Aplicação de classificador ingênuo de bayes para diagnóstico de doenças cardiovasculares

*Note: Sub-titles are not captured in Xplore and should not be used

1st Gabriel da Silva Soares
Centro de Informática
UFPE
Recife, Brasil
gss12@cin.ufpe.br

2nd Gabriel Ferreira da Silva
Centro de Informática
UFPE
Recife, Brasil
gfs4@cin.ufpe.br

3rd José Janailson de Arruda Cunha
Centro de Informática
UFPE
Recife, Brasil
jjac@cin.ufpe.br

4th Pedro Rodrigues Domingues da Silva
Centro de Informática
UFPE
Recife, Brasil
prds@cin.ufpe.br

***Abstract*—Neste trabalho buscamos aplicar o classificador ingêneo de bayes para ajudar uma pessoa, com um certo conjunto de atributos, a diagnostica-lo com uma doença cardiovascular. Usaremos uma base da dados da UCI machine Learning repository - 'heart disease data set' - e, através dela , extrairemos todos os resultados e conclusions contidos neste documento.**

***Index Terms*—doenças cardiovasculares, teorema de bayes, classificador de bayes, Conjunto de dados, machine learning**

I. Introdução

As doenças cardiovasculares são uma das principais causas de morte em todo o mundo. Por isso, o diagnóstico precoce é essencial para aumentar as chances de tratamento e recuperação. Nesse contexto, a aplicação de técnicas de aprendizado de máquina pode ser uma ferramenta útil para auxiliar no diagnóstico dessas doenças. Um exemplo é o classificador ingênuo de Bayes, que pode ser utilizado para identificar padrões e relacionamentos entre os sintomas apresentados pelo paciente e as doenças cardiovasculares. Nesta área, o uso do classificador ingênuo de Bayes tem se mostrado promissor, permitindo uma avaliação mais rápida e precisa do diagnóstico, possibilitando tratamentos mais eficazes. Neste contexto, este tema se propõe a discutir a aplicação do classificador ingênuo de Bayes para o diagnóstico de doenças cardiovasculares.

A detecção precoce da da doença renal crônica, em conjunto com tratamento terapêutico adequado para o retardamento do progresso da DCV pode reduzir o sofrimento dos pacientes, bem como minimizar os recursos financeiros associados ao tratamento. Portanto, torna-se de vital importância um mecanismo que auxilie os médicos, que trabalham na área básica de saúde, de modo que eles consigam utilizar alguns parâmetros para conseguir elaborar uma previsão acerca da probabilidade daquele paciente possuir DCV.

Identify applicable funding agency here. If none, delete this.

II. Metodologia

A. Dataset

A base de dados escolhida para a análise contém 4 bancos de dados sobre diagnóstico de doenças cardíacas. Todos os atributos têm valor numérico. Os dados foram colhidos do quatro locais a seguir:

1. Cleveland Clinic Foundation (cleveland.data);
2. Instituto Húngaro de Cardiologia, Budapeste (hungarian.data);
3. V.A. Centro Médico, Long Beach, CA (long-beach-va.data);
4. Hospital Universitário, Zurique, Suíça (suíça.data);

Cada banco de dados tem o mesmo formato de instância. Enquanto os bancos de dados têm 76 atributos brutos, apenas 12 deles são realmente usados. Os atributos presentes na base de dados sao:

1)ID:Identificação do paciente
2)AGE:Idade do paciente
3)SEX: sexo sendo:
valor 1: masculino;
valor 2: feminino;
4)CP: tipo de dor no peito sendo:
valor 1: angina típica;
valor 2: angina atípica;
valor 3: nenhuma dor de angina;
valor 4: assintomático;
5)TRESTBPS: pressão arterial em repouso em mmHg
6)CHOL: colesterol sérico em mg/dl

7)FBS: açúcar no sangue em jejum $>$ 120 mg/dl sendo:
valor 1: verdadeiro;
valor 0: falso;
8)RESTECG: resultados eletrocardiográficos em repouso sendo:
valor 0: normal;
valor 1: tendo anormalidade da onda;
valor 2: mostrando hipertrofia ventricular;
9)THALACH: frequência cardíaca máxima alcançada
10)EXANG: angina induzida por exercício
11)OLDPEAK: depressão de ST (atividade elétrica do coração) induzida por exercício em relação ao repouso
12)SLOPE: a inclinação do pico do segmento ST do exercício sendo:
Valor 1: ascendente;
Valor 2: plano;
Valor 3: descendente;
13)NUM: diagnóstico de doença cardíaca (estado angiográfico da doença)
valor 0: $>$ 50% estreitamento do diâmetro;
valor 1: $>$ 50% estreitamento do diâmetro;

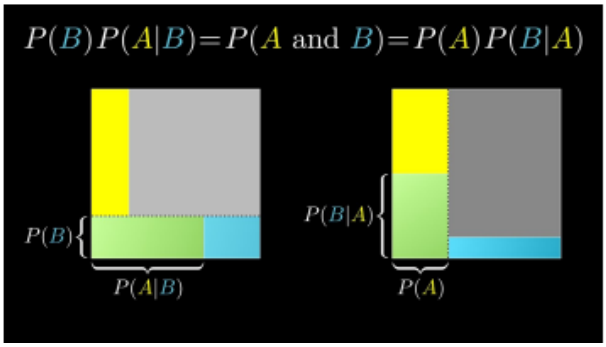
B. Classificador Probabilístico

Nesta seção, iremos apresentar o classificador Ingênuo de Bayes, que sera utilizado no desenvolvimento do projeto, e o teorema de Bayes, que é a base para o classificador.

