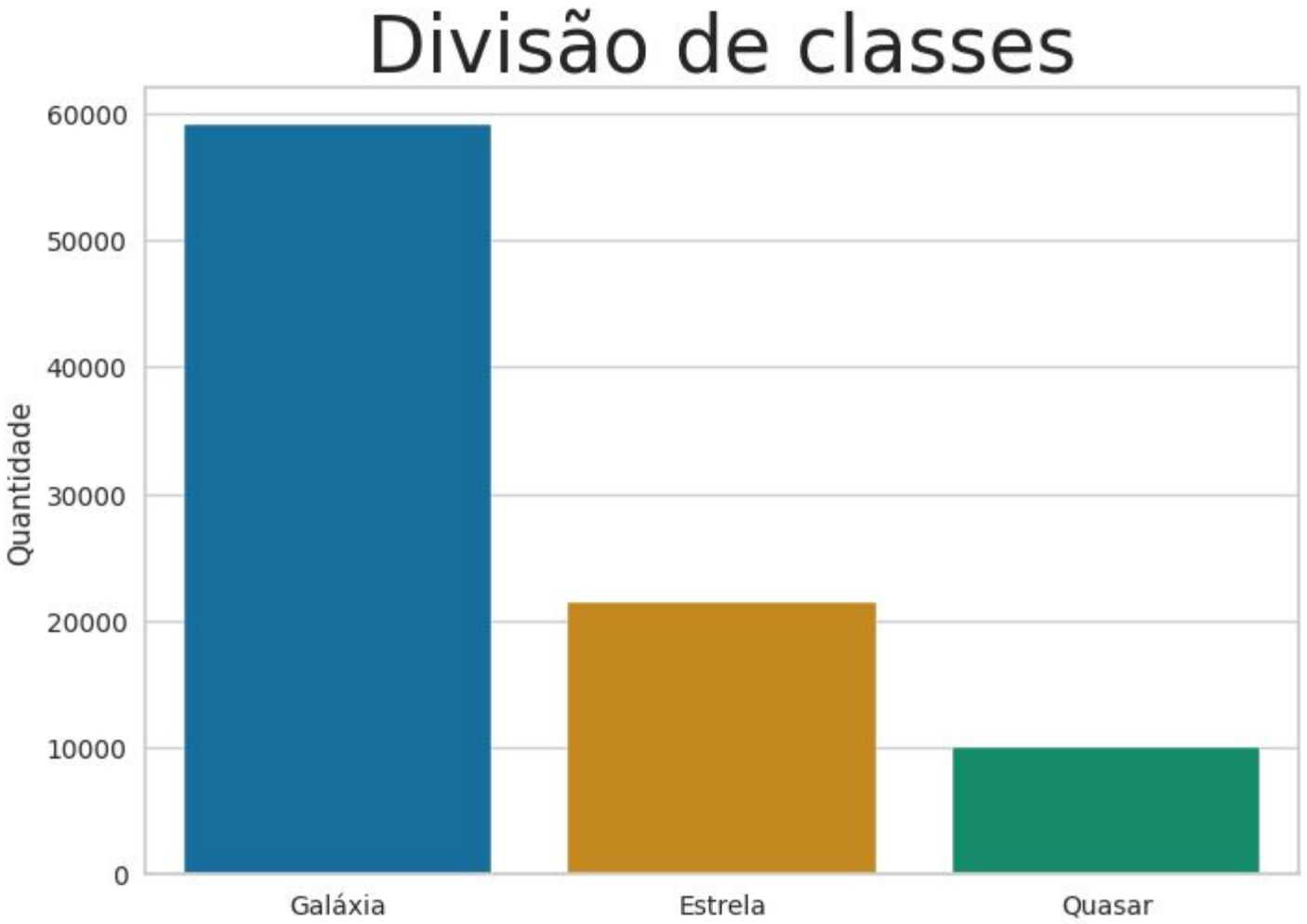
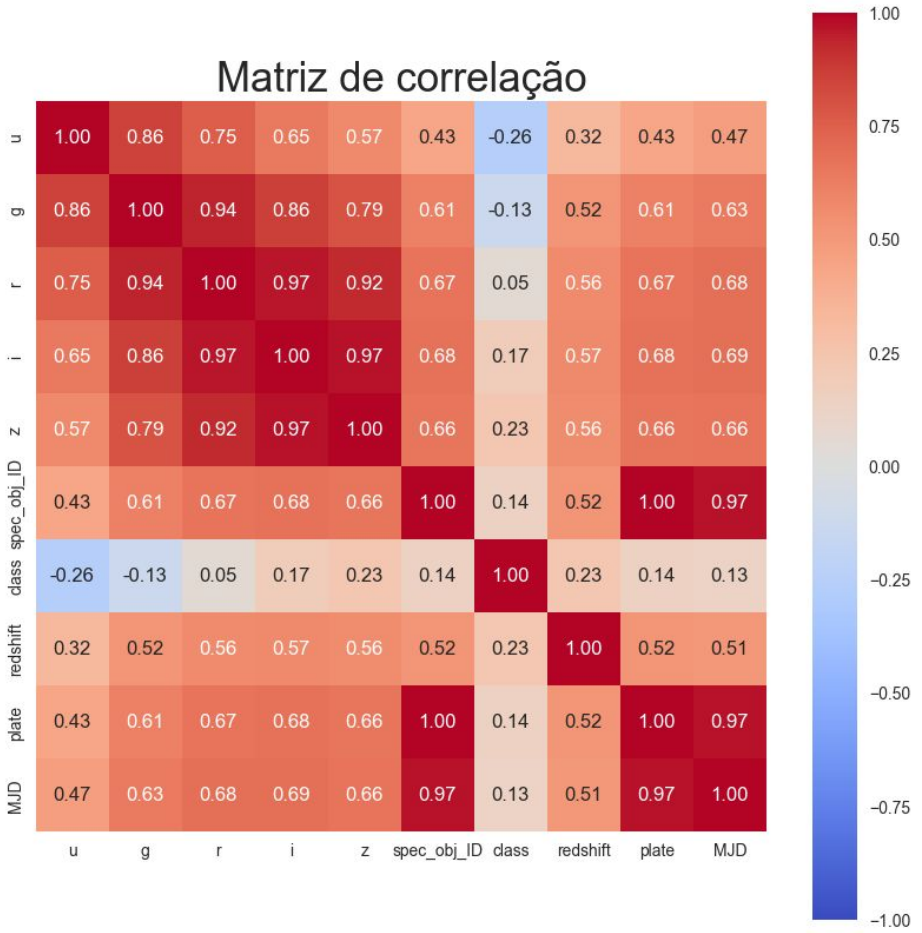
Machine Learning para classificação de corpos estelares

