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

FINAL PROJECT - INTERMEDIATE PROGRESS REPORT

cloud technologies/services used:

Google Cloud Platform for storing the heart disease dataset and deploying the application in GCP.

Team information: members, how the collaboration has been done (if applicable):

The team collaborated by assigning roles and responsibilities based on individual strengths and completing assigned tasks on a daily basis. This ensured that everyone was working efficiently and effectively towards the project objectives.

Things that have been tried/accomplished:

- Converted the dataset which is in xl format to csv using pandas.
- Preprocessed the dataset according to Random Forest Classifier, and splitting it into training and testing datasets.
- Implemented feature transformer namely Vector Assembler for feature transformation.
- Evaluating the model using various evaluation metrics such as accuracy to determine their performance on the testing dataset.
- Demonstrated the effectiveness of PySpark for performing heart attack prediction on heart disease dataset, making it a valuable tool for healthcare industries to improve their services by better understanding the serious causes of heart diseases in their patients.

Things to do:

Need to write queries, show visualizations and deploy the project in cloud**.**

Challenges and/or comments:

Data Constraints: The data related to heart disease has certain limitations in terms of its size and the types of features included, unlike data collected from hospitals or research institutes which may have more diverse and extensive data.

Accuracy: Although the machine learning algorithms utilized in the project are established and commonly used, there is no assurance that they will yield precise predictions for all feature types. As a result, the reliability of the prediction analysis findings could be limited..

Time Constraints: Even though data models consisting of various features that capture a comprehensive health profile of individuals could provide more detailed information, obtaining access to such data is often a challenging and time-consuming task due to privacy concerns.

Resources: The project might necessitate substantial computational resources, such as high-performance computing clusters, to effectively analyze and process the publicly accessible dataset. If such resources are not readily available, this could pose a limitation.