**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team members name and email:** |
| Name: Sachin Pandey  Email: sp422593@gmail.com |
| **The GitHub Repo link.** |
| Github Link :- https://github.com/Kingslayersach/Capstone-Project-3-Cardiovascular-Risk-Prediction/tree/main |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| In this Capstone Project our task is to build a model to predict the health sate of a patient whether he/she is vulnerable to CHD (coronary heart disease) in order to alert them about their health and get proper treatment. We are provided with data\_cardiovascular\_risk.csv which includes different information.We have performed following steps in order to solve the problem statement:  * Loading the dataset, Cleaning the dataset, Removing the outlier and Handling the Null/NaN values. * Handling Null Values. * Understanding data visualisation using different plots. * Correlation heatmap. * Resampling. * Splitting the data into train and test. * Model training. * Hyperparameter tuning. * Model testing using Regression metrices.   **Loading the dataset and Cleaning the dataset:**  First step is to load the dataset into the Dataframe then we have to clean the dataset by deleting the duplicate values in the dataset, then we have to delete the outlier from the dataset, then we have to handle the Null/NaN values in the dataset.  **Handling the Null/NaN values:**  While handling the Null/NaN values in the dataset we have to replace the values with mean, median and mode values, i.e. if Any Categorical data is empty then it will be replaced by the mode value and if Any Numerical data the it will be replaced either by mean or by median.  **Visualising the data:**  We have plotted different types of charts and plots like we have used boxplots to to depict the pickup distribution, then we have used barcharts to show the Univariate and Bivariate analysis.  **Correlation heatmap:**  We have used correlation heatmap to understand the relation between different features and to select the best features out of all features.  **Resampling:**  Resampling is a methodology of economically using a data sample to improve the accuracy and quantify the uncertainty of a population parameter. Resampling methods, in fact, make use of a nested resampling method.  **Splitting Data into train and test:**  By using the correlation heatmap firstly we will create a dataset and then we will divide it into two parts(Train and Test).  **Model Training:**  After the Splitting of Train and Test data we will now train our Model on the provided Train data which will further evaluated on the Test data.  **Hyperparameter Tuning:**  GridSearchCV is a library function that is a member of sklearn's model\_selection package. It **helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.**  **Model Testing using Classification metrices:**  Model is been tested using different Classification metrices like using Logistic Regression model, K Nearest Neighbour (KNN) model, Decision Tree model, Random Forest, XGBoost Classifier,  **Conclusion:**  On performing different operations on the dataset we can conclude that :   * the people who are on BPmeds are tend to have CHD. * The chance of getting CHD drastically increases after **60**+ age. * Diabetic patients are also prone to CHD as **86**% people got CHD who were diabetic. * Systolic BP of more than 150 is really dangerous and in case of more than **250** person can die due to stroke. * High blood pressure is also not good for CHD as more than **70%** got CHD who had high BP and systolic blood pressure and diastolic BP. |