Introdução

Há décadas a humanidade vem aprimorando as tecnologias e estudos para classificar estrelas, essa pesquisa gera resultados importantes para entendermos nosso universo e encontrar fenômenos que podem ser favoráveis para a evolução da humanidade.
Utilizando o dataset da SDSS (Sloan Digital Sky Survey) que classifica corpos celestes, nosso objetivo é utilizar as características espectrais de cem mil observações para sermos capazes de classificar observações futuras com algoritmos de machine learning e manipulação de dados.

Fundamentos Teóricos e Metodológicos

A Análise Exploratória de Dados (AED) é um conjunto de técnicas usadas para visualizar e resumir as principais características dos dados, frequentemente com gráficos.

Para aumentar a precisão de nossos diagnósticos e ter uma visão mais clara dos dados, utilizamos a técnica de IQR para remover os outliers do conjunto de dados, onde ela utiliza estatísticas descritivas simples (quartis) para identificar os pontos que estão fora de uma faixa considerada "normal" e define os outliers como valores que estão fora de uma faixa aceitável ao redor da mediana.

Para classificação, utilizamos os modelos de XGBoost e Floresta aleatória, onde o primeiro modelo funciona baseado em árvores de decisão, que utiliza a técnica de boosting para combinar várias árvores fracas em um modelo forte. Ele otimiza a função de perda por meio de gradiente descendente, inclui regularização para evitar overfitting, é eficiente e altamente preciso, sendo amplamente usado para classificação e regressão.

Ademais, o segundo modelo constrói várias árvores de decisão, onde cada árvore é treinada com um subconjunto aleatório do conjunto de dados, e cada árvore faz uma previsão individual, e a saída final é determinada por um processo de votação no caso de classificação, ou pela média no caso de regressão.

Aplicação

Para dar prosseguimento com a ADE foi feita uma mudança na variável qualitativa “class”, atribuindo números para cada uma das três opções e facilitando as análises com apenas variáveis numéricas.

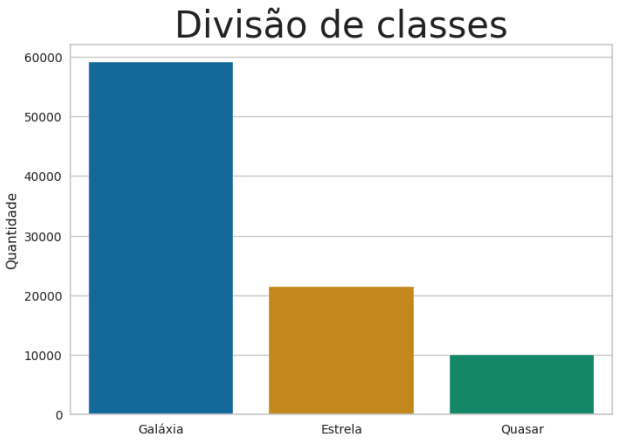

Imagem 1

No pré-processamento, foi realizada a remoção de outliers do dataset usando o IQR, onde através de um loop todas as variáveis foram verificadas, e aproximadamente 14% das observações eram outliers. Ademais, para a seleção de features foi feito um teste de correlação com o método de Pearson entre a variável classe e as demais, onde aquelas com correlações marginais foram desconsideradas, 10 variáveis foram selecionadas para os próximos processos e 8 removidas. O dataset não possui dados faltantes, duplicados ou nulos, o que facilitou a etapa de pré-processamento.

