Predição de diabetes com Naive Bayes: análise de dados médicos e demográficos de pacientes

Lucas Inojosa
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
lims@cin.ufpe.br

Nathalia Barbosa
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
nfab@cin.ufpe.br

Rafael Barros
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
rsb7@cin.ufpe.br

Sarah Melo
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brasil
slm2@cin.ufpe.br

***Abstract*—Este projeto tem como objetivo utilizar o algoritmo Naive Bayes para analisar o Diabetes prediction dataset, que consiste em dados médicos e demográficos de pacientes, juntamente com seu status de diabetes (positivo ou negativo). O dataset inclui características como idade, gênero, índice de massa corporal (IMC), hipertensão, doença cardíaca, histórico de tabagismo, nível de HbA1c e glicose no sangue. O objetivo é criar um modelo de aprendizado de máquina que possa prever o diabetes em pacientes com base em seu histórico médico e informações demográficas. Esse modelo pode ser útil para profissionais de saúde na identificação de pacientes que possam estar em risco de desenvolver diabetes e no desenvolvimento de planos de tratamento personalizados. Além disso, o dataset pode ser usado por pesquisadores para explorar as relações entre vários fatores médicos e demográficos e a probabilidade de desenvolver diabetes. As análises descritivas e inferenciais serão realizadas para avaliar a qualidade dos dados e verificar a adequação do modelo proposto. Espera-se que este projeto contribua para aprimorar a detecção precoce de diabetes e melhorar a qualidade de vida dos pacientes.**

***Index Terms*—Naive Bayes, Diabetes prediction, aprendizado de máquina.**

I. Objetivos

O objetivo deste projeto é analisar a chance de uma pessoa desenvolver diabetes dado algumas condições usando o algoritmo de classificação Naive Bayes. Para isso, está sendo analisada uma base dados de pessoas portadoras ou não-portadoras da doença levando em consideração fatores como índice de massa corporal (IMC), presença de doenças cardíacas, histórico de tabagismo, idade, gênero, nível HbA1c e nível de glicose no sangue. O uso do Naive Bayes permitirá identificar padrões entre essas variáveis, possibilitando, assim, uma melhor compreensão dos fatores de risco para o desenvolvimento da doença.

II. Justificativa

O algoritmo Naive Bayes é um classificador probabilístico frequentemente utilizado em Aprendizado de Máquina que usufrui do teorema de Bayes. Ele é um modelo simples e de fácil implementação que funciona bem na maioria dos casos, dessa forma iremos utilizar ele associado a algumas bibliotecas do Python para fazer nossas análises.

Escolhemos essa base de dados porque ela está bem organizada e contém informações relevantes para a análise que estamos realizando.

III. Metodologia

Nossa metodologia consiste em uma forma de determinar, através de análises estatísticas do nosso conjunto de dados, características não conhecidas do paciente através de campos conhecidos. Para este fim, utilizaremos o classificador ingênuo de Bayes.

A. A Base de Dados

A base de dados selecionada compreende informações sobre indivíduos com e sem diabetes, em nove campos, conforme a lista a seguir:

1) Gênero (gender)
2) Idade (age)
3) Hipertensão (hypertension)
4) Doença cardíaca (heart_disease)
5) Histórico de tabagismo (smoking_history)
6) Índice de Massa Corporal (IMC) (bmi)
7) Nível de HbA1c (HbA1c_level)
8) Nível de glicose no sangue (blood_glucose_level)
9) Diabetes (diabetes)

A maioria dos campos apresenta valores numéricos, exceto por dois campos: gênero e histórico de tabagismo. Para realizar a análise, é necessário converter esses campos em valores numéricos. Essa conversão pode ser feita atribuindo valores numéricos para cada opção no campo gênero e histórico de tabagismo, por exemplo, 0 para feminino e 1 para masculino no campo gênero, e 0 para nunca fumante, 1 para ex-fumante, 2 para fumante atual, etc, no campo histórico de

tabagismo. Dessa forma, os campos serão representados por valores numéricos, permitindo a análise dos dados por meio de técnicas estatísticas e de aprendizado de máquina.

