Aprendizado de Máquina Aplicado à Predição de Doenças Cardiometabólicas com Utilização de Indicadores Metabólicos e Comportamentais de Risco à Saúde

Alan Lopes de Sousa Freitas
Universidade Estadual de Maringá
Maringá, Paraná, Brazil
alanlopes4@gmail.com

Ana Silvia Degasperi Ieker
Universidade Estadual de Maringá
Maringá, Paraná, Brazil
anasilviaieker@hotmail.com

Heloise Manica Paris Teixeira
Universidade Estadual de Maringá
Maringá, Paraná, Brazil
hmpteixeira@uem.br

Josiane Melchiori Pinheiro
Universidade Estadual de Maringá
Maringá, Paraná, Brazil
jmpferreira@uem.br

Wilson Rinaldi
Universidade Estadual de Maringá
Maringá, Paraná, Brazil
wrinaldi@uem.br

ABSTRACT

Cardiometabolic diseases, developed throughout the worker's life, such as hypertension, diabetes, dyslipidemia and obesity are among the main causes of death and are associated with modifiable and controllable risk factors. The general objective of this study was to apply supervised Machine Learning techniques and to compare their performance to predict the risk of developing cardiometabolic disease from servers working at the School Hospital of south in Brazil. We sought to map the characteristics of individuals who are more likely to develop cardiometabolic diseases. The machine learning models evaluated were Naive Bayes, Decision Tree, Random Forest, KNN, Logistic Regression and SVM. The results obtained in the experiments showed that some supervised machine learning models produce a good classification, depending on the attributes and hyperparameters used.

KEYWORDS

Aprendizado de Máquina Supervisionado, Doença Cardiometabólica, Obesidade

1 INTRODUÇÃO

As doenças cardiometabólicas, desenvolvidas ao longo da vida, como hipertensão, diabetes, dislipidemia e obesidade, estão entre as principais causas de morte no mundo e são associadas a fatores de risco modificáveis e passíveis de controle. Estima-se que alguns fatores de risco sejam os responsáveis por mais de 40% da mortalidade global: hipertensão arterial (13%), tabagismo (9%), glicemia elevada (6%), inatividade física (6%), sobrepeso e obesidade (5%)[15] podendo esses fatores serem encontrados de forma isolada ou associados entre si.

Fatores de estilo de vida inadequado tendem a se agrupar com outros, causando efeitos aditivos na saúde[29] e complicações ainda mais graves. O estilo de vida inadequado é responsável por até dois terços das mortes cardiovasculares e está associado a um aumento substancial da mortalidade em muitas outras condições. O aumento crescente de desafios potencializados pela maior expectativa de vida, estilo de vida inadequado, desigualdade em acesso a serviços de saúde, aumento de despesas e escassez de profissionais de saúde, demandam ferramentas tecnológicas preditivas robustas, que possam mapear as principais condições de risco associadas ao estilo de vida, permitindo uma mudança no mesmo e evitando futuros problemas de saúde [9].

A Inteligência Artificial (IA) aplicada à medicina é um recurso importante na solução de problemas de atenção à saúde e vem se tornando um componente essencial da informática médica. O aumento crescente de desafios potencializados pela maior expectativa de vida, estilo de vida inadequado, desigualdade no acesso aos serviços de saúde, aumento de despesas e escassez de profissionais de saúde, demandam atualizações tecnológicas, com valores de previsões robustas e cenários alternativos que possam mapear futuros problemas de saúde e opções para modificar essas trajetórias [9].

Com a crescente demanda de métodos computacionais que possam contribuir com diagnósticos e otimizar o tempo dos profissionais de saúde, diversas pesquisas relacionadas com a aplicação de Aprendizado de Máquina (AM) à saúde têm sido desenvolvidas [7, 11].

O AM é uma sub-área de pesquisa em IA, que tem como objetivo desenvolver técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento automaticamente [22]. Além disso, possui capacidade de aprender a partir de grandes volumes de dados e gerar hipóteses úteis [2], que podem ajudar o profissional de saúde no diagnóstico, na prevenção e na busca por tratamentos mais eficazes. O modelo de AM possui todas as informações necessárias sobre um problema, como, quais dados constituem a análise e também o que se espera que ele produza como conhecimento [6]. Além disso, o AM também pode potencialmente prever a probabilidade de futuros sujeitos terem doenças específicas, dado o rastreamento precoce a partir dos dados de exames físicos e laboratoriais de rotina.

No contexto da presente pesquisa, algoritmos de aprendizado de máquina podem contribuir para a predição de tais doenças de acordo com as características do indivíduo. No entanto, deve-se selecionar entre vários algoritmos de aprendizado de máquina aquele que mais se encaixa no contexto do problema. Diante disso, o principal objetivo da pesquisa foi a avaliação e comparação de desempenho de diferentes técnicas de AM aplicado à predição de doenças cardiometabólicas. A principal motivação deste trabalho é o estudo de técnicas de redução de dimensionalidade dos dados a fim de

selecionar os atributos mais importantes, estudar e implementar técnicas de aprendizado de máquina supervisionado, comparando-as e avaliando o desempenho de cada uma com diferentes métricas para identificar a melhor para o problema.

Este artigo apresenta um estudo de 6 modelos de algoritmos de AM supervisionado (*Naive Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors* (KNN), *Logistic Regression* e *Support Vector Machines* (SVM)) para classificar servidores do Hospital Escola do Sul do Brasil que possuem ou não doença cardiometabólica, de acordo com variáveis comportamentais (tempo de tabagismo, consumo de bebida alcoólica (em anos) e número de refeições por dia) e laboratoriais (pressão arterial sistólica, pressão arterial diastólica, frequência cardíaca em repouso, colesterol total, hdl - colesterol bom, ldl - colesterol ruim, triglicerídeos e glicemia).

