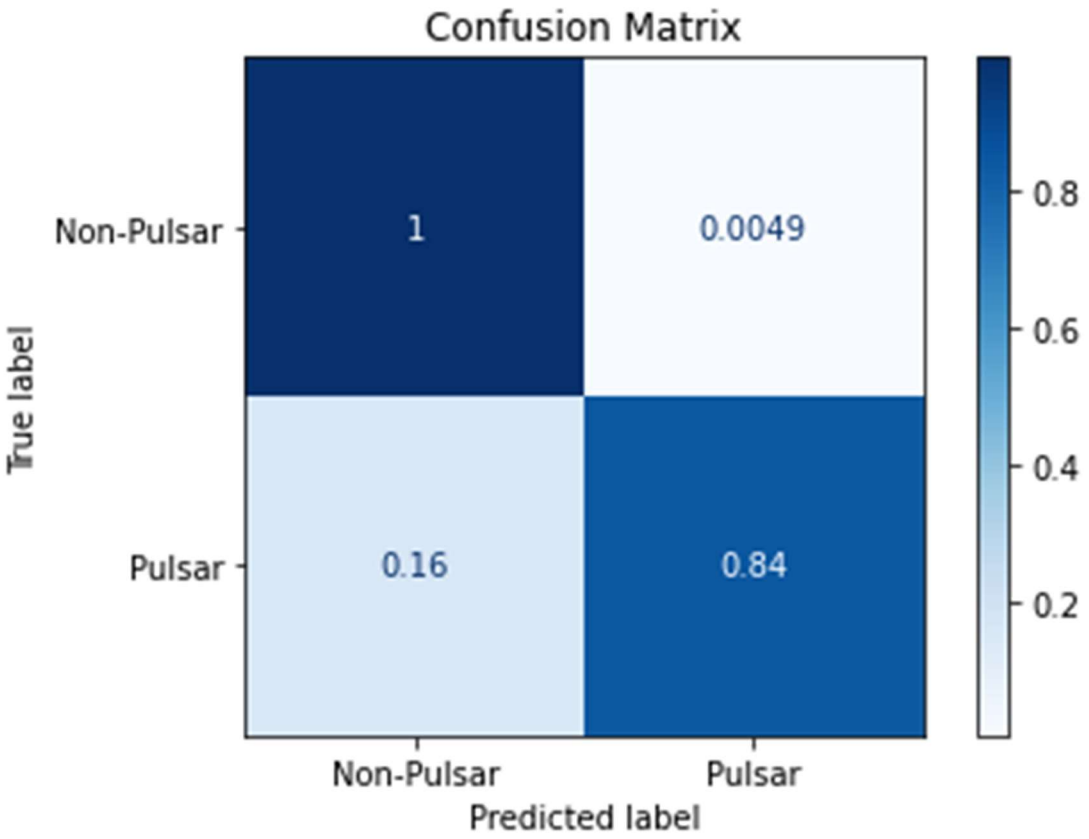
Fadi Younes 261001877

MAIS 202

# Progress and Preliminary results for Prediction Model of Pulsar Stars

The goal of this project is to predict whether or not a star is a pulsar using a random forest classifier. This method is used since it averages on many different decision trees that are trained on different parts of the training set, thus reducing the variance of the predictions and avoiding overfitting. The data set used for the prediction of pulsars is still the HTRU2 data set from the Center for Machine Learning and Intelligent Systems of UC Irvine. No changes were made to the initial data set. The data set contains data on 17 898 stars with 8 data points each and classification label on whether or not the star is a pulsar. The 8 data points for each star are the mean of the integrated profile, the standard deviation of the integrated profile, the excess kurtosis of the integrated profile, the skewness of the integrated profile, the mean of the DM-SNR curve, the standard deviation of the DM-SNR curve, the excess kurtosis of the DM-SNR curve, and the skewness of the DM-SNR curve. The classification label is 0 if the star is not a pulsar and 1 if the star is a pulsar. This is all contained in a single CSV file. Each row has the data on a single star, the first 8 columns are the data points and the 9th column is the classification label. The pulsars are already scattered randomly throughout the data set but the data is rambled another time before being used just to be sure. No data was deleted or modified and all 17 898 stars were used. To implement the random forest classifier, the *Scikit-learn* library was used as well as the usual *Numpy*, *Pandas* and *Pyplot* libraries. 1000 trees were used in the forest since this number would get a higher accuracy than smaller numbers of trees such as 50 or 100 that are closer to the usual number of trees used for this application. The first ¾ of the 17 898 stars are used as training data with the last quarter used as testing data. More stars are used in training than in testing to make sure the model is well trained and to minimise the error. The model was tested by passing the data about the remaining ¼ of the 17 898 stars through the random forest classifier and comparing the predictions about the classification of each star to the real class of each star. The model is very slightly underfitting since the accuracy of the model on the testing data is of 98% which is really good but still not perfect. No major challenges were encountered when implementing the model. Only the best number of trees to use was some what challenging to determine but it was finally determined by trial and error. The metrics testing performance of the model are a precision of 93.8%, an accuracy of 98.3% and a recall of 84.2%. The confusion matrix relative to the number of stars for each class is also given. The model is thus very good at telling when a star is not a pulsar: only 0.5% of false positives. However, the model is not as good at recognising pulsars: 16% of false negatives. This greater difficulty at recognising pulsars compared to the ease with which the model can recognise non-pulsars is probably due to the small number of pulsars compared to the overall number of stars in the data set: only 9% of stars in the data set are classified as pulsars. A bigger data set could help the model be better trained at recognising pulsars. The project is still on the path to success since the overall performances of the model are acceptable. The next step will be to fine-tune the model by systematically testing the model with a number of trees ranging from 10 to 10 000. The number of trees giving the best performance will be kept.




Confusion matrix: On the upper-left corner: gives the percentage of non-pulsars that were predicted to be non-pulsars (true negatives); On the upper-right corner: gives the percentage of non-pulsars that were predicted to be pulsars (false positives); On the lower-left corner: gives the percentage of pulsars that were predicted to be non-pulsars (false negatives); On the lower-right corner: gives the percentage of pulsars that were predicted to be pulsars (true positive).