Project Report

Stellar classification

Sara Gholamhoseinian
Data analysis course
Spring 2022

overview

in this project, we aimed to achieve three goals :

1. Find the relation between different features of stars dataset
2. Check the correlation between features
3. Plot the Hertzsprung-Russell diagram
4. Use machine learning methods and compare their accuracy

Part 1 : data exploration
In this part we analyzed the main features of the dataset. This dataset is described by 6 characteristics of stars :

1. Temperature
2. Radius
3. Luminosity
4. Absolute magnitude
5. Type
6. Spectral class

Generally, stars can be classified by too many different methods but in this method, we classified stars in 6 groups that each group, relates to the one special type :

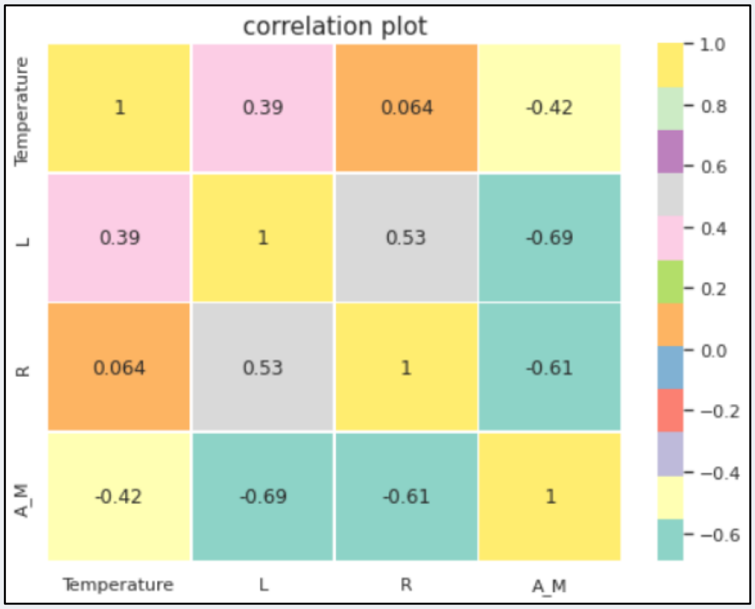
1. Brown dwarf : type 0
2. Red dwarf : type 1
3. White dwarf : type 2
4. Main sequence : type 3
5. Supergiant : type 4
6. Hypergiant : type 5

Properties of the dataset after the data exploration:

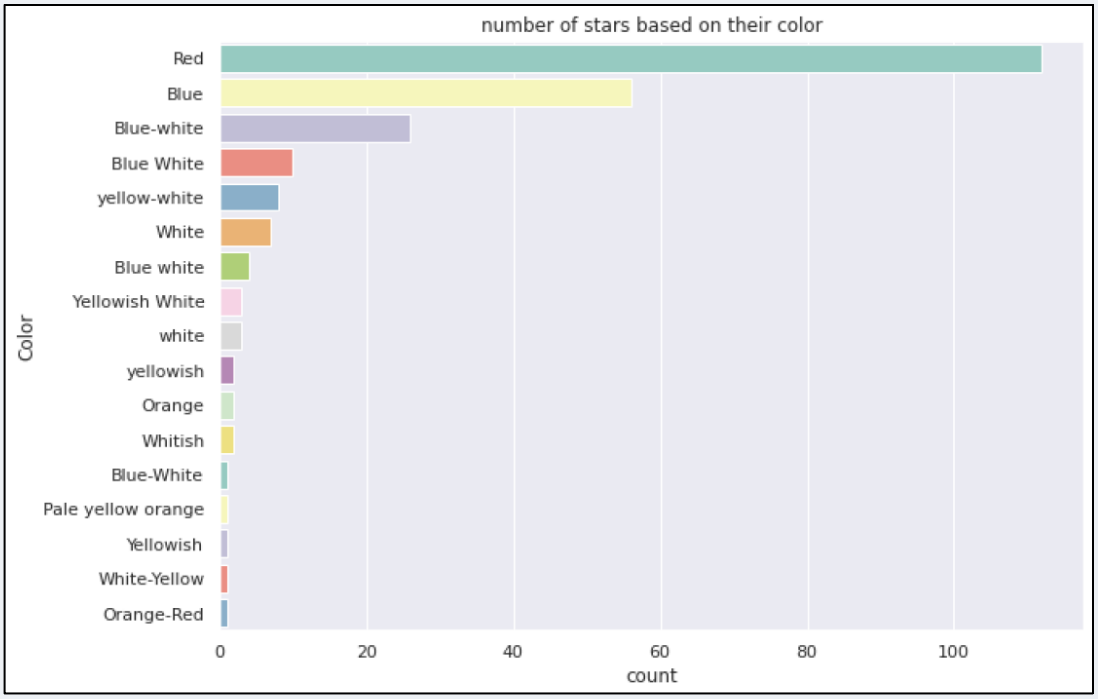
1. This dataset contains above 6 features about 240 stars.
2. It doesn't have any null or duplicated value.

