Instituto de Ciências Matemáticas e de Computação
(ICMC)

SCC0530 - Inteligência Artificial

Projeto de Aprendizado de Máquina

Leonardo Gonçalves Chahud- Nº USP: 5266649
Lucas Bichara - Nº USP: 11296738
Rafael Jun Teramae Dantas - Nº USP: 12563686

Sumário

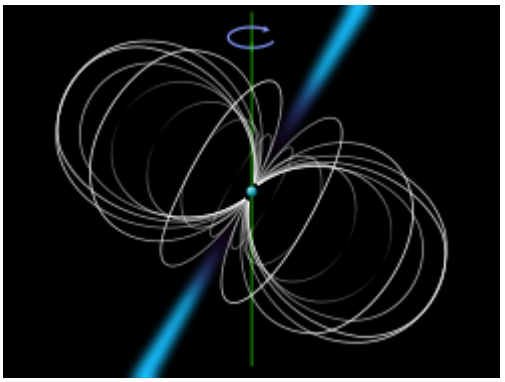
1. Dataset selecionado **2**
1.1. O que são pulsares? 2
1.2. Atributos do dataset 3
1.3. Gráficos dos dados da dataset 4
2. Técnicas utilizadas **5**
2.1. Pré-processamento: 5
2.2. Análise Exploratória de Dados: 5
2.2.1. Matriz de Correlação: 5
2.2.2. Divisão dos Dados: 5
2.2.3. Algoritmos de Classificação: 5
2.2.3.1. K-Nearest Neighbors (KNN): 5
2.2.3.2. Naive Bayes: 5
2.2.3.3. Árvore de Decisão: 5
2.2.3.4. Multilayer Perceptron (MLP): 6
2.2.4. Validação Cruzada: 6
3. Resultados **7**
4. Conclusão **8**
5. Fontes **8**

1. Dataset selecionado

O dataset utilizado traz exemplos de possíveis estrelas de nêutrons serem pulsares. O dataset traz ao todo 17898 exemplos de estrelas, sendo que 16259 não são pulsares e 1639 são realmente pulsares.

1.1. O que são pulsares?

Pulsares são um tipo raro de estrela de nêutrons que transforma sua energia rotacional em energia eletromagnética, sendo possível detectar do planeta Terra.

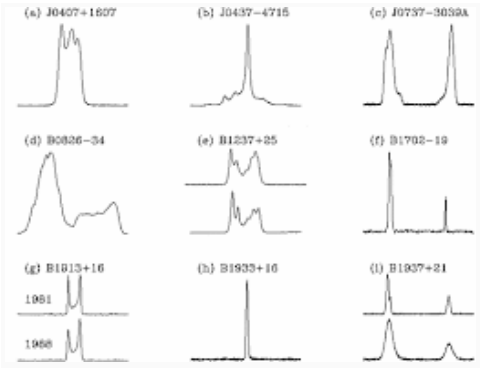


Visão esquemática de um pulsar. A esfera no centro representa a estrela de nêutrons, as linhas curvas indicam as linhas do campo magnético, os cones azuis indicam as zonas de emissão de luz e a reta verde representa o eixo de rotação da estrela.

O Pulsar de Caranguejo. Esta imagem combina informação óptica recolhida pelo Hubble (a vermelho) e imagens raio-X do Chandra (a azul).

1.2. Atributos do dataset

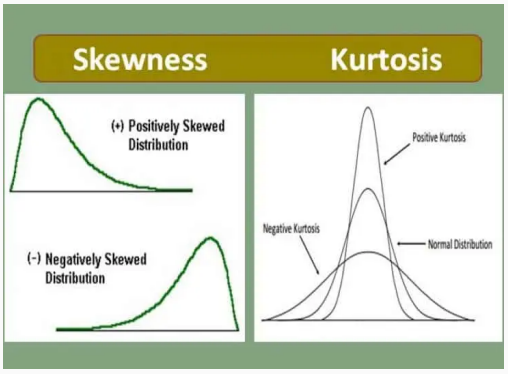
1. Mean of the integrated profile - média do perfil da estrela (única de cada estrela, como se fosse a digital do ser humano).



