Project Report

Title: Star Prediction

Data Mining and Warehousing

Hafiz Bilal Ahmed (B18101036)

About Dataset:

Stellar Classification uses the spectral data of stars to categorize them into different categories. The modern stellar classification system is known as the Morgan–Keenan (MK) classification system. It uses the old HR classification system to categorize stars with their chromaticity and uses Roman numerals to categorize the star's size.

In this Dataset, we will be using Apparent Magnitude and its distance from earth to Identify Giants and Dwarfs. Dwarfs is taken as 0 and Giants as 1 in Target Class

Source:

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

Column Description:

1. appmag Visual Apparent Magnitude of the Star
2. **distancefromearth** Distance Between the Star and the Earth
3. **errdistance** Error of distance
4. **colorindex** 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.
5. **absolutemag** Absolute magnitude of star
6. **TargetClass** Whether the Star is Dwarf (0) or Giant (1)

Preprocessing

Missing Values:

Missing values are found in column **appmag** and **distancefromearth** and those values are filled by taking **mean** of values of whole column.

Outliers:

Lower limit and upper limit is set on **appmag** column and values that are smaller than lower limit and greater than upper limit are discarded.
Lower limit and upper limit is set by using quantiles.
82 rows are found to be outliers.

Training Data is 70%

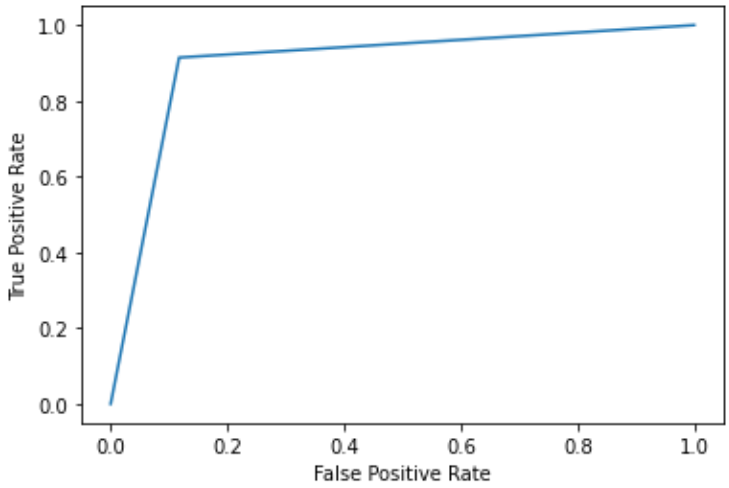
Testing Data is 30%

Model Evaluation:

Decision Tree—Gini Index

Accuracy= 0.8979400749063671

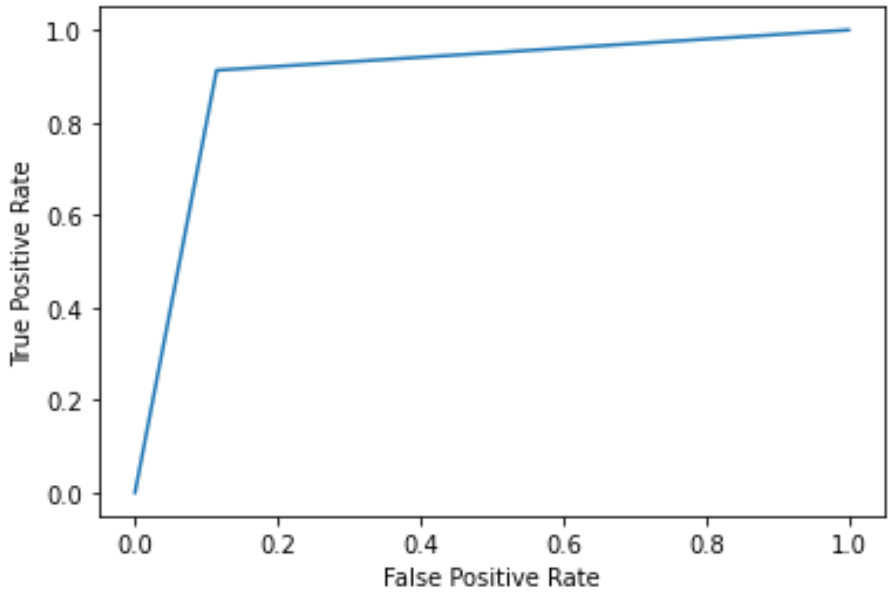
Roc Curve:



Decision Tree—Entropy

Accuracy= 0.898876404494382

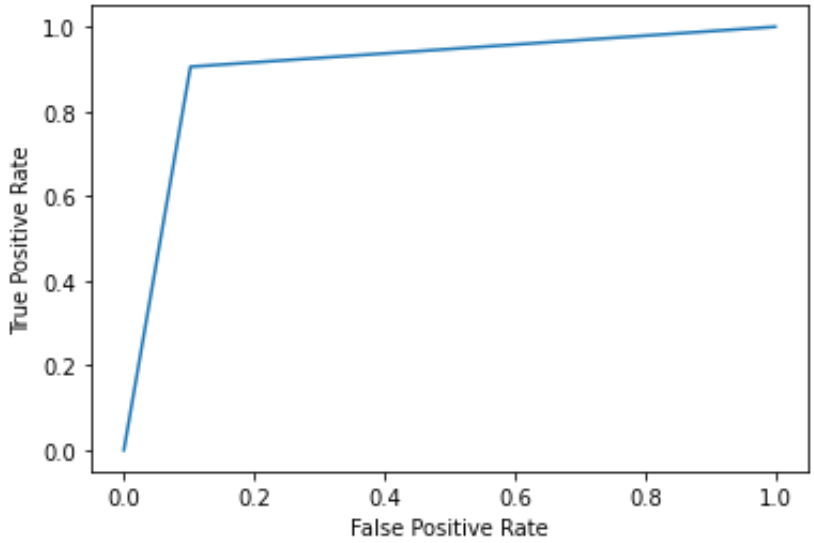
Roc Curve:



Logistic Regression

Accuracy= 0.901685393258427

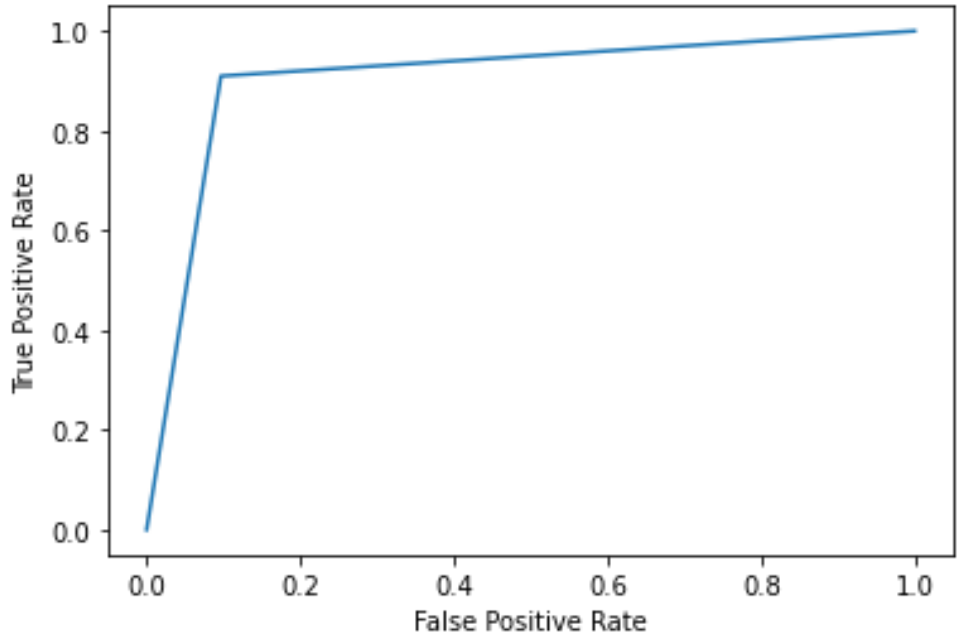
Roc Curve:



Random Forest

Accuracy= 0.9063670411985019

Roc Curve:



Naïve Bayes:

Accuracy= 0.8885767790262172

Roc Curve: