**Approach to Work**

1. **Problem Statement**

* **Written by Javaid and Vas**

If you could focus on using decision trees and clustering algorithms to determine group of data which are likely to default or be at risk to default on a loan

1. **Techniques Used for Cleansing:**We found out the missing value and the duplicates from the data and removed it. Moreover, we also found out the unique values in the data which might be non-related so we removed that as well. The techniques used were pretty simple and normal functions are used.
2. **Data Visualization**For Visualization we used matplotlib and seaborn libraries. Matplotlib is a versatile library for creating a wide range of static plots, while Seaborn provides a higher-level interface with enhanced aesthetics and easier integration with statistical plots.
3. **Data Algorithms**

In this task we used Random forest, Logistic regression and XGBoost.

Xg boost is used because it provides high predictive accuracy, handles missing data, reduces overfitting through regularization, is fast due to its parallelization, and performs well on large, complex datasets

Random Forest improve classification and regression accuracy by combining multiple decision trees to reduce overfitting and capture complex patterns in the data.

Logistic Regression is used for binary classification tasks, modeling the probability of a class label using a logistic function to predict outcomes between 0 and 1.

* **Decision-making assumptions**
* This project was done to find out the potential credit risk given their demographics, credit history, loan information.
* EDA showed heavily skewed data for age, income, employment length, so used the log transformation on it.
* Also found out class imbalance ratio was 27.7% which represents the lower class pretty well so didn’t make much changes to it.
* After running 3 classification models : Logistic Regression, Random Forest Classifier and XG Boost; we concluded that the best model was XG Boost with ROC AUC Score of 80.42
* After performing hyperparameter tuning on XG Boost, the best parameters showed an ROC AUC score of 80.73, which is not much of an improvement.
* Loan Grade has the highest affect on the credit risk. Following closely are home ownership and loan percentage of income, suggesting that these factors influence the risk of default on the loan heavily.