO, Higienópolis, 17 de Agosto de 2012. Disponível em: <http://www.abeso.org.br/noticia/custos-de-doencas-ligadas-a-obesidade-para-o-sus>. Acesso em: 02 de Outubro de 2018.

DA COSTA, I. F. A. F. et al. *Adolescentes: comportamento cardiovascular.* J Vasc Bras, 2017.

DALAKLEIDI, K. V. et al. *A Hybrid Genetic Algorithm for the Selection of the Critical Features for Risk Prediction of Cardiovascular Complications in Type 2 Diabetes Patients, 2013.*

DEGREGORY, K. W. et al. *A review of machine learning in obesity.* World Obesity Federation, 2017.

DOBBINS, C.; RAWASSIZADEH, R.; MOMENI, E. *Detecting Physical Activity within lifelogs towards preventing obesity and aiding assisted living.* Neurocomputing Press, 2016.

DIAS, P. C. et al. *Obesidade e Políticas Públicas: Concepções e Estratégias Adotadas pelo Governo Brasileiro.* Reports In Public Health. Universidade Federal Fluminense, Rio de Janeiro, 2017.

ESTEBAN, S. et al. *Development and Validation of Various Phenotyping algorithms for Diabetes Mellitus using data from Electronic Health Records.* Computer Methods And Programs in Biomedicine, 2017.

FAZIL, Q. A. A. A.; JAMALUDIN, U. K. *Investigation on the Relationship between Cholesterol and Blood Glucose Levels using Decision Tree Method in Healthy Subjects.* International Conference on Automatic Control and Intelligent Systems. p 161-166, 2017.

FEIJÓ, R. B. et al. *Estudos de hábitos alimentares em uma amostra de estudantes secundaristas de Porto Alegre*. Revista Pediatria. p. 257-62, 1997.

FERREIRA, Arthur Pate de Souza; SZWARCWALD, Célia Landmann and DAMACENA, Giseli Nogueira. *Prevalência e fatores associados da obesidade na população brasileira: estudo com dados aferidos da Pesquisa Nacional de Saúde*. Rev. bras. epidemiol. 2019, vol.22, 2013.

FERGUS, P.; HUSSAIN, A. J.; HEARTY, J. *A Machine Learning Approach to Measure and Monitor Physical Activity in Children.* Neurocomputing Letters, 2016.

FERNEDA, E. *Redes Neurais e sua Aplicação em Sistemas de Recuperação da Informação.* Revista Scielo, V. 35, n. 1, p. 25-30, 2006.

FRIZZARINI, C.; LAURETTO, M. S. *Proposta de um Algoritmo para Indução de Árvores de Classificação para Dados Desbalanceados.* Simpósio Brasileiro de Sistemas da Informação. Minas Gerais, p. 722-733, 2013.

HAROUS, S. et al. *.Hybrid Obesity Monitoring Model Using Sensors and Community Engagement.* United Arab Emirates University, 2017.

HAYKIN, S. Neural Networks - *A Comprehensive Foundation Prentice Hall.* New Jersey, 2nd edition, 1999.

HUANG, N.; OYANG, Y. *Microbial Abundance Patterns of Host Obesity Infered by the Structured Incorporation of Association Measures into Interpretable Classifiers.* IEEE International Conference on Bioinformatics and Biomedicine. p. 315-319, 2014.

JULIAN, N. R. et al. *Feasibility of a screening tool for obesity diagnosis in Mexican children from vulnerable community of Me'phaa ethnicity in the State of Guerrero, Mexico.* Global Medical Engineering Physics Exchanges, 2017.

KUMAR, P. S.; PRANAVI, S. *Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics.* International Conference on Infocom Technologies and Unmanned Systems. p. 508-513, 2007.

LANGLEY, P. *The Changing Science of Machine Learning.* Machine Learning, Vol. 2, Issue 3, p. 275-279, 2011.

LEE, B. J. et al. *Prediction of Footing Plasma Glucose Status Using Anthropometric Measures for Diagnosing type 2 Diabetes.* IEEE Journal of Biomedical And Health Informatics. vol. 18, nº 02, p. 555-561, 2014.

