# BYTE PANACHE COMPETITION; GROUP-S

FOUNDATIONS OF MODERN MACHINE LEARNING

## OBJECTIVES

Support Vector Machines (SVMs in short) are supervised machine learning algorithms that are used for classification and regression purposes. In this kernel, we are going to build a Support Vector Machines classifier to classify a Pulsar star. The dataset used for this project is 'Predicting a Pulsar Star'. We'll use:

1. Linear and Radial Basis Function kernel
2. Polynomial and sigmoid kernel
3. ROC-AUC performance evaluation
4. k-fold Cross validation

## MATERIALS & METHODS

The following methods were required to complete the research:

- Linear and Radial Basis Function kernel
- Polynomial and sigmoid kernel
- Receiver operating characteristic(ROC) curve
- Area under ROC curve
- k-fold Cross validation
- Stratified k-fold Shuffle split cross validation

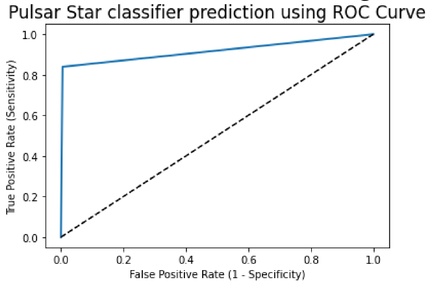
True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows: TPR = TP/(TP + FN)
False Positive Rate (FPR) is defined as follows: FPR = FP/(FP + TN)
AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). AUC ranges in value from 0 to 1. A model whose predictions are 100 percent wrong has an AUC of 0.0; one whose predictions are 100 percent correct has an AUC of 1.0

## INTRODUCTION

Pulsars 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. The class labels used are 0 (negative) and 1 (positive).

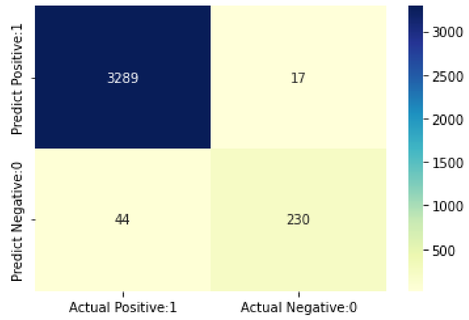
## RESULTS 2

**Figure 2:** ROC-AUC




The ROC-AUC Curve can be noticed to be approaching towards 1 (both along x axes). This gives us a fair idea as to how accurate the model is and how good of a job it has done at classifying the pulsar star data set. the ROC AUC curve gives an output accuracy of 0.9171 which is <1. The closer to 1, the better it is at classifying. An ROC curve is a graph showing the performance of a classification model at all thresholds. This curve plots two parameters, TPR vs. FPR at different classification thresholds. Lowering this classifies more items as positive, thus increasing FPs Tps.

## RESULTS 1




CONFUSION MATRIX:
we must explore alternative metrices that provide better guidance in selecting models. In particular, we would like to know the underlying distribution of values and the type of errors our classifer is making.
One such metric to analyze the model performance in imbalanced classes problem is Confusion matrix. This heatmap is plotted between, True positives, true negatives, False positives and False negatives
The result of the confusion matrix is:
[[3289 17]
[ 44 230]]

BOXPLOTS:
A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. On closer inspection, we can suspect that all the continuous variables may contain outliers.
Boxplot is used to visualise outliers in the variables of the given data.

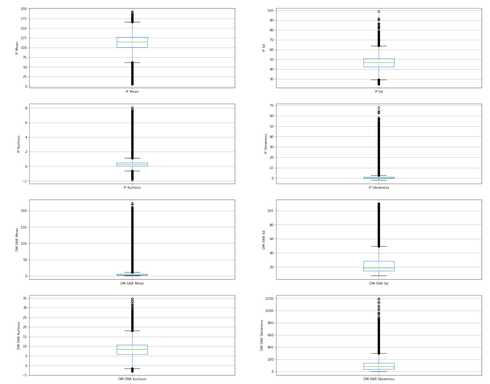


**Figure 1:** boxplot

## 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. In imbalanced datasets, Accuracy is an inadequate measure for quantifying predictive performance in the problem. So, we must explore confusion matrix that provide better guidance in selecting models.
- The ROC-AUC Curve can be noticed to be approaching towards 1 (both along x and y axes). This tells how good of a job it has done at classifying the pulsar star data set.
- We have obtained an average score of 0.9725 using the linear kernel and an average score of 0.978936 using RBF cross validation.

## REFERENCES

https://www.datacamp.com/community/tutorials/classification-scikit-learn-python
http://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/

## FUTURE RESEARCH

Some common applications of SVM are-

1) Face detection

2) Text and hypertext categorization

3) Classification of images

4) Bioinformatics – It includes protein classification and cancer classification.

## CONTACT INFORMATION

**Name** : SREEMAEE AKSHATHALA
**Email** : fmml20211004@ihub-data.iiit.ac.in