Aprendizagem Supervisionada - Problemas de classificação

SDSS Galaxy Classification DR18

Tópico L - Trabalho prático 2 - IA - Grupo A2_22

Félix Martins, up202108837
Pedro Lima, up202108806
Pedro Januário, up202108768

Problema

O Sloan Digital Sky Survey (SDSS) é um levantamento astronómico acerca do desvio para o vermelho de corpos celestes, que encontrou cerca de 1000 milhões de objetos, dos quais quase 3 milhões são galáxias.

O data set sobre o qual trabalharemos contém dados de imagens fotométricas de 100 mil dessas galáxias, classificadas como 'STARFORMING' ou 'STARBURST'.

Neste trabalho, utilizaremos os dados como parte de Aprendizagem Supervisionada, desde o pré-processamento (correção e filtragem) dos mesmos, à seleção de conjuntos de treino e de teste para modelos de aprendizagem, treinados segundo diferentes algoritmos.

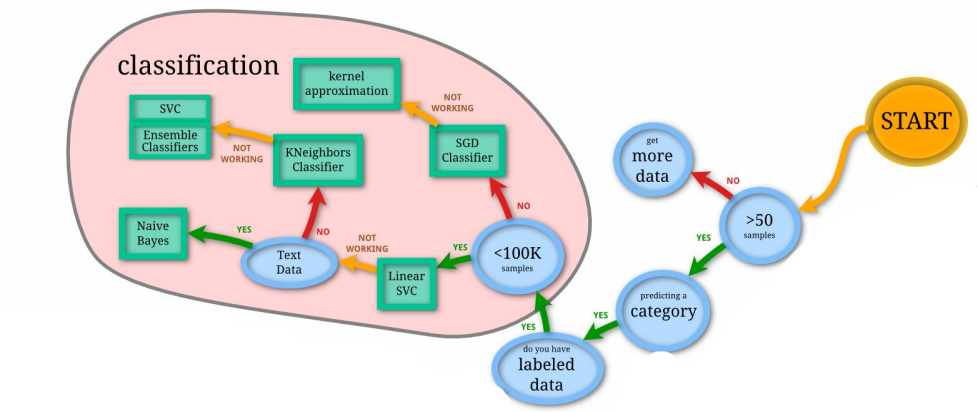
Material relacionado

- Website Oficial da *Sloan Digital Sky Survey* (https://www.sdss.org/)
- *Sloan Digital Sky Survey* na Wikipédia (https://pt.wikipedia.org/wiki/Sloan_Digital_Sky_Survey)
- *SDSS Galaxy Classification*, Bryan Cimo em Kaggle (https://www.kaggle.com/code/bryancimo/sdss-galaxy-classification)
- Biblioteca scikit-learn com explicação de algoritmos de aprendizagem supervisionada (https://scikit-learn.org/stable/index.html)

Ferramentas e Algoritmos

Algoritmos:

- Decision Tree
- K-NN: Nearest Neighbour
- SVM: Support vector machine
- Redes neuronais: MLPClassifier
- Random Forest



Auxiliar de escolha de algoritmos de aprendizagem supervisionada

Bibliotecas de Python: Pandas, Numpy, Matplotlib, Seaborn, Scikit-Learn

Implementação

Para implementar estes os modelos, utilizou-se a biblioteca scikit-learn, nomeadamente as funções DecisionTreeClassifier, KNeighborsClassifier, SVC, MLPClassifier e RandomForestClassifier.

Estratégia e Passos

O projeto foi dividido em 3 fases:

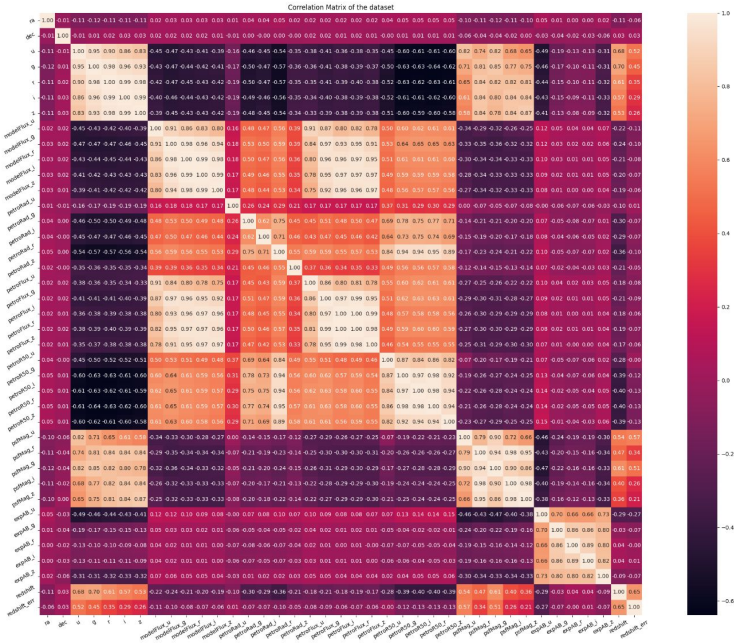
- Pré-processamento dos dados.
- Descoberta de quais os melhores parâmetros para cada modelo, através de Grid Search.
- Treino dos modelos finais, com os parâmetros descobertos, onde apenas avaliamos a performance de cada algoritmo.

Cada uma das fases encontra-se implementada e comentada num ficheiro Jupyter Notebook diferente.

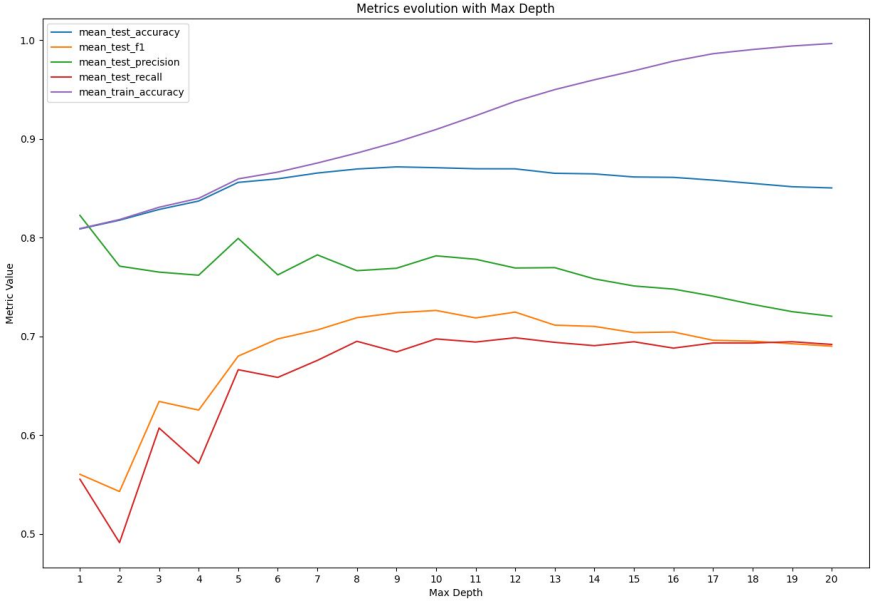
Pré-processamento dos dados

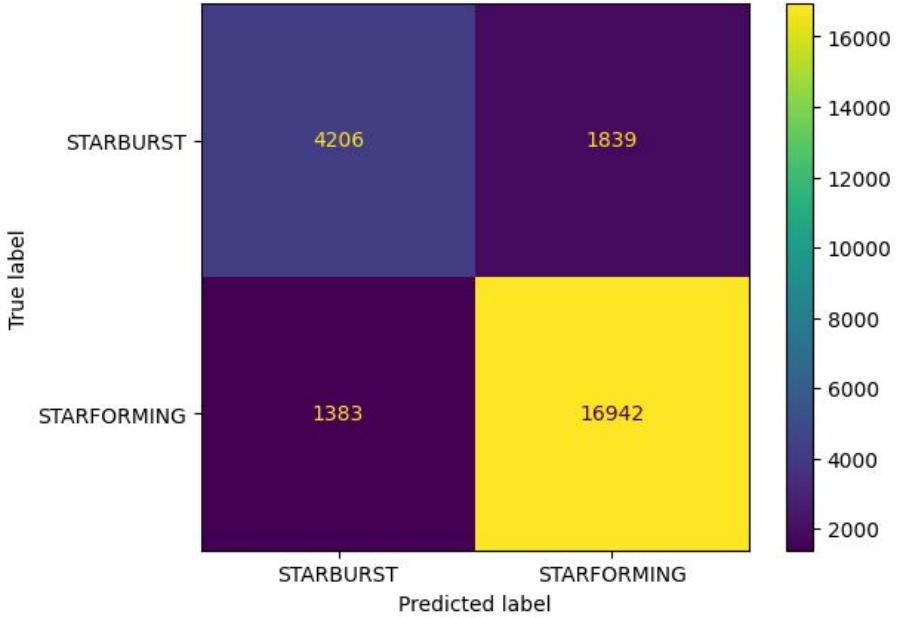
Foi levado a cabo o pré processamento dos dados, nomeadamente:

- Verificação da existência de valores nulos e linhas duplicadas (inexistentes);
- Procura de erros ortográficos entre os nomes possíveis para as subclasses de galáxias (resultado da classificação) (inexistentes);
- Eliminação de linhas que contêm algum valor igual a -9999, pois considerámos que tal valor é, presumivelmente, indicador de invalidez da medida;
- Procura de *outliers* nas colunas, eliminando linhas cujo valor de “redshift_err” seja superior a 1, dado que erros grandes põem em causa a fiabilidade dos dados da linha.
- Agregação de *features* fortemente correlacionadas, nomeadamente 5 conjuntos de 4 colunas cada, que tinham um elevado coeficiente de correlação de Pearson entre si, substituindo-se pelas suas médias.

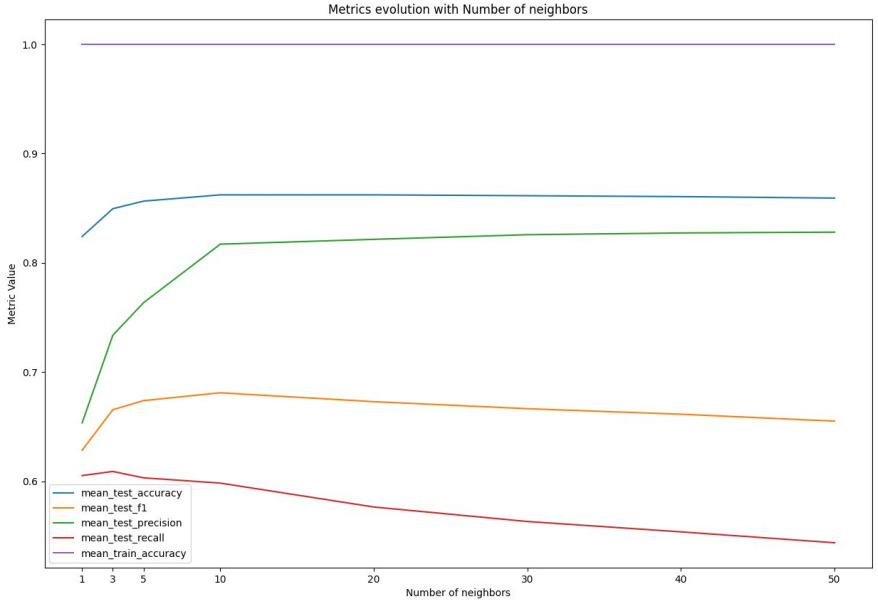


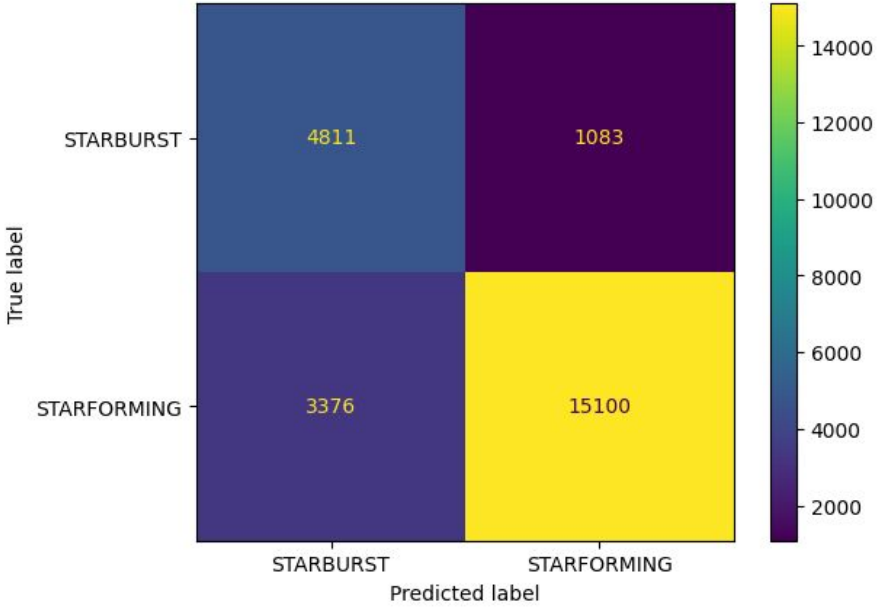
Avaliação e Comparação
Decision Tree



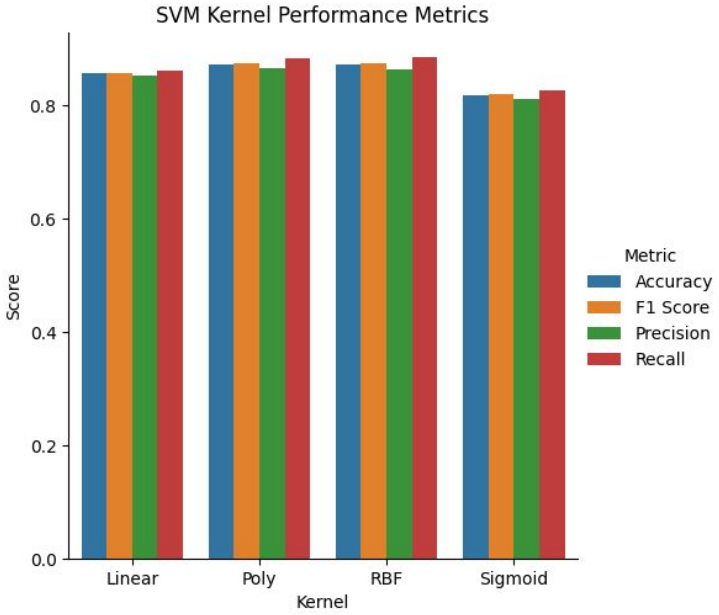


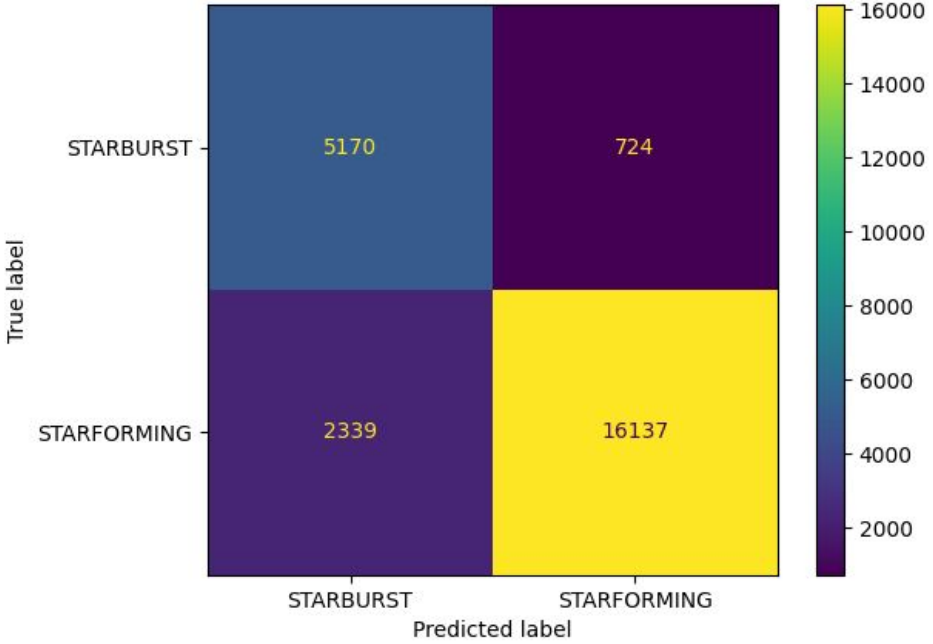
Avaliação e Comparação
K Nearest Neighbors



