PROJECT PROGRESS REPORT

Team Members:

1. Mohamed Ndiaye
2. Bill Muchero
3. Raghavendra Naidu Nayanuri
4. Bhavya Shri Meda

A. Progress so far

1. Completed Tasks

Data upload: After submitting our project proposal, we noticed that the specified dataset had already been cleaned (BRFSS2015). We acquired the original file from the CDC survey and performed the data preprocessing stage to master and better understand the techniques. The file we loaded had around 400,000 records and 330 features. To stay consistent with our study, we worked on getting the 22 features presented on BRFSS2015.

Data Preprocessing: The variables we retained for the study are the following: Diabetes_012, HighChol, Smoker, MentHlth, Education, Income

Here are the steps performed to prepare the data for the study:

1. Sectioning the data: By sectioning, we meant selecting features aligned with the subject we took. We also renamed the columns to make them more understandable.
2. Outliers of the target variable: We started by looking at the outliers on the target variable Diabetes_012. This variable should drive the focus on the data we are cleaning. We eliminated all the rows that could not provide a significant answer for the sake of the study.
3. Outliers of the other features: The process of removing the outliers is performed on the other features.
4. Discretization: This method reshaped the data into buckets or categories aligned with the BRFSS2015.

2. Link to the dataset

The data acquired was used to understand better the relationship between lifestyle and diabetes in the US. We have 21 features collected, and the target variable is whether a patient has diabetes, is pre-diabetic or is healthy (total of 22 columns). The link below presents the summary and characteristics of the dataset:

https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

The dataset we will use is accessible at the link below for downloading:

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

B. Challenges

We have encountered some challenges so far, such as:

1. Uploading the data: We had a hard time uploading the data on Colab because of the size (~500MB). To solve the issue, we uploaded the file to Drive and mounted it on Colab.
2. Exploring the data using visualizations: we used box plots on all the features simultaneously when they did not have the same scale of measurements, making it impossible to interpret. We had to run the box plots for each feature, and that's how we identified outliers.
3. Too many outliers: We discovered a significant portion of rows with answers that can't be used in the study. For example, Do you have diabetes? Answer: I'm not sure, or I refuse. These types of answers are irrelevant to our data.
4. Differences in background: We come from different backgrounds. Participating in the effort was difficult, but we managed to help out in any possible way, primarily by looking up tutorials and doing pre-research to streamline the process.

C. Collaboration

1. We have met once every 4-5 days to touch base, discuss our progress and individual challenges, and modify our plans accordingly.
2. Each member of our group contributes uniquely to the progress of our project. We have different backgrounds, i.e., Business, software, Medicine, etc., and we have divided our tasks in a way that leverages our strengths and encourages the rest of the group to learn and familiarize themselves with new skills and knowledge.

D. Next steps

1. Complete the data preprocessing: We need to run some descriptive statistics and more visualizations.
2. Run the algorithms: We already found tutorials for executing our data mining algorithms in Python and R. The plan is to divide and conquer; half of the team will work on running the algorithms in R, and the other half will run them in Python.
3. Plan: We will run the algorithms from June 3 to June 7, gathering and exploring the findings. On the week of June 10 to June 14, we will be working on the final report and submit it on time.
4. Potential challenges: There might be some challenges as the algorithms use packages that we don't know well, but we believe we can work them out and do research if necessary.