PREVER DIABETES ATRAVÉS DA MINERAÇÃO DE DADOS

Guilherme Almeida[1], Vinicius Santos[2] e Murilo Varges [3]

[1]Estudante do Curso de Engenharia da Computação, IFSP Câmpus Birigui
alves.g@aluno.ifsp.edu.br

[2]Estudante do Curso de Engenharia da Computação, IFSP Câmpus Birigui
vinicius.santos@ifsp.edu.br

[3]Doutorado em Ciência da Computação, professor do Instituto Federal de São Paulo, Câmpus Birigui, murilo.varges@ifsp.edu.br

Resumo

Este estudo buscou avaliar e comparar diferentes algoritmos de aprendizado de máquina para a classificação de dados de pesquisa médica. O objetivo foi identificar o classificador mais eficiente e determinar a métrica de desempenho mais adequada. Foram utilizados cinco algoritmos populares, incluindo árvores de decisão, Máquina de vetores de suporte, K-vizinhos mais próximos e redes neurais artificiais. Os dados foram pré-processados e divididos em conjuntos de treinamento e teste. A acurácia, precisão, recall e F1-score foram utilizados como métricas de desempenho. Após a análise dos resultados, verificou-se que o classificador Máquina de vetores de suporte apresentou o melhor desempenho em termos de acurácia. A acurácia também se mostrou a métrica mais confiável para avaliar o desempenho dos modelos. No entanto, é importante considerar outras variáveis, como o tamanho da amostra e a representatividade dos dados, em futuras pesquisas. Esses resultados contribuem para o avanço da aplicação de algoritmos de aprendizado de máquina na classificação de dados biomédicos.

Palavras-chave: classificação, algoritmos de aprendizado de máquina, métricas de desempenho, dados médicos.

PREDICTING DIABETES THROUGH DATA MINING

Abstract

This study sought to evaluate and compare different machine learning algorithms for classifying medical research data. The objective was to identify the most efficient classifier and determine the most appropriate performance metric. Five popular algorithms were used, including Decision Trees, Support Vector Machine, K-Nearest Neighbors and Artificial Neural Networks. The data were pre-processed and divided into training and test sets. Accuracy, precision, recall and F1-score were used as performance metrics. After analyzing the results, it was verified that the Support Vector Machine classifier presented the best performance in terms of accuracy. Accuracy also

proved to be the most reliable metric to evaluate model performance. However, it is important to consider other variables, such as sample size and data representativeness, in future research. These results contribute to the advancement of the application of machine learning algorithms in the classification of biomedical data.

Keywords: classification, machine learning algorithms, performance metrics, medical data

1. INTRODUÇÃO

1.1 Doença

A diabetes é uma doença crônica que afeta milhões de pessoas em todo o mundo, inclusive no Brasil. Ela é caracterizada pela incapacidade do organismo de regular adequadamente os níveis de glicose no sangue, o que pode levar a complicações sérias e até mesmo fatais se não for tratada corretamente. A compreensão da diabetes e seu manejo têm evoluído ao longo dos anos, graças aos avanços na pesquisa médica e científica.

1.2 A descoberta da diabetes

A descoberta da diabetes remonta a séculos atrás, quando médicos e pesquisadores começaram a observar os sintomas característicos dessa condição e a investigar suas causas e efeitos no corpo humano. Segundo Silva et al. (2018), relatos históricos indicam que os primeiros registros da doença datam de 1500 a.C., no antigo Egito, onde médicos já descreviam uma condição que apresentava sintomas semelhantes aos da diabetes.

1.3 Função do pâncreas e ação da insulina

Para entender como a diabetes funciona, é fundamental compreender o papel do pâncreas e da insulina no organismo. O pâncreas é um órgão localizado no abdômen e desempenha diversas funções, incluindo a produção de enzimas digestivas e a regulação dos níveis de glicose no sangue.

De acordo com a Sociedade Brasileira de Diabetes (SBD), a insulina é um hormônio secretado pelas células beta do pâncreas. Sua principal função é permitir a entrada da glicose nas células, onde ela é utilizada como fonte de energia. Quando os níveis de glicose no sangue estão elevados, o pâncreas libera insulina para facilitar a absorção da glicose pelas células.

1.4 Mineração de Dados para prevenção de diabetes

A mineração de dados pode ser uma ferramenta útil para analisar e extrair insights de uma base de dados contendo informações relacionadas à diabetes. No caso específico da base de dados fornecida, que inclui atributos como 'Número de Gestações', 'Glucose', 'Pressão Arterial', 'Espessura da Pele', 'Insulina', 'IMC', 'Função Pedigree Diabete', 'Idade' e 'Resultado', a mineração de dados pode ajudar a identificar padrões e relações entre essas variáveis, fornecendo informações valiosas no contexto da diabetes.

GUILHERME ALMEIDA
VINICIUS SANTOS
MURILO VARGES

A análise exploratória dos dados permitirá visualizar e compreender a distribuição dos valores em cada atributo. Por exemplo, é possível realizar gráficos de dispersão ou histogramas para examinar a relação entre 'Número de Gestações' e 'Glucose', ou entre 'IMC' e 'Resultado'. Essas visualizações podem destacar possíveis correlações entre os atributos e indicar a presença de outliers ou dados faltantes.

Além disso, a mineração de dados pode ajudar a construir modelos preditivos para prever o resultado da diabetes com base nos atributos fornecidos. Técnicas como classificação, regressão ou árvores de decisão podem ser aplicadas nesse contexto. Ao treinar esses modelos com a base de dados disponível, é possível gerar previsões para novos casos, auxiliando na identificação precoce da diabetes ou no monitoramento do risco de desenvolvimento da doença.

Outra abordagem que a mineração de dados pode oferecer é a identificação de regras de associação. Essas regras permitem descobrir associações frequentes entre os atributos. Por exemplo, pode-se identificar que um determinado conjunto de valores em 'Glucose', 'Pressão Arterial' e 'IMC' está fortemente associado a um resultado positivo para diabetes. Essas regras de associação podem fornecer insights sobre os fatores que influenciam a ocorrência da doença.

É importante ressaltar que a aplicação da mineração de dados na área da diabetes requer um conhecimento especializado em ambas as áreas. É necessário compreender as particularidades da doença e suas variáveis, bem como as técnicas e algoritmos de mineração de dados adequados para extrair conhecimento relevante. Além disso, é fundamental garantir a proteção e a privacidade dos dados, seguindo as diretrizes éticas e legais para o uso de informações pessoais sensíveis.

Em resumo, a mineração de dados pode ser uma ferramenta poderosa para analisar e explorar uma base de dados relacionada à diabetes, permitindo a identificação de padrões, a construção de modelos preditivos e a descoberta de regras de associação. Essas análises podem contribuir para a compreensão da doença, a identificação de fatores de risco e o desenvolvimento de estratégias de prevenção e tratamento mais eficazes.

1.5 Sobre os valores da Base de dados

Os valores da base de dados fornecida possuem correlação com a doença da diabetes, pois cada um dos atributos pode desempenhar um papel no desenvolvimento, no controle ou nos fatores de risco associados à diabetes. Vamos analisar brevemente a possível relação de cada atributo com a doença:

Número de Gestações: Mulheres que tiveram gestações anteriores têm um maior risco de desenvolver diabetes gestacional, que é uma forma temporária de diabetes que ocorre durante a gravidez. Esse atributo está mais relacionado à diabetes gestacional do que aos outros tipos de diabetes.

Glucose: A medida da glicose no sangue é um indicador fundamental para o diagnóstico e monitoramento da diabetes. Altos níveis de glicose podem indicar a presença da doença ou problemas no controle glicêmico.

Pressão Arterial: A hipertensão arterial é uma condição comum em pessoas com diabetes e está frequentemente associada a complicações cardiovasculares, como doenças cardíacas e acidentes vasculares cerebrais.

Espessura da Pele: Embora a relação direta da espessura da pele com a diabetes não seja clara, é possível que a obesidade, que pode ser indicada pela espessura da pele, seja um fator de risco para o desenvolvimento da doença.

Insulina: A insulina é o hormônio responsável por regular os níveis de glicose no sangue. A resistência à insulina, ou seja, a diminuição da sua ação no organismo, é um fator-chave no desenvolvimento do diabetes tipo 2.

IMC (Índice de Massa Corporal): O IMC é uma medida que relaciona o peso e a altura de uma pessoa. O excesso de peso e a obesidade estão fortemente associados ao desenvolvimento do diabetes tipo 2. Um alto IMC indica um maior risco de desenvolver a doença.

Função Pedigree Diabete: Esse atributo pode representar a predisposição genética para a diabetes. A história familiar de diabetes é um fator de risco importante para o desenvolvimento da doença.

Idade: A idade avançada é um fator de risco para a diabetes tipo 2. Conforme envelhecemos, a capacidade do corpo de usar a insulina de maneira eficaz pode diminuir, aumentando o risco de desenvolver a doença.

É importante destacar que a correlação entre esses atributos e a diabetes não é linear e pode variar de acordo com o tipo e o estágio da doença, bem como outros fatores individuais. Para uma análise mais aprofundada e precisa, é necessário aplicar técnicas de mineração de dados, como análise estatística, construção de modelos preditivos e identificação de padrões, na base de dados específica que você possui. Isso permitirá uma compreensão mais completa das correlações e relações entre os atributos e a doença da diabetes.

2. OBJETIVOS GERAIS

Aplicar os conhecimentos adquiridos na disciplina de mineração de dados em um problema real de previsão de diabetes, utilizando técnicas de seleção, pré-processamento e transformação de dados, visualização, análise descritiva, análise de grupos, classificação e estimação/regressão.

2.1 Objetivos específicos

- Seleção e pré-processamento de dados;
- Normalização e redução de dados;

- Análise descritiva de dados - Visualização;
- Análise descritiva de dados - Medidas;
- Análise de grupos;
- Classificação - KNN;
- Classificação - SVM;

3. MATERIAIS E MÉTODOS

Nesta seção, apresentamos os materiais utilizados e a metodologia adotada para realizar o estudo de previsão de diabetes utilizando técnicas de mineração de dados. Descrevemos a fonte da base de dados, as etapas de pré-processamento, normalização e redução de dados, bem como as análises descritivas, a análise de grupos e a classificação por meio dos algoritmos K-NN e SVM. O ambiente de desenvolvimento utilizado foi o Visual Studio Code (VSCode), e a linguagem de programação adotada foi o Python.

3.1 Base de Dados

A base de dados utilizada neste estudo foi obtida do repositório Kaggle [1]. A base de dados consiste em informações de pacientes relacionadas a variáveis clínicas, como número de gestações, níveis de glicose, pressão arterial, espessura da pele, insulina, índice de massa corporal (IMC), função pedigree de diabetes, idade e resultado. A base de dados foi selecionada por sua relevância na previsão de diabetes e seu potencial para aplicação de técnicas de mineração de dados.

3.2 Pré-processamento de Dados

O pré-processamento dos dados foi realizado utilizando a linguagem de programação Python juntamente com as bibliotecas pandas e numpy. Nessa etapa, foram realizadas técnicas de limpeza e tratamento de dados, incluindo a identificação e tratamento de valores ausentes, a remoção de outliers e a normalização de valores discrepantes. Essas técnicas são fundamentais para garantir a qualidade dos dados utilizados nas análises subsequentes.

3.3 Normalização e Redução de Dados

Após o pré-processamento, os dados foram submetidos à etapa de normalização, visando a equalização das escalas dos atributos. Para isso, utilizou-se a biblioteca numpy para aplicar técnicas de normalização, como a normalização por escala mínima e máxima (MinMax) e a normalização por média e desvio padrão (Z-Score). Além disso, a técnica de Análise de Componentes Principais (PCA)

foi empregada para reduzir a dimensionalidade dos dados, visando a eliminação de redundâncias e a extração dos principais componentes explicativos.

3.4 Análise Descritiva de Dados - Visualização e Medida

A análise descritiva dos dados foi realizada para obter insights sobre as características dos pacientes e identificar possíveis padrões ou tendências. Utilizando a biblioteca matplotlib, foram geradas visualizações gráficas, como histogramas, gráficos de dispersão e gráficos de setores, a fim de representar as distribuições de frequência, a relação entre os atributos e outras informações relevantes presentes na base de dados.