B. O Classificador Probabilístico

1) Teorema de Bayes: O Teorema de Bayes foi desenvolvido pelo pastor e matemático inglês Thomas Bayes (1702-1761). O manuscrito que descrevia o teorema não foi publicado por Bayes, tendo sido extensivamente editado por Richard Price antes de ser submetido a Royal Society. A publicação do teorema só se deu após a morte de Bayes, em 1973 em Philosophical Transactions.

O teorema é representado pela Equação 1, como segue abaixo.

$$P(Y|X) = \frac{P(X|Y) \cdot P(X)}{P(Y)} \tag{1}$$

Em que:

- $P(X)$ é a probabilidade de um evento X acontecer;
- $P(Y)$ é a probabilidade de um evento Y acontecer;
- $P(X|Y)$ é a probabilidade de um evento X acontecer, dado que Y aconteceu;
- $P(Y|X)$ é a probabilidade de um evento Y acontecer, dado que X aconteceu.

2) Classificador de Bayes Ingênuo: O classificador de Bayes Ingênuo (ou Naive Bayes) é um algoritmo que gera uma tabela de probabilidades a partir de uma técnica de classificação de dados, que tem como finalidade categorizar objetos de acordo com a probabilidade do objeto pertencer a determinada categoria. Considera-se que os atributos são independentes — o que não ocorre na prática — por isso é chamado ”ingênuo”.

IV. Cronograma de atividades

semana1	semana2	semana3	semana4	semana5
1	1	0	0	0
0	2	2	0	0
0	0	3	3	3
0	0	0	0	4

Nosso cronograma de atividades consiste em 5 conjuntos de atividades distintas realizadas em cada semana até o prazo final do projeto:

1) Implementar Biblioteca de Probabilidade
   A biblioteca de probabilidade será uma biblioteca que deverá realizar as seguintes funções (não necessariamente distintas):
   a) $P(Event, Condition)$: função que retorna um número associado à probabilidade de um evento simples ocorrer, dada uma única condição.
   b) $P([Event_1, Event_2, ..., Event_n], Condition)$: função que retorna um número associado à probabilidade de um evento composto ocorrer, dada uma única condição. **Esta função supõe que os eventos são independentes entre si**.
   c) $P([Event_1, Event_2, ..., Event_n])$: função que retorna a probabilidade total de um evento composto ocorrer, considerando todo o espaço amostral.
2) Classificar
   Nessa segunda etapa, implementaremos o classificador genericamente através de uma função:

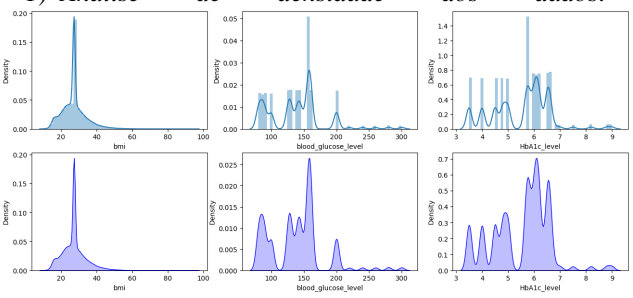
   Bayes Classifier$([Event_1, Event_2, ..., Event_n], Campo)$: função que retorna o a classe mais provável do campo especificado, dado os eventos já ocorridos.
3) Implementar no Banco
   Nesta Parte, os esforços serão direcionados para os nossos Objetivos com o banco de dados em si. Como as funções mais complexas já foram implementadas, a segunda etapa do nosso projeto visa utilizar estas funções para obter algumas previsões do nosso banco de dados.
4) Revisão Geral Do Projeto
   Por último, serão feitas revisões gerais do projeto, ajustando funções, modificando e melhorando implementaçõs para aumentar a eficiência do projeto.