Intro

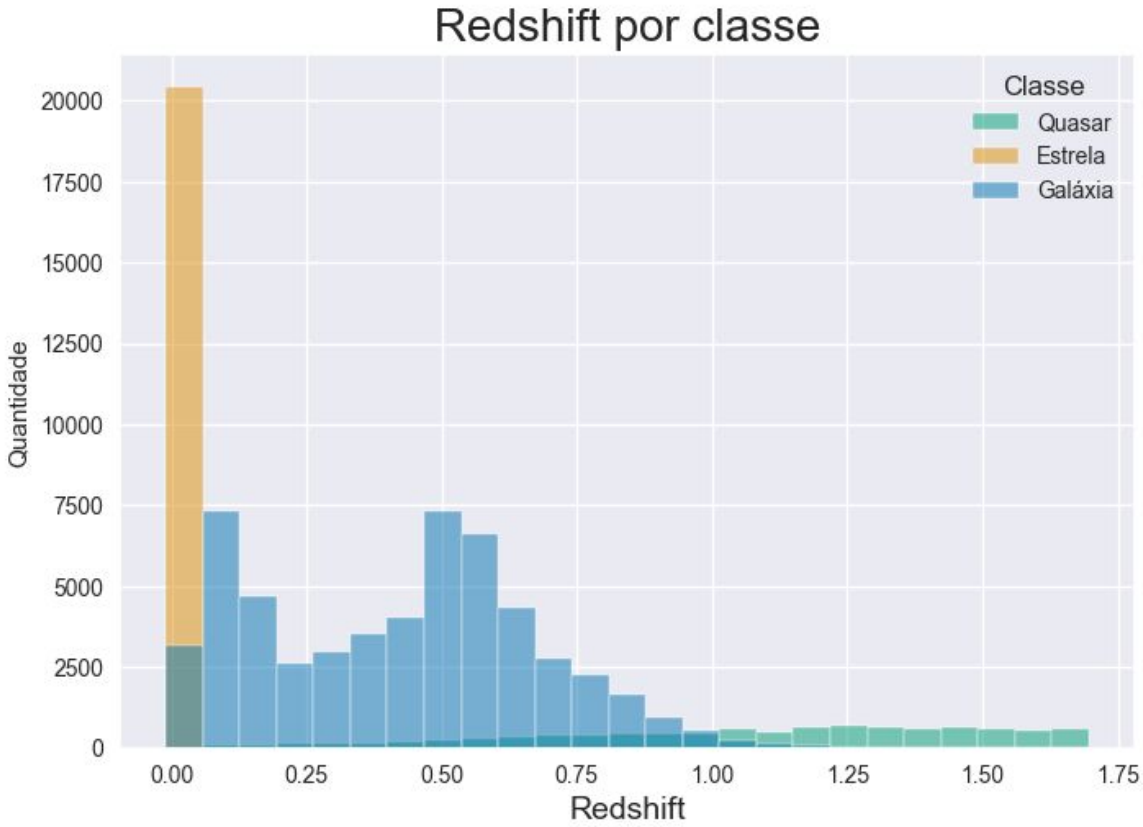
Há décadas a humanidade vem aprimorando as tecnologias e estudos para classificar estrelas, essa pesquisa gera resultados importantes para entendermos nosso universo e encontrar fenômenos que podem ser favoráveis para a evolução da humanidade. Utilizando o dataset da SDSS (Sloan Digital Sky Survey) que 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.


Divisão de classes
Quantidade
60000
50000
40000
30000
20000
10000
0
Galáxia
Estrela
Quasar

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.

