UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno:

1. Introdução

O dataset dado pelo professor Stellar Classification Dataset tem como uma das maiores pesquisas astronômicas e em seu principal conceito, ele apresenta informações importantes sobre as pesquisas espaciais, mais especificamente estrelas, galáxias e quasares, sendo cada qual direcionada a sua classe específica e sendo ainda seu objetivo, identificar objetos e construir modelos de classificação nas características físicas observadas. Suas principais características são as coordenadas dos objetos no céu, magnitude fotométrica, deslocamento que fornece uma ideia de distância e velocidade relativa do objeto e a classe alvo definindo o tipo de objeto estelar. Com base na característica apresentada, ele tenta prever a classe do objeto, procura identificar os atributos que terão maior impacto na classificação, explora também o redshift ou as magnitudes fotométricas em classes diversas.

2. Observações

Na verdade é agradecer ao professor, pois a prova foi mais leve que a passada e também por passar menos questões pois querendo ou não isso ajuda a entender um pouco mais da prova, por mais que tenhamos dificuldade na programação.

3. Resultados e discussão

1) Questão:

O primeiro passo na questão 1, é fazer a análise dos dados apresentados pelo dataset Stellar Classification Dataset como ler os dados e selecionar as colunas relevantes, com o intuito de apresentar dados limpos e prontos para o segundo passo.

O segundo passo é fazer a análise exploratória dos dados, distribuindo as classes analisadas utilizando o comando ***countplot***. O dado gerado vai permitir a visualização do balanceamento das classes do dataset, e em caso de desbalanceamento, efetuar a correção para não afetar o desempenho.

No terceiro passo, faze-se a preparação dos dados para a modelagem dividindo os dados em treino e teste sendo 70% para treinamento e 30% para teste para garantir a preservação de classes e não cause a repetição de dados do dataset.

No quarto passo, foi aplicada a técnica de análise para reduzir a dimensionalidade e garantir a preservação da variância dos dados, melhorando a eficiência computacional e minimizando a alta correlação de duas ou mais variáveis para que a qualidade dos resultados não sejam afetadas.

Normalizar os dados → Calcular os principais componentes → Selecionar os componentes que preservam 95% da variância → Levar os dados transformados para o novo espaço dimensional reduzido.

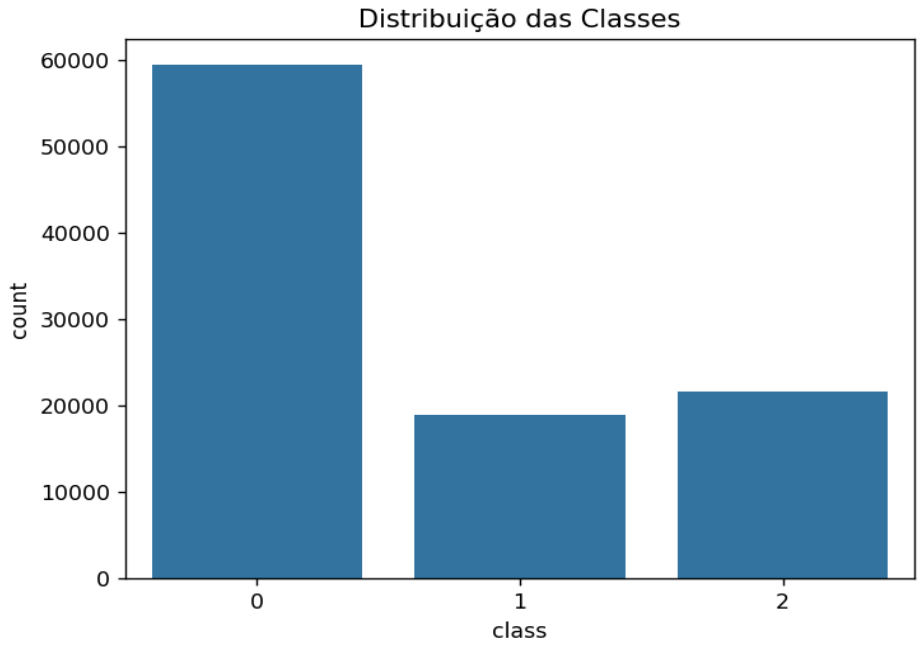
No quinto passo, realiza-se a otimização e treinamento utilizando o comando ***GridSearchCV*** que tem como sua finalidade de encontrar os melhores hiperparâmetros para o KNN. Sendo os melhores parâmetros "metric, euclidean, n_neighbours, weights e uniform".