1) Teorema de Bayes: O teorema de Bayes recebe esse nome por ter sido criado pelo pastor e matemático inglês Thomas Bayes (1702-1761), ele foi o primeiro a fornecer uma equação que permitia que novas evidências atualizassem a probabilidade de um evento a partir do conhecimento a priori(ou a crença inicial na ocorrencia de um evento). O manuscrito de Bayes so foi publicado após a morte de Thomas, sendo editado significativamente por Richard Price antes disso. E hoje é usado para o cálculo da probabilidade de um evento dado que outro evento ja ocorreu, o que é chamado de probabilidade condicional.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

A equação é a forma mais simples do teorema. P(A—B) representa a probabilidade do evento A ocorrer dado que o evento B ja foi observado, consequentemente P(B) precisa ser diferente de zero.Podemos classificar as probabilidades contidas no teoremada seguinte maneira: • Probabilidades marginais: A probabilidade de um evento independentemente do restante. Ex: P(A) e P(B). • Probabilidade conjunta: A probabilidade de dois ou mais eventos ocorrerem simultaneamente. Ex: P(A e B). • Probabilidade condicionada: A probabilidade de um ou mais eventos dada a ocorrencia de outro evento. Ex:P(A—B), P(B—A). Então, o teorema de Bayes trata de probabilidades condicionais, visto que, na equação temos a probabilidade de A condicionada pelo evento B. Nele, P(A) e P(B) sao as chamadas probabilidades a priori e P(A—B) e P(B—A) sao as probabilidades a posteriori. Outra forma de visualizar o teorema de Bayes e como a probabilidade conjunta de A e B, que pode ser simbolizada como P(A B).Na equação, temos a representação visual da igualdade.
P(B) · P(A—B) = P(A B) = P(A) · P(B—A)



Podemos reescrever o teorema utilizando a igualdade em

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2) Classificador Ingenuo de Bayes: O classificador Ingênuo de Bayes ou Naive Bayes e um popular classificador probabilístico usado frequentemente na area de aprendizagem de máquina (Machine Learning). Ele recebe o nome de ingênuo, pois desconsidera a correlação entre as variáveis, ou seja, trata cada uma de forma independente. Uma das suas aplicações é a análise de texto de acordo com a frequência das palavras usadas, e comumente utilizado na classificação de e-mails como spam. Por ser muito simples e rapido, possui um desempenho relativamente maior do que outros classificadores. Além disso, o Naive Bayes precisa de um pequeno número de dados de teste para concluir classificações com uma boa precisão. Dessa forma, teremos:

$$P(C|A) = \frac{P(C)P(A|C)}{P(A)}$$

$$P(C|A) = \frac{P(C)\prod_{i=1}^{n} P(a_i|C)}{P(a_i, ..., a_n)}$$

C. Aplicação

Para a criação do modelo, que será utilizado para a análise dos dados, sera utilizada a linguagem Python no ambiente do Google Colaboratory. A principal biblioteca a ser utilizada sera a Pandas, que por sua vez é baseada em duas bibliotecas de Python: matplotlib e NumPy. Essa biblioteca é utilizada para a manipulação e análise de dados, utilizando matplotlib para a

visualização gráfica e NumPy para as operações matemáticas. Outra biblioteca que pode ser usada e Scikit-Learn, uma biblioteca de Machine Learning que inclui vários algoritmos de classificação, regressão e agrupamento, foi projetada exatamente para interagir com bibliotecas númericas e científicas como NumPy e SciPy.

III. Análise exploratória de dados

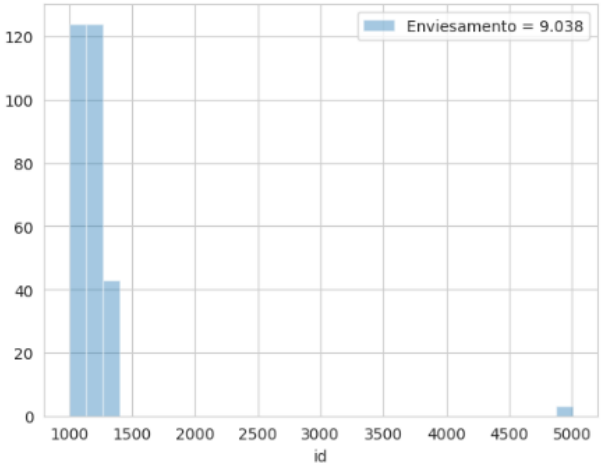
1) Pré-processamento dos dados: Os dados coletados devem passar por pré-processamento para remover informações irrelevantes, corrigir erros e preencher dados faltantes.
2) Seleção de características: A partir dos dados pré-processados, selecionar as características mais relevantes para o diagnóstico de doenças cardiovasculares, como pressão arterial, níveis de colesterol, índice de massa corporal, histórico familiar, entre outros.