2. Standard deviation of the integrated profile - desvio padrão.

3. Excess kurtosis of the integrated profile - achatamento da curva.

4. Skewness of the integrated profile - assimetria em relação a média.



5. Mean of the DM-SNR curve - média da pontuação SNR (signal-to-noise ratio) em função de DM (dispersion measure)

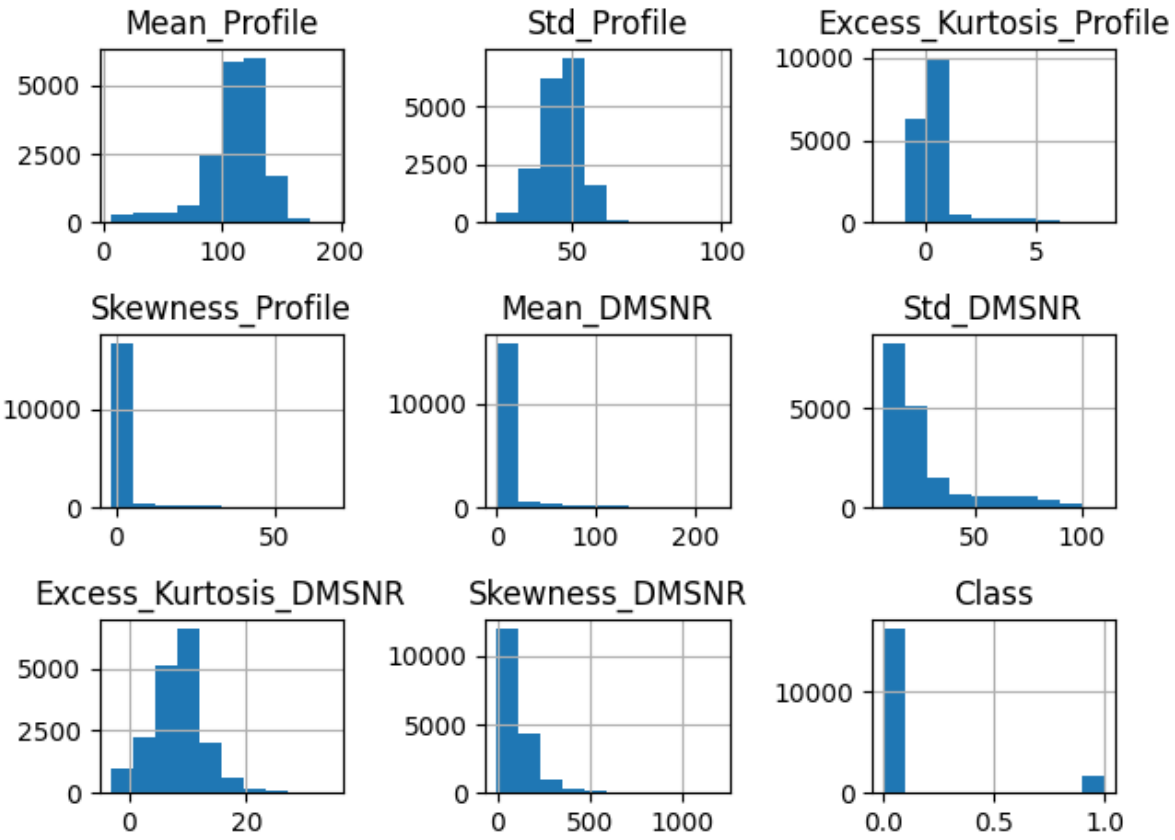
6. Standard deviation of the DM-SNR curve - desvio padrão.

7. Excess kurtosis of the DM-SNR curve - achatamento da curva.

8. Skewness of the DM-SNR curve - assimetria em relação a média.

9. Class - 0 se não for um pulsar, 1 caso contrário.

1.3. Gráficos dos dados da dataset



2. Técnicas utilizadas

2.1. Pré-processamento:

Antes de aplicar os algoritmos de classificação, foi realizada uma etapa de pré-processamento dos dados. Primeiro verificou se não haviam dados ausentes e depois, aplicou-se uma técnica para padronizar os atributos, que transforma os dados de modo que eles tenham média zero e variância unitária. Isso é importante porque muitos algoritmos de aprendizado de máquina são sensíveis à escala dos atributos. A padronização foi realizada usando a classe StandardScaler do módulo sklearn.preprocessing.

2.2. Análise Exploratória de Dados:

Foi realizada uma análise exploratória dos dados para entender melhor suas características e identificar possíveis padrões. Foram utilizadas diversas técnicas de visualização, incluindo histogramas, gráficos de dispersão, box plots, gráficos de barras e matriz de correlação.

