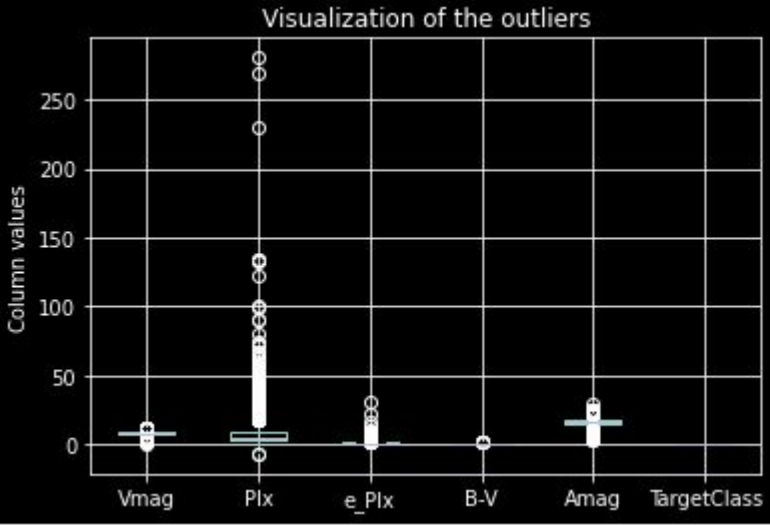
Machine Learning

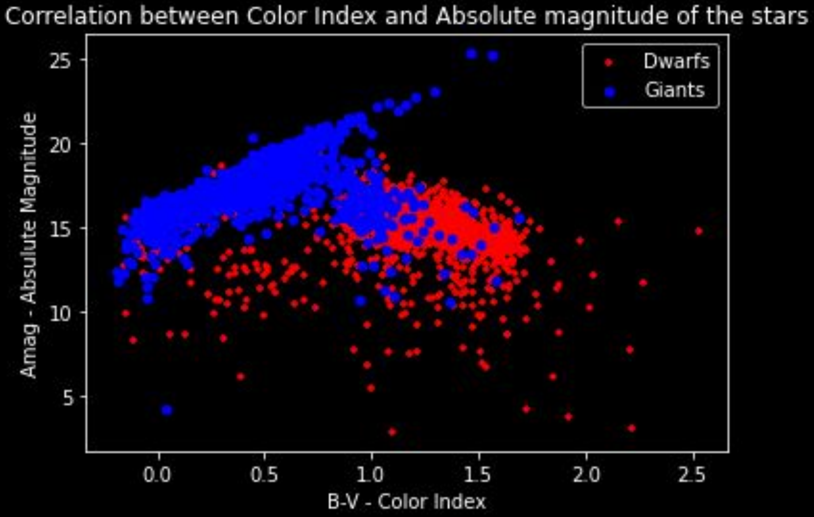
Identifying Giants and Dwarfs through Machine Learning

Dataset Features

Vmag	Visual Apparent Magnitude of the Star
Plx	Distance between the Star and the Earth
e_Plx	Standard error of Plx
B-V	Color Index
SpType	Spectral Type
Amag	Absolute Magnitude of the Star

Outlier Identification and Dataset Visualization





Feature Clipping

Removed samples that are 1.5 times greater than the standard deviation of "e_Plx", according to its values in that column

Feature Normalization

Using the Min-Max Normalization formula:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

3-way-split: Creation of Train, Dev and Test Sets

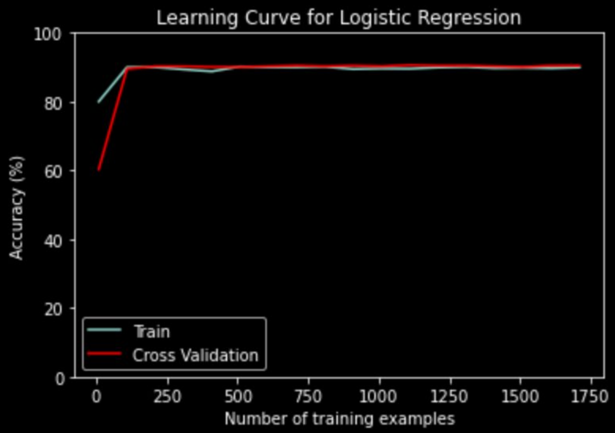
```
from sklearn.model_selection import train_test_split

x_train, x_cv, y_train , y_cv =
train_test_split(x_data,y_data, test_size = 0.20)

x_train, x_test, y_train , y_test =
train_test_split(x_train,y_train, test_size = 0.20)
```

Logistic Regression

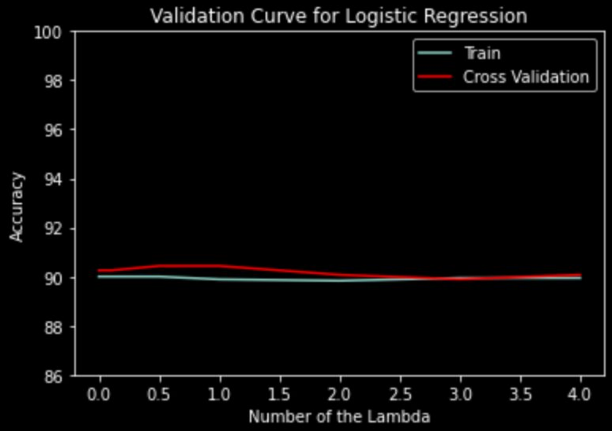
Cost Function and Gradient Descend with Ridge Regression

Accuracy, Confusion Matrix and F1-Score as performance mechanism evaluators

Learning Curve



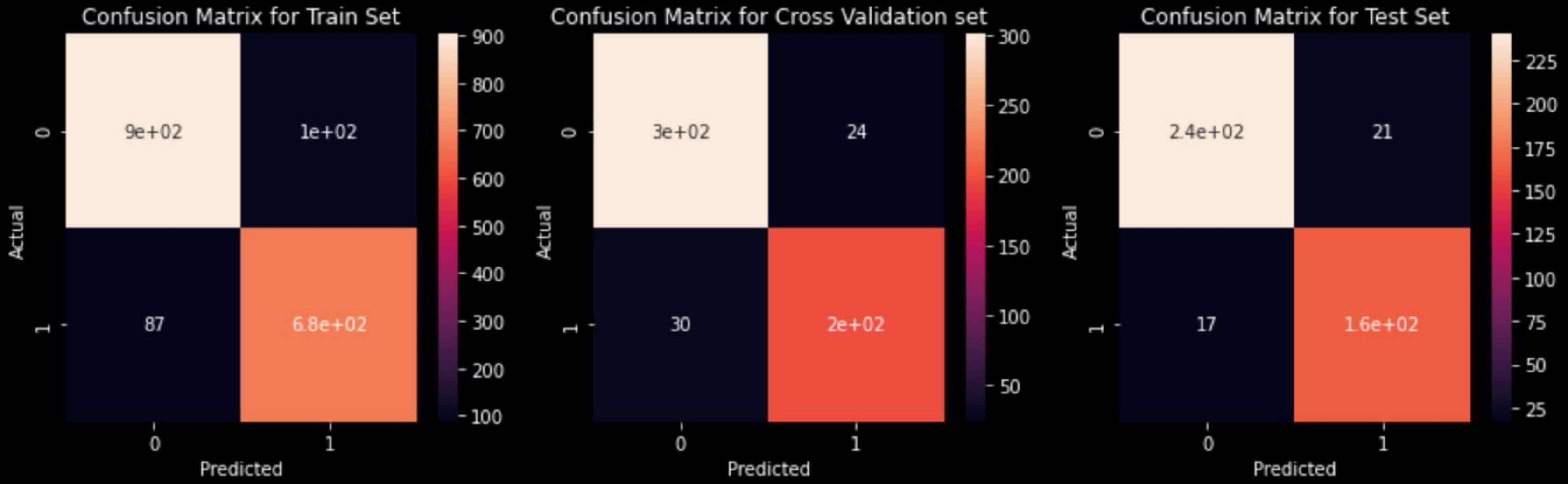
Validation Curve



Logistic Regression (Running Example)

```
Best Lambda:  0.5

Accuracy of Train Data:  90.0 %

Accuracy of Cross Validation:  90.433 %

Accuracy of Test Data:  87.585 %


F1-Score of Train Data: 88.632 %

F1-Score of Cross Validation Data: 88.3 %

F1-Score of Test Data: 84.68
```

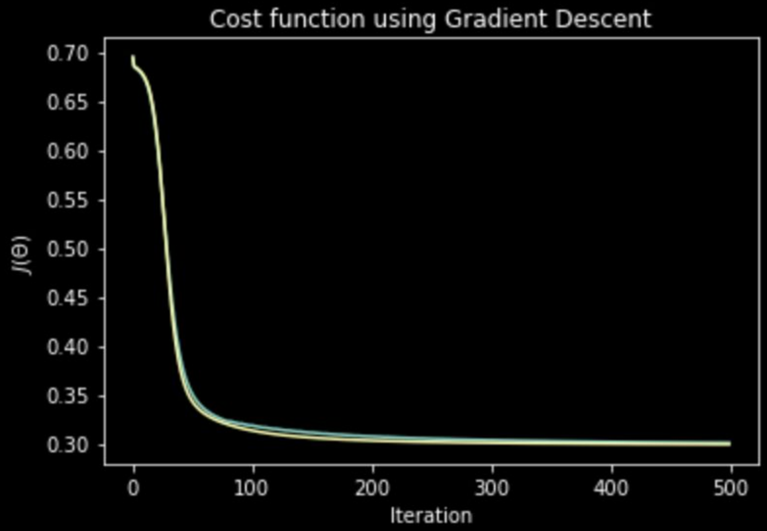
Logistic Regression (Confusion Matrix)