Part 2 : data visualization
This part includes many different plots which describes the relation between different features.

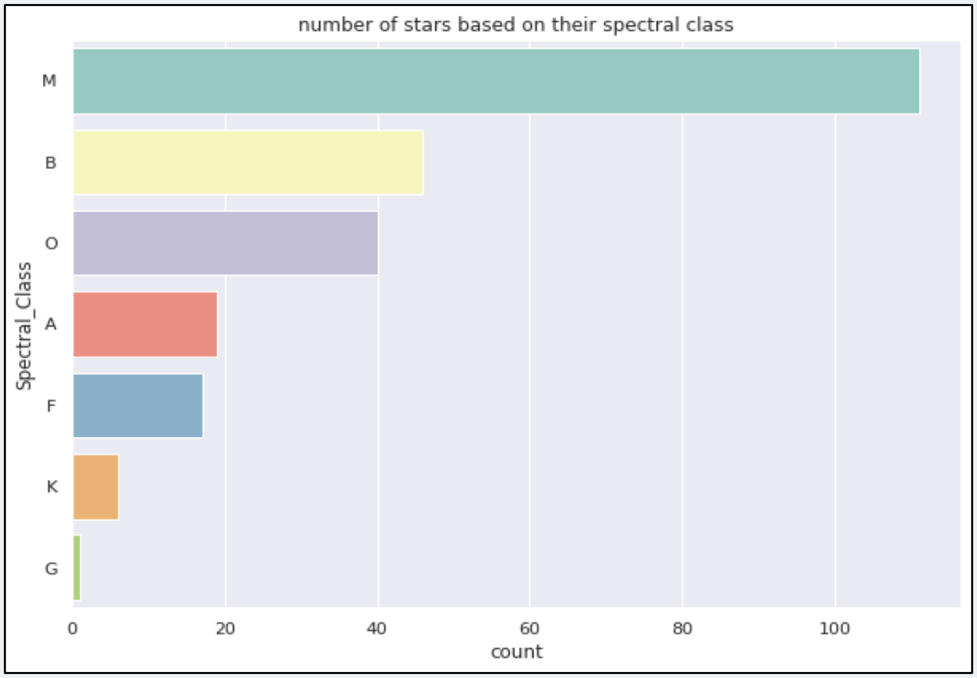
1. Correlation plot :
   We gained following results from the correlation plot:
   - Absolute magnitude is anti-correlated with other features
   - Radius and temperature are approximately uncorrelated
   - Luminosity is correlated with temperature and radius



