CapstoneProject Submission

Classification Model

DONE BY:

Praveen. S

(Cohort Everest)

Email id: spr87654@gmail.com

Github link:

https://github.com/prav87654/Classification-Cardiovascular-risk-prediction

Google Drive **link:https://drive.google.com/drive/folders/1hYu9VaV_NPnLLRaa6D_jr8LbxdmVrCfg?usp=sharing**

SUMMARY

Coronary heart disease **(CHD)** is the most common type of heart disease.Our experiment can help understand what could be the chance of diagnosing for CHD using logistic regression of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the current trend.

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

The dataset provides the patients' information. It includes over **4,000** records and **15** attributes.

For the next step,we performed data wrangling over the raw data and concluded what are the features or columns should be used or dropped.Since the dataset consist of 12% null values,we have dropped those rows.We proceeded with the dropping of features.I have dropped the "id"and"education" column as it was found irrespective of the target variable.

I had done the EDA and Data Visualization with the remaining features and found the independent features are not correlated with the target variable.Then Performed feature selection using Backward Elimination and also handled class imbalance using SMOTE.

For selecting the best models, I had imported pycharet to compare different classification models and found K-nearest neighbour and Support vector machines models had higher accuracy by a micro margin and proceeded with the test train split and started to fit the models.

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting and underfitting which was achieved by Grid search cv.

. In all of these models our accuracy revolves in the range of 82 to 83%.and had calculated the evaluation metrics for each model.