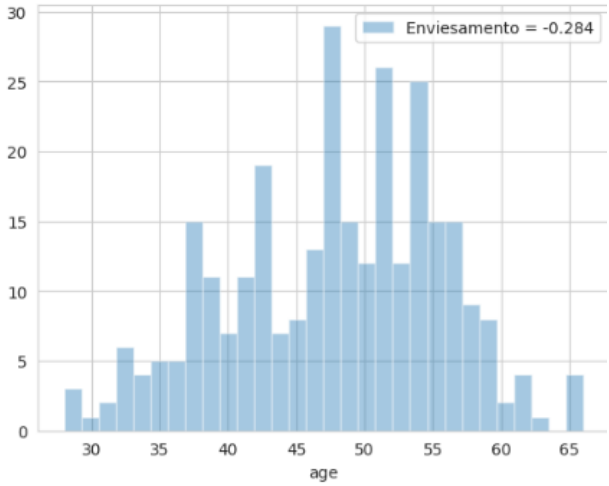
A. Estatística Descritiva

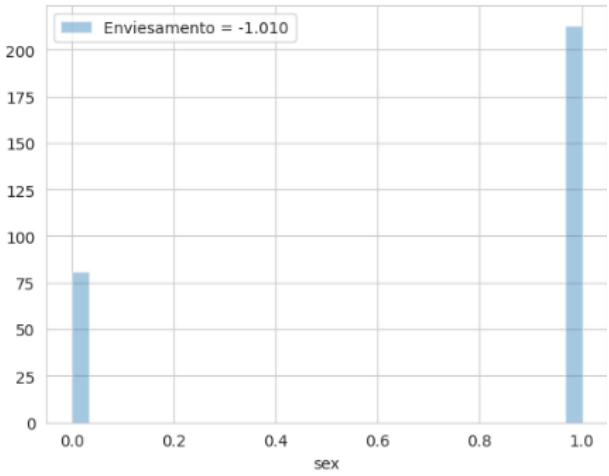
Devido a inconsistência dos dados tivemos de fazer um forte tratamento nos dados: 1) Identificar e eliminar outliers 2) Diminuir o enviesamento dos dados 3) Eliminar dados nulos ou ausentes Apos esse tratamento nossa base diminuiu consideravelmente de 294 observações para apenas 94. Além disso duas colunas,thal e cal, foram abandonadas por conterem em sua maioria dados ausentes.

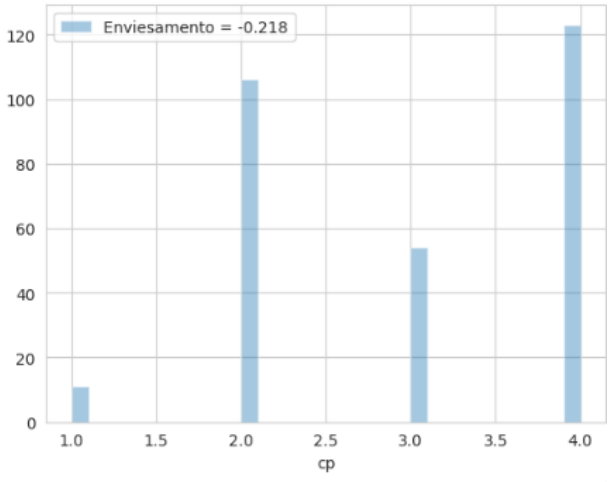
B. Vizualização de dados

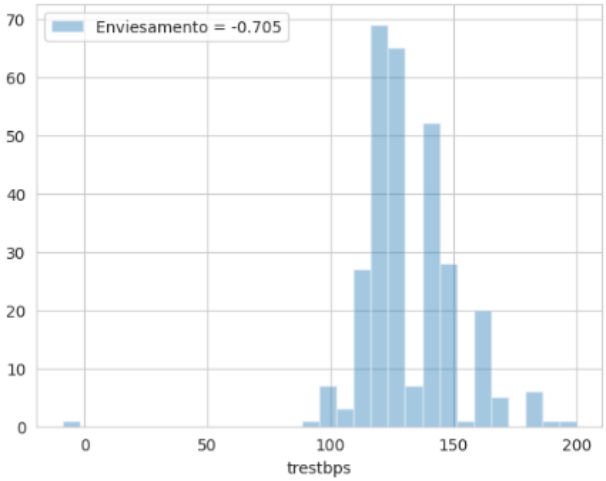
1) Graficos Univariados: Nessa etapa iremos dar uma olhada na distribuicão dos diferentes recursos desse conjunto de dados.

