Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
❖ Sahil pardeshi(8623879021.sp@gmail.com) Contribution: 1) Data understanding 2) Data visualization 3) Bivariate analysis 4) Decision Tree classifier 5) K-nearest neighbor classifiers 6) Hyperparameter tuning on KNN ❖ Kirtesh verma(kirteshverma12345@gmail.com) Contribution: 1) Data understanding 2) Handling Null and missing values 3) Performing EDA 4) Removing Outliers 5) Logistic regression 6) Support vector machine 7) Hyperparameter tuning on SVM ❖ Pravin Bejjo(praveen.bejo.pb@gmail.com) Contribution: 1)Data understanding 2)Data visualization 3) Multivariate analysis 4) Handle imbalance data using SMOTE technique

5) Random forest

6) XGBoost classifier

7) XGBoost(feature importance)

Please paste the GitHub Repo link.

Git Hub Link: **https://github.com/Sahilpardeshi1/cardiovascular-risk-prediction**

Google Drive link:
https://drive.google.com/drive/u/0/folders/1SM37K68GpFc5cJLUcXM-gLS3_cLA663Q

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Heart disease is the leading cause of death in the world . The term "heart disease" refers to several types of heart conditions. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The EHDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established.

The datasets provide the patients information. It includes over 3,390 records and 17 attributes.The classification is to predict the 10 year risk prediction of CHD future coronary heart diseases.variables each attribute is a potential risk factor. There are both demographic, behaviors and medical risk factors.Cardiovascular disease (CVD) is defined as any serious, abnormal condition of the heart or blood vessels(arteries, veins). Cardiovascular disease includes coronary heart disease (CHD), stroke, peripheral vascular disease, congenital heart disease, endocarditis, and many other conditions.A type of disease that affects the heart or blood vessels. The risk of certain cardiovascular diseases may be increased by smoking, high blood pressure, high cholesterol, unhealthy diet, lack of exercise, and obesity.

The first step in the exercise involved exploratory data analysis where we tried to dig insights from the data in hand. It included univariate and multivariate analysis in which we identified certain trends, relationships, correlation and found out the features that had some impact on our dependent variable.On next step we have clean and perform modifications. We checked for missing values and outliers and removed irrelevant features. We also do Feature Engineering and one hot encoding for the categorical features.On next step we have handled the imbalance data using SMOTE technique. On last step we have used Machine learning algorithms like Logistic regression, Decision tree, Random Forest, K-Nearest Neighbour, XGBoost, Support Vector Machine. We did hyperparameter tuning and evaluated the

performance of each model using various metrics. The best performance was given by the XGBoost and Random Forest model with accuracy 93% and 89%, and F1_score was 93% and 89%, Precision was 96% and 91%, and Recall was 90% and 87% respectively. The most important model predictions were Age, Heartrate, totchol, BMI, Education, avgBP, and Glucose.In the project we have seen men have more heart diseases than women in the world.