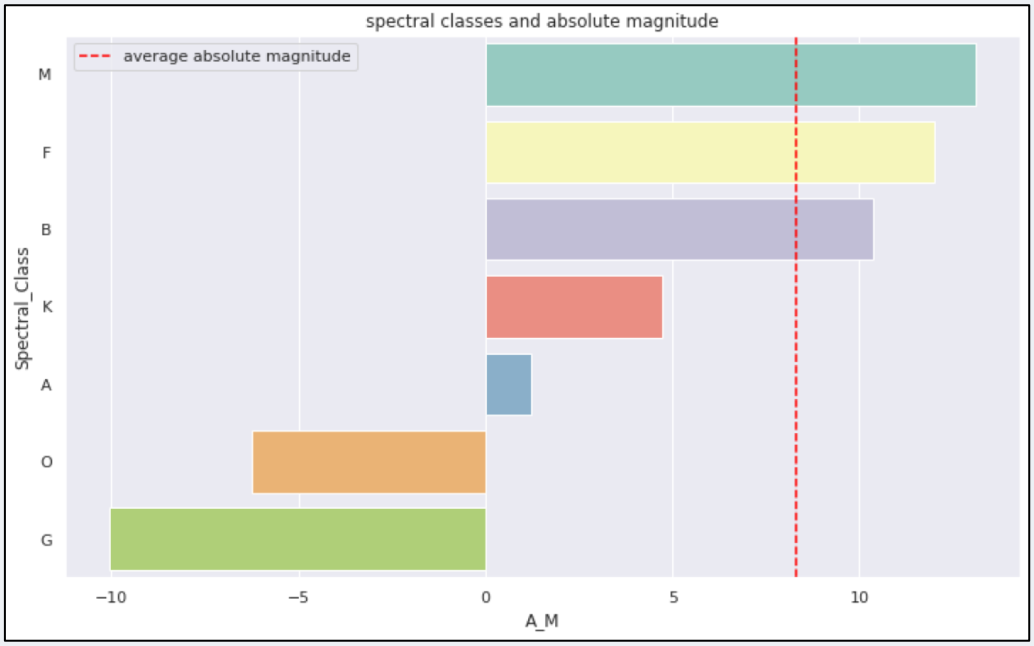
2. Count plots : we counted the number of stars according to their qualitative quantities including the spectral class and color.



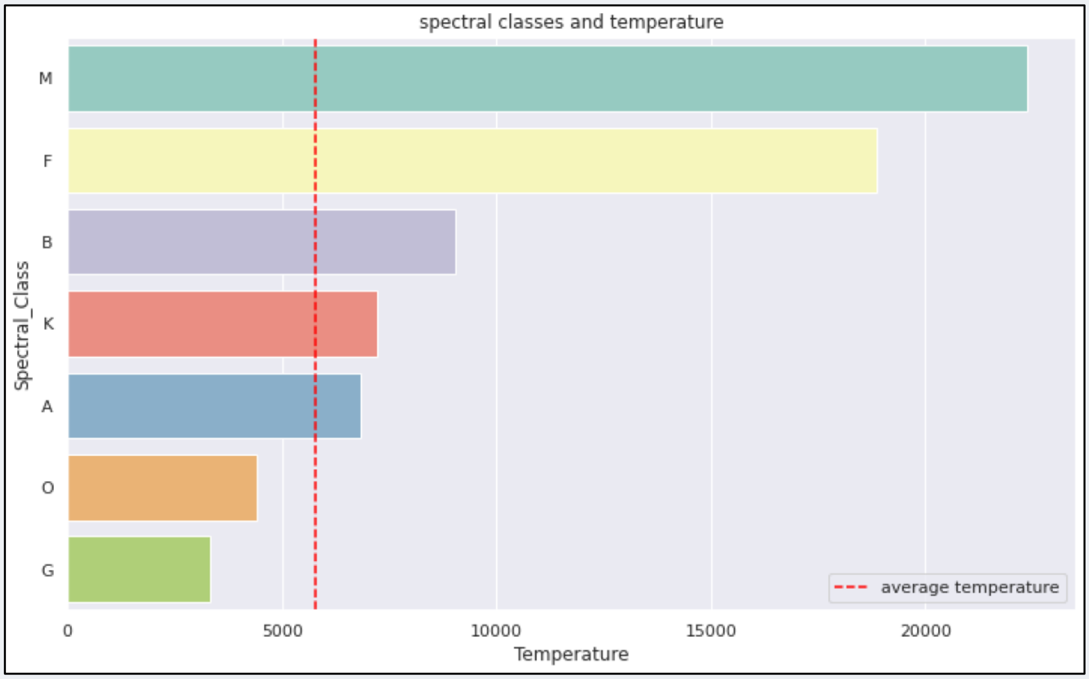
- According to the above plot, we figured out that the most of stars are red and Blue stars and Blue-White stars are at the second and the third place.