Foram utilizadas medidas estatísticas descritivas, como média, mediana, desvio padrão e variância, para resumir as características dos dados. Essas medidas fornecem informações sobre a tendência central, dispersão e variabilidade dos atributos. Além disso, quartis, percentis e coeficientes de correlação também foram utilizados para descrever a distribuição dos dados e identificar relações entre os atributos. Essas medidas complementam as análises visuais e fornecem uma descrição quantitativa dos dados utilizados no estudo de previsão de diabetes.

3.5 Análise de Grupos

A análise de grupos foi realizada utilizando o algoritmo K-means. Antes da aplicação do algoritmo, os dados foram submetidos novamente ao pré-processamento e à normalização. Para facilitar a visualização dos resultados, utilizou-se a técnica de PCA para reduzir a dimensionalidade dos dados a dois componentes principais. Dessa forma, foi possível plotar os resultados dos agrupamentos em um espaço bidimensional.

Além disso, medidas de avaliação foram utilizadas para quantificar a qualidade dos agrupamentos obtidos. Foram considerados coeficientes de forma, homogeneidade e outras métricas relevantes para verificar a consistência e a separação dos grupos formados.

3.6 Classificação - KNN

Para a tarefa de classificação, foi empregado o algoritmo K-NN (K-vizinhos mais próximos). Antes da classificação, os dados passaram por uma etapa adicional de pré-processamento e normalização, garantindo a consistência e a qualidade dos dados. Em seguida, a base de dados foi dividida em conjuntos de treinamento e teste, utilizando os métodos holdout (70% para treinamento e 30% para teste) e cross-validation com k=10.

Após a divisão dos dados, o algoritmo K-NN foi aplicado para classificar os pacientes em relação à presença ou ausência de diabetes. Métricas como matriz de confusão, acurácia e F1 Score

foram utilizadas para avaliar o desempenho do modelo classificador, fornecendo informações sobre a precisão, a taxa de acerto e o equilíbrio entre precisão e recall.

3.7 Classificação - SVM

Outro algoritmo utilizado para a tarefa de classificação foi o SVM (Support Vector Machine). Assim como no caso do K-NN, os dados passaram pelas etapas de pré-processamento e normalização antes da aplicação do algoritmo SVM. Também foram utilizados os métodos holdout e cross-validation para dividir a base de dados em conjuntos de treinamento e teste.

O algoritmo SVM foi aplicado para classificar os pacientes em relação à presença ou ausência de diabetes, e métricas como matriz de confusão, acurácia e F1 Score foram utilizadas para avaliar o desempenho do modelo classificador, fornecendo informações sobre a capacidade de discriminação e a eficácia da classificação.

4. RESULTADOS OBTIDOS

Para poder atingir o objetivo geral e específico nessa etapa do projeto foi separado em subtópicos com os resultados de cada tática de análise de dados seguindo com o proposto nos objetivos específicos.

4.1 Pré-processamento de Dados

Para poder iniciar o processo de limpeza dos dados temos que ter a dimensão dos conteúdos da base e podemos verificar ela perfeitamente na figura abaixo.

Figura 1. df.info() da base de dados

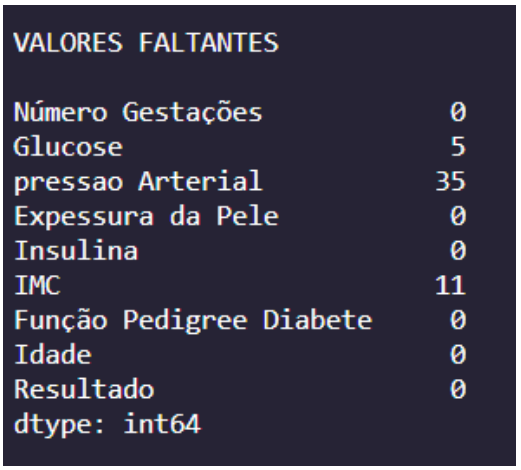
```
INFORMAÇÕES GERAIS DOS DADOS

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Número Gestações        768 non-null    int64
 1   Glucose                 763 non-null    float64
 2   pressao Arterial        733 non-null    float64
 3   Expessura da Pele       768 non-null    int64
 4   Insulina                768 non-null    int64
 5   IMC                     757 non-null    float64
 6   Função Pedigree Diabete 768 non-null    float64
 7   Idade                   768 non-null    int64
 8   Resultado               768 non-null    int64
dtypes: float64(4), int64(5)
memory usage: 54.1 KB
None
```

Fonte: produzida pelo próprio autor, 2023

GUILHERME ALMEIDA
VINICIUS SANTOS
MURILO VARGES

Tendo em vista o detalhamento da base de dados, precisa-se analisar dados faltantes e nessa etapa de pré-processamento verificamos esses dados e aplicamos métricas para tratá los, segundo McKINNEY 2012 [2] para uma base desbalanceada o correto aplicar a média pois segundo ele a média é a medida mais apropriada quando os dados seguem uma distribuição normal, com isso foi verificado a presença de dados faltantes e chegamos no resultado da figura 2.

Figura 2. df.isnull().sum() da base de dados



Fonte: produzida pelo próprio autor, 2023

Aplicando a moda nos dados faltantes e gerando um novo arquivo para ser utilizado nas análises já tratado é denominado de: diabetesClear.data, com isso foi feio a limpeza de dados e terminado a etapa de pré-processamento.

4.2 Normalização e Redução de Dados

A normalização é o processo de escalar os dados para que todos estejam na mesma escala. Isso é importante porque algumas variáveis podem ter uma escala muito maior do que outras e podem ter um impacto desproporcional na análise. Existem diferentes métodos de normalização, mas um dos mais comuns é a normalização Min-Max, que ajusta os valores para um intervalo entre 0 e 1. Para a base dados foi escolhido a normalização Min-Max que surtirá efeito positivos na análises das próximas técnicas e algoritmos, como a base é muito extensa e seria inviável mostrar os dados normalizados aqui segue um exemplo de alguns dados parciais normalizados.