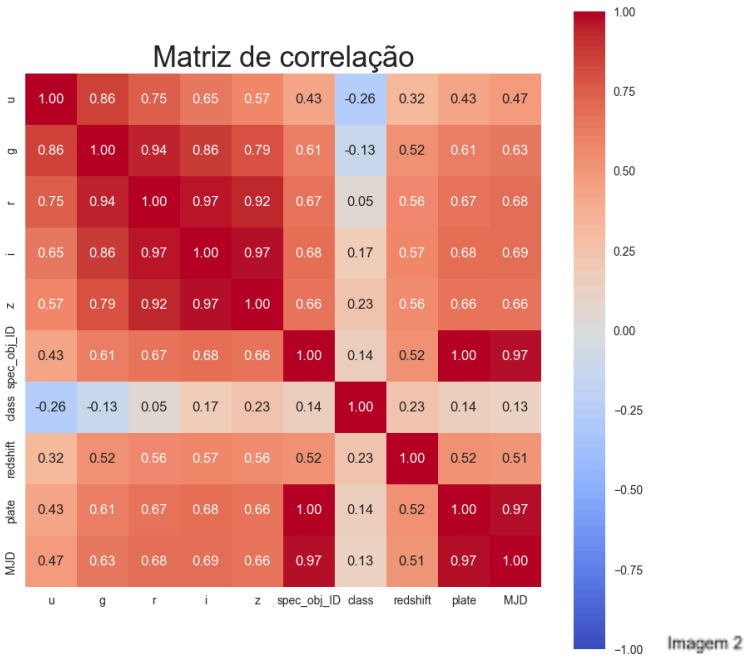

Imagem 2

Outrossim, os dois modelos de machine learning foram executados e aprimorados com o método de boosting e tuning presentes no Pycaret. Percebe-se que a variável “redshift” foi a mais importante para o algoritmo de classificação com mais de 70%. O redshift indica a mudança nas ondas de luz de um objeto que se afasta, ou seja, quanto maior o valor, maior a distância do ponto de observação.

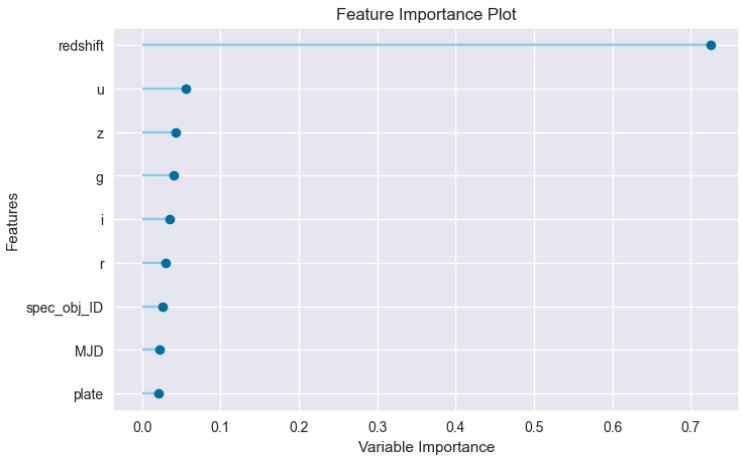

Imagem 3

As demais variáveis relevantes possuem uma característica em comum de serem filtros fotométricos para observação dos corpos celestes, sendo essas: “u”,”z”, e “g”. É notável que o corpo celeste “Quasar” foi o mais difícil de classificar corretamente, o que se mostra no recall baixo em comparação às demais.

