Capstone Project Submission

Cardiovascular Risk Prediction

Team Member's Name, Email and Contribution:

1. Rishika Rai
rishikarai70@gmail.com

- Model training and model cross validation.
- Slide work and documentation work for all trained models
- Slide work and documentation work for all trained models
- Data exploration with DataPrep library. Detailed report creation and briefing data insights and univariate analysis.

2. Aman Guleria
amansingguleria@gmail.com

- Finding out missing values. Data cleaning and plotting comparison graph.
- Data Preprocessing and setting matrices for score evaluation.
- Creating function for model and error plot.
- Hyperparameter tuning with randomised search cv.

3. Saurabh Aradwad
saurabhdilip95@gmail.com

- Importing a csv file and github admin work. Basic colab work for package imports and data inspection.
- Data summary work for slides and documentation.
- Outlier Dealing for target variable and latitude and longitude data.
- Final colab commenting work and document review.
- Work in abstraction, Data inspection and visualisation with Seaborn.
- Model comparison, finalising model and drafting results.
- Slide work - for plots and their explanations. Work on document submission & Colab final review.

GitHub Repo link.

Link:- SaurabhAradwad/Cardiovascular_Risk_Prediction_presentation (github.com)

Summary

Heart disease is a major health problem in India. According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death in India, accounting for more than 2.5 million deaths each year.

This classification case study will help us to understand important factors which contribute to heart related diseases. Supervised machine learning's classification model we have created for the prediction of whether the patient has a 10-year risk of developing coronary heart disease.

The given data set provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioural and medical risk factors.

Models Used for classification-

- Logistic Regression
- Random Forest
- XGBoost
- Support Vector Machine

The best performing model is Support Vector Machine algorithm.

Important Conclusions:

- **Risk** of **Cardiovascular heart disease** is almost **equal** between **smokers and non-smokers** and the same goes with **gender. It is** pretty much the same for both ***male*** and ***females***.
- **Correlation** obtained is **very poor** for this dataset but still due to **tuned parameters** and **strong** classification **algorithms** model **efficiency** obtained is about **93%**.
- The top **contributing features** in predicting the **ten year risk** of developing Cardiovascular Heart Disease are **'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'**.
- The **Support vector machine** with the **radial kernel** is the **best performing** model in terms of **accuracy** and the **F1 score** and Its **high AUC-score** shows that it has a **high true positive rate**.
- **Balancing** the dataset by using the **SMOTE technique** helped in **improving** the **models' sensitivity**.
- With **more data**(especially that of the minority class) **better models** can be built.