Deliverable 3

Project Idea:

I would like to create a web app that the astronomers can use to predict the existence of a pulsar star given the 8 features. They could use the prediction to further investigate and confirm the discovery of the star.

Link to the initial testing (code, also available on my github page):
https://colab.research.google.com/drive/1x6hevsBIp60kVpYxgwDZx2ZVssrSCaiQ

Link to the final model (code, also available on my github page):
https://colab.research.google.com/drive/1PzRWb3YcGX5XWr_Nf927OuFSqp9O54iI

Dataset:

Dataset: UCI Machine Learning Repo - https://archive.ics.uci.edu/ml/datasets/HTRU2 [1]
Reasons for choosing this:

- It contains 17,898 instances and 1,639 are real pulsar examples. They have been checked by human annotators.
- The data contains 9 variables (8 features and 1 label).
- The variables are:
    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
    9. Class
- The values had to be converted to their respected data types (floats and integers) before inputting them to the model (old approach, used in the initial testing)
- I searched for missing values and tried to fix them using sklearn's impute module
- I have since changed to pandas (in the final model) to deal with the data as it is a better choice

[1] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656

[1] R. J. Lyon, HTRU2, DOI: 10.6084/m9.figshare.3080389.v1.

Final Training Results:

- I trained models such as Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM) and Decision Tree Classifier (DTC)
- After validating the data and tuning the hyperparameters (using grid search), I have concluded that SVM has the best results. The results were compared by looking at the accuracy scores and the confusion matrix
- Therefore, I have decided to use SVM as the model for this project

Final Demonstration Proposal

- I have saved the trained model using the 'pickle' module on Python
- I plan on informing the viewers on the basic physics behind pulsar stars
- I plan to build a basic HTML form to take the features as input and return the prediction from my trained model
- I will use the Flask framework to handle the HTTP requests and probably Google Cloud to host a website (Getting a domain is under consideration as well)

[1] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656

[1] R. J. Lyon, HTRU2, DOI: 10.6084/m9.figshare.3080389.v1.