Heart Failure Prediction Using Machine learning

Abstract: This project includes applications of traditional machine learning techniques on the detection of the heart failure dataset. Heart failure prediction is essentially important since it is considered to be one of the deadliest diseases in the world. With the advancement in automation statistical Machine learning techniques have proved to be working best on the early detection of heart diseases. We have utilized the dataset from the UCI website which has been trained on the statistical ML models to analyze the performance of each of them and report the best model for the prediction.

Introduction:
The dataset describes various attributes of the heart disease prediction which will be used to predict the outcome and also includes a set of the features that are essential for the prediction and analysis. Namely some features are age, sex, diabetes, creatinine levels, blood pressure and also the smoking attribute relative to a particular gender.

We have been through some the essential topics in this project:

1. Exploratory data analysis (EDA) - the process of going through a dataset and finding out more about it.
2. Model training - create model(s) to learn to predict a target variable based on other variables.
3. Model evaluation - evaluating the models predictions using problem-specific evaluation metrics.
4. Model comparison - comparing several different models to find the best one.
5. Model fine-tuning - once we've found a good model, how can we improve it?
6. Feature importance - since we're predicting the presence of heart disease, are there some things which are more important for prediction?
7. Cross-validation - if we do build a good model, can we be sure it will work on unseen data?
8. And finally the analysis will be reported out here.

Data Analysis(Methodology):

Firstly, after taking the dataset we have done some initial preprocessing which includes finding any missing values. There have been none of them and next we went on with the Exploratory Data Analysis(EDA) which is an essential part of the project wherein we can visualize the features how they are varying and also some of the data imbalancing have been done from the preprocessing steps. Generally we use the concept of undersampling and upsampling for normalization. We have used a novel technique called SMOTE (Synthetic Minority Oversampling Technique) which has been previously explored but since it is not a regular technique which avoids the code reduction.

Thereafter, we have received the normalized data. Now we will be going ahead with training our data on the machine learning models which will require the splitting of the data into train and

test splits. We have taken the resampled data so that it would not misclassify any of the features and a correct analysis can be made. For the purpose of the model selection we have used some of the Supervised Machine Learning techniques like SVM, Logistic Regression, KNN, and Naive Bayes algorithm. SVM is selected as a good model in this project because it is highly useful when we have to create a clear margin of separation between classes. Thereafter we also have one of the get to start techniques in Machine Learning. It has performed best on the datasets for the initial analysis and also has been a great kick start for the next deep neural networks in building the architecture based on the analysis it has made initially. We have then observed that on every run there is new value of accuracy and other metrics showing different outcomes hence we have explored the concept of ensemble learning wherein we have combined the above methods to form a single model and calculate the average accuracy and other preformed parameters. This is the prior step for the tuning hyperparameters to find out which model works the best on the given dataset. For the purpose of this we have explored the concept of GridSearchCv and RandomSearchCV. We have defined the hyperparameters and have given the models to each of them thereafter we also analyzed the performance of each of the models and reported out some of the best params.

Results:
Some of the results that we have calculated are given below in the following table:

Model name	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Logistic Regression	81.7	85	82.3	78.57
SVM	81.7	85	81.8	78.57
Gaussian NB	78.04	67.5	78.2	88.09
KNN	81.7	82.5	75	71.4

The table given above shows how the various models have worked well for the dataset and how well they have been successful in classifying the disease. But after successfully running the algorithm multiple times we are able to see a change of these values hence using the concept of ensemble modeling which gives the results as Acc being: 81.7, precision: 82%, sensitivity: 75.7% and specificity: 85.71%. Thereafter, some of the hyperparameter tuning methods have been used on the models we have tested and the screenshots given show how the models work for both the GridSearchCv and RandomSearchCv.

Conclusions:
The models have worked well on the dataset, since these are some of the traditional machine learning models no layers can be changed or alteration of neural networks can be applied to the dataset to increase the performance on the dataset. In the future course our aim is to work on building the models from scratch and applying to the dataset which can assess better and hence evaluate the convolutional model can be done.

References:

1. https://arxiv.org/abs/1106.1813
2. https://doi.org/10.1016/j.csbj.2016.11.001

Github links:
MPS Sandhu: https://github.com/mpssandhu/Application_of_ML_in_heart_disease

Ankit Chhillar:
https://github.com/AnkitChhillar/Application-of-Machine-Learning-in-prediction-of-Heart-Disease

Anand Addepalli:
https://github.com/anandram9/Application-of-Machine-Learning-in-prediction-of-Heart-Disease.git