N**o último passo**, avalia-se o modelo no conjunto de teste gerando a acurácia dos gráficos e análises.

Portanto, é possível se analisar que na questão da AV1 o melhor valor de **k** encontrado foi 7, sendo a melhor métrica de distância Manhattan. Enquanto na AV2, o melhor valor de **k** encontrado foi 5 e a melhor métrica de distância foi Euclidean e isso deve-se a utilização do GridSearchCV que devido a sua otimização combinada e maior amostra de outros hiperparâmetros, oferece uma melhor parametrização do modelo, dando uma significativa melhora na acurácia.

Critério	AV1	AV2
Melhor métrica	Manhattan	Euclidean
Melhor K	7	5
Tipo de ponderaçã	Não testado	Distance
Acurácia	85%	87%

(Figura 1: comparações entre AV1 e AV2)



(Figura 2: gráfico do resultado da primeira questão.)

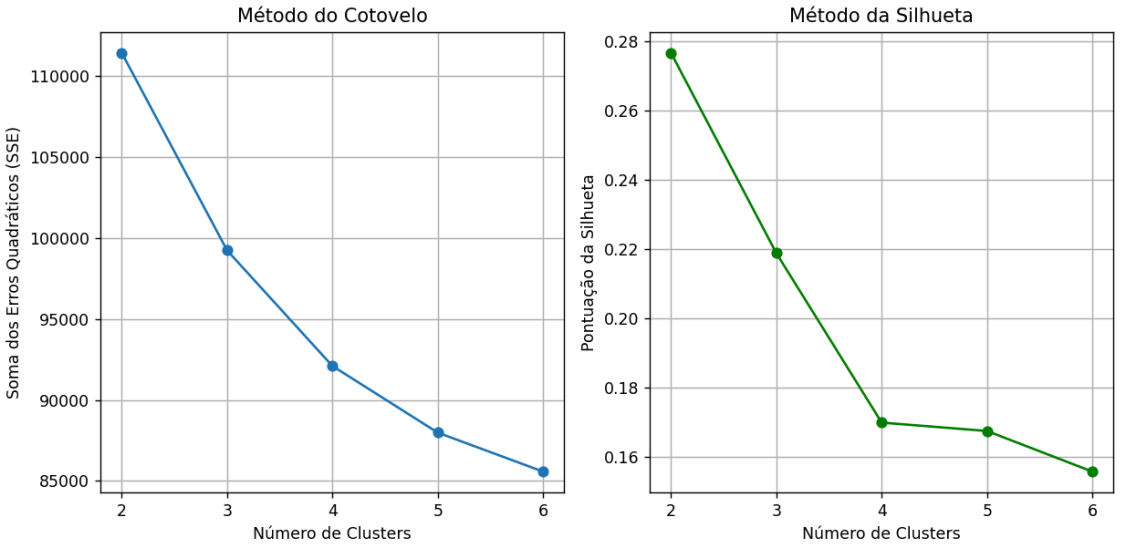
```
C:\Users\vitor\Downloads\IA.BLACK\IA--SI\IA-BLACKZIN.SI\Scripts\python.exe C:\Users\v

Melhores parâmetros: {'metric': 'euclidean', 'n_neighbors': 26, 'weights': 'uniform'}
Acurácia no teste: 0.6602
[[15932  1317   585]
 [ 3567  1988   133]
 [ 3892   701  1885]]
              precision    recall  f1-score   support

           0       0.68      0.89      0.77     17834
           1       0.50      0.35      0.41      5688
           2       0.72      0.29      0.42      6478

    accuracy                           0.66     30000
   macro avg       0.63      0.51      0.53     30000
weighted avg       0.66      0.66      0.63     30000


Process finished with exit code 0
```

(Figura 3: resultado do fim da execução do código da questão 1.)

