**Machine Learning Classification Project Technical Documentation**

**Overview**

This project involves building a machine learning model to classify instances into two classes based on a set of input features. The problem is a binary classification problem, which means we are trying to predict the presence or absence of a particular outcome based on the input features. In this project, we are using a dataset of individuals' health status and lifestyle habits, including age, gender, smoking status, blood pressure, and cholesterol levels, to predict the presence or absence of cardiovascular disease.

**Data Description**

The dataset contains 3390 records with 14 variables including the target variable, "CVD". The variables are as follows:

\* id: Unique identifier for each observation.

\* age: Age of the participant in years.

\* education: Level of education completed by the participant, with values ranging from 1 (less than high school) to 4 (college graduate).

\* sex: Gender of the participant, with values of "F" (female) or "M" (male).

\* is\_smoking: Smoking status of the participant, with values of "NO" (not currently smoking) or "YES" (currently smoking).

\* cigsPerDay: Number of cigarettes smoked per day by the participant.

\* BPMeds: Whether the participant is taking blood pressure medication, with values of 0 (not taking medication) or 1 (taking medication).

\* prevalentStroke: Whether the participant has a history of stroke, with values of 0 (no history) or 1 (history of stroke).

\* prevalentHyp: Whether the participant has a history of hypertension, with values of 0 (no history) or 1 (history of hypertension).

\* diabetes: Whether the participant has diabetes, with values of 0 (no diabetes) or 1 (diabetes).

\* totChol: Total cholesterol level in mg/dL.

\* sysBP: Systolic blood pressure in mmHg.

\* diaBP: Diastolic blood pressure in mmHg.

\* BMI: Body mass index, calculated as weight (kg) / height (m)^2.

\* heartRate: Resting heart rate in beats per minute.

\* glucose: Fasting blood glucose level in mg/dL.

\* TenYearCHD: Ten-year risk of coronary heart disease, with values of 0 (low risk) or 1 (high risk).Data Preprocessing

Before building a machine learning model, we need to preprocess the data to ensure that it is in the right format for the model. The preprocessing steps include:

Handling missing data: We will check for missing data and handle it appropriately, such as imputing missing values or dropping records with missing values.

**Handling outliers**: We will identify and handle outliers, such as replacing them with the mean values of the feature.

**Scaling features:** We will scale numerical features to have a similar range of values, such as using Standard scaling.

**Encoding features:** We will encode categorical features, such as using label encoding.

**Model Building**

We will build a binary classification model to predict the presence or absence of cardiovascular disease based on the given features. The following steps will be taken:

\*Split the dataset into training and testing sets.

\*Build a binary classification model using an appropriate algorithm, such as logistic regression, SVC

ane random forest.

\*Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and F1 score.

\*Tune the model's hyperparameters using techniques such as grid search to improve its performance.

**Conclusion**

In this project, we have built a machine learning model to predict the presence or absence of cardiovascular disease based on individuals' health status and lifestyle habits. By preprocessing the data appropriately and building an accurate predictive model, we can identify individuals at high risk of developing cardiovascular disease and recommend appropriate interventions to prevent or manage it.