Figura 3. MinMaxScaler().fit_transform() da base de dados

GUILHERME ALMEIDA
VINICIUS SANTOS
MURILO VARGES

	Número Gestações	Glucose
0	0.352941	0.670968
1	0.058824	0.264516
2	0.470588	0.896774
3	0.058824	0.290323
4	0.000000	0.600000
5	0.294118	0.464516
6	0.176471	0.219355
7	0.588235	0.458065
8	0.117647	0.987097
9	0.470588	0.522581

Fonte: produzida pelo próprio autor, 2023

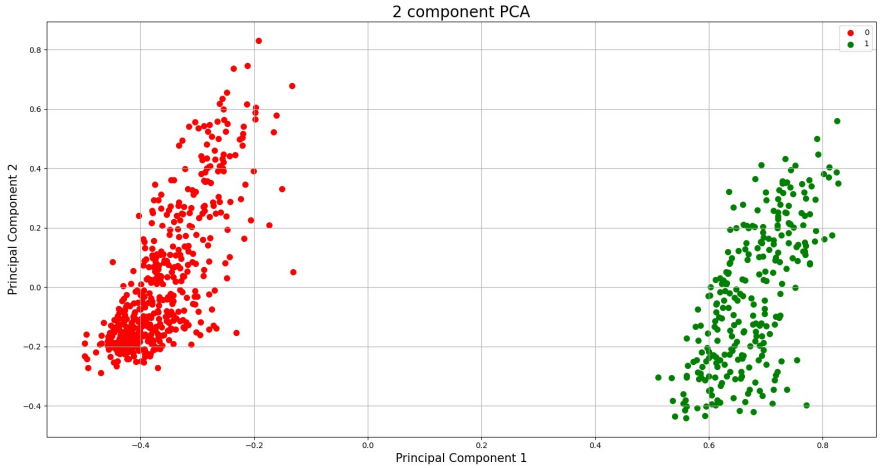
4.2.1 Redução de dados

A redução de dados é o processo de reduzir a dimensionalidade dos dados para eliminar variáveis desnecessárias ou redundantes. Isso pode ajudar a reduzir o tempo de processamento e melhorar a precisão do modelo. Uma técnica comum de redução de dados é a Análise de Componentes Principais (PCA), que reduz as variáveis para um número menor de componentes que capturam a maior parte da variação dos dados.

O PCA (Principal Component Analysis) é um método utilizado em análise de dados para reduzir a dimensionalidade de um conjunto de variáveis correlacionadas. Ele funciona transformando as variáveis originais em novas variáveis não correlacionadas, chamadas de componentes principais, que explicam a maior parte da variação presente nos dados. Dessa forma, é possível simplificar a análise e identificar relações de dependência entre as variáveis de uma maneira mais clara. O PCA é uma técnica bastante utilizada em diversas áreas, como na análise de dados de imagem, por exemplo, para essa base de dados foi feito o PCA utilizando a normalização Min-Max e tivemos ótimos resultados pois como demonstra a figura 4 a distribuição esta bem dimensionada sendo possível utilizar alguns algoritmos classificadores e de regressão para saber melhor a acurácia da base e o quanto ela pode ser assertiva.

Figura 4. Análise do Gráfico PCA da base de dados

GUILHERME ALMEIDA
VINICIUS SANTOS
MURILO VARGES



Fonte: produzida pelo próprio autor, 2023

4.3 Análise Descritiva de Dados Visualização e Medição

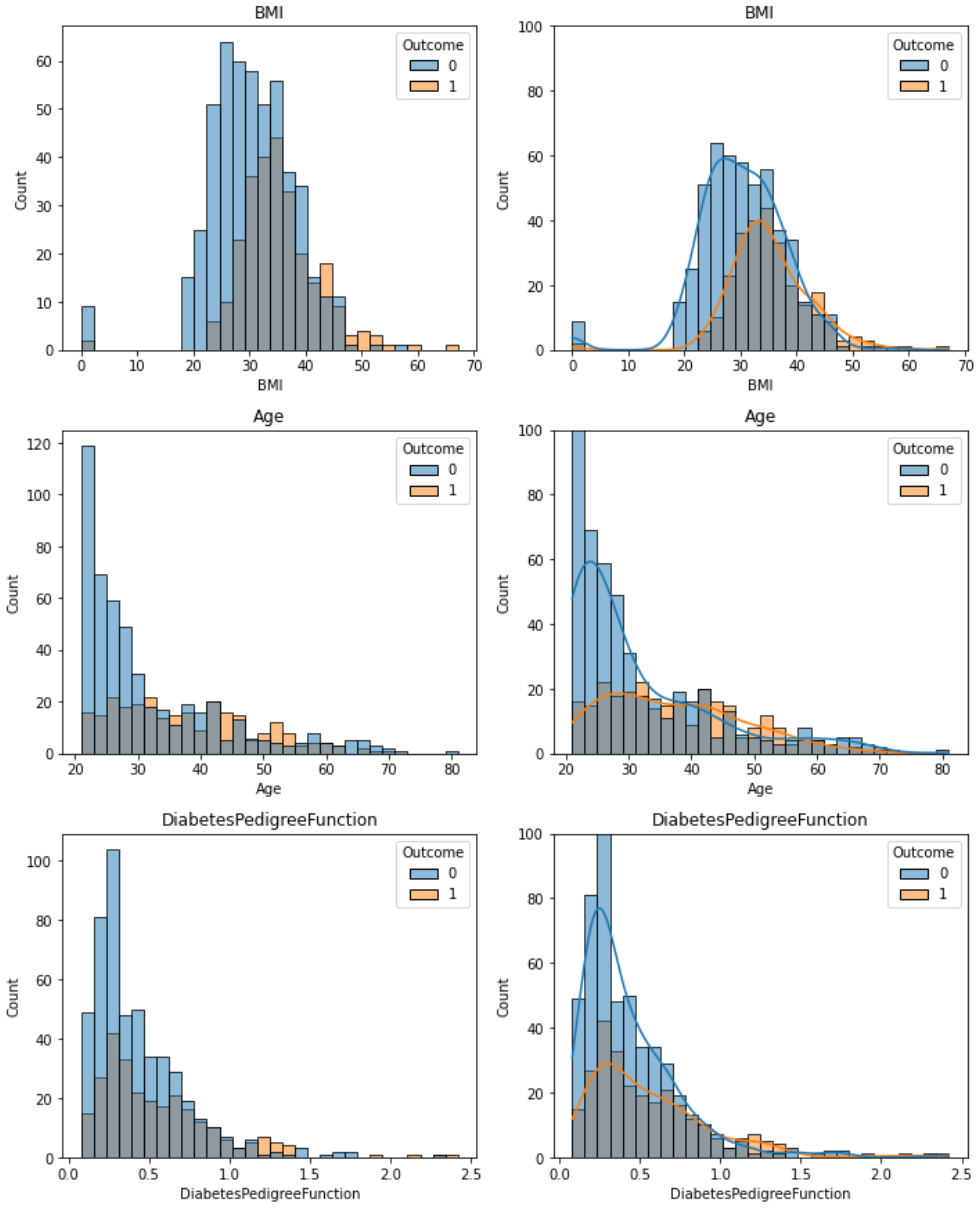
A análise descritiva de dados é uma técnica muito utilizada em mineração de dados, que tem como objetivo analisar e entender os dados de um conjunto de dados para obter informações relevantes. Essa técnica permite entender melhor as características dos dados, identificar padrões, tendências e anomalias, além de verificar a qualidade dos dados.

