Capstone Project Submission

Instructions:

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

Team Member's Name, Email and Contribution:

Name - Vinay Kumar

Email – vkvinay6923@gmail.com

Contribution – 1. Loading and Understanding Data
2. Exploratory Data Anaysis
    a. Multi-variate Analysis
3. Feature Engineering
4. Train-test split
5. Models
    a. Models Training
    b. Error Plots
6. Hyper parameter Optimization
7. Conclusion

Name – Nikhil Lamje

Email – Nikhillamje123@gmail.com

Contribution – 1. Loading and Understanding Data
2. Exploratory Data Anaysis
    a. Univariate
    b. Bivariate
    c. Multi-variate Analysis
3. Feature Engineering
4. Train-test split
5. Models
    a. Models Training
    b. Error Plots
6. Hyper parameter Optimization

Please paste the GitHub Repo link.

Github Link:- https://github.com/vkvinay0627/Cardiovascular-Risk-Prediction

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 dataset, we have Cardiovascular Risk Prediction data on residents of the town of Framingham, Massachusetts. in goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). Our DataFrame has 3389 records and 17 columns

First, we started with data wrangling, checking all the null values that we did have in our data, and then we went for the outliers. All the independent columns in our data set were in ranges, and for dependent variables we performed a SMOTE (Synthetic Minority Oversampling Technique) to remove the skewness.

After that, we performed exploratory data analysis and did univariate, bivariate, and multivariate analyses and saw some interesting relationships between the variables. For example, we found that People between the ages of 35 and 50 are less likely to develop chronic health problems.

Using a heat map, feature selection was done on certain independent features that appear to be highly co-related to one another. Then, on categorical features such as sex, is_smoking, one hot encoding was performed. This makes the data suitable for splitting into tests and training before importing them to machine learning algorithms.

We used four machine learning models, which are: LogisticRegression, GaussianNB, KNeighborsClassifier, DecisionTreeClassifier, on test and training data and evaluated them on the basis of performance metrics like Precision, Recall and confusion matrix. We found that LogisticRegression and KNeighborsClassifier, were best on data on which SMOTE is performed, which had recalls of 0.72 and 0.54, for LogisticRegression and KNeighborsClassifier respectively.

Because our issue statement involves medical concerns, we must emphasise the recall (false negative) performance matrix above other performance matrixes such as accuracy, F1, and so on. After that, we performed hyperparameter optimization on LogisticRegression and KNeighborsClassifier to fine-tune the results for the lowest possible error and we achieved Recall value of 0.80 for KNeighborsClassifier, which is the best we got.