**FINAL PROJECT PLAN**

**TITLE:**

**HEART RISK ANALYSIS USING BIG DATA AND MACHINE LEARNING TECHNIQUES**

**NAMES AND ROLES:**

**Sai Prudhvi Charan Pothumsetty - 16343752**

* Identify and obtain the heart disease dataset.
* Verify data for completeness and consistency.
* Preprocess the dataset according to the random forest classifier.
* Splitting the dataset into training and testing datasets

**Manchireddy Kavya Reddy - 16342232**

* Perform feature transformation using Vector Assembler.
* Compare and select the best-performing feature extraction method.

**Sai Deeksha Enukonda - 16343736**

* Train the machine learning model using the training dataset.
* Evaluate the models using various evaluation metrics such as accuracy, precision, recall, and F-1 score.
* Select the best-performing model based on its evaluation metrics.
* Use the selected model for random forest analysis on new data.
* Show the visualizations using matplotlib and seaborn or by using power BI.

**MOTIVATION OR PURPOSE:**

The prediction of heart attack is a critical issue in the field of healthcare. The use of machine learning techniques has been proposed as a means to accurately predict the occurrence of heart attacks. In this study, we developed a machine learning model for heart attack prediction using a dataset containing various risk factors. The dataset was preprocessed according to Random forest classifier. We then trained different machine learning algorithms including random forest to predict heart attack. The performance of each algorithm was evaluated using metrics such as accuracy. This study demonstrates the potential of machine learning techniques in predicting heart attacks and could help healthcare providers to identify individuals at risk of heart attack and take preventive measures.

The implementation was done using PySpark in Google Colab, allowing us to perform Heart Disease prediction on the entire heart disease dataset in a reasonable amount of time.

This project provides a useful tool for heart disease prediction on heart disease data using PySpark and machine learning algorithms. It demonstrates the effectiveness of PySpark for processing large datasets and extracting insights from social media data.

**UTILIZED CLOUD TECHNOLOGY:**

1. Cloud Storage-(Azure): Used to store the Crime data set