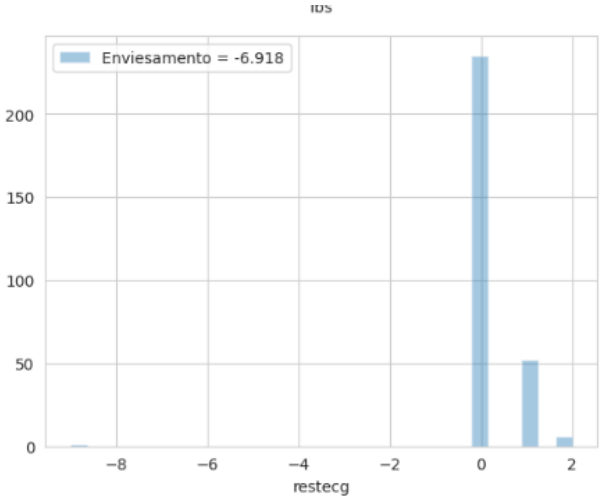


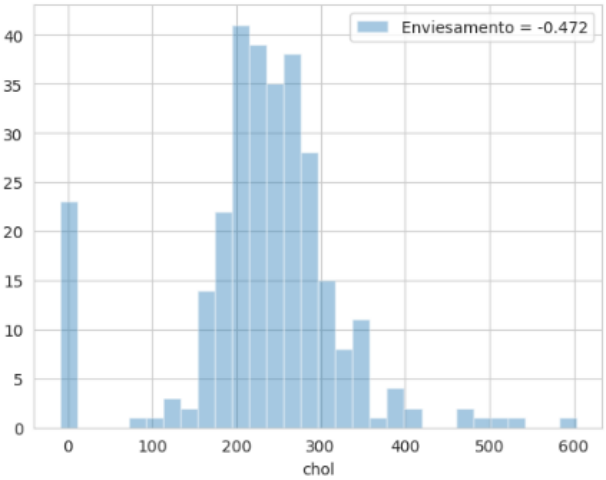


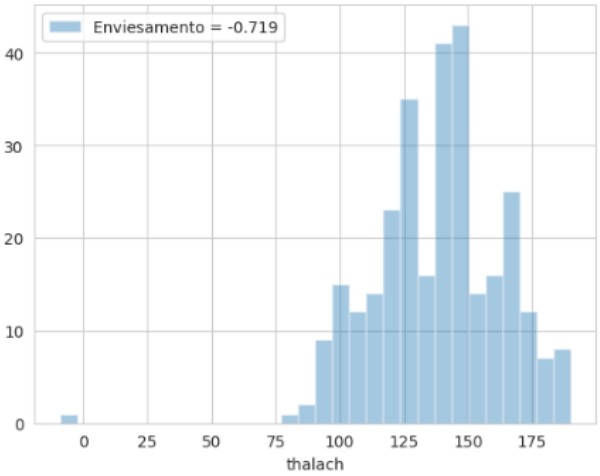


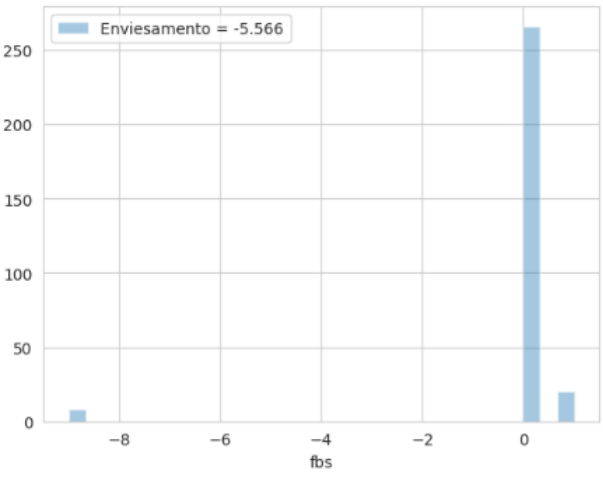


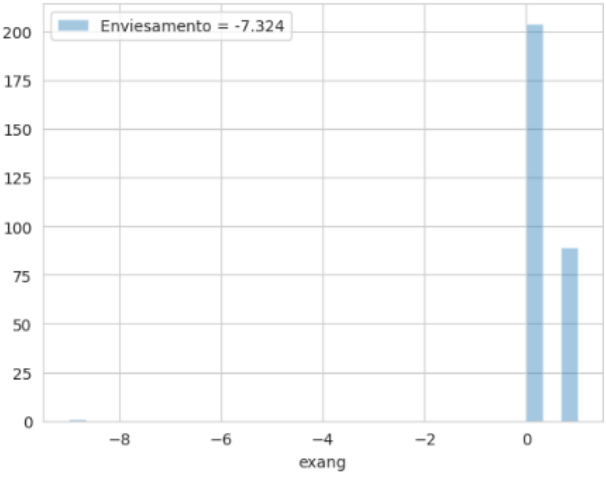

Enviesamento = -0.705
trestbps

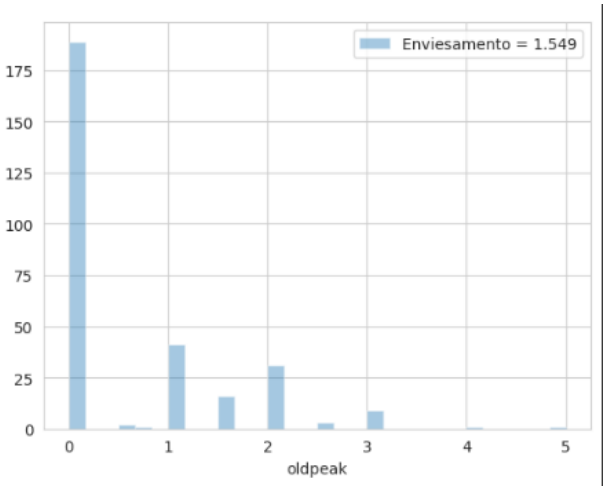

Enviesamento = -6.918
restecg

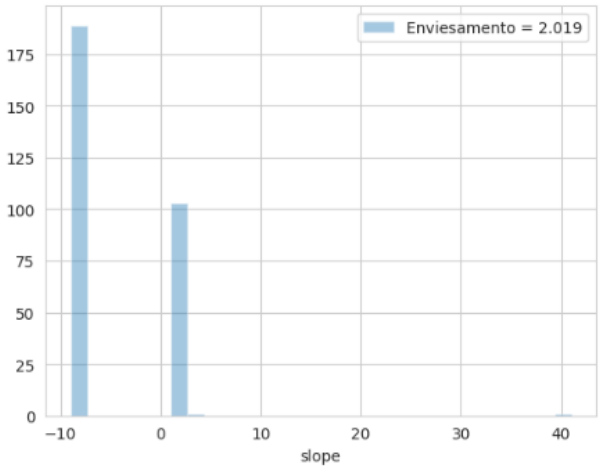

Enviesamento = -0.472
chol

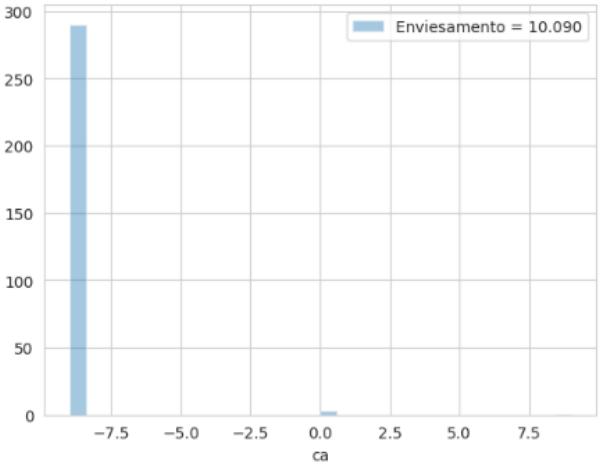

Enviesamento = -0.719
thalach

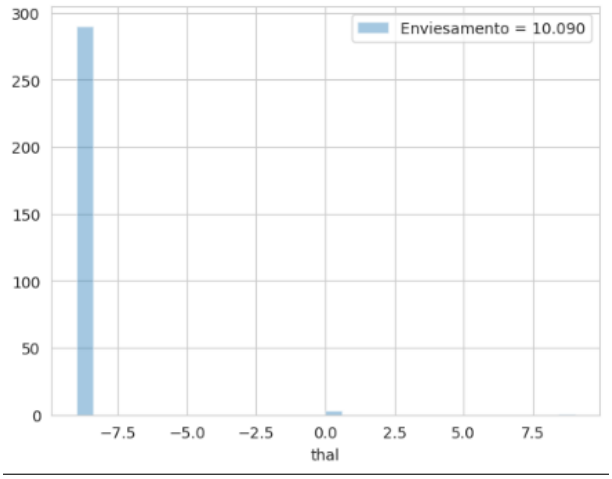

Enviesamento = -5.566
fbs


Enviesamento = -7.324
exang

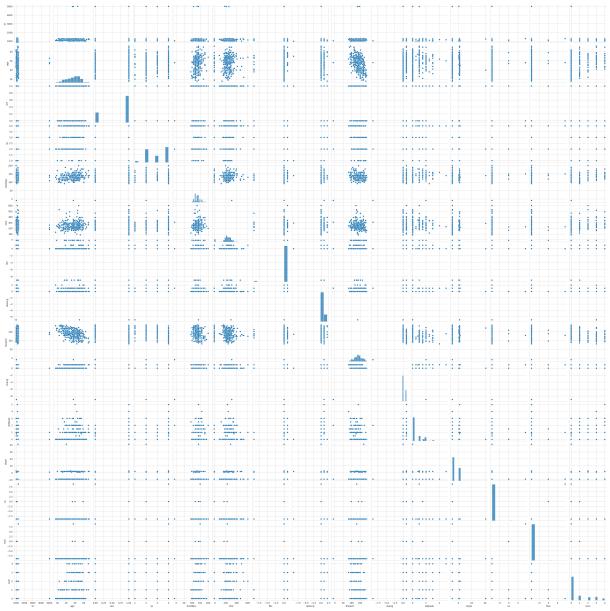






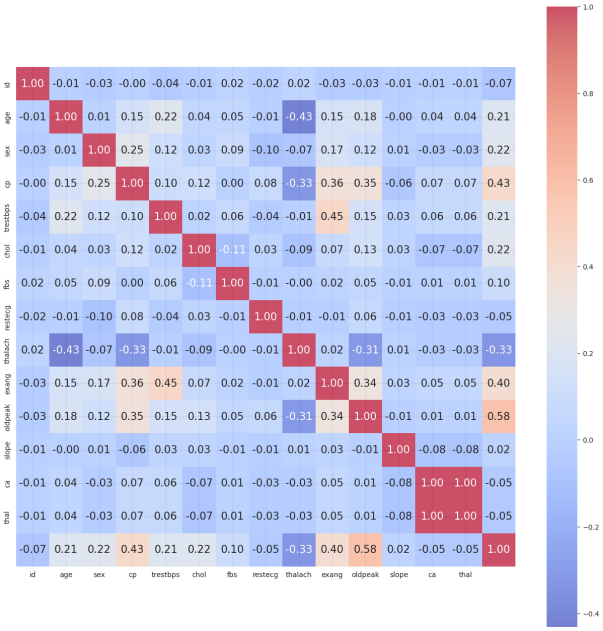


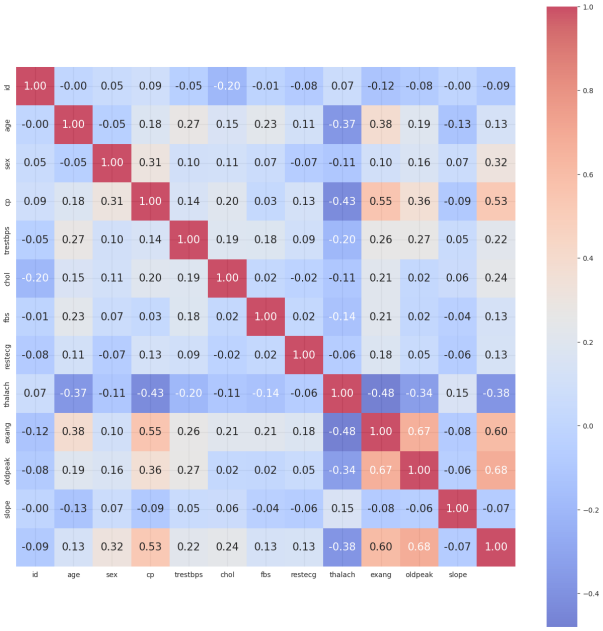