LORENA, A. C.; CARVALHO, A. C. P. L. F. *Uma Introdução às Support Vector Machines*. RITA, Volume XIV, Número 02, p. 43-67, 2007.

MEDEIROS, S. C. *Avaliação do Peso, Ingestão Alimentar e Atividade Física em Adolescentes de uma Escola Particular em Almada.* Dissertação de Mestrado. Faculdade de Medicina da Universidade de Coimbra - FMUC, 2014.

MERCALDO, F.; NARDONE, V.; SANTONE, A. *Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques.* International

Conference on Knowledge Based and Intelligent Information and Engineering Systems. p. 2519- 2528, 2017.

MONARD, M. C.; BARANAUSKAS, J. A. *Conceitos sobre Aprendizado de Máquina.* Sistemas Inteligentes, Cap. 04, Vol. 1, p. 39-56, 2003.

MONTÃNEZ, C. A. C. et al. *Machine Learning Approaches for the Prediction of Obesity using Publicly Available Genetic Profiles.* Neurocomputing Press. P. 2743-2750, 2017.

MORIM, D. S. *Uso de detectores de dimensões variáveis aplicadas na detecção de anomalias através de sistemas imunológicos artificiais.* Dissertação(Mestrado em Ciência da Computação) - Universidade do Estado do Rio de Janeiro. Rio de Janeiro, 2009.

NGUYEN, D. T. et al. *SwallowNet: Recurrent Neural Network Detects and Characterizes Eating Patterns.* Percom Workshop on Pervasive Health Technologies, 2017.

NIRMPLA, D. M.; APPAUV, B. S.; SWATHI, U. V. *An Amalgan KNN to Predict Diabetes Mellitus.* IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology. p. 691-695, 2013.

OLIVEIRA, T. R. P. R.; CUNHA, C. F.; FERREIRA, R. A. *Characteristics of adolescents assisted in obesity outpatient service: know them to intervene.* Nutrire: rev. Soc. Bras. Alim. Nutr. São Paulo, v. 35, n. 2, p. 19-37, 2010.

OLIVERA, A. R.; ROESLER, V.; IOCHPE, C. *Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado.* Repositório Digital - Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2016.

RIBEIRO, A. C. C. *Diagnosis of Diabetes Type II Efficient Coding and Vector Machine Support.* 52f. Dissertação de Mestrado - Universidade Federal do Maranhão, São Luís, 2016.

SANTANA, J. E. et al. *Aprendizagem de Redes Bayesianas para Prevenção do Câncer Cervical.* XIV Congresso Brasileiro de Informática em Saúde, 2014.

SANTOS, I. C. C.; ALVES, R. L. S. *Estudo para Identificação de Promotores Procarióticos através de Classificação Bayesiana.* VIII Encontro Regional de Matemática Aplicada e Computacional, 2008.

SCHMIDT, M. I. et al. *Doenças crônicas não transmissíveis no Brasil: carga e desafios atuais.* The Lancet, págs. 61-74, 2011.

SIMON, P. *Too Big to Ignore: The Business Case for Big Data.* Wiley, 2013.

SOUTO, M. C. P. et al. *Técnicas de Aprendizado de Máquina para Problemas de Biologia Molecular.* Minicursos de Inteligência Artificial, Jornada de Atualização Científica em

Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, p. 103-152, 2003.

VAISHALI. R. et al. *Genetic algorithm based feature selection and MOE fuzzy classification algorithm and Pima Indians Diabetes dataset.* IEEE International Conference on Bioinformatics and Biomedicine, 2017.

VARGAS, R. R. *Uma nova forma de calcular os centros dos Clusters em algoritmos de agrupamento tipo fuzzy c-means.* Tese(Doutorado em Ciência da Computação) - Universidade Federal do Rio Grande do Norte. Natal, 2012.

WANG, C. et al. *Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach.* Diabetes research and clinical practice. p. 111-118, 2013

ZHU. Q.; AKKATI, A.; HONGWATTANAKUL, P. *Risk feature assessment of readmission for diabetes.* IEEE International Conference on Bioinformatics and Biomedicine. p 538-543, 2016.