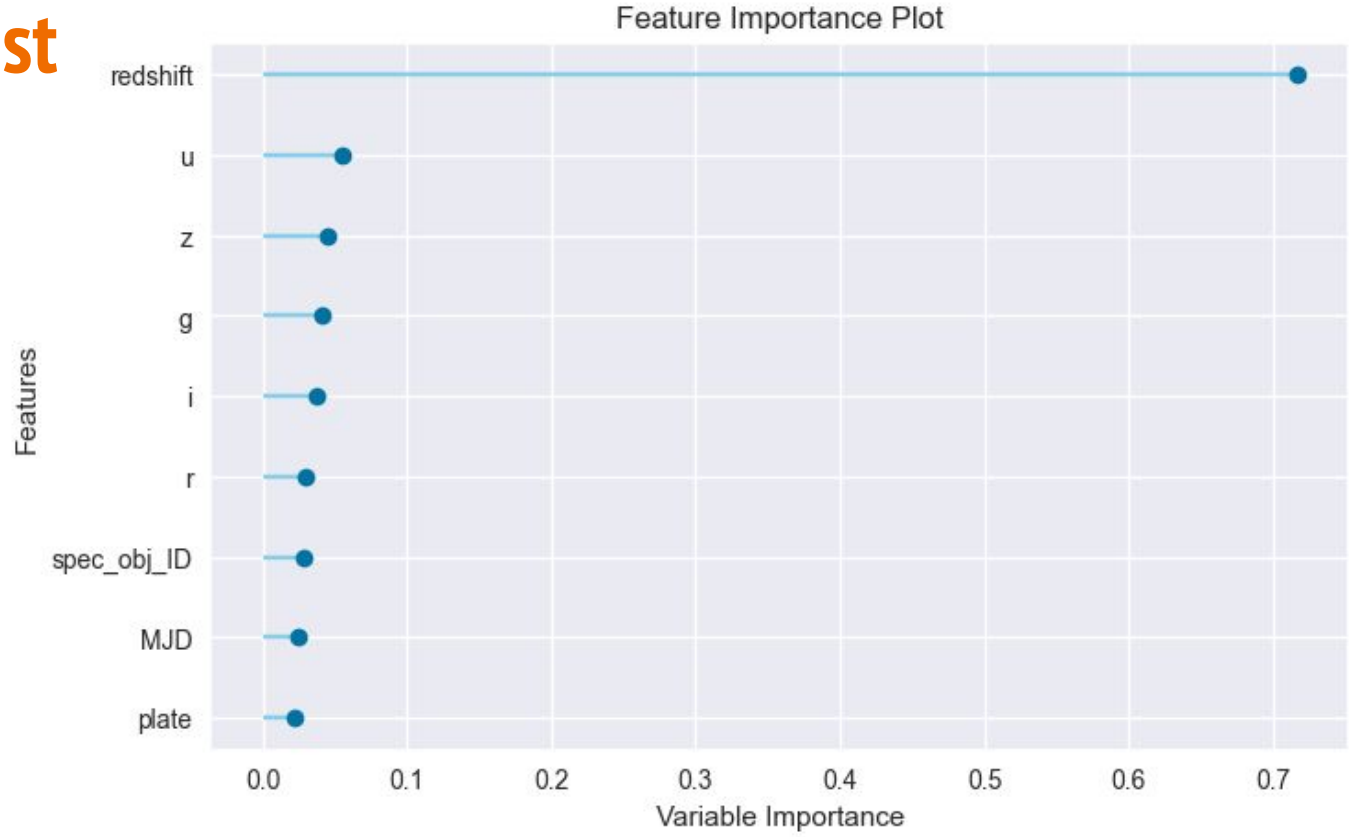


Os dois modelos de machine learning foram executados e aprimorados com o método de boosting e tuning. Percebe-se que a variável “redshift” foi a mais importante para o algoritmo de classificação com mais de 70%. O redshift indica a mudança nas ondas de luz de um objeto que se afasta, ou seja, quanto maior o valor, maior a distância do ponto de observação.



Random Forest

As demais variáveis relevantes possuem uma característica em comum de serem filtros fotométricos para observação dos corpos celestes, sendo essas: “u”,”z”, e “g”