2) Gráficos Multivariados: Agora, vamos desenhar o pair-plot para examinar visualmente a correlação entre as feature.



C. Tratamento das Correlações entre as Features

Como o classificador ingênuo de Bayes considera que as features são independentes entre si, é necessário observar se essa condicão é respeitada entre as features da base utilizada. Por isso, iremos desenhar o heatmap das correlações, para analiza-las.





Depois de eliminar alguns dados nulos vemos que as variveis tiveram seus valores de correlação alterado. vemos por exemplo que nossa variavel de interesse 'num' que representa o diagnostico de doença cardíaca, tem uma relação alta com:

cp: que representa tipo de dor no peito

exang: angina induzida por exercicio

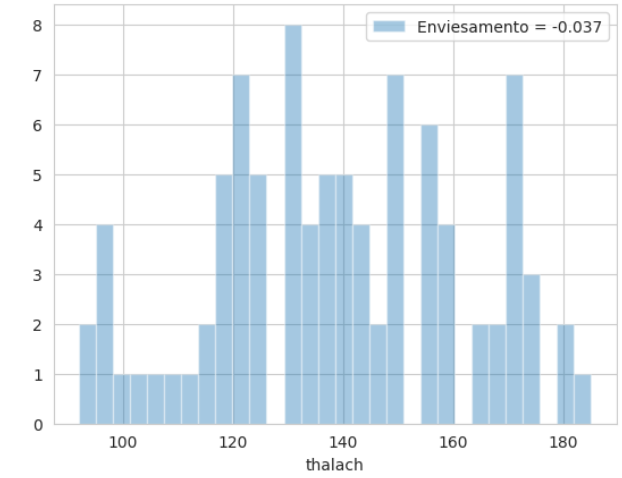
oldpeak: depressão induzida por exercicio relativa ao descanso e tem uma anticorrelação com

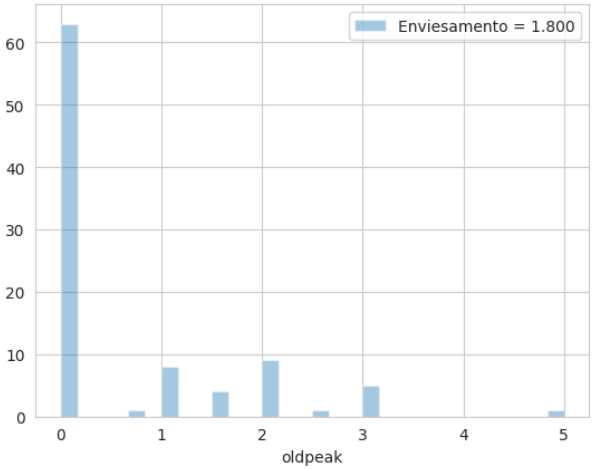
thalach: maxima taxa cardiaca atingida

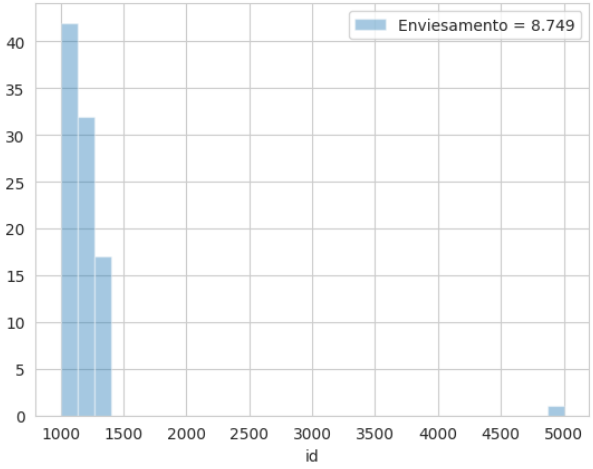
Aqui podemos tira algumas conclusões, Atributos com maior correlação com diagnostico positivo: – Cp: tipo de dor no peito – Exang : angina induzida por exercício – oldpeak : depressão induzida por exercício relativa ao descanso Atributos com maior anti correlação com diagnostico positivo: – thalach: máxima taxa cardíaca atingida