Neural Networks

NN Cost Function with BackPropagation

Gradient Descent with Adaptative Learning Rate and Momentum



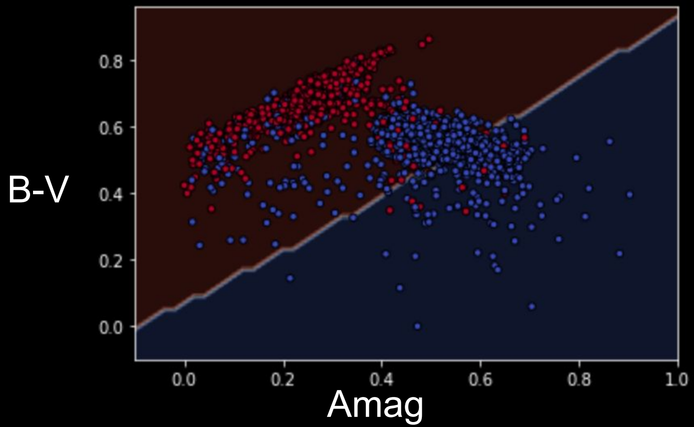
```
Lambda= 0.5
alpha= 2.1
momentum = 0.01

Lambda_alt = 0.5
alpha_alt = 1.2
momentum_alt = 0.5

Train Set Accuracy: 90.113 %
Test Set Accuracy: 88.262 %
```

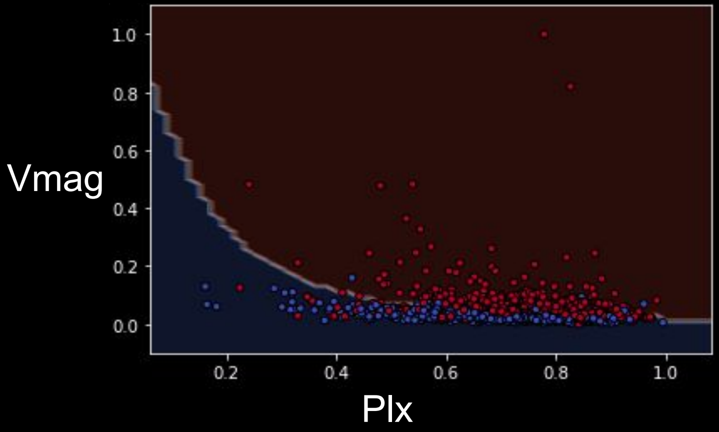
SVM

Training examples to points in space

Divides area in two spaces, when something is to be classified, appears in the area representing the classification

Best results:





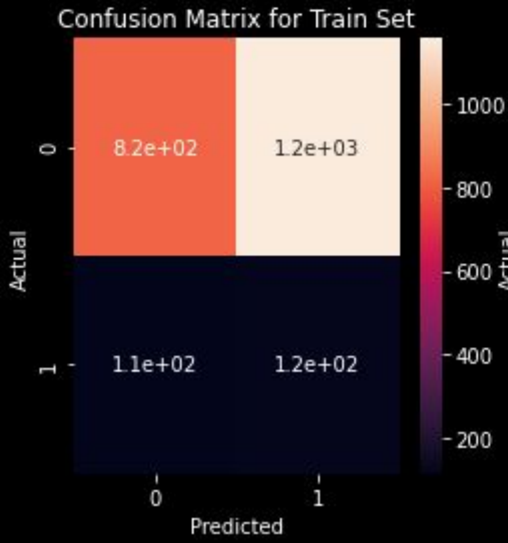
KFold

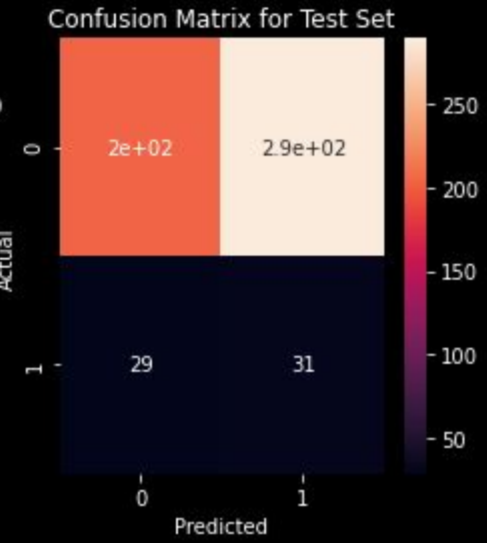
Used on SVM Cross Validation

5 splits

Average Balanced Accuracy

Average F1 Score


Confusion Matrix for Train Set


Confusion Matrix for Test Set

Conclusions

Logistic Regression, Neural Networks and SVM all have around 90% accuracy

Regarding this project, and considering the resources necessary to implement all the methods, the best option is the Logistic Regression, as it provides almost the same results consuming the least resources.

Authors

José Moreira Nº79671

Bruno Aguiar Nº80177