V. Resultados

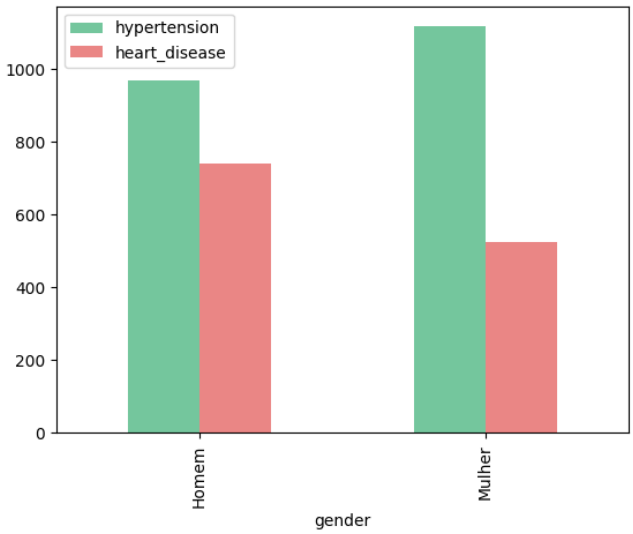
A. Plot de gráficos

O processo de análise de dados pode se tornar bastante complexo, especialmente quando se trata de grandes conjuntos de informações. A visualização gráfica desses dados é uma técnica muito útil para facilitar a compreensão dos padrões e tendências presentes nos mesmos. Neste sentido, o uso de ferramentas como seaborn e matplotlib se torna imprescindível para criar gráficos claros e concisos, que possam ser facilmente interpretados por profissionais de diferentes áreas de conhecimento. A seguir, apresentaremos alguns gráficos gerados a partir da base de dados selecionada, com o objetivo de ilustrar alguns dos principais insights obtidos através da análise exploratória dos dados.

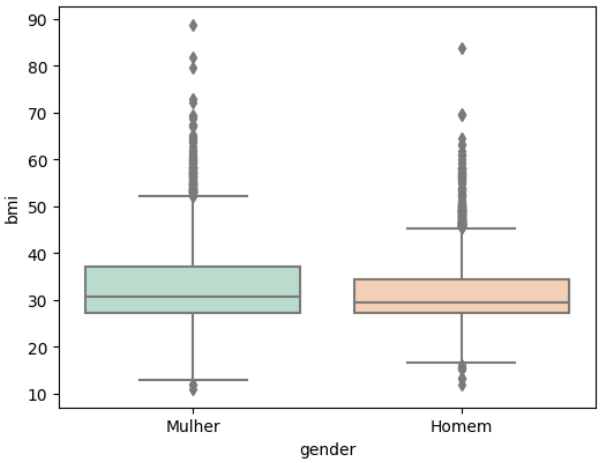
1) Análise de densidade dos dados:



2) Presença de Hipertensão e Problema Cardíaco em Diabéticos por Gênero:



3) Índice de Massa Corporal dos Diabéticos por Gênero:

