Group 2: Carson McBroom, Morgan Starkey, Victor Rodriguez-Ortiz

ChE 4230

Final Project

Preliminary Report

For management of this project, we plan to use a github repository to maintain a consistent version of our code and files. This will also allow for us to use branching to make changes to our code without breaking or overwriting our main branch. All files in the github repository will be sorted into relevant folders to keep file management clean and make everything as easily accessible and locatable as possible. The folder hierarchy will be broken down into each part of the assignment having a folder with the files pertaining to that section of the project being placed within those folders. Actual editing of the code will be done using a combination of VS Code editor and jupyter notebook. VS Code allows easy editing and creation from our local machines while jupyter notebooks give the benefit of being able to run one section of code extremely easily without having to use computational power to run the rest of the code.

In order to do the machine learning section, a combination of both supervised and unsupervised learning will be used. The data will be preprocessed using standardization to bring all the data to the same scale and then normalized so that magnitude differences are clearer. We plan to use 60 percent of the data as training data with the remaining 40 percent being split evenly between validation data and test data. The lower percentage of training data was chosen in order to avoid overfitting where our model scores highly on training data but poorly on test data.

For the supervised learning section, entropy and gain of information calculations will be utilized to best determine the defining criteria. The calculations will be run two extra steps down the decision tree in each case in order to avoid a greedy approach and try to find the solution that yields the highest information gain of all. This will all be done within an artificial neural network that will be used to complete the decision tree and also to build a fault classification model.

For the unsupervised section, data mining will be utilized in order for our model to learn and predict future outcomes. We also plan to use dimensionality reduction and clustering to find key patterns within the given data. Dimensionality reduction will be used to find patterns in the upper dimensions of the data and then clustering will be applied to group samples that show some similarity. Other than that, we are unsure of what will be done in this section as we have not learned it yet.

The tentative internal due dates we have set for each section will be March 11 for the supervised learning section, and we will have the unsupervised learning section complete the week of the final due date. The plan is for each of us to work evenly on every deliverable, but, because of outside commitments, we understand that that isn’t the most realistic of goals. Despite this, we all do plan to put maximum effort in to splitting work evenly and not leaving it all to one or two people.