Buscou-se com a aplicação de AM modelar matematicamente um conjunto de dados de modo que, futuramente, seja possível prever a ocorrência de doença cardiometabólica em indivíduos ainda não submetidos ao modelo.

O artigo está organizado da seguinte forma: a seção 2 apresenta conceitos introdutórios sobre aprendizado de máquina. Na seção 3 são descritos trabalhos correlatos e na seção 3 os procedimentos metodológicos para o desenvolvimento de cada etapa da pesquisa. A seção 4 apresenta os resultados e discussão, detalhando as métricas de cada algoritmo utilizado. Por fim, a seção de considerações finais finaliza o artigo apontando os principais achados da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção introduz conceitos sobre AM em Inteligência Artificial, que tem como objetivo desenvolver técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento automaticamente [22]. Algoritmos de AM permitem que o computador tome decisões com base em conhecimento de dados prévios e em dados utilizados pelo usuário, assim o computador tem a habilidade de tomar decisões que podem resolver problemas [21]. Com essa capacidade de tomar decisão, o computador pode resolver diversos problemas como, por exemplo, o reconhecimento de fala, reconhecimento de imagens e robótica.

De acordo com [8], o AM pode ser dividido em duas categorias: **preditivas** e **descritivas**. As preditivas, baseando-se nos valores de seus atributos de entrada, tem como objetivo encontrar uma função que possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo. Para isso, utilizam algoritmos de **aprendizado supervisionado**, ou seja, há a presença de um "supervisor externo" que conhece a saída desejada para cada exemplo. Para as descritivas, o objetivo é explorar e descrever um conjunto de dados. Como elas não sabem qual é a saída esperada, os algoritmos utilizados aqui não fazem uso do atributo de saída (variável de interesse), ou seja, utilizam o paradigma de **aprendizado não supervisionado**.

[24] destaca que o AM pode ser usado em aplicações como: aprendizado por agrupamento (também conhecido como regras de associações), regressão e classificação. O primeiro tem como objetivo descobrir relações entre variáveis de um grande conjunto de dados, onde a descoberta não seria algo fácil de se fazer manualmente. O segundo é utilizado para prever quantidades, utilizando uma técnica estatística. Por exemplo, dado um conjunto de atributos de dados, a saída da regressão será dada por valores numéricos.

No Aprendizado não-supervisionado o objetivo é categorizar os dados de entrada em conjunto, buscando semelhanças nos dados de entrada. Portanto, os resultados variam de acordo com as variáveis de entrada. Já no aprendizado supervisionado, é realizado um treinamento em um conjunto de dados com as classes corretas e a saída é o modelo que foi aprendido para responder corretamente todas as entradas.

Além desses dois tipos de aprendizagem, há ainda dois outros: aprendizagem por reforço e aprendizagem semi-supervisionada. O primeiro trata-se de tentar aprender qual a melhor ação a ser tomada guiando-se por uma função de avaliação de semelhança que mede quão boa é a solução inferida. O segundo assume que nem todos os exemplos do conjunto de treinamento são rotulados, isto é, há uma grande quantidade de exemplos não rotulados com uma pequena quantidade de exemplos rotulados.

No aprendizado supervisionado, o modelo de AM possui todas as informações necessárias sobre um problema, como, quais dados constituem a análise e também o que se espera que ele produza como conhecimento [6]. Na presente pesquisa foi utilizado a aprendizagem supervisionada, pois há uma base de dados com as características do indivíduo e também o rótulo de saída (doença). Com essas informações é possível que o modelo tenha um objetivo para o treinamento.

Os principais algoritmos do aprendizado de máquina supervisionado são: K-Vizinhos-Mais-Próximos, Regressão Linear, Redes Neurais e Máquinas de Vetores de Suporte, entre outros.

De acordo com [19], os modelos preditivos de regressão linear propõem-se a verificar se existe uma relação entre duas ou mais variáveis. Para isso, o algoritmo de regressão linear, sendo considerado um dos mais básicos de estatística, tenta achar uma equação de reta entre duas ou mais variáveis, sendo uma dependente e as outras independentes.

O problema de regressão logística (RL) tem como objetivo predizer a probabilidade de um determinado exemplo E pertencer a uma determinada classe de resposta y. De acordo com [18] a variável de resposta do conjunto de treinamento é modelada por meio da distribuição binomial que leva como parâmetro p a probabilidade de ocorrência de uma determinada classe. Vale a pena ressaltar que a probabilidade deve ser limitada $[0, 1]$. A RL recebe um número qualquer e, por meio de uma função logística que descreve uma curva em formato de S, transforma-o em um outro número dentro do intervalo $]0, 1[$. Por meio do valor gerado, o modelo estima a probabilidade p de uma resposta y, baseada em um conjunto de variáveis independentes de tamanho n, no qual $P(y|x)$ segue uma distribuição de Bernoulli com uma probabilidade de sucesso $P(x)$ [17].

O Método K-Vizinhos-Mais-Próximos (*K-Nearest Neighbors - KNN*), ao invés de usar uma função ajustada para a predição dos dados, utiliza preditores armazenados e a resposta de interesse para a predição da resposta de um novo exemplo [12]. Na etapa de treinamento, o KNN armazena os preditores e a resposta de interesse para que a predição seja realizada através do conjunto de treinamento. Assim, seja x^* o novo exemplo representando um

vetor p-dimensional de mensurações para os preditores de interesse, $x^* = (x_1^*, ..., x_p^*)^T$, a predição da resposta é feita através das observações individuais do conjunto de treinamento mais próximas à x^*. Para problemas de classificação, a resposta predita será representada pela classe mais frequente, observada na vizinhança [28]. Segundo [27], a letra 'K' no início do nome do algoritmo, significa a quantidade vizinhos mais próximos que irão compor a vizinhança. Além disso, se faz necessário definir a medida de distância responsável por identificar os k exemplos mais próximos do conjunto de treinamento.