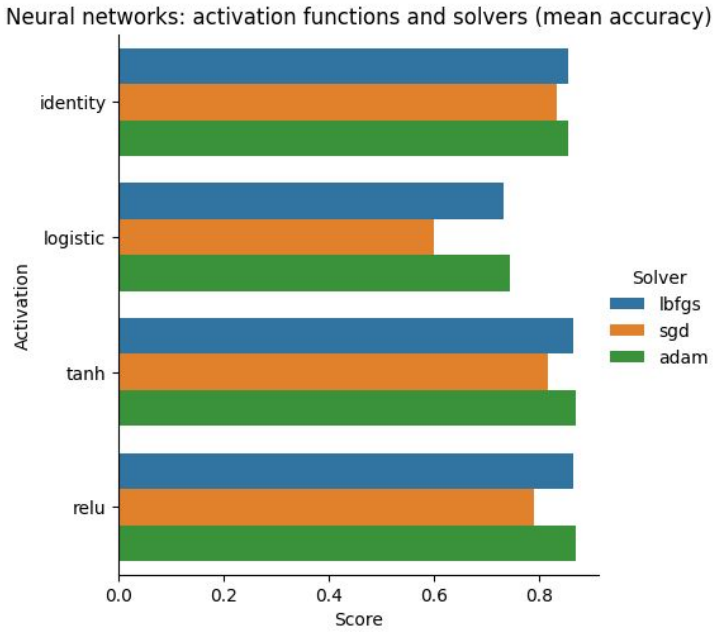


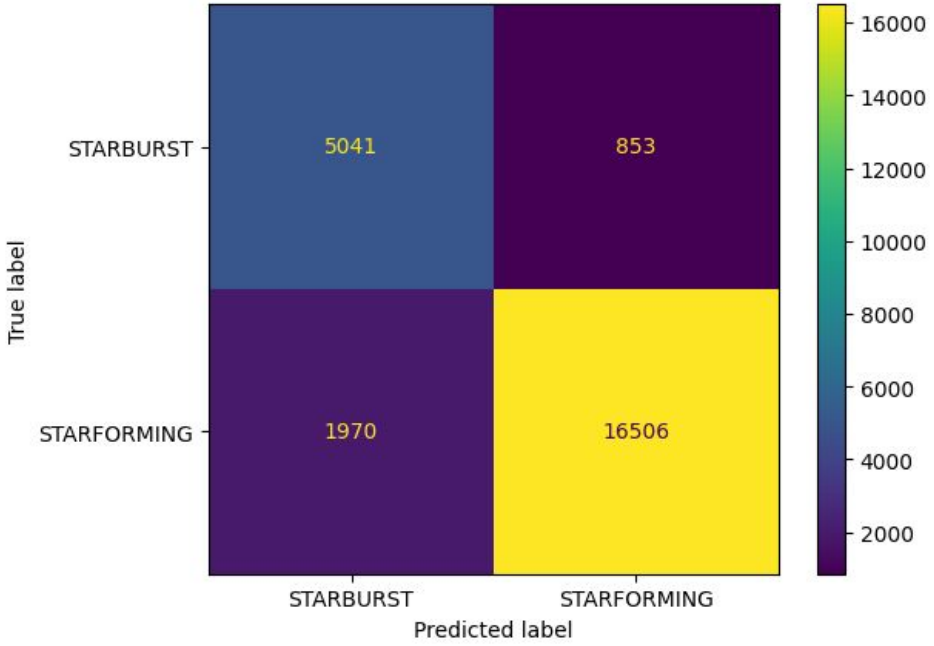
Avaliação e Comparação
Support Vector Machine