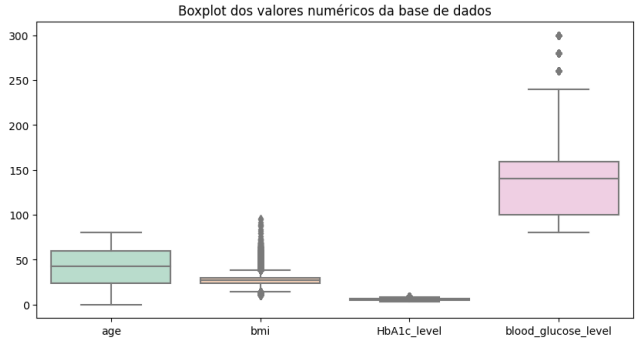


B. Identificação de Outliers através do Boxplot

O boxplot é um gráfico que fornece uma visualização da distribuição dos dados e é frequentemente usado para identificar a presença de outliers. Ele é uma ferramenta útil na análise exploratória de dados, pois ajuda a identificar valores extremos que podem afetar significativamente as conclusões que podem ser tiradas a partir dos dados. Ele consiste em um retângulo que representa o intervalo interquartil (IQR) dos dados, ou seja, a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) da distribuição. Dentro do retângulo, uma linha horizontal representa a mediana dos dados. Fora do retângulo, linhas horizontais (chamadas de ”bigodes”) se estendem até os valores mínimo e máximo dos dados que não são considerados outliers. Qualquer valor que esteja fora do intervalo dos bigodes é considerado um outlier e é representado por um ponto individual no gráfico. Esses pontos são frequentemente considerados incomuns ou anômalos em relação aos demais dados na distribuição.

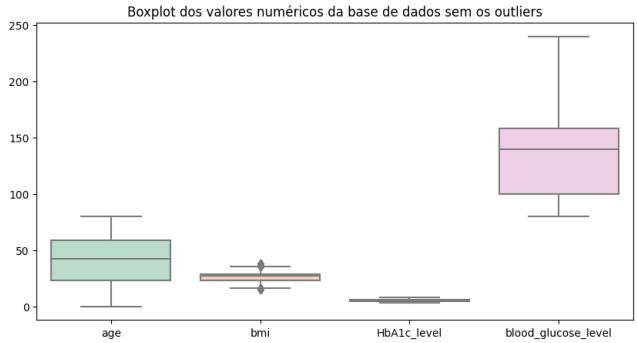
Ao interpretar um boxplot, é fundamental considerar a posição e a quantidade de outliers, bem como a forma da distribuição dos dados. Quando a distribuição é muito assimétrica, pode indicar que a média não é uma medida adequada de tendência central, e que os outliers podem ter um impacto maior na análise. É importante prestar atenção a esses detalhes para tirar conclusões mais precisas a partir do boxplot.

Abaixo está o boxplot dos dados numéricos da base de dados analisada. O uso dessa ferramenta permitiu identificar possíveis outliers na distribuição dos dados, o que pode ser útil na hora de entender melhor a natureza dos dados e as relações entre as variáveis.



1) Remoção de Outliers: Para garantir uma análise mais precisa e confiável dos dados, é importante remover valores anômalos que podem distorcer as estatísticas e métricas utilizadas em técnicas de aprendizado de máquina, como o Gaussian Bayes. Para isso, uma base de dados secundária foi criada, a qual desconsidera os valores anômalos observados na base original.

A remoção desses dados foi realizada utilizando a regra do 1,5 IQR, a qual define limites superior e inferior baseados no intervalo interquartil (IQR) dos dados. Valores que estiverem acima ou abaixo desses limites são considerados outliers e removidos do conjunto de dados. Dessa forma, os dados restantes representam de maneira mais fiel a distribuição original, garantindo uma análise mais precisa e confiável. Essa técnica é comumente utilizada em análise exploratória de dados e pode ser aplicada em diferentes tipos de conjuntos de dados. Abaixo pode-se observar o novo boxplot, sem os valores anômalos:



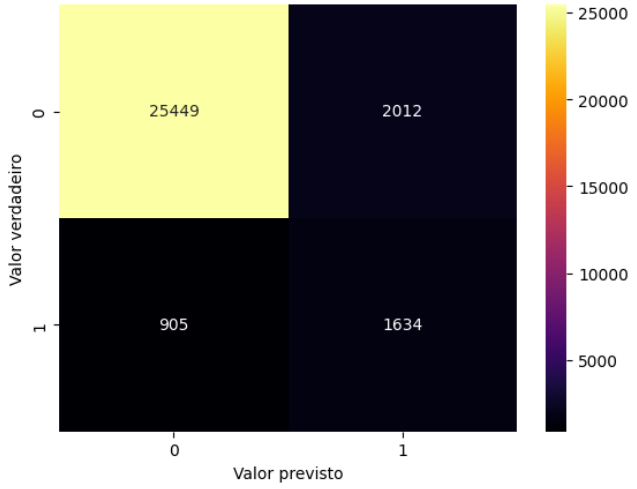
C. Aplicação do Gaussian Bayes

Após o pré-processamento dos dados do conjunto de dados, é hora de aplicar o algoritmo Gaussian Bayes e avaliar sua precisão, tanto com os dados brutos quanto sem a presença de outliers.