De acordo com [12], há dois tipos de árvores: Árvores de regressão quando a resposta é contínua, e a árvore de classificação quando a resposta é categórica. O uso de árvores é geralmente utilizado quando a relação entre os preditores e a resposta de interesse é não linear e complexa. Os dois tipos de árvores, particionam o espaço dos preditores em N regiões distintas e disjuntas e após ajusta um modelo simples para a predição da resposta de interesse de cada região. As regras utilizadas para o particionamento do espaço dos preditores é conhecida como árvore de decisão. Qualquer que seja a alteração no conjunto de treinamento, pode resultar em mudanças na estrutura da árvore ou em suas regras. Essas mudanças podem alterar a interpretação do modelo ajustado. A árvore de decisão completa gerada é complexa e induz sobre o ajuste do conjunto de treinamento e ao erro de generalização para novos exemplos. Uma solução para este problema é o uso da poda, onde o tamanho da árvore é diminuído resultando é uma subárvore menos complexa e com performance preditiva melhorada [12, 16].

O *Random Forest* (ou Floresta Randômica) sorteia aleatoriamente preditores do conjunto de treinamento com o objetivo de reduzir a correlação entre as árvores agregadas (floresta). Para isso, o *random forest* obtêm C conjuntos de dados de tamanho n e, para cada conjunto, gera uma árvore de decisão [12, 16].

Uma floresta é a combinação de várias árvores de decisão, usando o método de *bagging*. O método de *bagging* diz que quando combina-se vários modelos de aprendizado, obtém-se um aumento no desempenho geral do modelo. Sendo assim, o *Random Forest* combina várias árvores de decisão para gerar uma maior acurácia em suas predições [3]. Após a obtenção de C conjuntos de dados de tamanho n, o algoritmo de árvore de decisão é aplicado, resultando em C modelos preditivos [12, 16].

Uma das limitações do algoritmo é a grande quantidade de árvores geradas, pois quanto maior o número de árvores geradas, maior será a lentidão para a realização das predições. Embora o algoritmo possa ser rápido para o conjunto de treinamento, ele apresentará lentidão quando aplicado aos testes. Sendo assim, uma predição com maior acurácia requer mais árvores geradas [1].

O *Support Vector Machine* utiliza um algoritmo denominado classificador de margem máxima (*maximal margin classifier* em problemas de classificação. Por ele não estimar, diretamente, probabilidades, mas sim a classe de resposta de interesse para uma nova observação, acaba sendo diferente dos demais algoritmos [14]. Se os dados podem ser perfeitamente separados, um número infinito de hiperplanos candidatos à fronteira de decisão pode ser definido, onde o classificador de margem máxima irá considerar a localização da nova observação com relação à fronteira de decisão para a realização da classificação [14]. A separação dos dados só é possível se existir um hiperplano. No entanto, para a maioria dos problemas reais, não é possível definir esse hiperplano, pois pode haver sobreposição de algumas observações de diferentes categorias da variável de resposta [16].

3 TRABALHOS RELACIONADOS

Com a demanda de métodos que possam facilitar diagnósticos e otimizar o tempo dos profissionais de saúde, diversas pesquisas relacionadas com aplicação de AM à saúde tem sido desenvolvidas.

No estudo de [30] são comparados modelos gerados por quatro algoritmos de AM para a predição de doença cardiovascular em um espaço de tempo de 10 anos. Esse estudo segue as recomendações do Código Americano de Cardiologia. A base de dados utilizada é de uma coorte prospectiva de 378.256 pacientes. Como resultado, foi constatado que o modelo de AM de redes neurais apresentou o melhor desempenho dentre todos os modelos analisados com 355 predições corretas adicionais de doença cardiovascular quando comparado ao modelo baseado nas recomendações do Código Americano de Cardiologia.

Na pesquisa de [11], a meta foi a compreensão de determinantes de saúde usando AM para a construção de modelos preditivos para ocorrência de doenças em 1 e em 5 anos a partir do *baseline*. No estudo foram utilizados dados da pesquisa social *Understanding Society* de 6800 indivíduos. Com base nos resultados obtidos, construir modelos preditivos a partir do uso de dados de saúde apresentou o melhor resultado dentre os testes aplicados, com uma acurácia de predição de 71%. A atividade física e a presença de algumas condições de saúde foram fortes preditores individuais.

No estudo apresentado em [23], o objetivo foi desenvolver modelos preditivos de diabetes não diagnosticada a partir de uma base de dados de 12.447 adultos entrevistados para o Estudo Longitudinal de Saúde do Adulto (ELSA). Obteve-se como resultado 11% de frequência de diabetes não diagnosticada. Entre os 403 indivíduos do conjunto de dados de testes que tinham diabetes não diagnosticada, 274 foram identificados.

4 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa foi desenvolvida utilizando a base de dados obtida do projeto: "Perfil de fatores de risco para as Doenças Crônicas não transmissíveis e programa de exercício físico em servidores públicos de um Hospital Escola do Sul do Brasil", aprovado pelo Comitê de Ética em Pesquisa sob parecer número 1.766.685/2016.

A base de dados utilizada no estudo contém 560 amostras coletadas entre os anos de 2016 e 2018 sobre o estado de saúde de servidores que atuam no Hospital Escola. Os atributos foram coletados por meio de uma avaliação antropométrica, análise documental do periódico do servidor e uma junção de questionários validados.

Para os experimentos realizados neste estudo, foram utilizado uma máquina com sistema operacional Linux Mint 20 com 16 GB de memória RAM DD4, processador Intel Core i7 8th gen e 1 Terabyte de armazenamento em disco. Além disso, foi utilizado a linguagem de programação Python (versão 3.7.7) por ser considerada uma linguagem popular e conhecida para o aprendizado de máquina. Referente as ferramentas utilizadas, destacam-se a ferramenta Anaconda (versão 1.7.2) e a Spyder (versão 4) por fornecerem um ambiente de desenvolvimento que facilita recursos de edição e depuração. Já para

as bibliotecas, utilizou a *Scikit-learn* por possuir características que tornam o código manutenível e seguir convenções de nomenclatura de funções técnicas de AM.