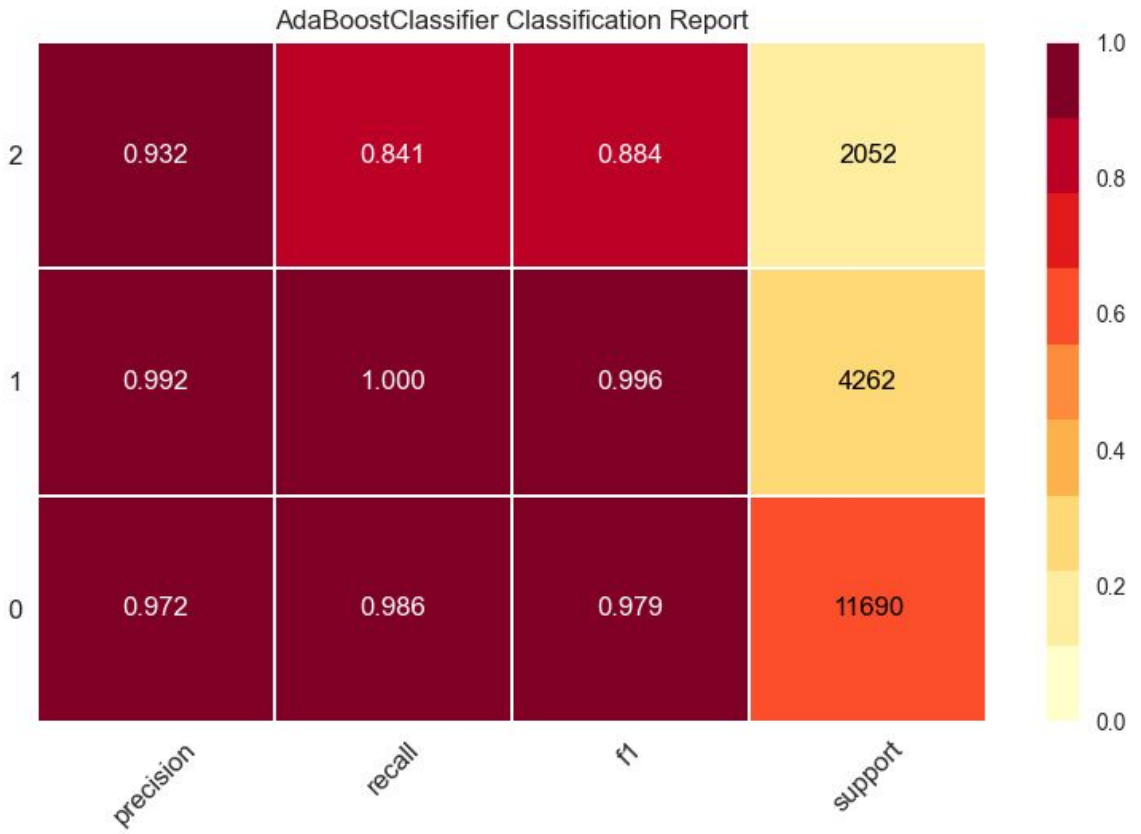
Random Forest

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.