APÊNDICE I - TABELA DE AVALIAÇÃO DE ALGORITMOS ATRAVÉS DE CRITÉRIOS DEFINIDOS

Algoritmo	Critério 1	Critério 2	Critério 3	Critério 4
Floresta Aleatória	SIM	SIM	SIM	SIM
Redes Neurais Artificiais	SIM	SIM	SIM	NÃO
Algoritmo K-means	SIM	SIM	SIM	NÃO
Algoritmo KNN	SIM	SIM	NÃO	NÃO
Árvore de Decisão	SIM	SIM	SIM	NÃO
Máquina de Vetor de Suporte(SVM)	SIM	SIM	SIM	SIM
Árvore de modelo logístico	SIM	SIM	SIM	NÃO
Rede Neural múltiplas camadas	SIM	SIM	SIM	NÃO
Classificador de Bayes	SIM	SIM	NÃO	NÃO
Hoeffding Tree	SIM	SIM	SIM	NÃO
Rede Bayesiana	SIM	SIM	SIM	NÃO

Legenda:

- SIM - Atende ao critério estabelecido;
- NÃO - Não atende ao critério estabelecido;

Critérios:

- Critério 1 - Utilização no contexto de obesidade;
- Critério 2 - Atualidade da pesquisa em que o algoritmo foi aplicado. Foram consideradas pesquisas com até oito(8) anos da data de sua publicação;
- Critério 3 - Algoritmo deve realizar tarefas de predição de dados;
- Critério 4 - Algoritmo deve ter precisão nos resultados das pesquisas com uma taxa maior que 90%;

ANEXO I - FORMULÁRIO PARA REALIZAÇÃO DE ENTREVISTAS COM PROFISSIONAIS DE SAÚDE

Aplicação de técnicas de Machine Learning em estudos de Obesidade

Este formulário pertence a uma pesquisa realizada pelo Pesquisador Paulo César Oliveira Brito, com título "UMA FERRAMENTA DE APRENDIZADO DE MÁQUINA PARA PREVISÃO E ANÁLISES DE DADOS EM ESTUDOS DE CASOS DE OBESIDADE". Deve ser respondido por profissionais de saúde que atendam aos seguintes critérios:

CRITÉRIO 1 - Profissional possui formação na área da saúde;

CRITÉRIO 2 - Profissional possui experiência comprovada na área de estudo obesidade;

CRITÉRIO 3 - Profissional possui pós-graduação ou curso de atualização na área de estudo obesidade;

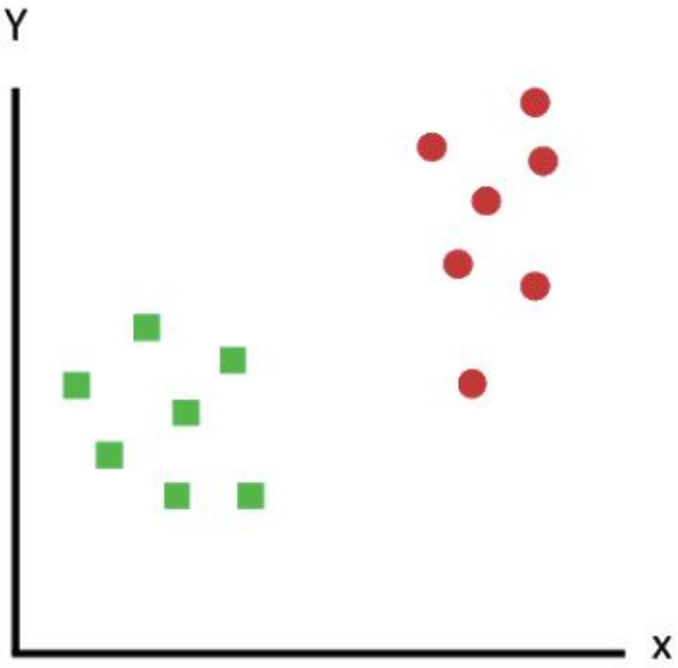
CRITÉRIO 4 - Profissional atua ou já atuou na área da saúde pública;

