Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Name: -** OM PRAKASH PRADHAN
   Email ID: - omprakashdoverr@gmail.com

   Contribution: -

   - EDA: -
     - Visualization of numerical features
     - Numerical features vs dependent variable
     - Correlation matrix
   - Feature engineering
   - Feature selection
   - Data preparation
   - Model fitting and testing –
     - Logistic Regression
     - Decision Tree
     - KNN
   - Feature importance

2. **Name: -** RUCHIKA NAYAK
   Email ID: - nayakruchika1999@gmail.com

   Contribution: -

   - EDA :-
     - Visualization of categorical features
     - Categorical features vs dependent variable
     - Multicollinearity
   - Feature engineering
   - Feature Selection
   - Data preparation
   - Model fitting and testing-
     - Random Forest
     - XGBoost
     - Support Vector
   - Feature importance

Please paste the GitHub Repo link.

GitHub Link:- https://github.com/Ruchika810/ML-Classification

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

Cardiovascular risk is the probability of suffering in the future from a clinical cardiovascular event. Coronary Heart Disease is a type of disease that develops when the arteries of the heart cannot deliver enough oxygen rich blood to the heart. Many features affect or cause our heart problems. In this classification project our goal is to predict whether the patients have any risk of coronary heart disease based on their medical records.

In our first step of EDA we performed cleaning of our dataset. Then we performed numerical features visualization and analyzation by taking single feature a time and drew some insights using various plots. Then we compared those numerical features with dependent column and shown it with the help of boxplot. Then for categorical features analysis we have used pie chart. For continuous features we have done log transformation and visualize it with the help of subplots. Then to show the overall relationship in between the features we have used correlation matrix.

After performing EDA, we have found that some features are highly correlated so we add them up and used their average as a new feature. For null values, KNN imputation was used. In next step we have removed some features which are not so important. Label encoding was also done for categorical features. Then after the whole dataset is prepared for model fitting we split up the data into train-test datasets. Then after the data is prepared, here comes the main purpose starting with logistic regression where we got the recall score of 78% for test dataset. Except logistic regression, Decision Tree, Random Forest, XGBoost and KNN is also used to find out the best recall score and best model for our dataset.

So overall we found that Logistic Regression, Random Forest and Support Vector Machine are giving a good overall balanced result. We also observed that features like age, cigsPerDay played a very important role in this 10 year risk CHD.