A Figura 1 ilustra as principais etapas realizadas no desenvolvimento da pesquisa. A base de dados original anonimizada era composta por dois arquivos, um com dados de 2016 e outro com dados de 2018. Primeiro, realizou-se a filtragem dos dados com o objetivo de se obter as características mais relevantes com base no conhecimento de um especialista na área de saúde. Posteriormente, estes foram convertidos para o formato Comma-separated values (CSV) e então, realizou-se a análise dos dados aplicando os algoritmos de redução de dimensionalidade, detecção de *outliers* e de valores faltantes, e somente após isso, foram gerados os modelos preditivos.

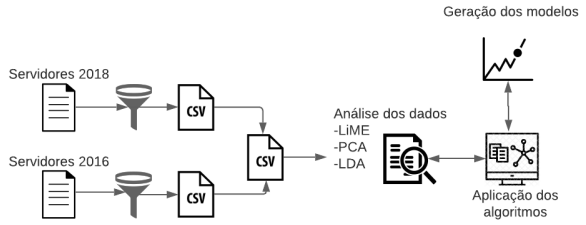

Figura 1: Metodologia de desenvolvimento da análise dos dados.

O pré-processamento dos dados, etapa em que os dados originais são modificados para melhorar o desempenho dos algoritmos [18], foi aplicado para todos os atributos da base de dados original, selecionando os dados mais relevantes e significativos para a pesquisa, simplificando informações redundantes e preenchendo informações inexistentes para possibilitar a aplicação dos algoritmos de AM supervisionado.

Os atributos que não se aplicam ao contexto da pesquisa foram retirados com base na análise de um profissional especializado, restando os seguintes: idade do servidor, jornada de trabalho semanal em horas, horas de sono semanal, soma das barreiras para AF, tempo de tela, soma de atividade física na semana, quantidade de refeições que o servidor faz por dia, tempo (em anos) de fumante, tempo (em anos) de alcoolismo, pressão arterial sistólica, pressão arterial diastólica, frequência cardíaca em repouso, cintura no momento da entrevista, índice de massa corporal, mudança de comportamento nos últimos 3 meses, colesterol total, colesterol bom, colesterol ruim, triglicerídeos, glicemia e se o servidor possui uma doença cardiometabólica.

Para tratar os atríbutos com valores faltantes (Tabela 1) foi utilizada a estratégia de imputação, por meio da média (para o atributo pesoAvaliado e cinturaAvaliada), mediana (pra o atributo alturaAvaliada) e do algoritmo KNN (para os demais atributos com dados faltantes), nenhum indivíduo ou atributo foi removido da base. Os *outliers* (valores extremos que aparentemente não seguem o mesmo padrão dos outros) foram identificados por meio de *boxplot* e tratados individualmente.

Tabela 1: Atributos com valores ausentes

Atributo	Quantidade
ps	60
pd	23
fcReps	30
pesoAvaliado	48
alturaAvaliada	40
cinturaAvaliada	13
imcAvaliado	143
tri	50
gli	56
ldl	53
hdl	50
colesterol	50

Após as etapas de pré-processamento, observou-se que a base de dados estava desbalanceada. Dos 560 registros da amostra, 148 eram de servidores com alguma doença cardiometabólica, 251 de servidores sem nenhuma doença e 161 com alguma doença sem ser cardiometabólica. O uso de dados desbalanceados pode prejudicar o desempenho de algoritmos de aprendizado de máquina [13]. A solução adotada para o desbalanceamento foi a aplicação de novos dados para a classe minoritária (cardiometabólica). Para isso, foi utilizada a técnica denominada *Synthetic Minority Oversampling Technique* (SMOTE), a qual tem como objetivo a geração de registros sintéticos (artificiais) para a classe de interesse minoritária a partir dos registros já existentes, com o intuito de aumentar o número de amostras das classes minoritárias e, dessa forma, melhorar a generalização dos classificadores [10].

Muitos algoritmos de AM possuem os chamados parâmetros de sintonização ou hiperparâmetros, que são parâmetros que podem influenciar no desempenho do modelo. Esses parâmetros são ajustados ao executar o algoritmo, pois não há como estimá-los através do treinamento ou otimizá-los por validação cruzada [18]. Nos experimentos realizados neste estudo a variação dos hiperparâmetros foi aplicada utilizando uma ferramenta chamada *Grid Search* para a escolha da melhor combinação de hiperparâmetros possíveis.

Quando há uma base de dados grande a melhor abordagem segundo [14] é dividir aleatoriamente a base em três partes: treinamento, validação e teste. No entanto, quando a base é pequena e não pode ser dividida, destacam-se as técnicas de reamostragem para aproximar o conjunto de validação por meio da reutilização de observações do conjunto de treinamento. Nesta pesquisa foi utilizada a técnica de validação cruzada *k-fold*, uma técnica comumente utilizada em AM. Essa técnica divide aleatoriamente a base de treinamento em k partes de tamanhos aproximadamente iguais, onde $k - 1$ partes formam o conjunto de treinamento e a parte restante forma o conjunto de testes. Esse processo é repetido até que todas as partes tenham participado tanto do treinamento quanto do teste, onde as k estimativas resultantes são utilizadas para o cálculo da média e do erro padrão [18]. Essa técnica de validação é utilizada com o objetivo de selecionar um entre vários algoritmos de AM, pois resulta em uma boa estimativa em termos de erro de predição e resultado esperado [26].