2) Questão:

Na primeira etapa, é realizado o carregamento de dados, fazendo uma leitura desses dados, logo após foi realizada a remoção da variável alvo permitindo uma análise supervisionada, e por último a normalização dos dados padronizando as variáveis.

Na segunda etapa, foi feito o agrupamento de e a avaliação dos resultados, explorando a segmentação de dados por meio do clustering MiniBatchKMeans.
Dentro desse ponto primeiramente foi realizada a definição do intervalo de clusters testando de 2 a 6 clusters.
Segundamente a execução desses MiniBatchKMeans para cada valor de k.
Por terceiro, a avaliação do modelo utilizando a soma dos erros quadráticos e por fim a visualização desses resultados por meio da Silhueta e o Cotovelo.



(Figura 4: Gráficos do resultado da questão 2.)

```
2 clusters: 0.2767
3 clusters: 0.2188
4 clusters: 0.1699
5 clusters: 0.1675
6 clusters: 0.1557

Process finished with exit code 0
```

(Figura 5: resultado da execução do código da questão 2)

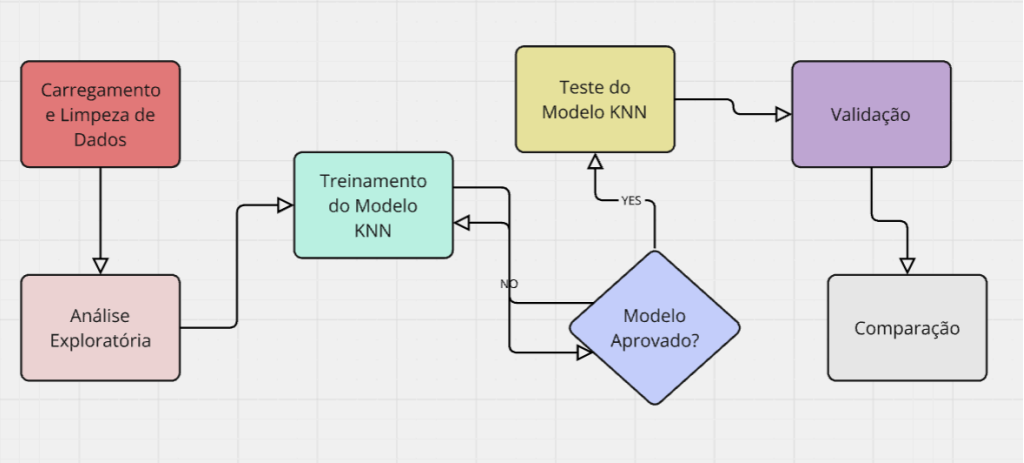
3) Questão:

Na primeira etapa, foi realizado o carregamento e o processamento dos dados, executando a filtragem e leitura dos dados para garantir a normalização dos dados apresentados pelo dataset.

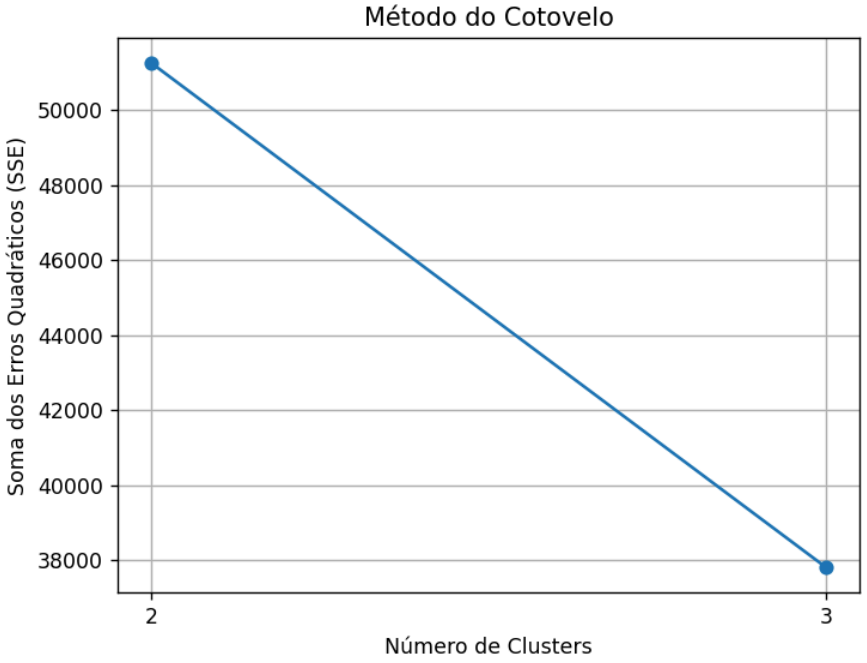
Na segunda etapa, foram selecionados os atributos mais relevantes, executando o treinamento do modelo Lasso com o valor de penalização ***alpha= 0.1,*** foi identificado os atributos mais relevantes analisando os coeficientes do modelo e por último a seleção dos maiores atributos que causam maior impacto.