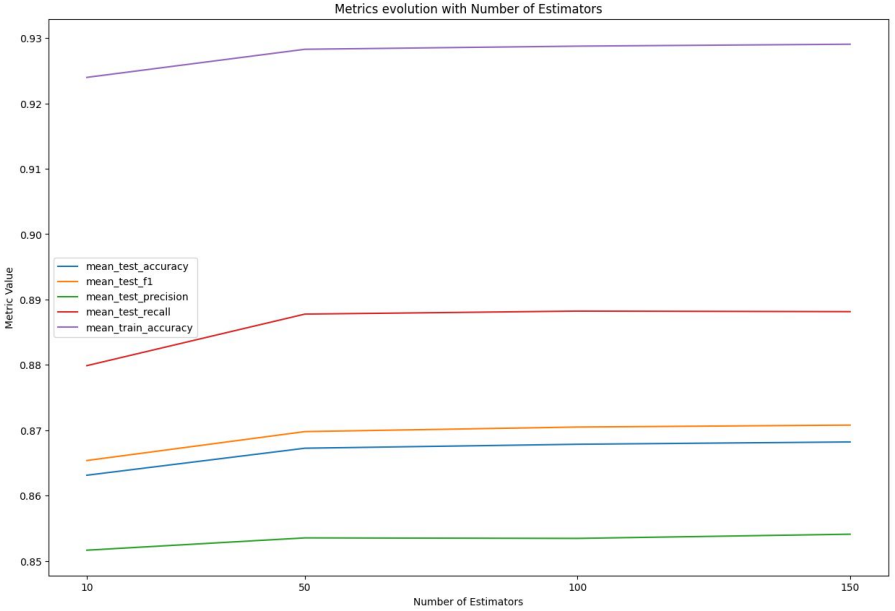


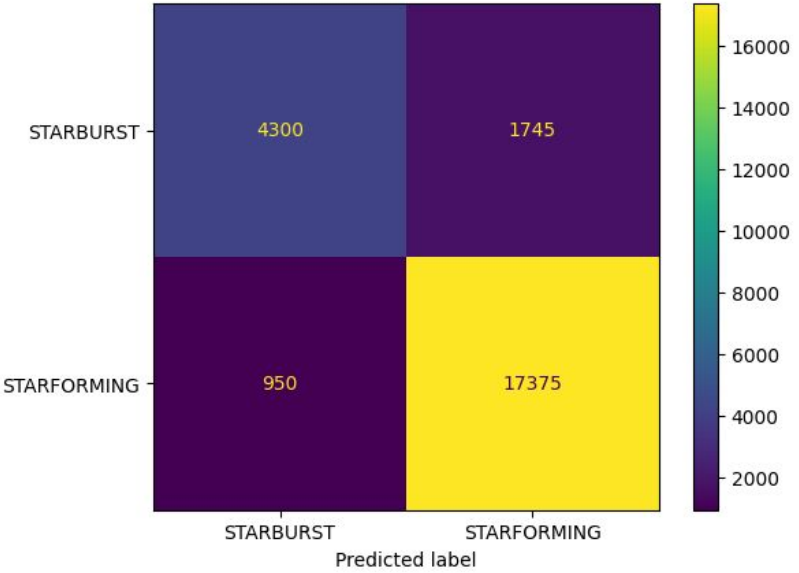
Avaliação e Comparação
Neural Network





Avaliação e Comparação
Random Forest





Conclusões

Todos os classificadores têm valores idênticos para os indicadores de *performance*.

Na *accuracy* e precisão, destaca-se o *Random Forest*.

Já quanto a *recall*, o classificador vencedor é *Neural Networks* e, em F1, SVM.