A visualização de dados é uma ferramenta essencial para a análise descritiva de dados, pois permite a representação gráfica dos dados de uma forma que facilita a compreensão e a interpretação das informações contidas no conjunto de dados. Com a visualização, é possível identificar padrões e tendências que não seriam detectados com apenas a análise dos dados em formato tabular.

Como forma de obter uma melhor visualização dos dados foi gerado um gráfico de histograma para alguns tipos de dados da base, com esse histograma pode-se observar a relação de resultados positivos (1) e negativos (0), na coluna de outcome.

Essa relação dá uma dimensão do que os valores de cada coluna têm em correlação com a coluna principal de resultados.

Figura 5. Análise do Gráficos de Histograma nas colunas 'BMI', 'Age', 'DiabetesPedigreeFunction' da base de dados



Fonte: produzida pelo próprio autor, 2023

As medidas estatísticas são ferramentas fundamentais na análise descritiva de dados. As medidas de tendência central, dispersão, posição relativa e associação são exemplos de medidas estatísticas utilizadas na análise descritiva de dados.

Medidas de tendência central: São medidas que indicam o ponto central de um conjunto de dados. As três medidas de tendência central mais comuns são a média, a mediana e a moda.

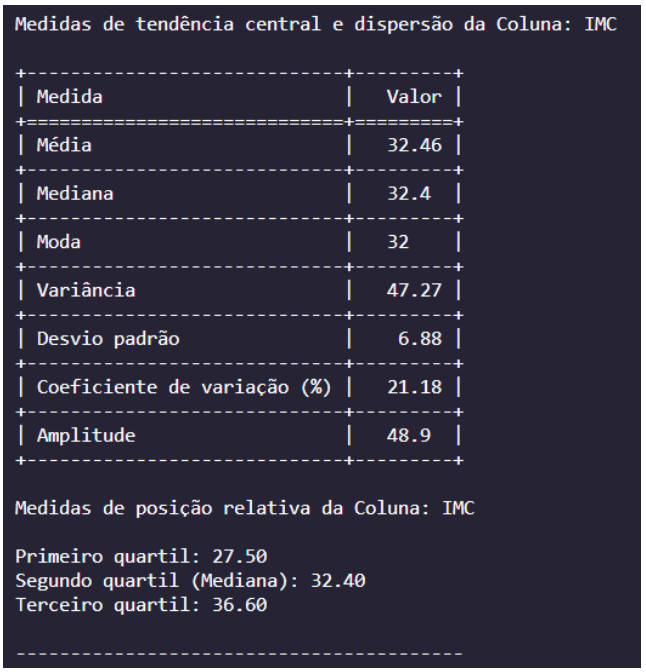
Medidas de dispersão: São medidas que indicam o grau de variação dos dados em relação a uma medida central. As medidas de dispersão mais comuns são o desvio padrão, a variância e o coeficiente de variação.

Medidas de posição relativa: São medidas que indicam a posição de um valor em relação aos demais valores de um conjunto de dados. As medidas de posição relativa mais comuns são o percentil e o quartil.

Medidas de associação: São medidas que indicam a relação entre duas ou mais variáveis. As medidas de associação mais comuns são a correlação e a covariância.

Como forma de experimento foi feito os conceitos de médias na coluna IMC e na imagem abaixo conseguimos verificar a saída dela.

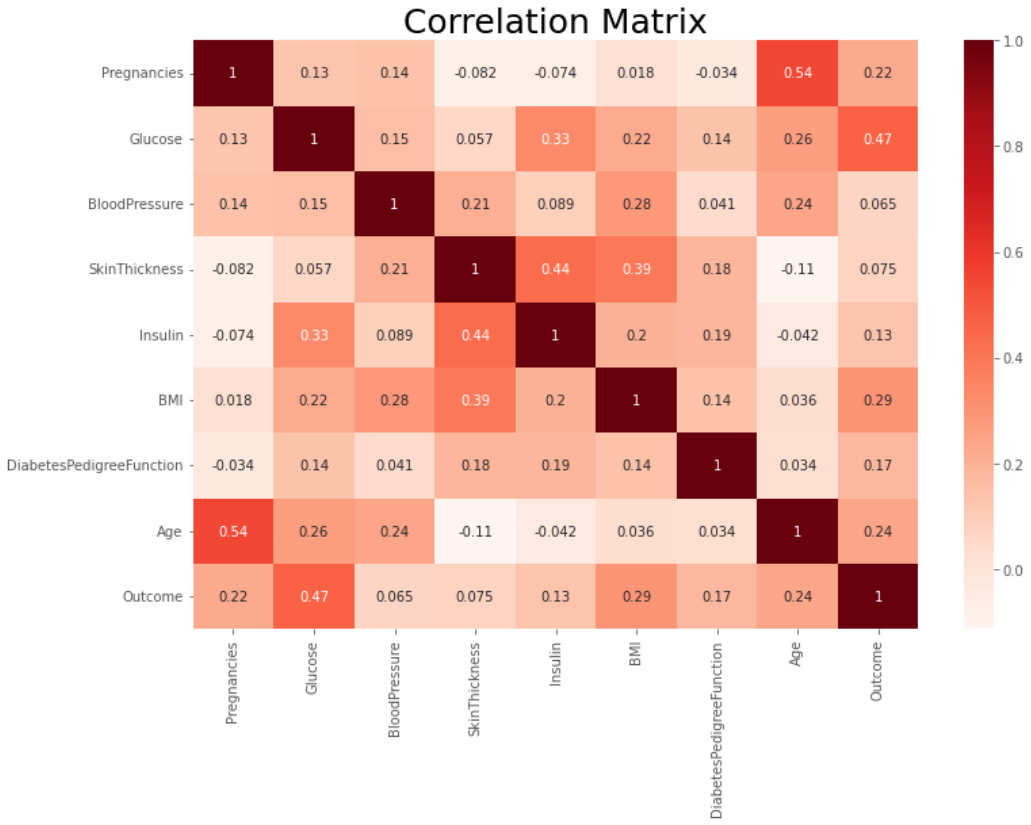
Figura 6. Análise Descritiva de Medidas da coluna IMC



