PREDICTION OF HEART DISEASE

UNIVERSITY COLLEGE DUBLIN | DONE BY - DIWAKAR MOHAN(22200318) & K.SAKETH SAI NIGAM(22201204)

BACKGROUND

- [1]A major source of death and morbidity, heart disease is a serious worldwide health concern. According to the WHO (Rath et al., 2021) 17.8 million people die from heart disease globally per decade. The evaluation of the patient's medical history, physical examinations, laboratory testing, stress tests, and cardiac catheterisation are frequently combined in the identification and diagnosis of heart disease. These diagnostic techniques, however, could not always give a complete picture or might need for intrusive treatments.
- Furthermore, heart disease sometimes manifests without symptoms or with mild signs, making it challenging to recognise those who are at risk or who are in the early stages of the disease.

OBJECTIVE

- Given the challenges in diagnosing cardiac illness, there is a growing need for innovative, trustworthy techniques to improve early detection and risk stratification. The goal is to develop prediction models that, uses a variety of clinical data and risk variables to accurately detect cardiac illness at an early stage.

DATASET

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

TABLE.1 HEART-DISEASE.CSV

- The original data came from the Cleaveland data from the UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+Disease
- **Features of the Dataset:-**

1. age - age in years
2. sex-(1=male;0=female)
3. cp - chest pain type
   - 0: Typical angina: chest pain related decrease blood supply to the heart
   - 1: Atypical angina: chest pain not related to heart
   - 2: Non-anginal pain: typically esophageal spasms (non heart related)
   - 3: Asymptomatic: chest pain not showing signs of disease
   - 4. trestbps - resting blood pressure (in mm Hg on admission to the hospital) anything above 130-140 is typically cause for concern
5. chol - serum cholesterol in mg/dl

serum = LDL + HDL + .2 * triglycerides above 200 is cause for concern

6. fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) '>126' mg/dL signals diabetes
7. restecg - resting electrocardiographic results
   - 0: Nothing to note
   - 1: ST-T Wave abnormality
     - can range from mild symptoms to severe problems
     - signals non-normal heart beat
   - 2: Possible or definite left ventricular hypertrophy
     - Enlarged heart's main pumping chamber
8. thalach - maximum heart rate achieved
9. exang - exercise induced angina (1 = yes; 0 = no)
10. oldpeak - ST depression induced by exercise relative to rest looks at stress of

heart during exercise unhealthy heart will stress more

11. slope - the slope of the peak exercise ST segment
12. ca - number of major vessels (0-3) coloured by flourosopy
13. thal - thalium stress result
14. target - have disease or not (1=yes, 0=no) (= the predicted attribute)

TECH STACK

- The language used to code is **PYTHON**.
- Libraries used for Data Modelling & Analysis is **SCIKET LEARN, TENSOR FLOW**.
- Libraries used for Data Manipulation is **NUMPY, PANDAS**.
- Libraries used for Data Visualisation is **SEABORN, MATPLOTLIB**.

EDA

DECIDING SUITABLE EVALUATION METRICS FOR THE MODEL

- If we are dealing with a balanced dataset, accuracy might be a suitable methodology for evaluation.
- [3]If we are dealing with an imbalanced dataset, one class significantly outnumbers the other(s), making the model prone to biased predictions. The standard accuracy metric might not be suitable as it could be misleading due to the class imbalance.
- In such cases Precision, Recall (Sensitivity or True Positive Rate),F1-score,.. etc.
- **The dataset we have is nearly balanced, with an equal proportion of persons suffering from disease and those who do not.**
- **Hence, Accuracy and Sensitivity both will be a proper evaluation metric**

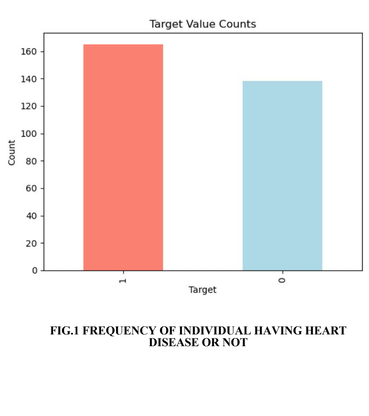

FIG.1 FREQUENCY OF INDIVIDUAL HAVING HEART DISEASE OR NOT

CHECKING FOR MISSING VALUES

- Checking for missing values in a dataset is an essential step in the data analysis process. Dealing with missing values appropriately is crucial for building accurate and reliable models.
- It is necessary to handle missing values using suitable measures like replacing the missing values with mean , median, mode or using matrix completion method, or by just removing the missing values.
- **The dataset we have considered does not have missing values so we can proceed with data for modelling**

ANALYSING THE UNDERLYING DISTRIBUTION OF DATA FOR SELECTING SUITABLE MODELS:

- [2]Knowing the distribution of the data helps in selecting appropriate modelling techniques. For example, if the data is highly skewed, certain models might not be suitable or may require transformations.
- From the above output we can infer that the mean and the median values are almost equal for all the numerical variables which states that the data is normally distributed.
- **Since the data is normally distributed logistic, Random forest are more suitable models. But For this analysis we are considering other models like KNN which is more suitable for skewed data.**

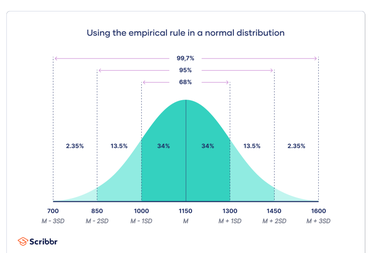

FIG.2 NORMAL DISTRIBUTION CURVE

FEATURE SELECTION:-

1. Is SEX a important Feature?

- From the above output we can infer that the individuals diagnosed with heart disease or not differs based on sex.
- So Sex can be considered as a important factor for predicting heart disease.