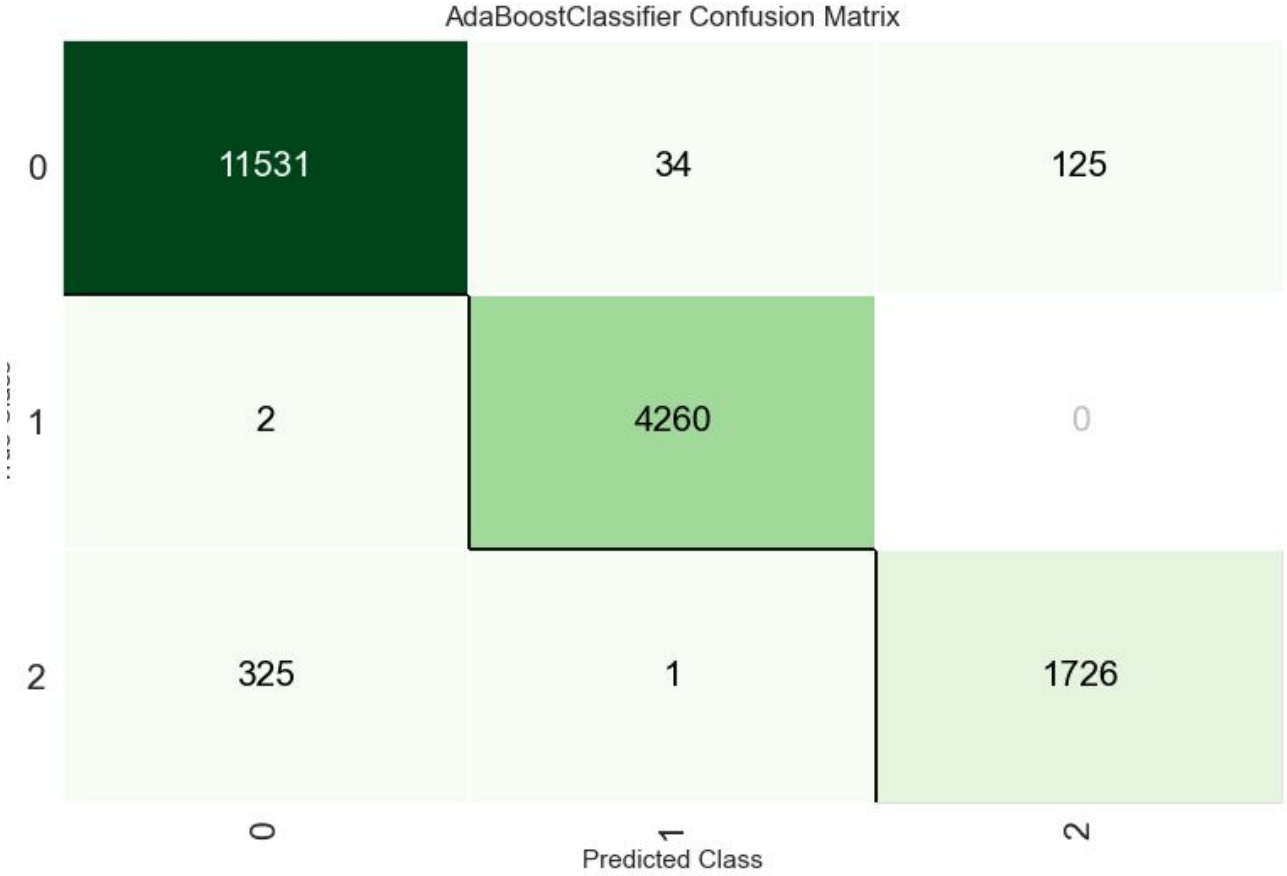
0 - Galáxia

1 - Estrela

2 - Quasar



Random Forest

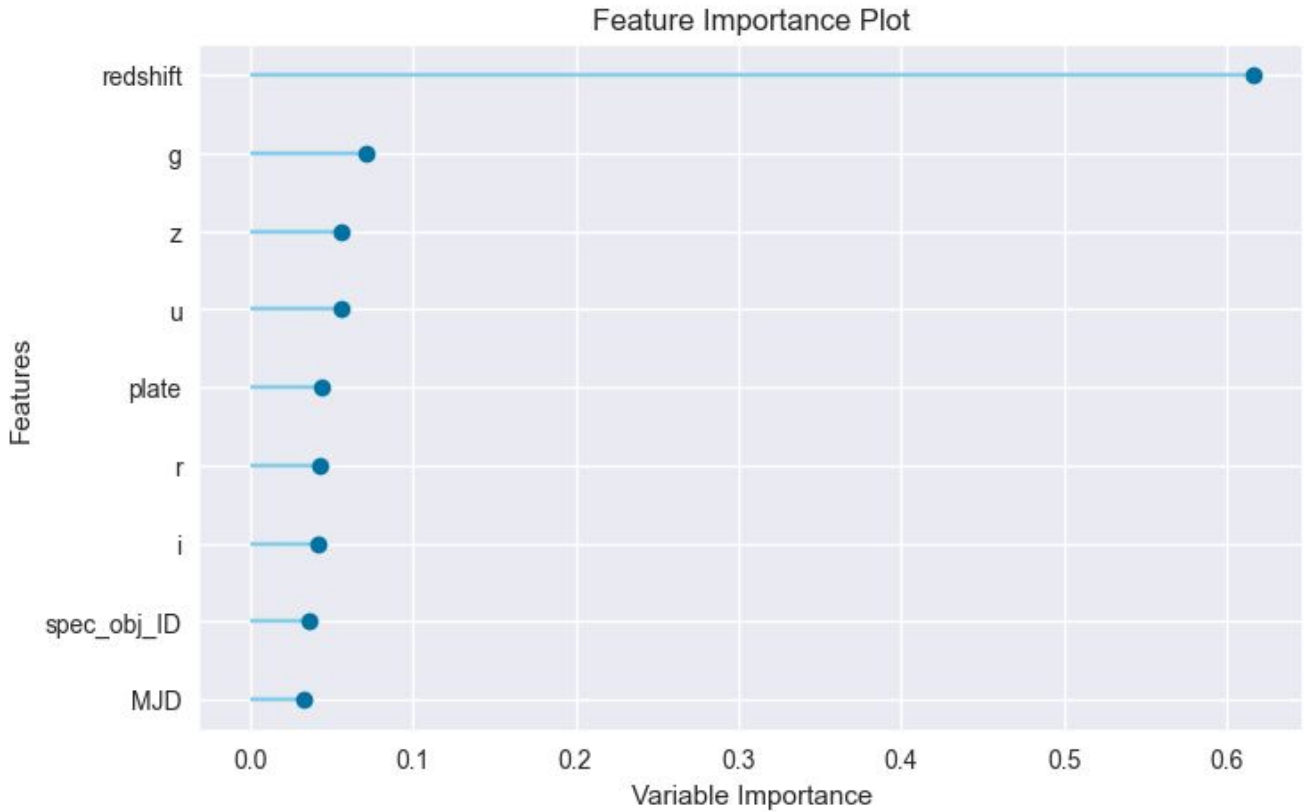