- Above plot shows that most of the stars are in the spectral class M and then B and O are the most common spectral classes.



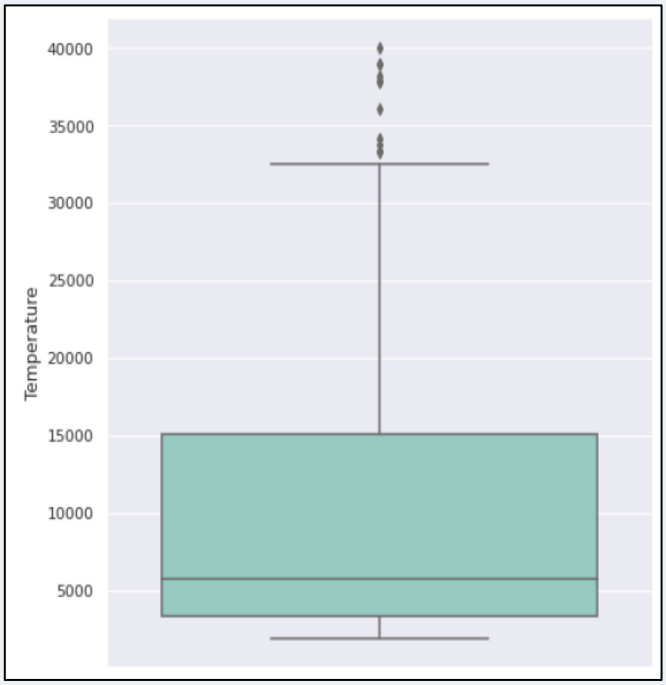
- Above plot shows the relation between absolute magnitude and the spectral class. Classes O and G have negative absolute magnitude and the highest absolute magnitude belongs to the class M.

