CSE299.1 , Group 5, Al Imran(2122071642), MD. Mukzanul Alam Nishat(2212445042), Khondkar Sayif Ali(2111323642), Arman Hossain Nawmee(2221395042), MD Araf UI Haque Dhrubo(2021493042)

Weekly Progress Report

Over the past week, we have been particularly focusing on data preprocessing and standardization for cardiovascular disease prediction. We began by extracting the Cardiovascular Disease Dataset from a compressed .zip file to access the necessary data. Once extracted, we loaded the dataset into a Pandas DataFrame and conducted an initial exploratory data analysis (EDA). This means checking for missing values, reviewing the dataset's shape and structure, and generating basic statistical summaries to understand the distribution of features. Additionally, we used Seaborn and Matplotlib to visualize key variables, such as plotting the chestpain, restingBP, serumcholestrol, restingrelectro etc. distribution to assess any underlying patterns. Later, we proceeded with data standardization, which is a crucial step for optimizing machine learning performance. We first separated features (X) and the target variable (Y), where the target represents whether a patient has cardiovascular disease. To ensure model performance, we split the dataset into training and testing subsets using an 80-20 ratio, ensuring stratification to maintain class distribution consistency. We then computed the standard deviation of the dataset to evaluate feature variability. Using Scikit-learn's StandardScaler, we fit the scalar to the training data and transform both training and testing data into standardized formats. Lastly, we performed class distribution analysis on the target variable, to verify whether our dataset is balanced or requires further handling for class imbalance in future steps. Next, we plan to implement this data set in a linear regression model to analyze predictive accuracy and based on that we will perform further tuning/optimization on our data set.