2.2.1. Matriz de Correlação:

Foi gerada uma matriz de correlação para analisar a relação entre os atributos do conjunto de dados. A matriz de correlação é uma tabela que mostra os coeficientes de correlação entre todos os pares de atributos. Essa técnica é útil para identificar atributos que estão altamente correlacionados e podem ser redundantes para o modelo de classificação.

2.2.2. Divisão dos Dados:

Os dados foram divididos em atributos (variáveis independentes) e classes (rótulos) usando o método iloc do módulo pandas. Essa etapa é necessária para treinar e testar os algoritmos de classificação.

2.2.3. Algoritmos de Classificação:

Foram utilizados quatro algoritmos de classificação para resolver o problema de identificação de pulsares: K-Nearest Neighbors (KNN), Naive Bayes, Árvore de Decisão e Multilayer Perceptron (MLP).

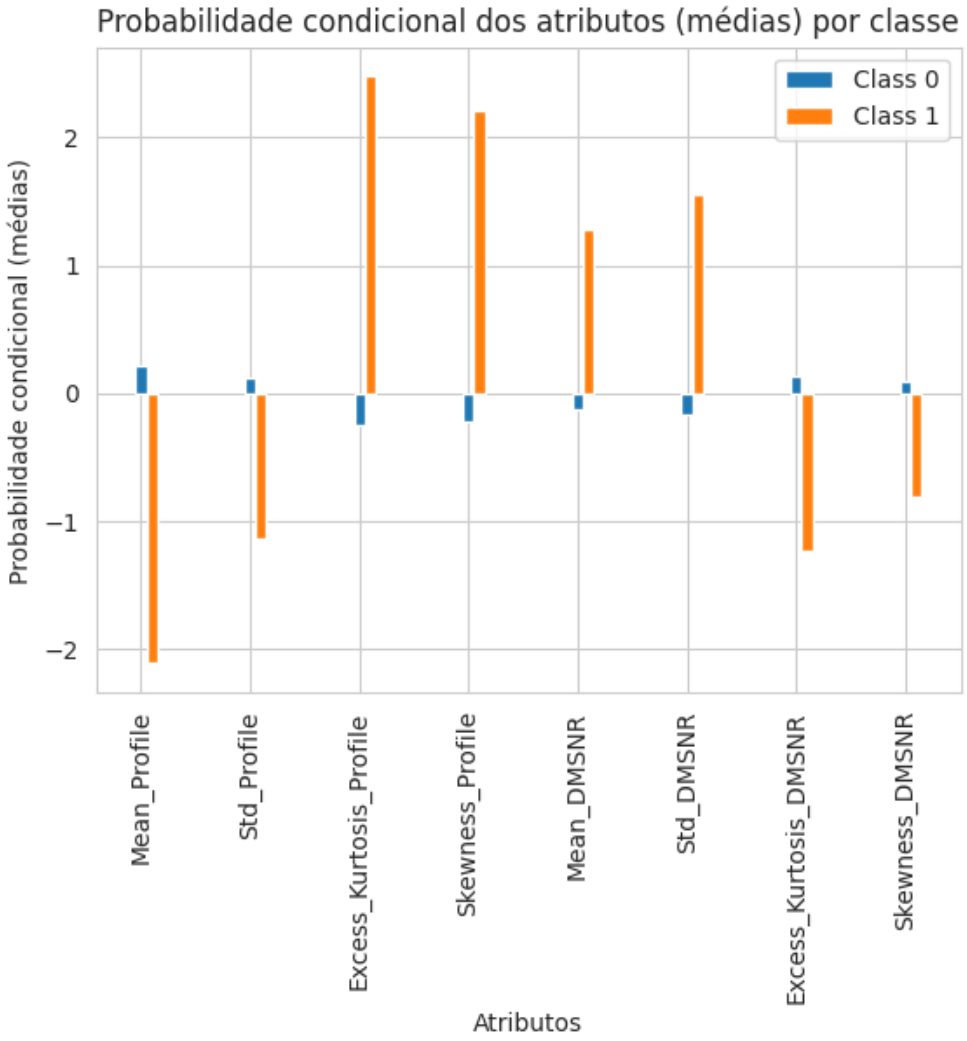
2.2.3.1. K-Nearest Neighbors (KNN):

O algoritmo KNN é um método de classificação baseado em instâncias, ou seja, ele classifica uma amostra com base nas classes das amostras vizinhas mais próximas. O valor de K, ou seja, o número de vizinhos considerados, foi definido como 5.