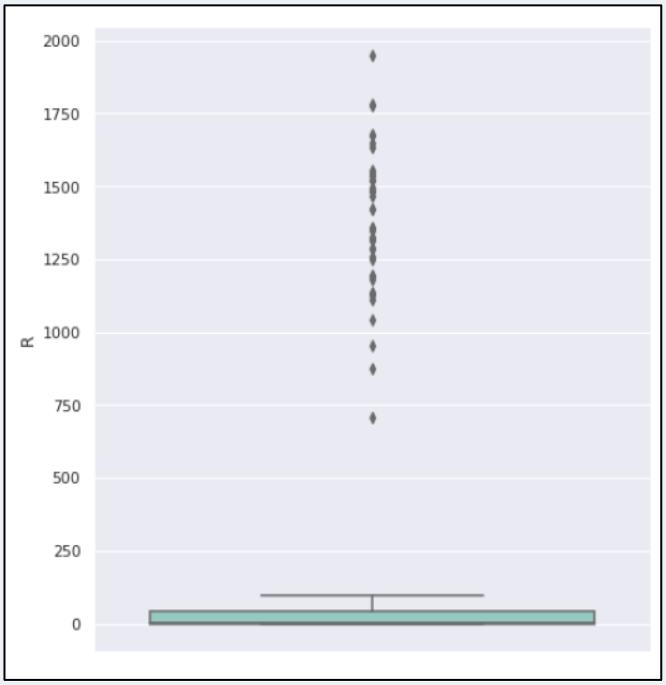


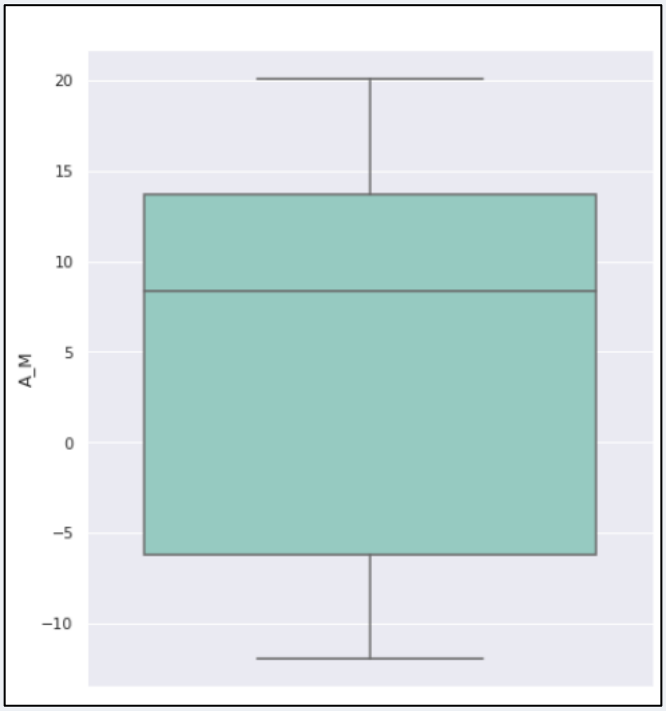
- The fourth plot shows that the spectral class M has the highest temperature among all of the classes.

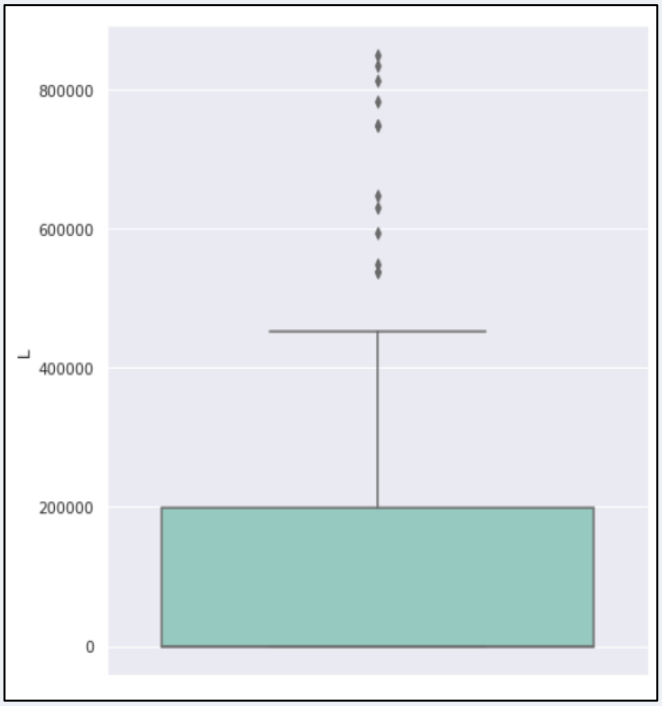
There are a lot of different plots in the project notebook that shows the relation between other quantities, but these four are the most important plots because they compare the spectral class with other features.

