CS 4320: Machine Learning

Assignment: Hyper Parameter Search

Use the <u>March 2021 Playground Series</u> data set at Kaggle. Use hyper parameter search with cross validation to create a decision tree classification model (such as `sklearn.tree.DecisionTreeClassifier`), to obtain the best `F1` score possible.

It is expected that you will use the Titantic hyper parameter search with cross validation decision tree source code as a starting point for your code development.

Create a report that includes the data exploration plots and analysis, which hyper parameters were used in the search, the range or set of values used for each hyper parameter, the hyper parameters selected, the number of cross validation sets, the `F1` cross-validation score obtained, the training `F1` score of the model when trained on all training data, and finally, the `F1` score of the model on the testing data.

Include a comparison of the three `F1` scores, interpret the meaning of these comparisons.

Required Steps

- Download your data.
- Explore and analyze your data.
- Split the data 80%/20%, for training/testing.
- Write (or modify) a Python program using sklearn to process and fit the training data to decision tree models with a hyper parameter search and cross validation.
- *AFTER* finding your best fit model, measure the model's `F1` score on your test data.
- Create a report with the contents mentioned above.
- Commit and push your code in the git repository.
- Submit the report (as PDF) to Canvas.