Na terceira etapa, foi realizada a redução de dimensionalidade, testes de diferentes números de clusters e a avaliação dos modelos visando reduzir os dados, encontrar o melhor agrupamento e apresentação do modelo Cotovelo e Silhueta, sendo que o cotovelo apresentou 3 clusters proporcionando uma boa redução no SSE e a Silhueta apresentou 2 clusters com valores mais altos indicando uma separabilidade dos grupos.

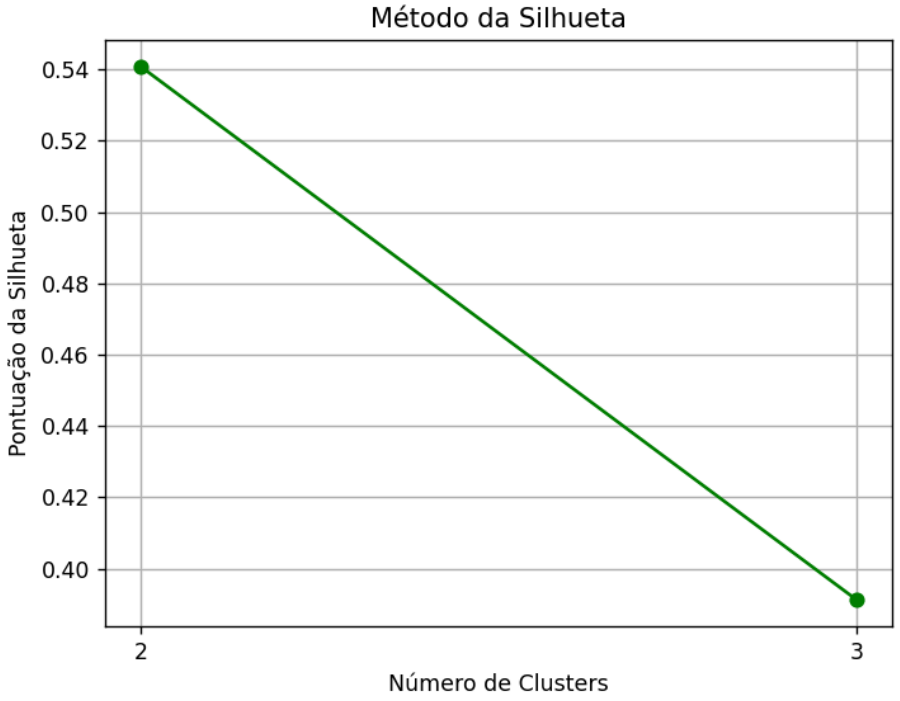
Na terceira etapa, é realizado o treinamento do MiniBatchKMeans com o valor de k = 3 e logo após a geração de um scatterplot colorindo cada ponto respectivos a cada clusters permitindo a análise visual da segmentação dos dados.