Fonte: produzida pelo próprio autor, 2023

Como análise de valores também temos a matriz de correlação que é uma matriz quadrada que mostra a correlação entre pares de variáveis em um conjunto de dados. A matriz de correlação é frequentemente utilizada na análise exploratória de dados para entender a relação de dependência entre as variáveis e identificar padrões nos dados. Os valores da matriz variam entre -1 e 1, onde -1 representa uma correlação negativa perfeita, 0 representa nenhuma correlação e 1 representa uma correlação positiva perfeita entre as variáveis, veja o exemplo da mesma na figura abaixo.

Figura 7. Matriz de correlação da base de dados



Fonte: produzida pelo próprio autor, 2023

Em resumo, a análise descritiva de dados utiliza medidas estatísticas para resumir e descrever os dados, de forma a obter insights e compreender melhor as informações disponíveis. As medidas de tendência central, dispersão, posição relativa e associação são algumas das medidas estatísticas mais comuns utilizadas na análise descritiva de dados.

4.4 Análise de Grupos

A análise de grupo, também conhecida como análise de clustering, é uma técnica de mineração de dados que visa encontrar grupos de objetos semelhantes em um conjunto de dados.

Essa técnica se baseia em métodos de aprendizado não supervisionado, que buscam identificar estruturas subjacentes nos dados sem a necessidade de rótulos.

Existem diferentes métodos de análise de grupo, como análise hierárquica, k-means e mistura de gaussianas. Esses métodos utilizam medidas de distância para calcular a similaridade entre os objetos e agrupá-los em clusters.

A escolha da medida de distância depende do tipo de dados e do problema em questão. Além disso, é importante avaliar a qualidade dos clusters obtidos, considerando critérios como coesão interna e separação entre os grupos.

A análise de grupo é amplamente utilizada em áreas como biologia, marketing, finanças e ciência da computação para identificar padrões e estruturas em dados não rotulados, gerando insights e embasando a tomada de decisões.

Existem diversas ferramentas e técnicas para realizar análise de grupo, como análise hierárquica, k-means, mistura de gaussianas, análise de densidade e redução de dimensionalidade. A escolha da ferramenta depende das necessidades do projeto e da familiaridade do usuário.

O método k-means é uma técnica comum de análise de grupo. Ele agrupa os dados em k clusters, onde k é um valor pré-definido. O algoritmo do k-means envolve a inicialização dos centróides, a atribuição de objetos aos clusters mais próximos e a atualização dos centróides interativamente até a convergência.

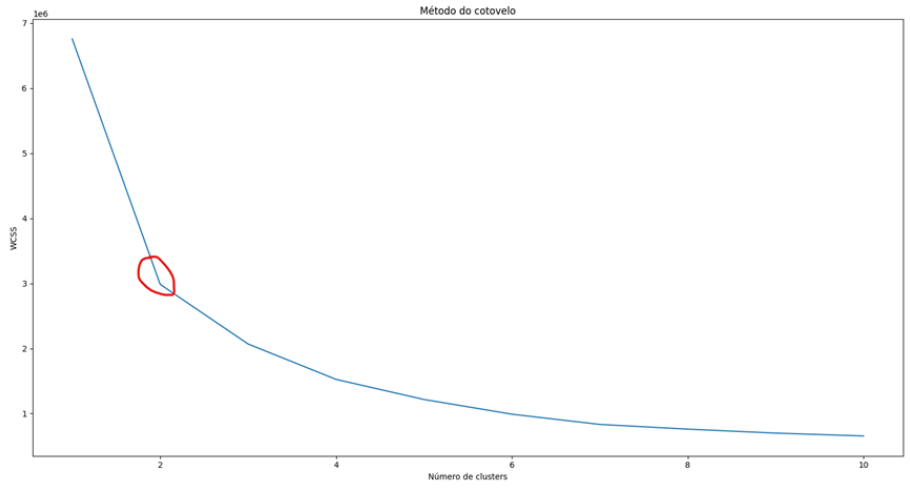
4.4.1 Kmeans

O k-means é eficiente em lidar com grandes conjuntos de dados, mas a qualidade dos clusters pode depender da escolha inicial dos centróides. É recomendado executar o algoritmo várias vezes com diferentes valores iniciais e selecionar a solução com a menor inércia, que é uma medida da dispersão dentro dos clusters.

O número ideal de clusters pode ser determinado usando o "método do cotovelo". Esse método envolve a execução do k-means para diferentes valores de k e a plotagem da inércia em relação ao número de clusters. O ponto de inflexão na curva indica o número ideal de clusters.

Para implementar o método do cotovelo, você pode usar bibliotecas como o scikit-learn e o matplotlib em Python. É possível ajustar o número máximo de iterações e outros parâmetros conforme necessário, um exemplo do método do cotovelo pode se observar na Figura 9, onde que para essa base ele determina que o número de clusters é 2.