3. Density plots
   In order to have a better point of view, we decided to plot the pdf of quantitative features. Also, we wrote a function to calculate their statistical properties including the average, median, variance, standard deviation, skewness, kurtosis, range, and cv. The following plots are box plots of these features :

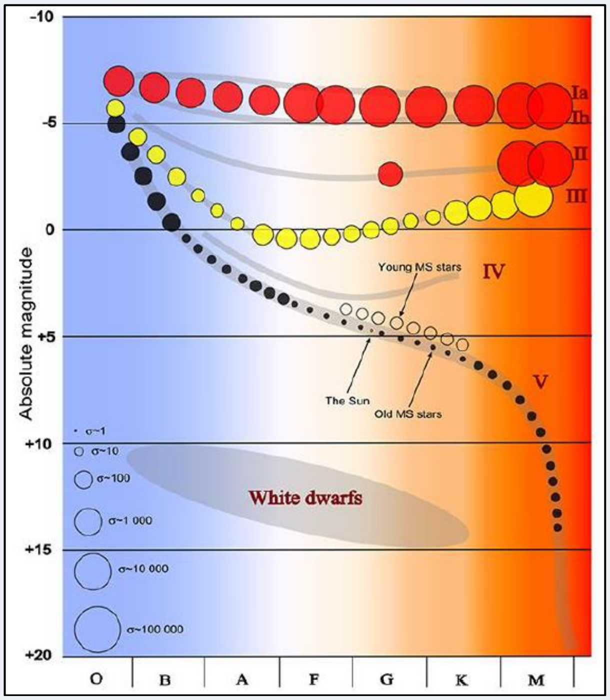

Temperature
40000
35000
30000
25000
20000
15000
10000
5000


R
2000
1750
1500
1250
1000
750
500
250
0

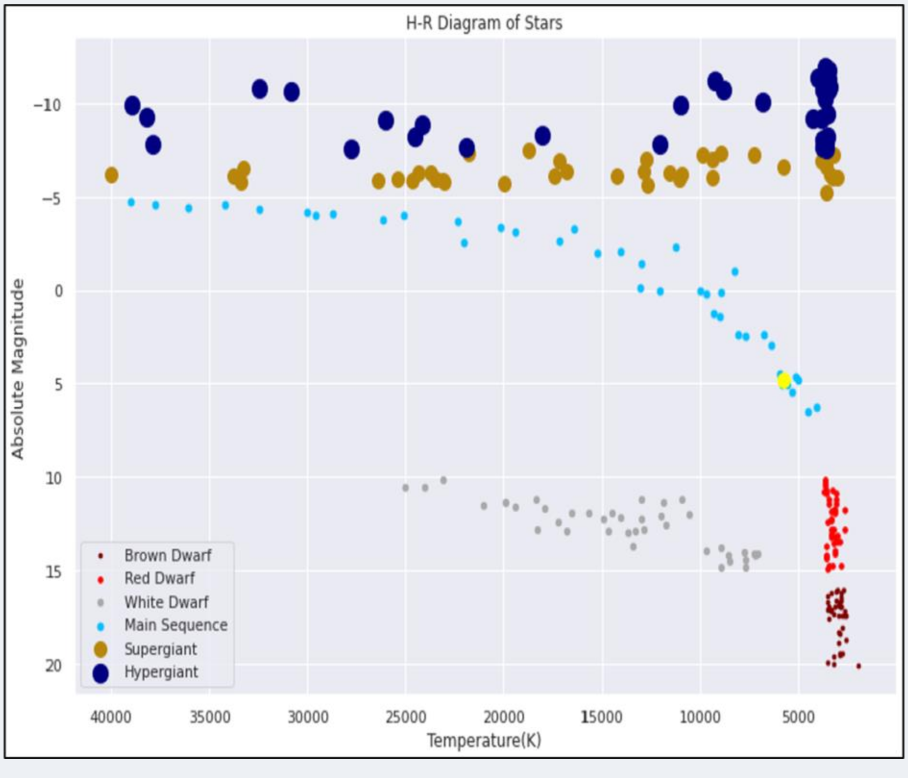

A_M
20
15
10
5
0
−5
−10


L
800000
600000
400000
200000
0

Part 3 : the H-R diagram
Here is the picture of H-R diagram from google :