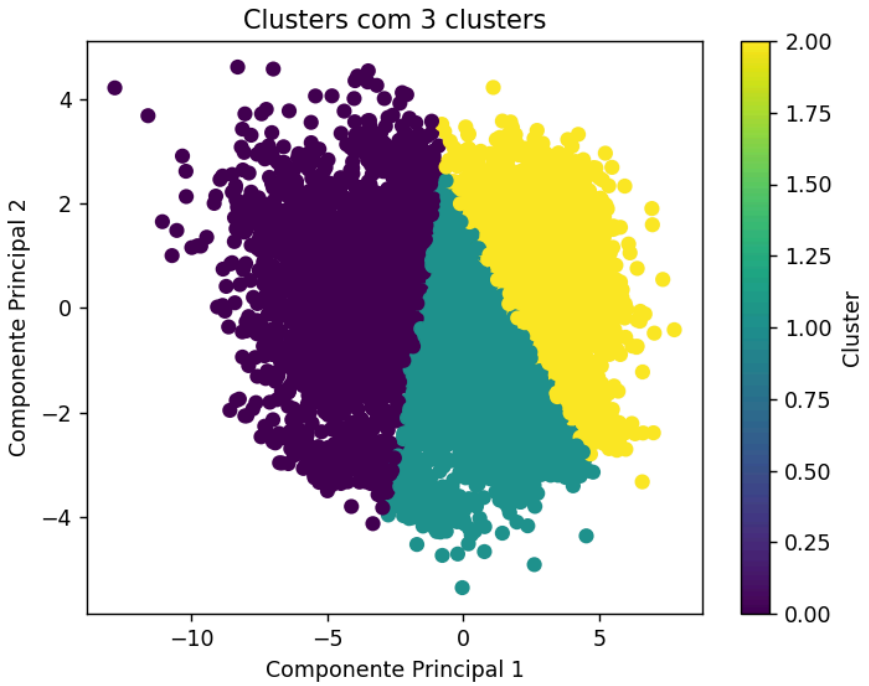
(Figura 6: fluxograma de estruturação da criação do código por etapas.)



(Figura 7: resultado gráfico do método cotovelo.)



(Figura 8: resultado gráfico do método silhueta.)



(Figura 9: resultado gráfico em cores.)

```
C:\Users\vitor\Downloads\IA.BLACK\IA--SI\IA-
Atributos mais relevantes selecionados pelo
2 clusters: 0.5408
3 clusters: 0.3914

Process finished with exit code 0
```

(Figura 10: resultado dae questão no fim da execução do código)

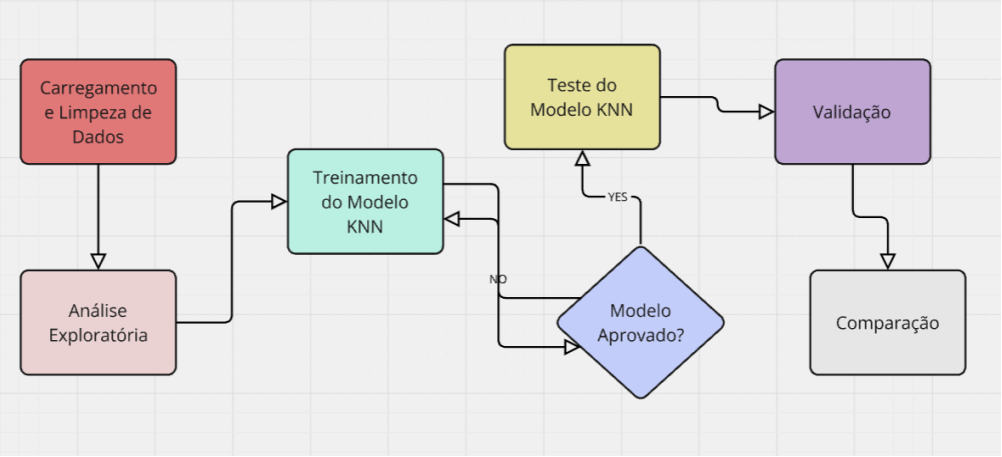
4) Questão:

Na primeira etapa, é realizado o processamento de dados que esta relacionado a filtragem na qual é selecionado os dados relevantes por meio da leitura dos dados no dataset e feita a normalização garantindo que as variáveis tivessem a mesma escala, deixando os dados mais organizados e equilibrados para alcancar uma maior eficiencia e também retirar o que não precisa ser utilizado.

