Fadi Younes 261001877

MAIS 202

Data Selection Proposal for Prediction Model of Pulsar Stars

The data set used for the prediction of pulsars is the HTRU2 data set from the Center for Machine Learning and Intelligent Systems of UC Irvine [1]. This data set was the recommended data set from the list of pre-approved projects list for deliverables in MAIS 202. This 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. This is all contained in a single CSV file which is easy to access and work with *Python* and the *Pandas* library. Each row has the data on a single star and the first 8 columns are the data points for each star and the 9th column has the information on whether or not the star is a pulsar. The pulsars are already scattered randomly throughout the data set. The first ¾ of the stars will be used as training data with the last quarter used as testing data. The goal is to predict whether or not a star is a pulsar using a random forest classifier. This method will be used since it averages on many different decision trees that were trained on different parts of the data set, thus reducing the variance of the predictions and avoiding overfitting [2]. This method has been shown to work best for such a problem of classifying stars compared to a logistic regression model, for example, or support vector machine which tends to over fit the training data [3]. After developing the model, the confusion matrix that reports for each two labels how many were predicted correctly and how many were predicted wrongly will be stated as will also be stated the precision and the recall of the model which are the fraction of correct pulsar predictions on all pulsar predictions and the fraction of pulsars on all, correct and wrong, pulsar predictions [4]. Finally, the model will be integrated in a simple landing-page web app where the user will input 8 numbers in pre-assigned slots to variables pertaining to data on stars and the user will get as an output a prediction in text on whether or not this star is a pulsar. This simple web app will be accompanied with scientific information on pulsars and brief explanations of the random forest classifier model used.

References

[1] HTRU2 Data Set, https://archive.ics.uci.edu/ml/datasets/HTRU2#, (Feb. 2022)

[2] Random forest, Wikipedia, https://en.wikipedia.org/wiki/Random_forest#Algorithm, (Feb. 2022)

[3] Predicting a Pulsar Star using different Machine Learning Algorithms, https://satyam5120.medium.com/predicting-a-pulsar-star-using-different-machine-learning-algorithms-d22ee8fc71b4, (Feb. 2022)

[4] Understanding Classification, CSC2019 - Introduction to Machine Learning, https://as595.github.io/classification/, (Feb. 2022)