0 - Galáxia

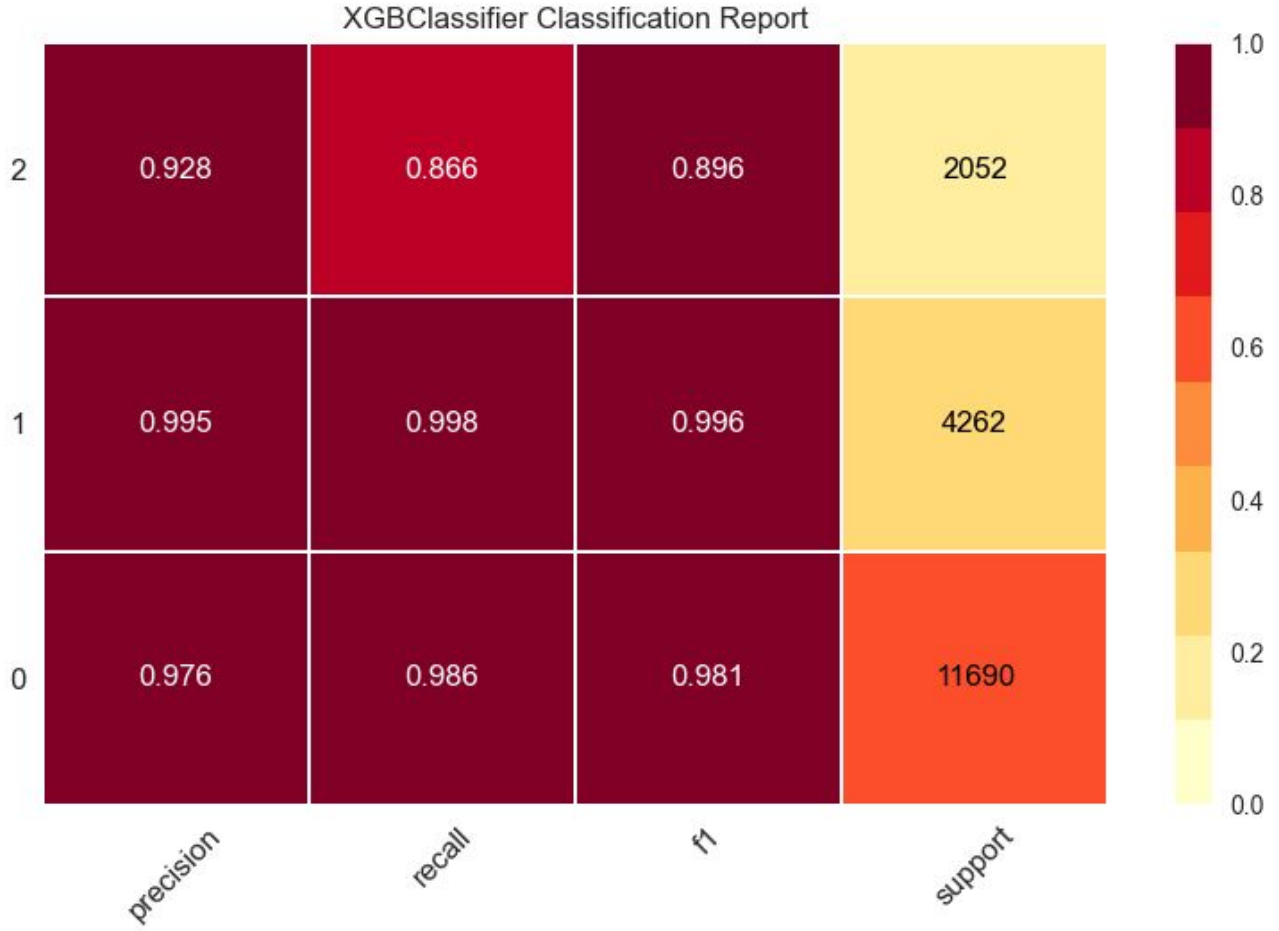
1 - Estrela

2 - Quasar

Xgboost


Feature Importance Plot
redshift
g
z
u
plate
r
i
spec_obj_ID
MJD
Features
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Variable Importance

