# BYTE PANACHE COMPETITION

## FOUNDATIONS OF MODERN MACHINE LEARNING

## OBJECTIVES

Each candidate is described by 8 continuous variables, and a single class variable. The first four are simple statistics obtained from the integrated pulse profile. The remaining four variables are similarly obtained from the DM-SNR curve . These are summarised below:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.

## METHODS

The following methods were required to complete the research:

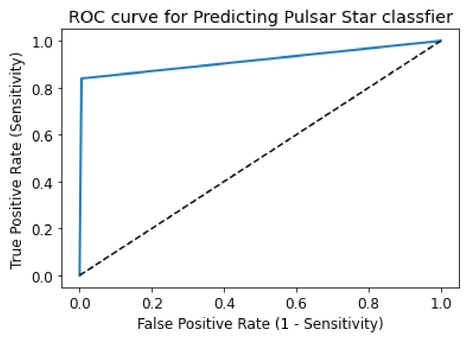
- Handle outliers with SVMs.
- SVM with default hyperparameters
- Compare model accuracy with null accuracy
- Confusion matrix
- Classification metrices
- ROC Curve
- Stratified k-fold Cross Validation with shuffle split.

## INTRODUCTION

PPulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Classification algorithms in particular are being adopted, which treat the data sets as binary classification problems. Here the legitimate pulsar examples form minority positive class and spurious examples form the majority negative class. The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. Each row lists the variables first, and the class label is the final entry. The class labels used are 0 (negative) and 1 (positive).

## RESULTS 2

ROC AUC is a single number summary of classifier performance. The higher the value, the better the classifier. ROC AUC of our model approaches towards 1. So, we can conclude that our classifier does a good job in classifying the pulsar star.




## RESULTS 1

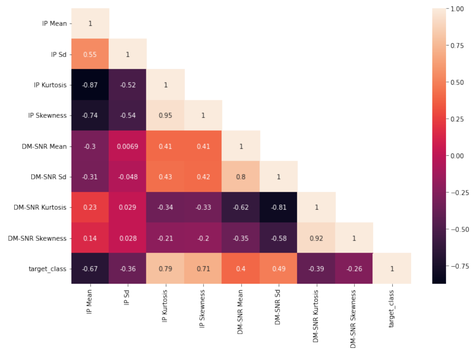



**Figure 1:** Heat map

As per our dataset , we can suspect that all the continuous variables may contain outliers.For better visualizing the outliers in the dataset we have used heatmap.

The boxplot here is drawn to visualise outliers in the dataset variables .The boxplot confirms that there are a lot of outliers in these variables.

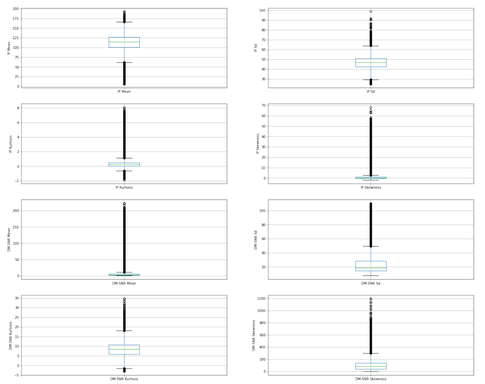


**Figure 2:** Figure caption

## CONCLUSION

- There are outliers in our dataset. So, as we increase the value of C to limit fewer outliers, the accuracy increased. This is true with different kinds of kernels.
- We get maximum accuracy with rbf and linear kernel with C=100.0 and the accuracy is 0.9832. So, we can conclude that our model is doing a very good job in terms of predicting the class labels.
- ROC AUC of our model is very close to 1. So, we can conclude that our classifier does a good job in classifying the pulsar star.
- We obtain higher average stratified k-fold cross-validation score of 0.9789 with linear kernel but the model accuracy is 0.9832. So, stratified cross-validation technique does not help to improve the model performance.

## REFERENCES


[1] Aurelien Gero. *Hands on Machine Learning with Scikit-Learn and Tensorflow*. O'Reilly Media, 2nd edition, 2013.

[2] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 1st edition, 2006.


## FUTURE RESEARCH

We can remove outliers, extend the model, increase the efficiency of models by using many ideas and methods to get the best model.

## CONTACT INFORMATION


**Web** www.iiit.ac.in

**Email** fmml20211022@ihub-data.iiit.ac.in

**Phone** +91-9054853880