Outline:

Our objective for this class project will be to analyze mortgage and debt data as it pertains to select demographics. These demographics include, but are not limited to age, location, income and education. It is our goal to determine how an individual is affected by debt given their classification and uncover any disparities between groups. After performing our analysis we will be able to conclude if and how wealth inequality is distributed throughout our population. The dataset we have chosen to run our algorithms on can be found on Kaggle.com and is provided by the Golden Oak Research group. It includes 40,000 records and 80 features, because of this we will be conducting feature selection to determine the most statistically significant statistics. Some of the features that we have perceived to be useful in this dataset include City, State, Population, High School Degree, Mean age, Mean Rent, Mean Income, Mean Mortgage, Home Equity and Debt to name a few. Looking at these features we can clearly see that they include summary statistics, which also include mean, median, standard deviation and sample size. The dataset will be cleaned to eliminate outliers, normalization will be performed on the features to map them in a lower dimensionality, generally between 0 and 1. K nearest neighbour algorithm will be used to classify the features and extract patterns from them. The backend of our analysis will be coded in Pycharm and PyQt will be used for generating the GUI. In order to gain a sufficient grasp of the material in this dataset, we will use its data dictionary, as well as online resources that describe the financial statistics included. Once finished with our analysis we can use classification rate as a performance measure.

Schedule:

We plan on meeting at a minimum of once a week throughout the lifespan of our project with additional collaboration opportunities through Github. Initially, our work will consist of preprocessing, as well as data discovery, feature selection and correlation analysis. Once completed we will then apply models to our dataset within the following weeks so that we can have enough time to train and for trial-error (depending on our results). After that, we will be taking the subsequent week to work on our group report, GUI and put together a coherent narrative of our analysis. Finally, we will dedicate the last week for last-minute changes and the preparation of the slides for class.

TASKS	DAYS NEEDED	DESCRIPTION OF TASK
Data Preprocessing	10 days	Reduce Dimensionality, Normalize,eliminate outliers, eliminate missing values if any.
Finding the strategically significant variables	5 days	Gini Index, Entropy
KNN	5 days	Fit the model using the strategically significant variables
Testing of model	2 days	Calculating the accuracy of the model
GUI-coding and linking	10 days	Coding of front end and linking it to back end
Testing of GUI	2 days	Ensuring that the GUI displays the outputs accurately
Report	2 days	
PowerPoint	2 days	