Para os experimentos deste trabalho foi utilizado um k=10. Isso significa que a cada iteração a base de treinamento foi dividida em 10 partes, 9 para o treinamento, 1 para o teste de predição e a cada final de iteração o resultado é armazenado. Ao final de todo o processamento, isto é, de todas as 10 iterações, os resultados são utilizados para calcular o desempenho médio do modelo.

Para cada algoritmo utilizado nesta pesquisa, utilizou-se a validação cruzada *k-fold* para a análise do desempenho de cada um, por meio da área sob a curva - *Area Under Curve Receiver Operating Characteristic* (AUC ROC). As curvas ROC apresentam taxa de verdadeiros positivos no eixo Y e taxa de falsos positivos no eixo X, ou seja, o canto superior esquerdo do gráfico é o ponto ideal (uma taxa de falsos positivos de zero e uma taxa de verdade de um). Uma área maior abaixo da curva geralmente é um resultado melhor, pois a curva tende a se distanciar da reta x = y (que representa o resultado dos modelos aleatórios). A inclinação das curvas ROC também é importante, pois é ideal maximizar a taxa de verdadeiros positivos e minimizar a taxa falsos positivos [25].

Seis modelos de aprendizado de máquina foram usados para classificar os dados e comparar os resultados: *Naive Bayes*, *Decision Tree*, *Random Forest*, KNN, *Logistic Regression* e SVM. Para cada modelo, foram realizados testes com as seguintes variações: a) uso da técnica de PCA antes do treinamento do modelo; b) uso da técnica de LDA antes do treinamento do modelo; c) Não utilização das técnicas de LDA e PCA antes do treinamento do modelo.

Cada algoritmo possui sua lista de hiperparâmetros individuais, onde cada parâmetro pode influenciar positivamente ou negativamente o desempenho do modelo gerado. Para encontrar a melhor combinação de hiperparâmetros de cada algoritmo, foi utilizada a ferramenta *Grid Search*. Esta ferramenta executa o algoritmo N vezes e em cada execução são utilizadas diferentes combinações de hiperparâmetros. O resultado gerado com cada combinação é armazenado e posteriormente utilizado para encontrar aquela combinação que obteve o melhor resultado possível. Como se trata de um problema de classificação o *Grid Search* foi configurado para obter a melhor combinação levando em conta a precisão de cada algoritmo. Para este estudo, o *Grid Search* foi aplicado às três variações citadas anteriormente (a - LDA, b - PCA e c - Nenhuma).

A melhor combinação de hiperparâmetro gerada pelo *Grid Search* foi utilizada no treinamento do modelo classificador, usando validação cruzada *k-fold* com k = 10. Ao final de todas as execuções foram avaliadas as métricas de acurácia, precisão, revocação e medida F1. Além disso, foram geradas matrizes de confusão para cada modelo treinado.

5 RESULTADOS E DISCUSSÃO

Os resultados obtidos nos experimentos mostraram que alguns modelos de aprendizado de máquina supervisionado produzem uma boa classificação dependendo dos atributos e dos hiperparâmetros utilizados. A escolha dos hiperparâmetros é uma tarefa que impacta diretamente no desempenho do modelo de classificação, por isso a sua definição foi apoiada pela ferramenta *Grid Search*. A Tabela 1 apresenta os melhores hiperparâmetros selecionados pela ferramenta para cada algoritmo.

Tabela 2: Hiperparâmetros dos melhores resultados de cada algoritmo

Algoritmo	Hiperparâmetros
Naive Bayes	nenhum
Decision Tree	criterion: entropy max_features: auto min_samples_leaf: 2 min_samples_split: 2
RandomForest	criterion: gini max_features: auto n_estimators:40
KNN	algorithm: brute metric: seuclidean n_neighbors: 10 p: 0.5
Logistic Regression	class_weight: balanced penalty: none solver: sag
SVM	C: 0.5 class_weight: balanced gamma: scale kernel: poly

A Tabela 2 apresenta os melhores resultados de cada algoritmo, levando em consideração as métricas de acurácia, precisão, revocação e medida F1.

Também foram geradas as matrizes de confusão e as curvas AUC ROC para cada modelo obtido. A matriz de confusão consegue apontar a quantidade de verdadeiro-positivo, falso-positivo, falso-negativo e verdadeiro-negativo, onde a diagonal principal apresenta os casos em que as classes foram corretamente preditas e a diagonal secundária os erros de predição [18]. A matriz de confusão é uma importante métrica para a avaliação de desempenho do modelos gerados pelo AM, pois é a partir dela que as outras métricas (precisão, acurácia, revocação, entre outras) são geradas.

A Figura 2 mostra a matriz de confusão gerada pelos resultados do modelo *Random Forest*, verificou-se que o classificador gerado apresentou uma alta taxa de acerto e uma baixa taxa de erro. O modelo classificou mais de 10% (49 servidores) como saudáveis sendo que eles possuem a doença e 60 como doentes sendo que não possuem a doença.

A métrica AUC ROC mede o quão bem o modelo pode diferenciar entre verdadeiro positivo e falso positivo, onde a área sob a curva indica a probabilidade do caso ser corretamente classificado [25]. Quanto mais a linha dos testes *k-fold* se afastar da linha central (linha pontilhada) mais preciso será o resultado do teste.

As Figuras 3 a 8 mostram as curvas ROC para todos os modelos gerados. Visivelmente pode-se verificar que a curva ROC gerada para o *Random Forest* é a que possui o formato com melhores resultados, se afastando mais da linha pontilhada e confirmando os melhores resultados para este modelo.