Xgboost

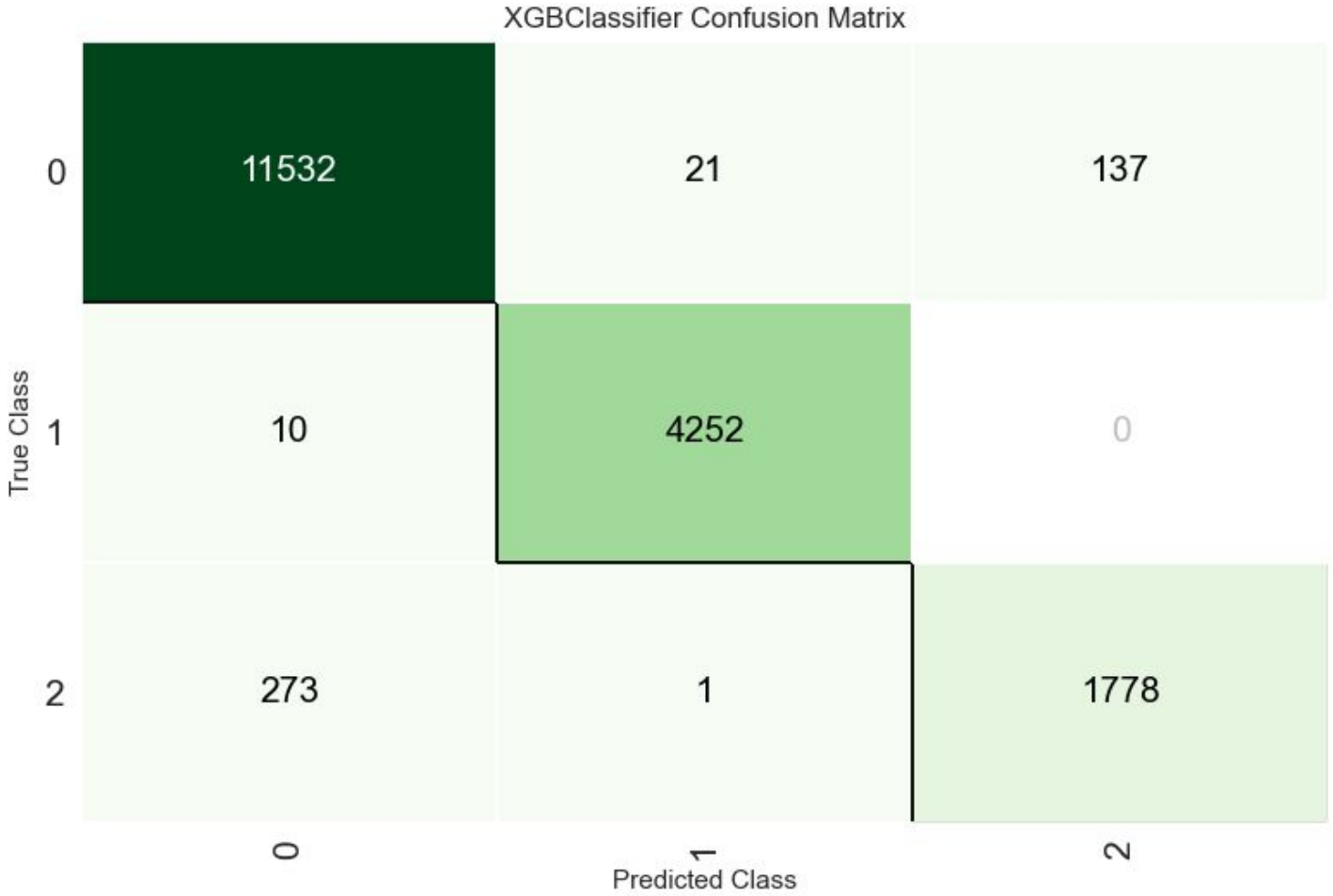


0 - Galáxia

1 - Estrela

2 - Quasar

Xgboost


XGBClassifier Confusion Matrix
True Class
0
1
2
11532
21
137
10
4252
0
273
1
1778
0
1
2
Predicted Class

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.

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.