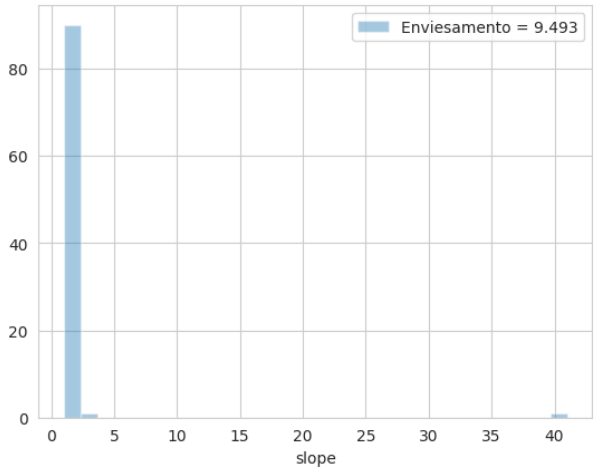
D. Localizando e Removendo os Outliers

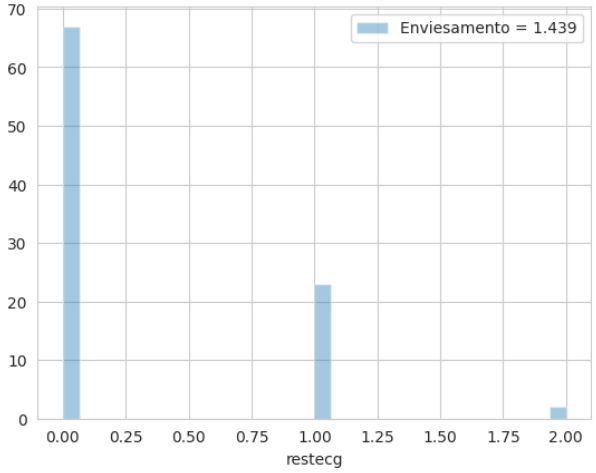
Nossa base diminuiu consideravelmente de 294 observações para apenas 94.Além disso duas colunas,thal e cal, foram abandonadas por conterem em sua maioria dados ausentes.