Ao considerar todas as métricas citadas acima, a quantidade de Falsos Negativos (FN) e Falsos Positivos (FP) e a curva AUC ROC

Tabela 3: Melhores resultados gerados com os modelos treinados

Algoritmo	Variação	Acurácia	Precisão	Revocação	Medida F1
Naive Bayes	PCA	0.66667	0.72403	0.53865	0.61773
Decision Tree	PCA	0.76812	0.78756	0.73430	0.76000
RandomForest	PCA	0.86957	0.87136	0.86715	0.86925
KNN	PCA	0.76691	0.72973	0.84783	0.78436
Logistic Regression	Nenhum	0.71498	0.72139	0.70048	0.71078
SVM	PCA	0.65580	0.81463	0.40338	0.53958

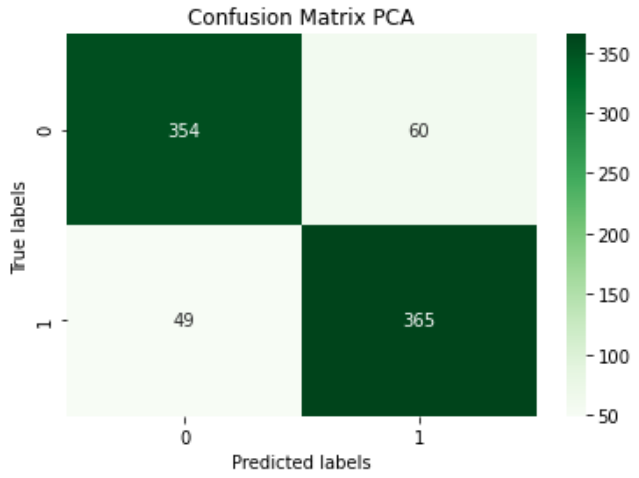

Figura 2: Matriz de confusão para o Random Forest utilizando PCA.

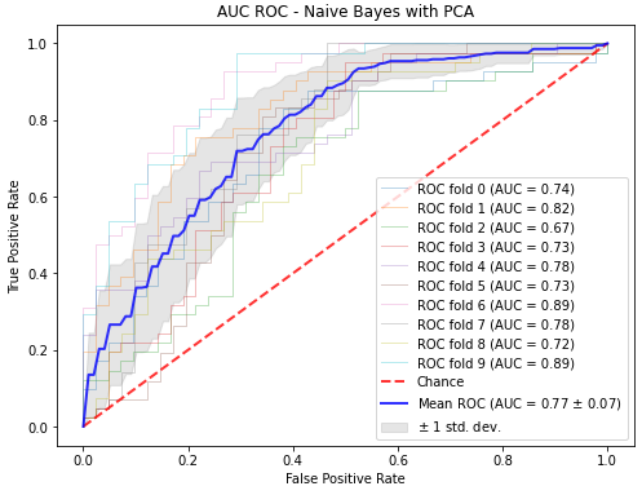

Figura 3: Métrica AUC ROC para o Naive Baiyes utilizando PCA.

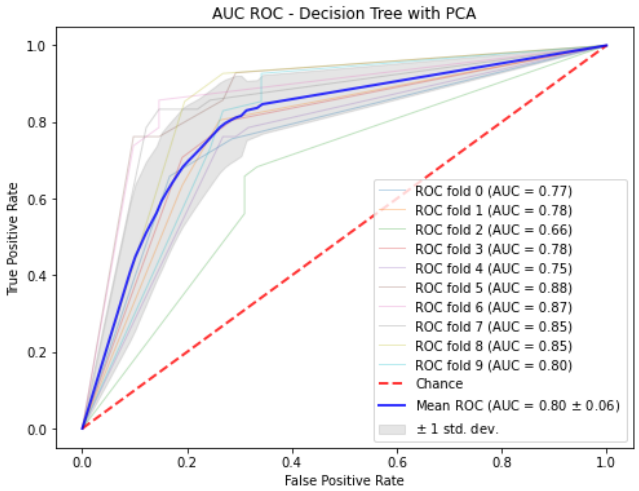

Figura 4: Métrica AUC ROC para o Decision Tree utilizando PCA.

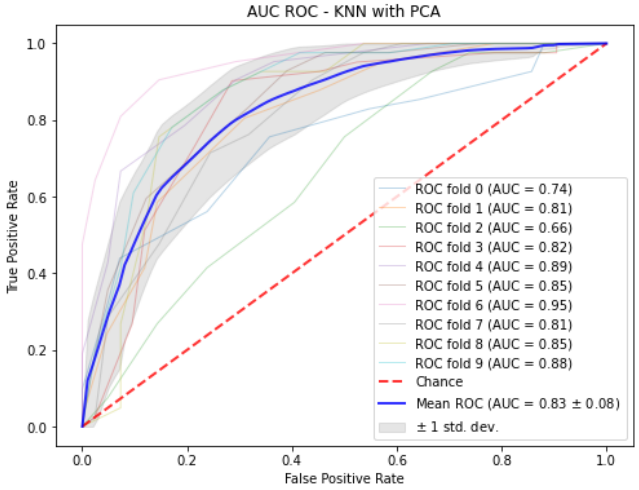

Figura 5: Métrica AUC ROC para o KNN utilizando PCA.

para cada algoritmo, foi possível concluir que o melhor algoritmo para a predição foi o *Random Forest* utilizando PCA.

Como ilustra o gráfico da Figura 8, todos os testes realizados pela validação cruzada para o *Random Forest* obtiveram resultados superiores ou iguais a 85%, obtendo uma média de 93%. Em alguns

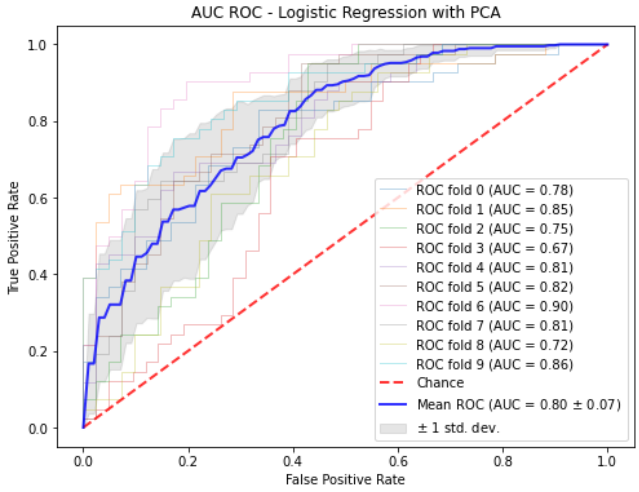

Figura 6: Métrica AUC ROC para o Logistic Regression utilizando PCA.