2.2.3.2. Naive Bayes:

O algoritmo Naive Bayes é um classificador probabilístico baseado no teorema de Bayes. Ele assume que os atributos são independentes entre si. Foi utilizado o classificador Naive Bayes Gaussiano, que assume uma distribuição gaussiana para os atributos.

Gráfico da probabilidade de decisão:



2.2.3.3. Árvore de Decisão:

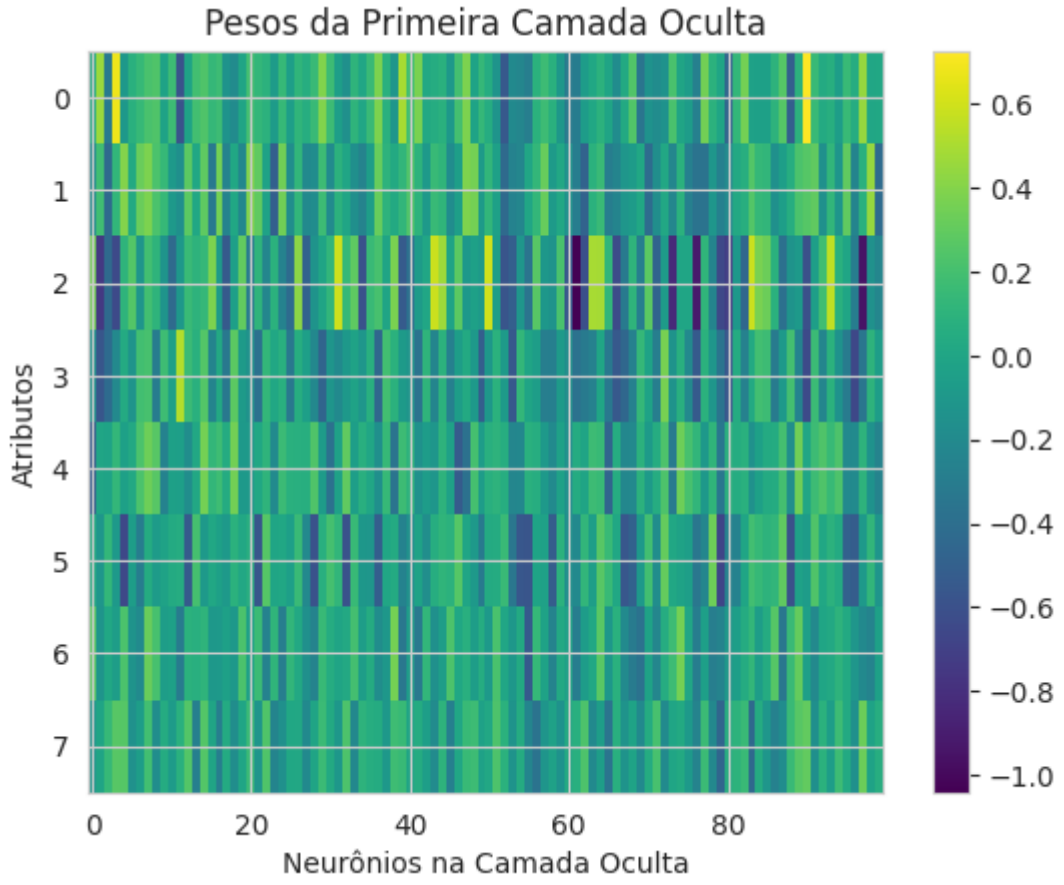
O algoritmo de Árvore de Decisão constrói um modelo de classificação em forma de árvore, onde cada nó interno representa uma decisão baseada em um atributo, e cada folha representa uma classe. Foi utilizado o classificador DecisionTreeClassifier.

Obs: o gráfico da árvore de decisão está no ipython notebook e enviado como arquivo .png.

2.2.3.4. Multilayer Perceptron (MLP):

O algoritmo MLP é uma rede neural artificial com várias camadas de neurônios. Foi utilizado o classificador MLPClassifier para treinar uma MLP com uma camada oculta.

Gráfico:



2.2.4. Validação Cruzada:

A técnica de validação cruzada foi aplicada para avaliar o desempenho dos algoritmos de classificação. A validação cruzada divide o conjunto de dados em k partes iguais (k-folds) e realiza k iterações, onde em cada iteração uma parte diferente é usada como conjunto de testes e as outras partes são usadas como conjunto de treinamento. Isso permite avaliar o desempenho dos algoritmos em diferentes conjuntos de treinamento e teste e obter uma estimativa mais confiável do desempenho real do modelo.