1) Com a presença de outliers: Considerando os outliers, tivemos os seguintes resultados:

- Acurácia: 0.90351
- Probabilidade de ser diabético: 0.085
- Probabilidade de não ser diabético: 0.915
- Média de cada feature para diabéticos: [5.24823529e-01 6.09465882e+01 2.45647059e-01 1.49058824e-01 2.39929412e+00 3.19883824e+01 6.93495294e+00 1.94094706e+02]
- Média de cada feature para não diabéticos: [5.91551913e-01 4.01151869e+01 5.89836066e-02 2.92349727e-02 2.33946448e+00 2.68871635e+01 5.39676066e+00 1.32852470e+02]

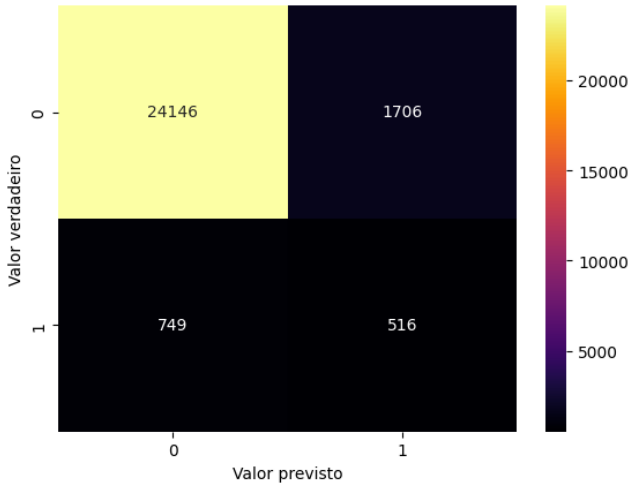
Esses resultados indicam que o algoritmo teve uma acurácia de 0.90351, o que significa que 90,35% das classificações foram corretas. Abaixo é possível observar a matriz de confusão:



2) Sem a presença de outliers: Desconsiderando os outliers, tivemos os seguintes resultados:

- Acurácia: 0.90825
- Probabilidade de ser diabético: 0.049
- Probabilidade de não ser diabético: 0.95
- Média de cada feature para diabéticos: [5.03448276e-01 6.19174638e+01 2.40266963e-01 1.53726363e-01 2.43114572e+00 2.93079889e+01 6.57915462e+00 1.67262291e+02]
- Média de cada feature para não diabéticos: [5.85700647e-01 4.03344889e+01 5.65128301e-02 2.94439529e-02 2.33975225e+00 2.61611935e+01 5.39756438e+00 1.32861815e+02]

Foi observado um aumento na acurácia, o que pode ser um bom sinal. Abaixo é possível observar a matriz de confusão:

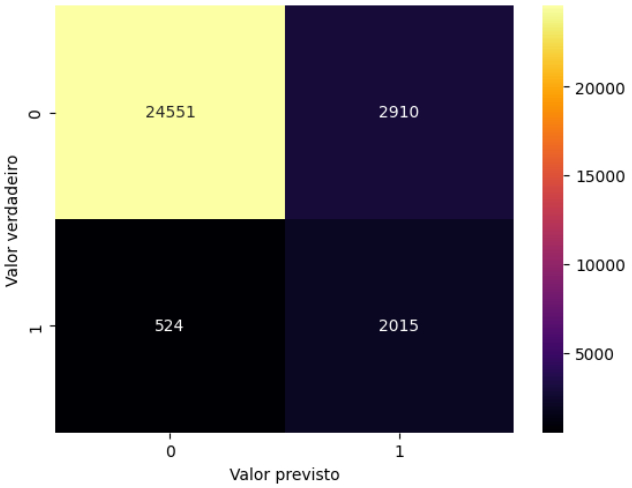


D. Uso da técnica SMOTE

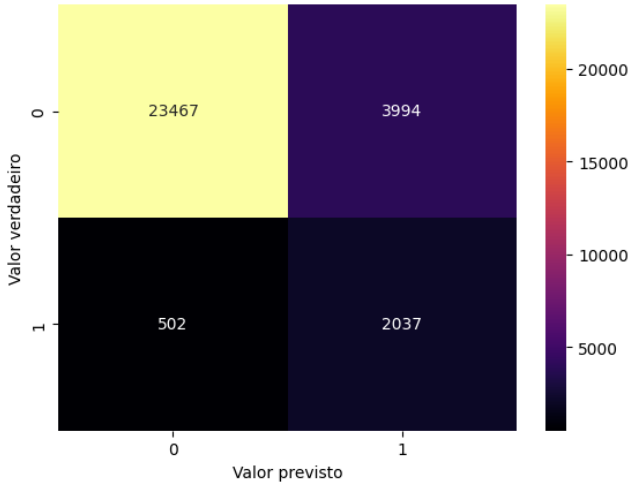
A técnica SMOTE para equilibrar classes é uma abordagem que tem sido amplamente utilizada em problemas de aprendizado de máquina. Ela aumenta a quantidade de dados da classe minoritária, incluindo dados sintéticos, a fim de reduzir o viés causado pela baixa disponibilidade de dados da classe minoritária. Nesta seção, descreveremos como aplicamos a técnica SMOTE à nossa base de dados para lidar com o desequilíbrio entre pessoas com diabetes e sem diabetes. Para avaliar o desempenho do modelo de classificação Gaussian Bayes, faremos uma comparação entre os resultados obtidos a partir da base de dados original e da base de dados secundária, que excluiu os outliers.

1) Com a presença de outliers: Primeiramente, aplicamos o SMOTE aos dados com outliers e obtivemos uma significativa redução na quantidade de falsos negativos, que passaram a representar apenas 20% das previsões quando a pessoa realmente possuía a doença. No entanto, a acurácia do modelo diminuiu devido ao aumento na quantidade de falsos positivos, uma vez que o SMOTE não melhora a qualidade dos dados, apenas aumenta a quantidade de dados.

Considerando a presença de outliers, a acurácia do modelo foi de 0.88653, como pode ser observado na matriz de confusão abaixo:



2) Sem a presença de outliers: Por outro lado, desconsiderando os outliers, tivemos uma acurácia de 0.85243. Abaixo está a matriz de confusão:



VI. Conclusão

Inicialmente, utilizando somente o algoritmo naive bayes, havia sido obtida uma acurácia de 90.3%, todavia quando observarmos os falsos negativos, ou seja pessoas diabéticas que teriam recebido o resultado negativo, havia uma alta quantidade muito alta. Como os falsos negativos são um problema grave para a nossa análise, optamos então para o uso da técina SMOTE para reduzir esse erros. Com isso, reduzimos a nossa acurácia para 88,6%, todavia reduzimos a quantidade de falsos negativos, obtendo uma acurácia de 80% para pessoas com diabetes.

References

[1] MEYER, Paul L. Probabilidade: aplicações à estatística. 2. ed. Rio de Janeiro: LTC, 2000. 546 p.

[2] STANFORD ENCYCLOPEDIA OF PHILOSOPHY. Bayes' Theorem. Disponível em: https://plato.stanford.edu/entries/bayes-theorem/. Acesso em: 02 mar. 2023.

[3] SACRAMENTO, Gabriel. Naive Bayes: Como funciona esse algoritmo de classificação. Disponível em: https://blog.somostera.com/data-science/naive-bayes. Acesso em: 02 mar. 2023.

[4] Diabetes prediction dataset, A Comprehensive Dataset for Predicting Diabetes with Medical and Demographic Data. Disponível em: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?datasetId=3102947sortBy=dateRuntab=profile. Acesso em: 24 abr. 2023.