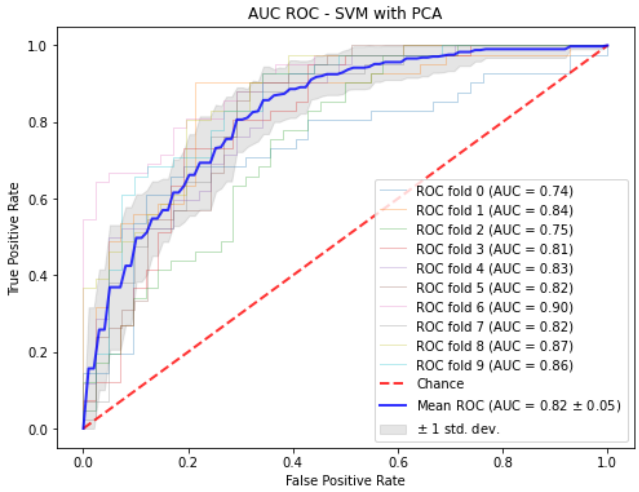

Figura 7: Métrica AUC ROC para o SVM utilizando PCA.

problemas da área da saúde, um balanço entre acurácia e interpretabilidade pode ser desejável na escolha do modelo preditivo final, com o objetivo de facilitar a sua adoção por profissionais de saúde [4]. Considerando a natureza do problema, classificação de doença cardiometabólica ou não, deve-se levar em consideração quando selecionar o modelo de aprendizado de máquina, não somente a precisão e acurácia, mas outras métricas que destacam os FN e os FP. Os FN para este modelo define o número de vezes que um servidor é classificado como saudável quando ele possui uma doença cardiometabólica na vida real. Para o campo da saúde, fazer essa afirmação é um erro grave, pois a pessoa não terá a oportunidade de realização de exames, uma piora em seu quadro de saúde, com complicações metabólicas e cardíacas e, possivelmente uma maior dificuldade para o tratamento. Já levando em consideração a baixa taxa de FPs, isto é, aquele servidor que foi classificado

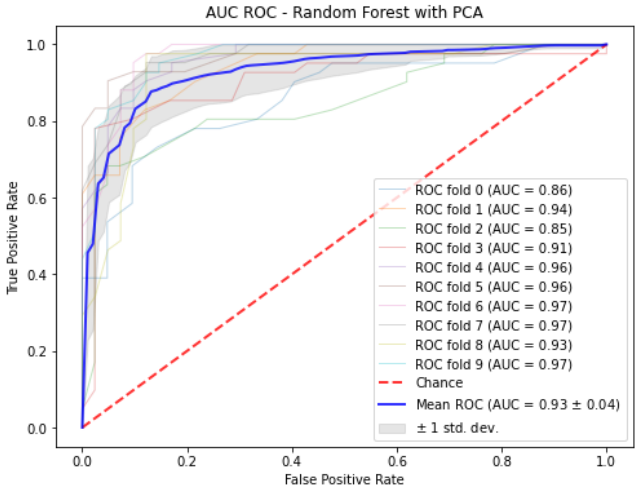

Figura 8: Métrica AUC ROC para o Random Forest utilizando PCA.

como possuindo doença cardiometabólica quando na vida real não possui, a métrica não demonstra um erro tão grave para saúde, pois possibilita que esse sujeito olhe para a sua saúde de um modo preventivo.

Levando em consideração as métricas do estudo, observa-se que o modelo gerado pelo *Random Forest* resultou em uma baixa taxa de FN e FP, resultando em uma classificação mais precisa. A partir da base de dados em questão fazer diagnóstico apenas com variáveis bioquímicas e pressóricas, como é utilizado normalmente na clínica médica, parece não ser suficiente para classificar o aparecimento de doenças cardiometabólicas. Além disso, modelos preditivos de doenças crônicas podem basear-se não só em fatores de risco modificáveis, mas também em características biológicas não modificáveis, como idade e sexo, que, embora contribuam para a desempenho preditivo do modelo, podem não ser relevantes em estratégias de prevenção ou controle [20]. Esse modelo foi treinado levando em consideração, além das variáveis bioquímicas, também as variáveis comportamentais como o tabagismo, alcoolismo, inatividade física, tempo de tela, dentre outros. Existem registros na base de dados em que as variáveis bioquímicas (ldl - colesterol ruim, hdl - colesterol bom, triglicerídeos e glicemia) resultaram em valores normais (valores que não afirmam determinada doença), demonstrando a importância do olhar ampliado e integral à saúde, levando em consideração os diversos indicadores do estilo de vida.

É importante destacar que mesmo um modelo preditivo com bom poder discriminatório e bem calibrado pode não se traduzir em melhores cuidados à saúde, pois uma predição acurada não diz o que deve ser feito para modificar o desfecho sob análise [5]. Portanto, este modelo enfatiza a ampliação do entendimento para os demais indicadores da saúde e a necessidade de um parecer clínico e individual.

Os resultados desta aplicação possui como principal limitação o fato de não ser levado em consideração o uso de remédios, ser uma amostra um tanto quanto pequena e com dados faltantes. Espera-se que, com o aumento da disponibilidade dos dados de indicadores

de saúde dos servidores da instituição, seja possível melhorar consideravelmente, o desempenho desses algoritmos.

