Classification

Presented by:
Amer Saleh

December 2021

I. Introduction

A star is an astronomical object consisting of a luminous spheroid of plasma held together by its own gravity. The nearest star to Earth is the Sun. Many other stars are visible to the naked eye at night, but due to their immense distance from Earth they appear as fixed points of light in the sky. The most prominent stars are grouped into constellations and asterisms, and many of the brightest stars have proper names. The observable universe contains an estimated 1022 to 1024 stars, but most are invisible to the naked eye from Earth.

II. Study Methodology

The methodology of this project is as follows, extracting data from Kaggle for a stars and it was more than 32,000 rows and 7 columns.

In the cleanup step, we dropped the Specter type column because it was categorical and useless, and we dropped missing values in the dataset and outliers.

We explored the data, applied comprehensive analysis methods to the data, and extracted important information from the data.

We will forecast type of stars if dwarf or giant, by color index and Visual Apparent Magnitude of the Star (Vmag), Distance Between the Star and the Earth (Plx), spectral type, B-V color index.

III. Data Description

The data set is provided in .csv format, contains information of Visual Apparent Magnitude

of the Star(Vmag), Distance Between the Star and the Earth (Plx), spectral type, B-V color index

The data set was extracted from Kaggle

Variables	Description
Vmag	Visual Apparent Magnitude of the Star(Vmag), like 8.25.
PLX	Distance Between the Star and the Earth i.e. 31.66.
E-plx	Standard Error of Plx .
B-V	B-V color index. (A hot star has a B-V color index close to 0 or negative, while a cool star has a B-V color index close to 2.0.
SpType	Spectral type.
Amag	Absolute Magnitude of the Star
Target Class	Whether the Star is Dwarf (0) or Giant (1)

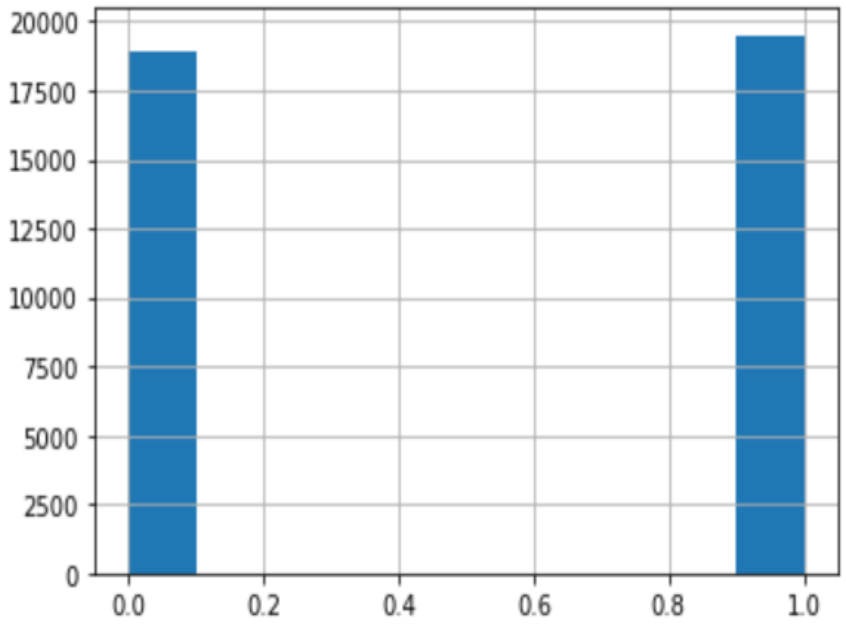
IV. Tools and Libraries:

- Python.
- Jupyter Notebook.
- PowerPoint.
- Excel
- NumPy.
- Pandas.
- Matplotlib
- sciket learn