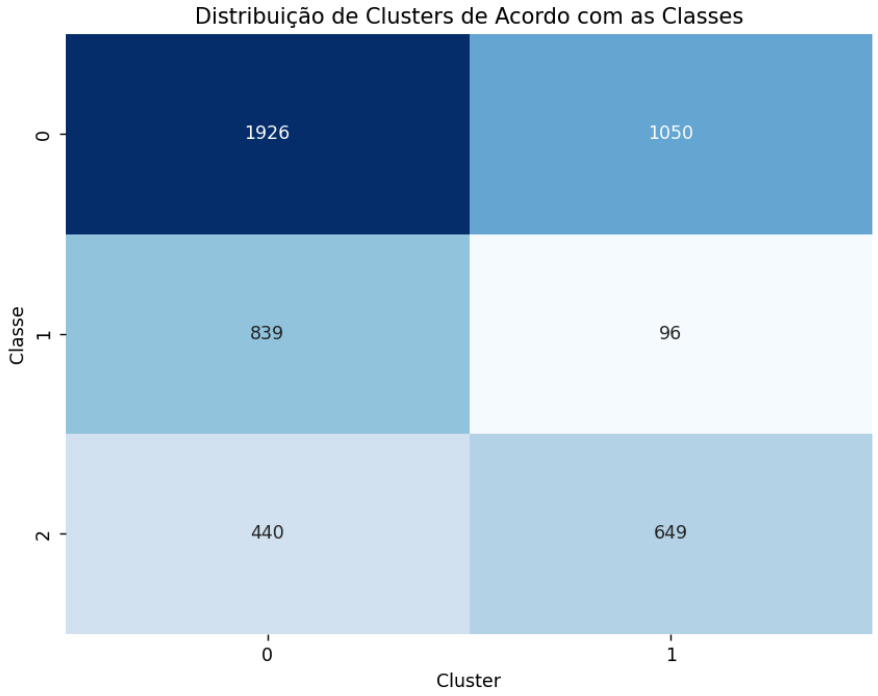
Na segunda etapa, onde acontece a divisão dos dados em teste e treino, é realizado o treinamento do modelo Lasso, é separado os atributos mais importantes e selecionados também os mais significativos, ou seja, com maior peso na classificação.

Na terceira etapa, e realizado a redução de dimensionalidade, teste para diferentes quantidades de clusters e a seleção dos melhores numeros tendo como base o método silhueta, sendo o maior valor relacionado como melhor agrupamento.

Na quarta etapa, e por último é feita as atribuições dos conjunto de dados originais dos clusters, a criação da tabela cruzada e a geração de um heatmap. Portanto, a tabela de contigência revelou algumas classes agrupadas dentro de específicops clusters, mas também, outras classes que apresentaram uma sobreposição entre diferentes clusters, sugerindo que outros atributos adicionais poderiam melhorar a eficiencia da segmentação.



(Figura 11: fluxograma de passos para execução do código.)



(Figura 12: resultado gráfico da questão 4.)

```
Distribuição de Clusters de Acordo com as Classes:
Cluster     0     1
Classe
0        1926  1050
1         839    96
2         440   649

Process finished with exit code 0
```

(Figura 13: resultado do fim da execução do código.)

4. Conclusões

Sim, mesmo da ajuda do professor em reduzir o trabalho, ainda sim tivemos que buscar nos esforçar para compreender a maneira de fazer as quetões, mas acredito que como comentado no trabalho passado, foi possível compreender algo sobre o assunto e absorver um pouco do conteudo, conforme as pesquisas e o desenvolver do mesmo. Ainda sim, sempre visando a possibilidade de fazermos melhor do que as vezes passadas.

5. Próximos passos

Acredito que esse projeto poderia continuar acontecendo, para que as outras turmas que virão, também possam ter a experiencia de lidar com o conteúdo na prática pura, pois esse tipo de matéria, em sua dificuldade apresentada para nós discentes que estamos tendo o primeiro contato é necessário esse contato complicado, mas desafiador, pois a todo instante o professor mostrou estar disponível sendo essa mais um motivo de colaboração para a aprendizagem dos discentes através desse projeto.