3. Resultados

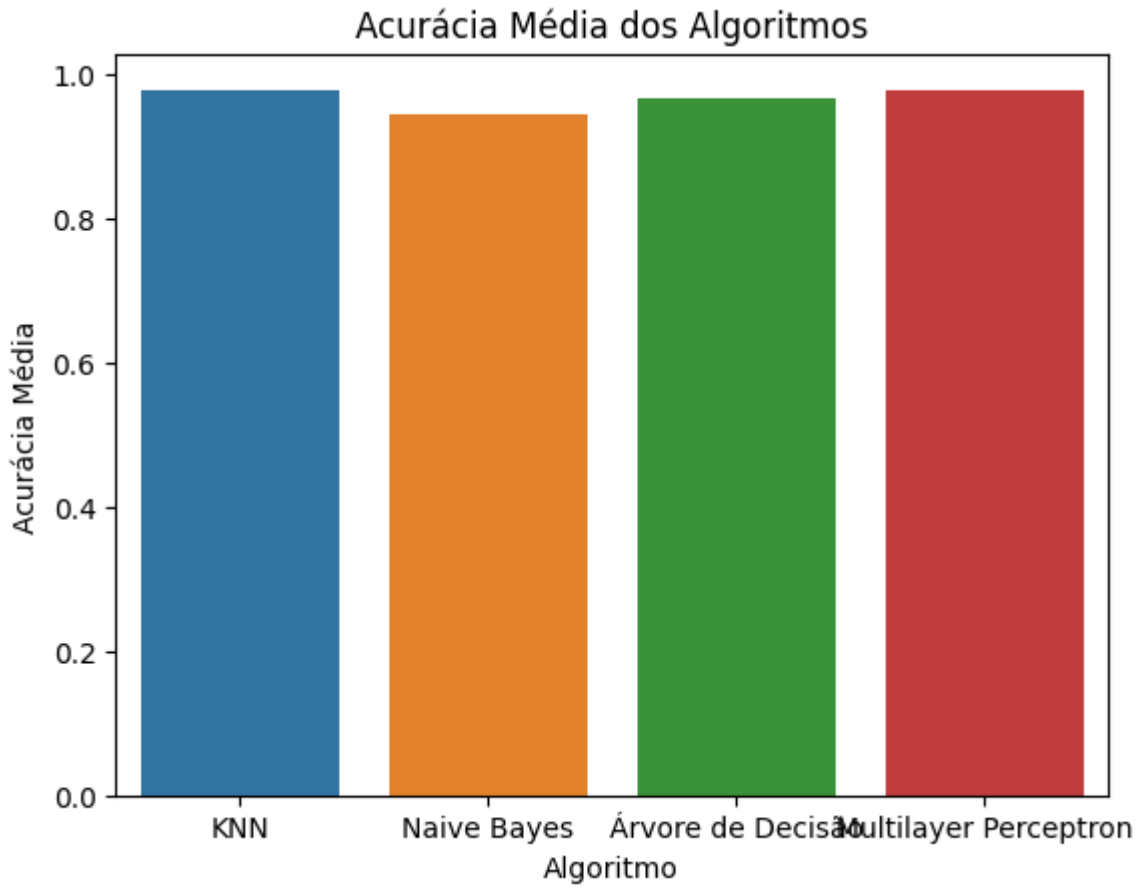
- Classifier: algoritmo utilizado;
- Accuracy: precisão do algoritmo;
- Std: standard deviation (desvio padrão).

```
Classifier: KNN
Accuracy: 0.9783765763063501 (Std: 0.005165108647875351)

Classifier: Naive Bayes
Accuracy: 0.9444648369639932 (Std: 0.02594499392247301)

Classifier: Árvore de Decisão
Accuracy: 0.9668666341719471 (Std: 0.00704083266013163)

Classifier: Multilayer Perceptron
Accuracy: 0.9796061889942884 (Std: 0.004566844813888393)
```



4. Conclusão

Os quatros algoritmos utilizados apresentaram precisões excelentes quando treinados neste dataset. Entretanto, o algoritmo Multilayer Percepton apresentou uma precisão um pouco melhor quando comparada com os outros algoritmos.

5. Fontes

1. https://pt.wikipedia.org/wiki/Pulsar
2. http://ipta.phys.wvu.edu/files/student-week-2017/IPTA2017_KuoLiu_pulsartiming.pdf
3. https://d1b10bmlvqabco.cloudfront.net/attach/jz8smbptoj35ra/i8xgc5x4yhoyo/k0fny104qt72/project.pdf