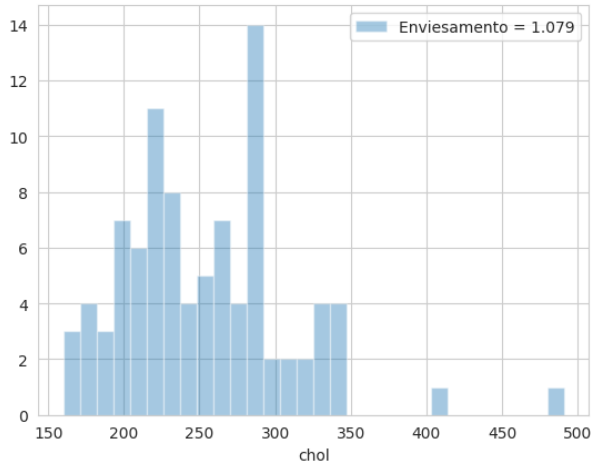


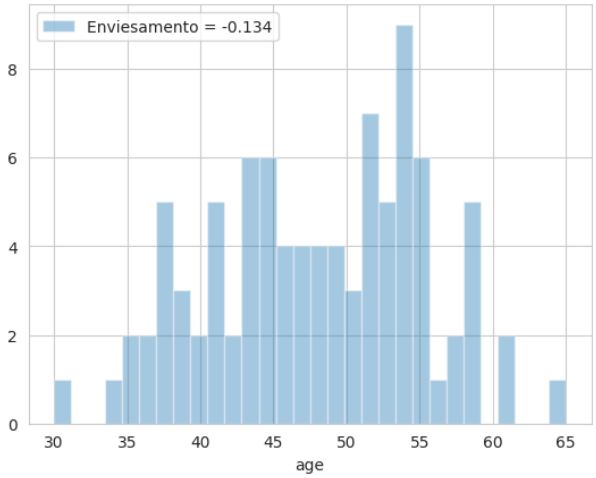


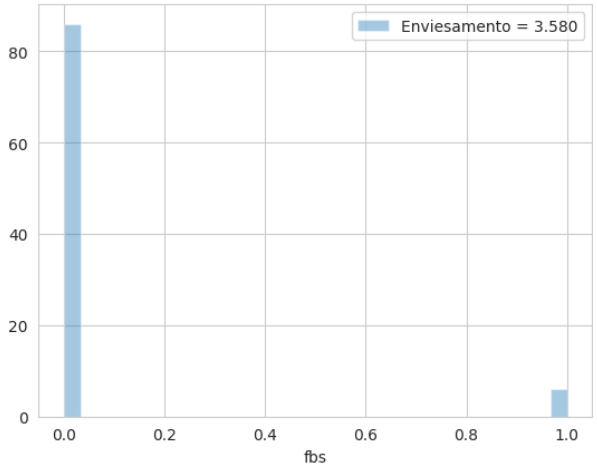


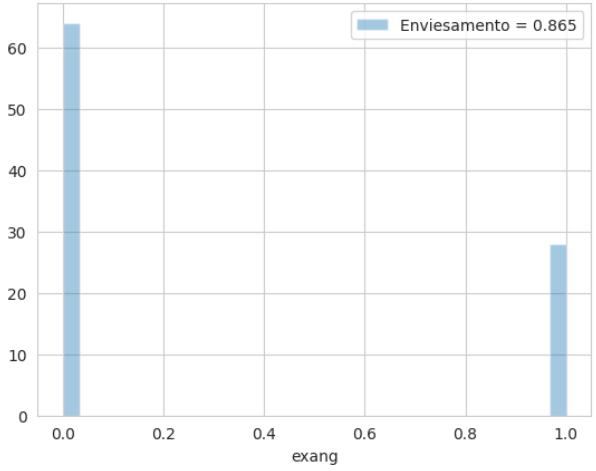

Enviesamento = 9.493
slope

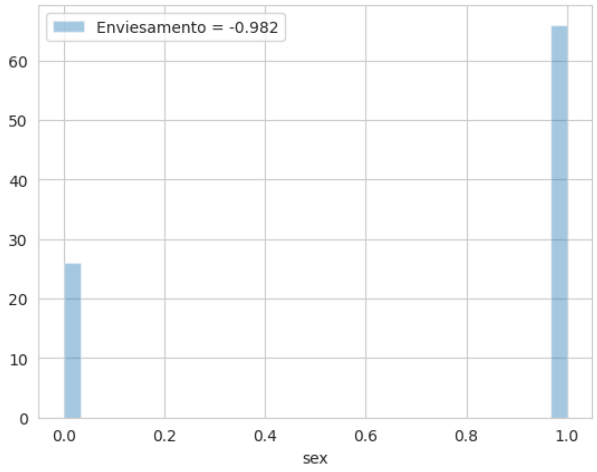

Enviesamento = 1.439
restecg

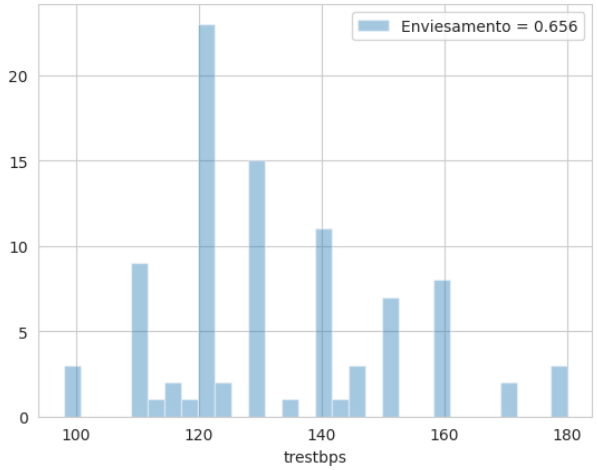

Enviesamento = 1.079
chol

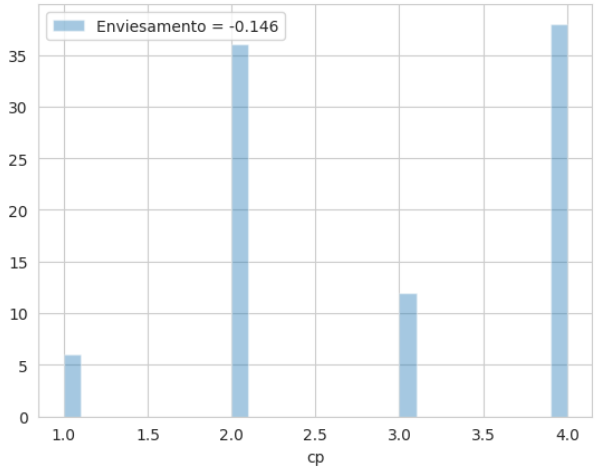

Enviesamento = -0.134
age


Enviesamento = 3.580
fbs


Enviesamento = 0.865
exang







E. Transformação de dados

Vamos examinar se uma transformação Box-Cox pode contribuir para a normalização de alguns recursos, já que como utilizaremos do algoritmo de classificação gaussian NB, precisamos de nossa base normalizada para uma melhor acurácia do modelo. Deve-se enfatizar que todas as transformações devem ser feitas apenas no conjunto de treinamento para evitar espionagem de dados. Caso contrario, a estimativa do erro de teste será tendenciosa. Após a aplicação da transformação ao de Box-Cox, fizemos um teste e printamos gráficos onde se é mostrado que o enviesamento de todas as características foi reduzido.

IV. Treinamento do modelo

Com os dados tratados, e feito o treinamento e a medição de acurácia do classificador de Bayes, será utilizado o algoritmo classificador GaussianNB, já que todos elementos do nosso dataset se tratam de valores numericos flutuantes em uma distribuição normal após a transformação de Box-cox. Dividimos os dados de forma que 80deles sejam usados para o treinamento e 20para os testes.

V. Análise dos resultados

Analisar os resultados obtidos para avaliar a efetividade do classificador Naive Bayes no diagnóstico de doenças cardiovasculares e identificar possíveis limitações e oportunidades de aprimoramento. Considerando os resultados obtidos, é possível que o modelo de classificação desenvolvido neste estudo possa ser implementado em práticas clínicas para auxiliar profissionais de saúde no diagnóstico de doenças cardiovasculares?

VI. Conclusões

Com base na análise da acurácia retornada pelo modelo proposto onde utilizamos o algoritmo gaussiano do classificador de naive bayes, primeiro, aumentamos o tamanho dos dados de treinamento para 95e diminuímos os de testes para 5obtemos uma acurácia de 0.60. Depois, trocamos a base de dados para switzerland.data que passou pelos mesmos métodos e obtivemos uma acurácia de apenas 0.20. Isso se deve por ela ter mais atributos ausentes, menos dados e de ter mais uma coluna com valores nulos.

VII. Referências

References

[1] Probabilidade Aplicações à Estatística - Paul L. Meyer - 2ª Edição
[2] Estatística Aplicada. 6ª Edição, por Ron Larson, Betsy Farber
[3] https://archive.ics.uci.edu/ml/datasets/heart+disease
[4] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/