Then we tried to plot this graph with accessible dataset and here is the result:

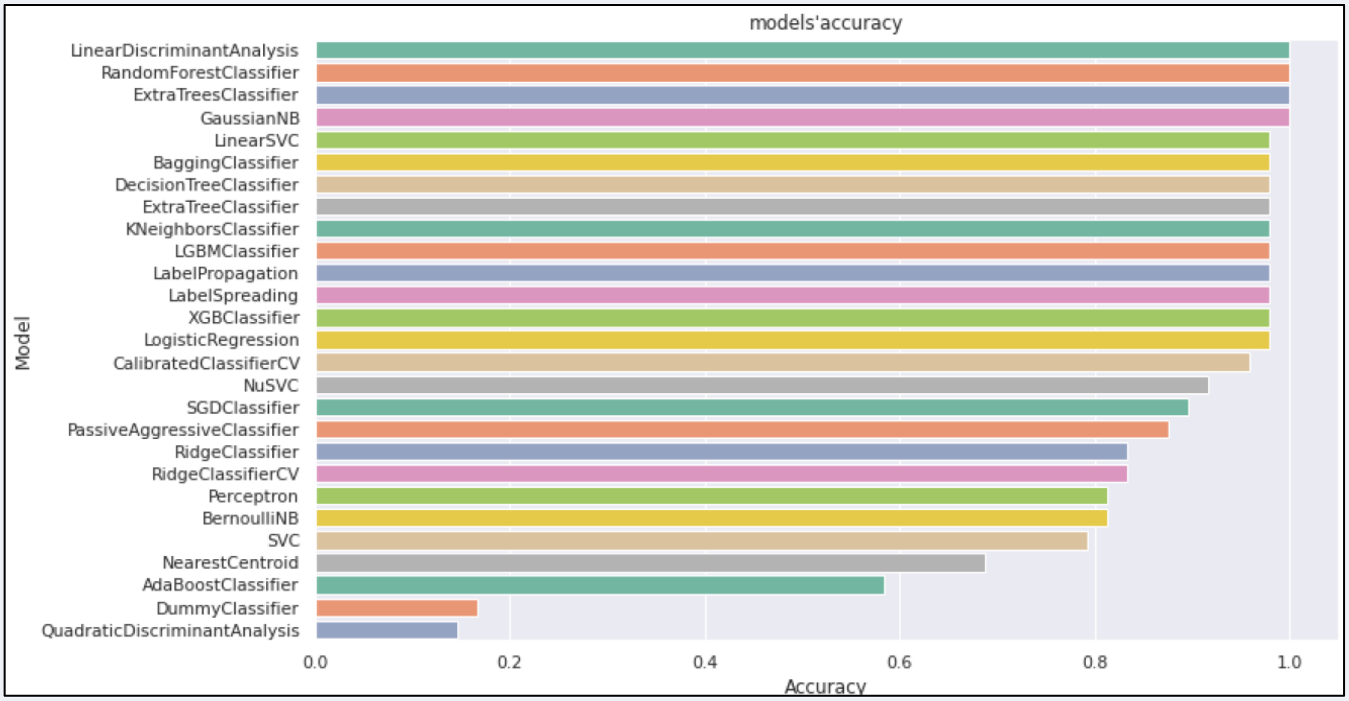


- It's clear that our dataset follow the H-R diagram so it's another proof for the accuracy of this model.

Part 4 : ML Tools

- In this part we tried to find the best machine learning method for training this dataset. 80% of dataset is for train and 20% for test.

We used lazypredict library to compare more than 20 different methods of ML on this dataset and here is the result:



According to the above plot, the most accurate methods for this dataset are Linear Discriminant Analysis, Random Forest Classifier, Extra Trees classifier, and Gaussian NB.