*Obrigatório

Em estudos de aplicação de Machine Learning, ou seja, aprendizado de máquina, a Idade, Altura, Peso e IMC foram considerados como fatores determinantes no processo de aprendizado do algoritmo que identifica se os dados de um novo paciente inserido no banco de dados é diagnosticado como obeso ou não-obeso.Tomando como base essa afirmação, estes resultados são significantes e/ou contribuem para o processo de diagnóstico do profissional de saúde em estudos de casos de obesidade? (Se possível, justificar sua resposta) *

Sua resposta

O seguinte gráfico apresenta a classificação de dados de pacientes como "obesos" e "não-obesos". Esta técnica de Machine Learning, denominada "Máquina de Vetor de Suporte" aprende sobre uma base de dados de pacientes previamente diagnosticados e, ao receber um novo dado como entrada de um novo paciente, classifica este paciente e apresenta os resultados na forma deste gráfico. A imagem abaixo auxilia o profissional de saúde no processo de tomada de decisão para novos estudos de casos de pacientes? (Se possível, justificar sua resposta) *



Sua resposta

ANEXO II - RESPOSTAS DOS ENTREVISTADOS AO FORMULÁRIO

- Respostas para a primeira pergunta

Em estudos de aplicação de Machine Learning, ou seja, aprendizado de máquina, a Idade, Altura, Peso e IMC foram considerados como fatores determinantes no processo de aprendizado do algoritmo que identifica se os dados de um novo paciente inserido no banco de dados é diagnosticado como obeso ou não-obeso.Tomando como base essa afirmação, estes resultados são significantes e/ou contribuem para o processo de diagnóstico do profissional de saúde em estudos de casos de obesidade? (Se possível, justificar sua resposta)

Sim

5 respostas

Sim. Para diagnóstico de obesidade utiliza-se o IMC.

1 resposta

Não, outros dados devem ser explorados

1 resposta

Sim, contribuem e sao significante. Mas nem sempre o suficiente e determinantes.

1 resposta

Sim, são dados significantes e podem ser mais explorados

1 resposta

- Respostas para a segunda pergunta

O seguinte gráfico apresenta a classificação de dados de pacientes como "obesos" e "não-obesos". Esta técnica de Machine Learning, denominada "Máquina de Vetor de Suporte" aprende sobre uma base de dados de pacientes previamente diagnosticados e, ao receber um novo dado como entrada de um novo paciente, classifica este paciente e apresenta os resultados na forma deste gráfico. A imagem abaixo auxilia o profissional de saúde no processo de tomada de decisão para novos estudos de casos de pacientes? (Se possível, justificar sua resposta)

Sim

6 respostas

Sim, mostra que os dados estão agrupados e diferenciados

1 resposta

Não.

1 resposta

Não

1 resposta

Sim, são é necessários mais recursos gráficos ecplicativos

1 resposta

ANEXO III - INFORMATIVO SOBRE A PLATAFORMA COLABORATORY

À plataforma Google Colaboratory é um ambiente de desenvolvimento que utiliza notebooks jupyter e não requer configuração inicial de tecnologias ou importação de bibliotecas e, é executado na nuvem.

É uma plataforma gratuita, disponibilizada pela Google, em que é possível realizar tarefas utilizando as seguintes funcionalidades:

- Escrever e executar códigos em Python;
- Salvar no Google Drive pessoal ou corporativo;
- Compartilhar análises com outros usuários Google;
- Trabalhar, estudar e pesquisar de forma colaborativa, em que uma ou mais pessoas acessam um mesmo arquivo do Colaboratory;
- Acessar poderosos recursos de computação científica;
- Suporte para Python 2.7 e 3.6;
- Aceleração de GPU grátis;
- Bibliotecas pré-instaladas;
- Suporta comandos bash;

Através das funcionalidades como as bibliotecas pré-instaladas no Colaboratory, foi possível executar as etapas desta pesquisa, utilizando os algoritmos de aprendizado de máquina, sem se preocupar com a pesquisa por bibliotecas e instalação em ambientes de desenvolvimento;