6 CONSIDERAÇÕES FINAIS

A presente pesquisa apresentou um estudo de 6 modelos de algoritmos de AM supervisionado para classificar indivíduos que possuem ou não doença cardiometabólica, de acordo com variáveis comportamentais e laboratoriais. Buscou-se modelar matematicamente um conjunto de dados para que possa prever a ocorrência de doença cardiometabólica em indivíduos ainda não submetidos ao modelo.

O desempenho de um modelo preditivo dependente da base de dados e da resposta de interesse, ou seja, do conhecimento relacionado ao problema e de variáveis informativas e com poder discriminatório. Após a avaliação de desempenho dos modelos selecionados através das métricas utilizadas, ficou evidente que o *Random Forest* representa uma escolha interessante como modelo final para a utilização em dados futuros.

Portanto, o uso do aprendizado de máquina supervisionado para classificação da existência da doença cardiometabólica pode auxiliar o servidor no autocuidado, ficando evidente quais variáveis influenciam para o desenvolvimento destas doenças. Sendo assim, este estudo sugere estratégias de enfrentamento aos comportamentos de risco, rastreamento preventivo de doenças cardiometabólicas e implementação de políticas institucionais com o foco em saúde do trabalhador.

REFERENCES

[1] Brunna de Sousa Pereira Amorim et al. 2019. **Uso de aprendizado de máquina para classificação de risco de acidentes em rodovias.** (2019).
[2] Pierre Baldi, Søren Brunak, and Francis Bach. 2001. *Bioinformatics: the machine learning approach*. MIT press.
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.
[5] Jonathan H Chen and Steven M Asch. 2017. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine* 376, 26 (2017), 2507.
[6] Vinícius Paes de Camargo. 2018. Aplicação de Aprendizado de Máquina Supervisionado para Predição de Custos de Medicamentos Após Cirurgia Bariátrica no SUS/PR.
[7] Hellen Geremias dos Santos, Carla Ferreira do Nascimento, Rafael Izbicki, Yeda Aparecida de Oliveira Duarte, Alexandre Dias Porto Chiavegatto Filho, et al. 2019. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. *Cad. Saúde Pública* 35, 7 (2019), e00050818.
[8] Katti Faceli, Ana Carolina Lorena, João Gama, André Carlos Ponce de Leon Carvalho, et al. 2011. **Inteligência Artificial: Uma abordagem de aprendizado de máquina**. (2011).
[9] Kyle J Foreman, Neal Marquez, Andrew Dolgert, Kai Fukutaki, Nancy Fullman, Madeline McGaughey, Martin A Pletcher, Amanda E Smith, Kendrick Tang, Chun-Wei Yuan, et al. 2018. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet* 392, 10159 (2018), 2052–2090.
[10] D Georgios, B Fernando, and L Felix. 2018. Oversampling for imbalanced learning based on K-means and SMOTE. *Inf. Sci.* 465 (2018), 1–20.
[11] Mark A Green. 2018. Use of machine learning approaches to compare the contribution of different types of data for predicting an individual's risk of ill health: an observational study. *The Lancet* 392 (2018), S40.
[12] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
[13] Haibo He. 2009. Member, IEEE, and Edwardo A. Garcia," Learning from Imbalanced Data,". *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1041–4347.
[14] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
[15] Ana Silvia Degasperi Ieker. 2017. *Nível de atividade física, perfil de saúde e fatores de risco para doenças crônicas em servidores de um hospital escola do sul do Brasil*. Master's thesis. Universidade Estadual de Maringá.
[16] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112. Springer.
[17] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.
[18] Max Kuhn, Kjell Johnson, et al. 2013. *Applied predictive modeling*. Vol. 26. Springer.
[19] Marcos Nascimento Magalhães and Antonio Carlos Pedroso de Lima. 2002. *Noções de probabilidade e estatística*. Vol. 5. Editora da Universidade de São Paulo.
[20] Luis J Mena, Eber E Orozco, Vanessa G Felix, Rodolfo Ostos, Jesus Melgarejo, and Gladys E Maestre. 2012. Machine learning approach to extract diagnostic and prognostic thresholds: application in prognosis of cardiovascular mortality. *Computational and mathematical methods in medicine* 2012 (2012).
[21] Thoralf Mildenberger. 2014. Stephen marsland: Machine learning. An algorithmic perspective. *Statistical Papers* 55, 2 (2014), 575.
[22] Maria Carolina Monard and José Augusto Baranauskas. 2003. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações* 1, 1 (2003), 32.
[23] André Rodrigues Olivera, Valter Roesler, Cirano Iochpe, Maria Inês Schmidt, Álvaro Vigo, Sandhi Maria Barreto, and Bruce Bartholow Duncan. 2017. Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes-ELSA-Brasil: accuracy study. *Sao Paulo Medical Journal* 135, 3 (2017), 234–246.
[24] Gregory Piatetsky-Shapiro. 1991. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases* (1991), 229–238.
[25] RC Prati, GEAPA Batista, and MC Monard. 2008. Curvas ROC para avaliação de classificadores. *Revista IEEE América Latina* 6, 2 (2008), 215–222.
[26] Sebastian Raschka and Vahid Mirjalili. 2017. Python Machine Learning: Machine Learning and Deep Learning with Python. In *scikit-learn, and TensorFlow*. Packt Publishing.
[27] Sebastian Raschka and Vahid Mirjalili. 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.
[28] Hellen Geremias dos Santos. 2018. *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*. Ph.D. Dissertation. Universidade de São Paulo.
[29] Sari Stenholm, Jenny Head, Mika Kivimäki, Ichiro Kawachi, Ville Aalto, Marie Zins, Marcel Goldberg, Paola Zaninotto, Linda Magnuson Hanson, Hugo Westerlund, et al. 2016. Smoking, physical inactivity and obesity as predictors of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. *International journal of epidemiology* 45, 4 (2016), 1260–1270.
[30] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one* 12, 4 (2017), e0174944.