BYTE PANACHE COMPETITION

FOUNDATIONS OF MODERN MACHINE LEARNING

OBJECTIVES

The objectives are, the main objective is to find the best machine learning model to predict the pulsar stars.

1. Predict the pulsar stars using various models like SVM, Random Forest Classifier etc.
2. Check the metrics of each model.
3. Plotting the confusion matrix.
4. Checking if the model overfits or underfits the data

MATERIALS & METHODS

The following formulas are used to find the accuracy, precision, recall, f1 scores:

- Accuracy = TP+TN/(TP+FN+TN+FP)
- PRECISION = TP/(FP+TP)
- RECALL = TP/(FN+TP)
- F1 SCORE = (precision * Recall *2) / (precision +Recall)

Then plotting the values accordingly with respect to the classification gives how accurate the classifier works when outliers exist.
The following libraries are required to complete the research:

- Numpy (Linear algebra)
- Pandas (data processing)
- Matplotlib (Data Visualization)
- seaborn (Statistical data visualization)
- sklearn.metrics (classification report)
- sklearn.preprocessing (Standard Scalar)
- sklearn.svm and different SVC kernals(Linear,polynomial,Sigmod)

INTRODUCTION

We predict the value of the , *pulsar stars*. 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.

RESULTS 2

These are the F1 scores of all the models we've tested

Model Name	C value	F1 Score
SVM with rbf kernal	1.0	0.98
SVM with rbf kernal	100.0	0.98
SVM with rbf kernal	1000.0	0.98
SVM with linear kernel	1.0	0.98
SVM with linear kernel	100.0	0.98
SVM with linear kernal	1000.0	0.98
SVM with polyn kernel	1.0	0.98
SVM with polyn kernel	100.0	0.98
SVM with sigmoid kernel	1.0	0.89
SVM with sigmoid kernel	100.0	0.89
Decision Tree Classifier	-	0.97
Random Forest Classifier	-	0.98

Table 1: Accuracy of Models

RESULTS 1

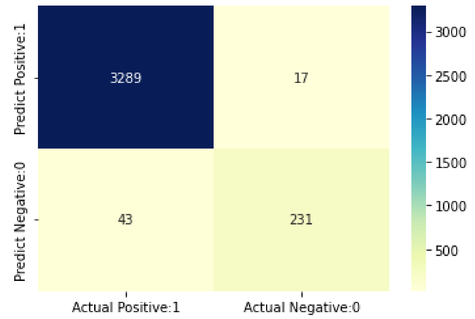

Figure 1: Confusion matrix for SVM

The Figure 1 displays the confusion matrix obtained from the SVM with a linear kernal when the Regularisation parameter (C) is 100. The confusion matrix is used to display the accuracy of prediction using an SVM with a linear filter. The F1-Score for this set-up is 0.98. The confusion matrix is similar to the confusion matrix of other models.

The Figure 2 displays the confusion matrix obtained from the Random Tree classifier, where the number of trees (n) is 200. The confusion matrix is used to to display the metrics of prediction using the random tree classifier. The F1-score for this setup is 0.98, we can notice it's performance similar to SVM models we've looked before.

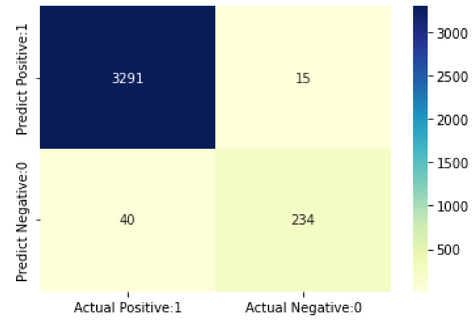

Figure 2: Confusion matrix for Random Forest Classifier

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.
- The accuracy of random forest classifier is similar to the SVM models using the rbf and linear kernal with C=100.0.

So, we can conclude that our model is doing a very good job in terms of predicting the class labels. But, this is not true. Here, we have an imbalanced dataset. Accuracy is an inadequate measure for quantifying predictive performance in the imbalanced dataset problem. So, we must explore confusion matrix that provide better guidance in selecting models.

REFERENCES

[1] Scikit-learn svm tutorial with python (support vector machines).

[2] Kernel functions-introduction to svm kernel and examples, Mar 2021.

[3] Support-vector machine, Mar 2022.

FUTURE RESEARCH

There are various different models available for machine learning. To predict the pulsar stars we can use different approches like *Artificial Neural Networks* and *Convelutional Neural Networks*.

The datasets are imbalanced, to resolve the data imbalance problem and overcome model overfitting we can try using *Data Resampling* or *Synthetic Minority Oversampling Technique* (SMOTE)

CONTACT INFORMATION

Name	Email
Nikhil Chagi	nikhilchagi69@gmail.com
Niranjan M	niranjan.m2199@gmail.com
Vikas Reddy	vikasvr2021@gmail.com