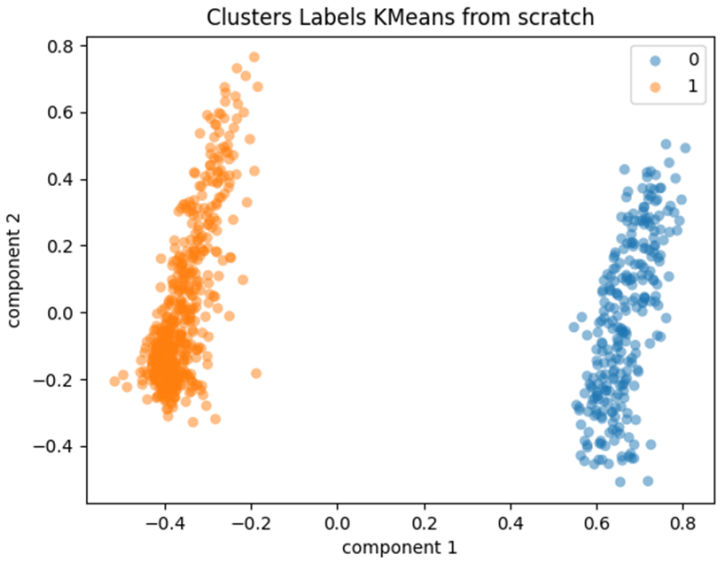
Figura 8. Método do Cotovelo para essa base de dados



Fonte: produzida pelo próprio autor, 2023

Usando a biblioteca from sklearn.cluster import KMeans e usando a distância euclidiana que é uma medida comumente usada para calcular a proximidade entre dois pontos em um espaço Euclidiano, no algoritmo K-means, a distância Euclidiana é usada para medir a dissimilaridade entre os pontos de dados e os centroides, ajudando a determinar quais pontos pertencem a cada grupo, podemos ver isso na figura 9, onde os pontos em azul parecem estar mais concentrados do que os pontos em laranja, sugerindo que os pontos azuis estão mais fortemente agrupados no caso as pessoas não diabéticas. No entanto, ainda há alguma sobreposição entre os clusters, com alguns pontos laranja parecendo estar próximos dos pontos azuis. Isso pode indicar que a separação entre os clusters não é completamente clara. Também podemos ver que os dados estão bastante dispersos, com pontos tanto na parte superior quanto na inferior do gráfico, sugerindo que existem diferentes combinações de valores nos recursos da base de dados. No geral, o comportamento do gráfico KMeans sugere que a base de dados pode não estar completamente agrupada e pode exigir mais análises para entender melhor as características dos dados e como eles podem ser agrupadas.

Figura 9. Análise do Kmeans pela distância euclidiana usando 2 clusters



Fonte: produzida pelo próprio autor, 2023

Em resumo, a análise de grupo é uma técnica valiosa para identificar grupos de objetos semelhantes em dados não rotulados. O método k-means é amplamente utilizado nessa análise, e o método do cotovelo que pode ser aplicado para determinar o número ideal de clusters.

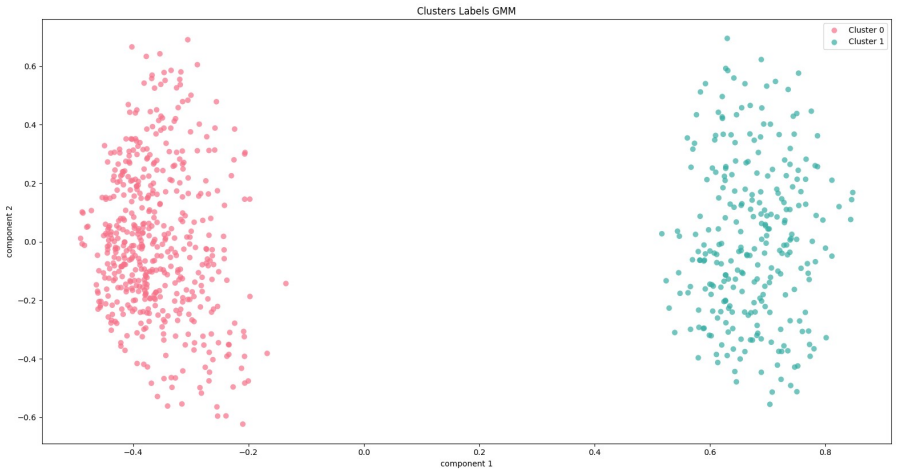
4.4.2 Gmm(Gaussian Mixture Model)

O algoritmo GMM (Gaussian Mixture Model) é um algoritmo de aprendizagem não supervisionada que é usado para modelar distribuições contínuas de dados. Ele é amplamente utilizado na mineração de dados por causa de sua capacidade de identificar subpopulações ocultas dentro de um

conjunto maior de dados. Mais especificamente, o GMM pode ajustar distribuições de probabilidade a dados multidimensionais, permitindo que sejam usados na modelagem de dados complexos.

O GMM é muito importante na mineração de dados porque pode ser usado para identificar claramente as distribuições subjacentes de dados que podem ser de interesse para análises adicionais. Por exemplo, um conjunto de dados com várias subpopulações distintas pode ter diferentes necessidades de análise devido à natureza dessas subpopulações. Nesse caso, a capacidade do GMM de identificar essas subpopulações pode ajudar o analista de dados a identificar quais subpopulações estão presentes e, em seguida, ajustar a análise em conformidade.

Figura 10. Análise do Gmm pela distância euclidiana usando 2 clusters



Fonte: produzida pelo próprio autor, 2023

Com base no gráfico, posso ver que existem dois clusters bem definidos. Cada cluster está representado por um conjunto de pontos próximos uns aos outros. Isso sugere que existem dois grupos distintos de dados na base de dados. Além disso, podemos ver que o primeiro eixo (componente 1) parece ser mais importante na separação dos dados em clusters do que o segundo eixo (componente 2). O comportamento do gráfico sugere que o modelo GMM conseguiu encontrar uma boa representação dos dados em dois clusters. Isso pode ser útil para entender melhor a estrutura dos dados e para tarefas de classificação de dados futuras.

Em resumo, o GMM é capaz de capturar a sobreposição entre grupos por meio da atribuição de probabilidades ponderadas para cada componente gaussiana, permitindo que os pontos de dados sejam associados a múltiplos grupos com diferentes graus de pertencimento.

4.5 Classificação

A classificação em mineração de dados refere-se ao processo de atribuir rótulos ou categorias a objetos ou instâncias com base em suas características ou atributos. O objetivo é construir um

modelo capaz de aprender padrões a partir de um conjunto de dados rotulados e usar esse modelo para prever a classe de novos dados não rotulados.