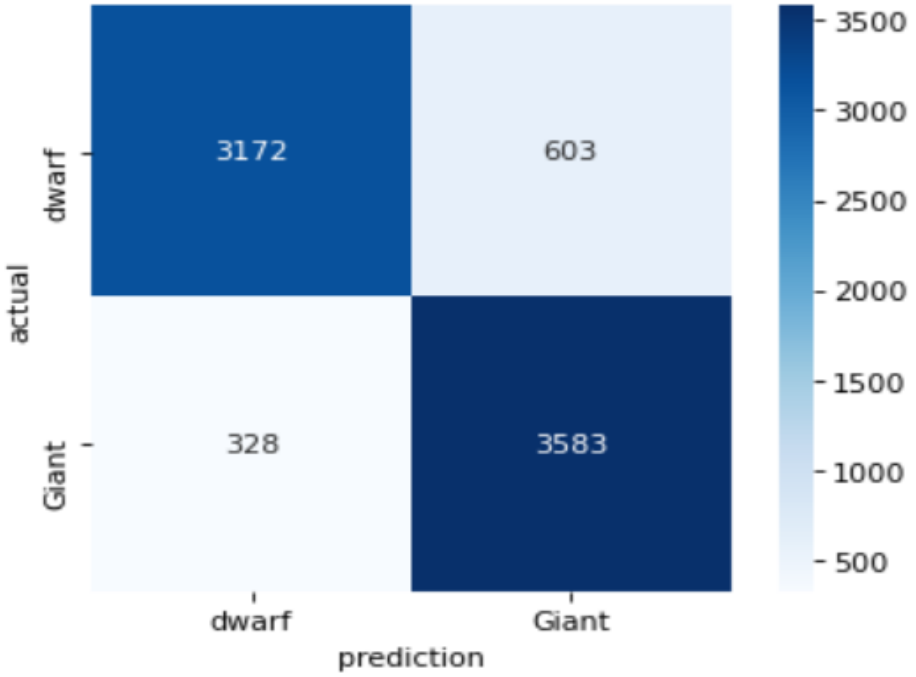
V. Classification Model

Model	Train	Test
Baseline	86%	84%
KNN	88%	88%
Logistic Regression (GS)	88%	87%
Voting(soft)	88%	89%
XGBoost	89%	89%

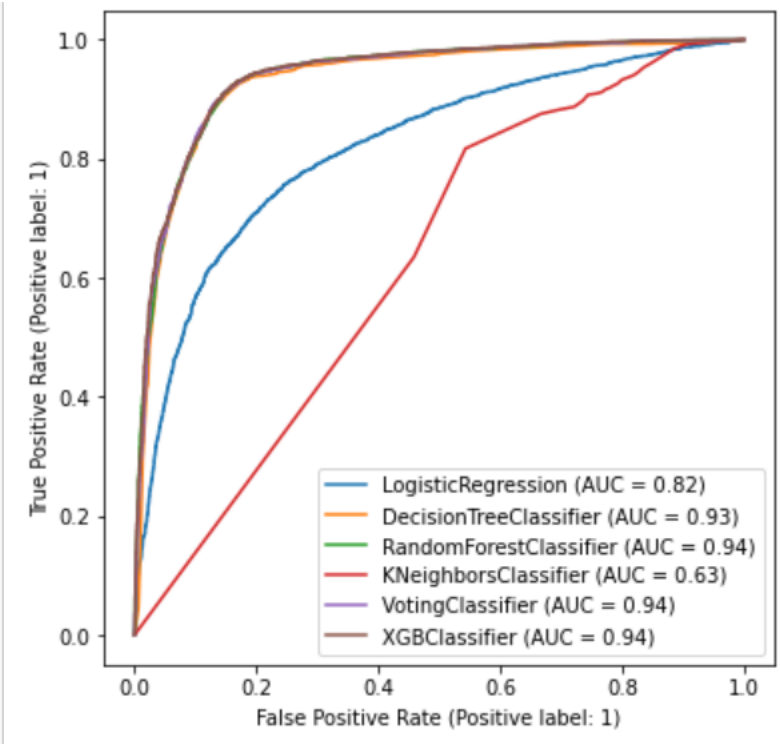
Target histogram :



Confusion matrix plot :



Models plot :



VI. Summary

- The results was very close to each other.
- We find The best model is **XGBoost Model.**

Reference :

https://www.kaggle.com/vinesmsuic/star-categorization-giants-and-dwarfs