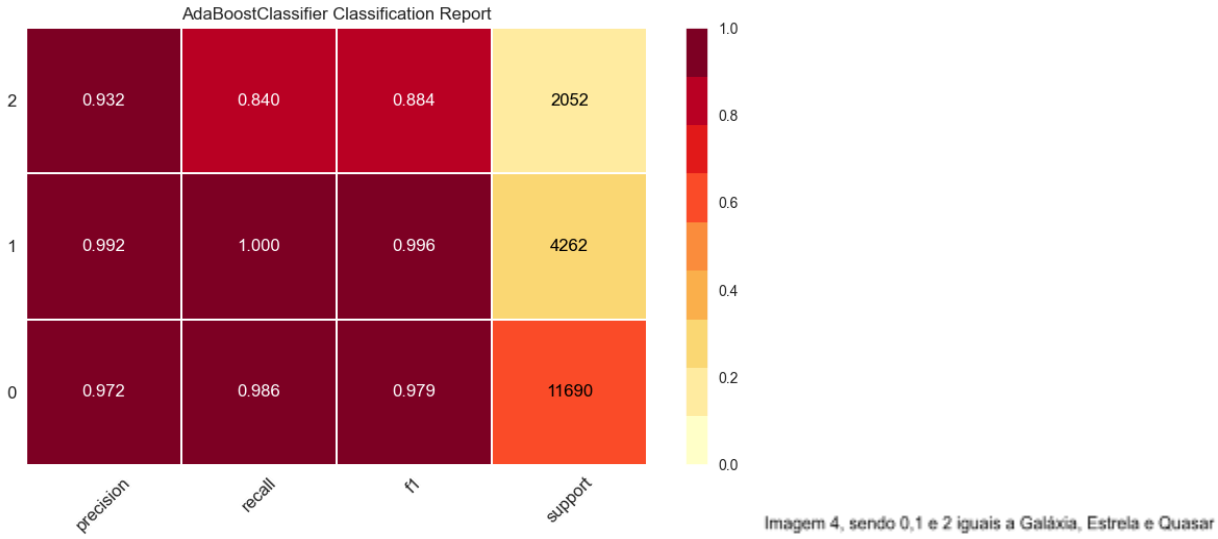

Imagem 4, sendo 0,1 e 2 iguais a Galáxia, Estrela e Quasar

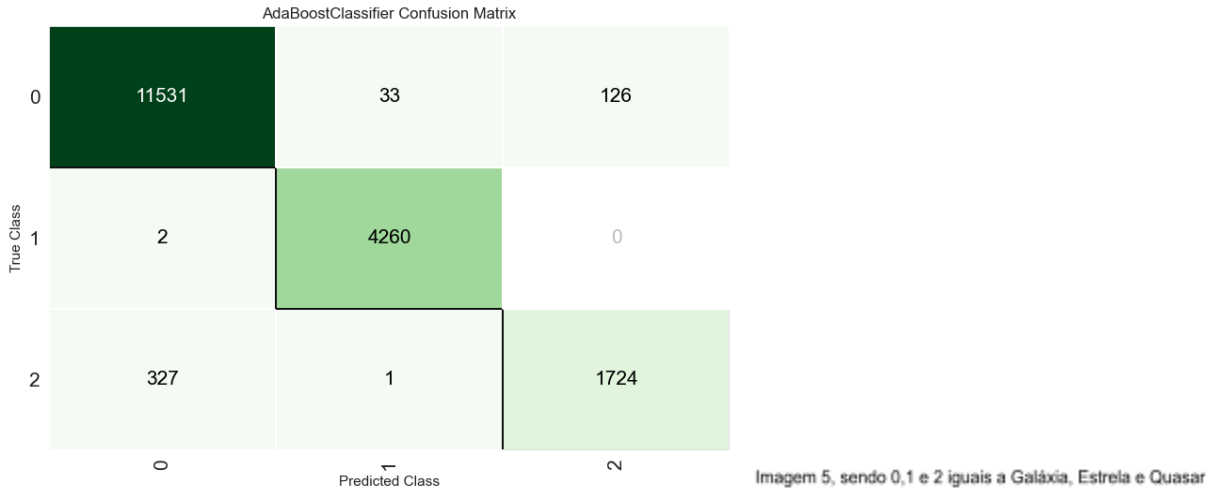

Imagem 5, sendo 0,1 e 2 iguais a Galáxia, Estrela e Quasar

Ambos modelos apresentaram uma acurácia superior a 95%, o treinamento com árvore de decisão e floresta aleatória se mostrou eficiente, as classes tinham um conjunto de filtros fotométricos com correlação alta, facilitando sua distinção com um treinamento adequado.

Conclusão

Ambos os modelos tiveram uma performance satisfatória e conseguiram, com uma boa precisão, fazer a classificação dos corpos celestes, sendo assim válidos para próximas classificações em surveys realizados pela SDSS. A eficiência na classificação dos corpos celestes pode contribuir na categorização e mapeamento de nosso universo, colaborando com diversas áreas no campo científico. Estrelas, galáxias e até os quasares podem ser reconhecidos rapidamente com o modelo bem treinado.

Contribuições da equipe

João Portela (50% de contribuição) - Contribuiu na limpeza dos dados, seleção de features e treinamento do modelo de floresta aleatória.

Raphael Passos(50% de contribuição) - Contribuiu na limpeza dos dados, identificação e remoção de outliers e treinamento do modelo de XGBoost.

Referências

- Gráficos feitos com bibliotecas do Python e R
- @fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Coletado[21/09/2024] de https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.
- Aprendizado de máquina: uma abordagem estatística, Izbicki, R. and Santos, T. M., 2020.
- Estatística e ciência de dados., Morettin, Pedro Alberto, and Julio da Motta Singer, 2022.