Aqui está um resumo dos métodos de classificação utilizados nesse projeto:

1. Árvore de Decisão: É um modelo que toma decisões com base em perguntas hierárquicas sobre os atributos dos dados. Ele constrói uma estrutura de árvore onde cada nó representa um atributo e cada ramo representa uma decisão ou resultado.

```
# Criar e treinar classificador de árvore de decisão
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
```

2. KNN (K-Nearest Neighbors): É um algoritmo que classifica uma instância desconhecida com base na maioria das classes das instâncias vizinhas mais próximas. O valor de K define o número de vizinhos considerados.

```
# Criar e treinar classificador KNN
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
```

3. SVM (Support Vector Machines): É um método que mapeia os dados em um espaço de alta dimensão para separar as classes por meio de um hiperplano de máxima margem. Ele busca encontrar o melhor hiperplano que separa os dados com a maior distância entre as classes.

```
# Crie e treine o classificador SVM
svm = SVC()
svm.fit(X_train, y_train)
```

4. Rede Neural MLP (Multilayer Perceptron): É um modelo de rede neural artificial que consiste em camadas de neurônios interconectados. Cada neurônio recebe entradas ponderadas, aplica uma função de ativação e passa os resultados para a próxima camada.

```
# Crie um classificador de rede neural com 2 camadas ocultas cada
uma com 4 neurônios
nn = MLPClassifier(hidden_layer_sizes=(4,4), max_iter=1000)
```

Aqui estão algumas métricas e métodos comuns de avaliação de modelos de classificação:

- Matriz de Confusão: É uma tabela que mostra a contagem de classificações corretas e incorretas feitas por um modelo. Ela compara as classes reais com as classes previstas.

```
Confusion Matrix:
```

```
Neural Network:
[[75 28]
 [10 41]]
Decision Tree:
[[81 22]
 [18 33]]
K-Nearest Neighbors:
[[72 31]
 [13 38]]
Support Vector Machine:
[[79 24]
 [ 9 42]]
```

- Acurácia: É a proporção de instâncias corretamente classificadas em relação ao total de instâncias. É uma medida geral de precisão, mas pode ser enganosa em conjuntos de dados desbalanceados.
  Neural Network: 0.76
  Decision Tree: 0.74
  K-Nearest Neighbors: 0.71
  Support Vector Machine: 0.79
- F1 Score: É uma medida que combina precisão e recall em uma única métrica. É útil quando há um desequilíbrio nas classes.
  Neural Network: 0.69
  Decision Tree: 0.63
  K-Nearest Neighbors: 0.64
  Support Vector Machine: 0.72
- Holdout (Treinamento 70% e Teste 30%): É uma técnica de avaliação em que os dados são divididos em um conjunto de treinamento e um conjunto de teste. O modelo é treinado no conjunto de treinamento e avaliado no conjunto de teste.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

- Cross-Validation (k=10): É uma técnica que divide os dados em k partes iguais (chamadas de folds). O modelo é treinado em k-1 folds e testado no fold restante. Esse processo é repetido k vezes e a média das métricas de avaliação é calculada.

- Balanceamento: nos nosso teste de classificação usamos a biblioteca from imblearn.over_sampling import SMOTE para balancear a base nos treinamentos

```
# Balanceamento de classe
oversample = SMOTE()
X_train_b,y_train_b= oversample.fit_resample(X_train,y_train)
```

Utilizando essa biblioteca from sklearn.metrics import accuracy_score, precision_score. Obtivemos os resultados das métricas exigidas no objetivo e elas seguem sendo demonstradas na Tabela 1 abaixo.

Tabela 1. Análise das métricas do sklearn para a base de dados.

Modelo	Acurácia	Precisão	f1_score
Árvore de Decisão	0.75	0.60	0.63
SVN	0.79	0.64	0.72
Rede Neural	0.76	0.60	0.69
KNN	0.72	0.55	0.64

Fonte: produzida pelo próprio autor, 2023

Considerando os modelos testados o SVN apresentou a melhor acurácia com cerca de 79%. Quando ampliamos a análise verificamos que o Recall apresentou valores de 0,77 para 0 = Não e 0,82 para 1= Sim. Isso significa que o modelo conseguiu identificar mais a Classe 1 que a classe 0. A análise do modelo em questão parece promissora se tratando de previsão. Além disso, antes de finalizar um diagnóstico de situação de saúde com base em modelos de Machine Learning, é essencial colocar um foco maior na interpretação da matriz de confusão como falsos positivos – falsos negativos podem ser arriscados.

Observação: Para esse estudo no tópico de classificação a base foi balanceada para ter uma melhor compreensão da classificação, conforme explicado no método SMOTE.

5. CONCLUSÃO

Com base nos resultados obtidos, conclui-se que o classificador SVN obteve o desempenho mais satisfatório em comparação aos demais classificadores avaliados. Além disso, a métrica de acurácia mostrou-se a mais adequada para avaliar o desempenho dos modelos, apresentando resultados consistentes e confiáveis. No entanto, é importante ressaltar que outros fatores, como o tamanho e a

representatividade da amostra de dados, podem influenciar os resultados e devem ser considerados em futuras pesquisas.

Em suma, este estudo exploratório utilizou técnicas de pré-processamento, análise descritiva e análise de grupos para analisar os dados disponíveis. Os resultados obtidos revelaram insights relevantes sobre o fenômeno em estudo, demonstrando a eficácia das técnicas aplicadas. Essas descobertas têm o potencial de contribuir para pesquisas futuras e aplicações práticas na área, melhorando a compreensão e possibilitando o desenvolvimento de soluções mais precisas. No entanto, é importante continuar validando os resultados e explorar outras abordagens para fortalecer as conclusões obtidas.

6. REFERÊNCIAS

[1] CHAUHAN, Aman. Predict Diabetes. Kaggle.com. Disponível em: <https://www.kaggle.com/datasets/whenamancodes/predict-diabities>. Acesso em: 22 fev. 2023.

[2] McKinney, Wes. Python for Data Analysis: Data Wrangling with Pandas, Numpy, and Ipython. 1st ed. Sebastopol, CA: O'Reilly Media, 2012.