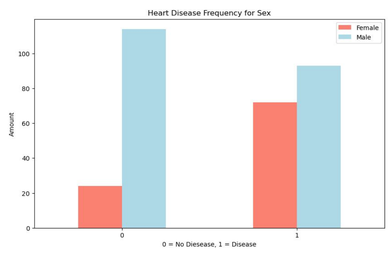

FIG.3 FREQUENCY OF INDIVIDUAL HAVING HEART DISEASE OR NOT GROUPED BY SEX

2. Is CP type, AGE & MAXIMUM HEART RATE an important Feature contributing to Heart Disease?

FIG.4 RELATION BETWEEN AGE AND MAX HEART RATE

- From the scatter plot we can infer that the maximum heart rate decreases with age.
- As the Age increases the maximum heart rate decreases. This states that there is a relationship between the age and the maximum heart rate.
- Understanding the relationship between age and maximum heart rate can reveal any interaction effects between these two predictors. Interaction effects occur when the combined effect of two predictors is different from their individual effects. Identifying such interactions can provide nuanced insights into how age and maximum heart rate jointly impact heart disease risk.

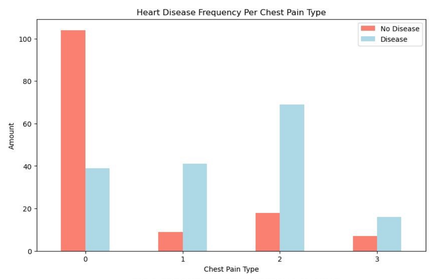

FIG.5 FREQUENCY OF INDIVIDUAL HAVING HEART DISEASE OR NOT GROUPED BY CHEST PAIN TYPE

- From the above output we can infer that Different chest pain types are associated different number of people diagnose with heart disease.
- Chest pain type 1 & 2 are associated with heart diseases more as compared to other types. This states the Heart disease diagnostic is associated with the chest pain type.
- Therefore Chest pain type is an important factor to be considered while predicting heart disease.

EXPLORING MULTI COLLINEARITY PROBLEM:-

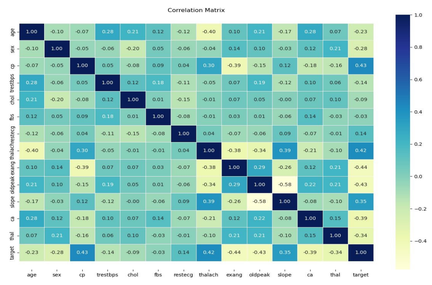

FIG.6 CORRELATION PLOT

- 0.8 is a commonly used threshold for detecting multicollinearity. From the above output we can infer that there is not high relationship between the variables and hence they are independent variables. So the problem of multicollinearity doesn't exist.

MODELS

MODELS

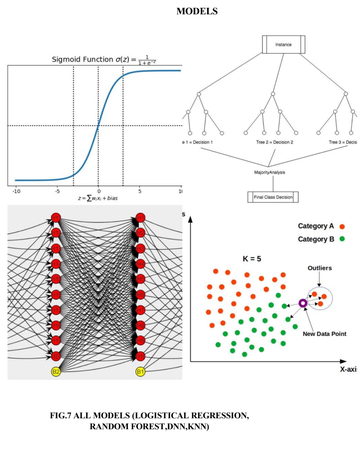

FIG.7 ALL MODELS (LOGISTICAL REGRESSION, RANDOM FOREST,DNN,KNN)

MODELS USED:-

- LOGISTICAL REGRESSION
- RANDOM FOREST REGRESSION
- K-NEAREST NEIGHBOUR (KNN)
- DEEP NEURAL NETWORKS (DNN)

MODEL IMPLEMENTATION

- TRAIN - TEST - SPLIT :- 80% TRAINING the data and 20% for TESTING.
- FITTING the data for Logistical Regression, Random Forest Regression, KNN.
- Designing DNN using Tensor flow Package.
- Comparing all the models based on accuracy. Therefore the best model we got is "LOGISTICAL REGRESSION" of all the models used with accuracy of 0.885.

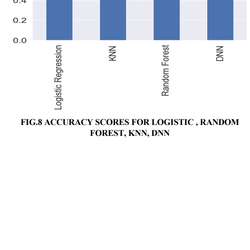

FIG.8 ACCURACY SCORES FOR LOGISTIC , RANDOM FOREST, KNN, DNN

HYPER PARAMETER TUNING

- Employing RANDOM SEARCH CV for LOGISTICAL REGRESSION and RANDOM FOREST REGRESSION.
- [4] Employing GRID OF NEIGHBOURS for K-NEAREST NEIGHBOUR (KNN)
- Employing GRID OF VALUES for DEEP NEURAL NETWORKS (DNN)
- Comparing all the values based on accuracy to get the Optimised Model

K – NEAREST NEIGHBOUR TUNING

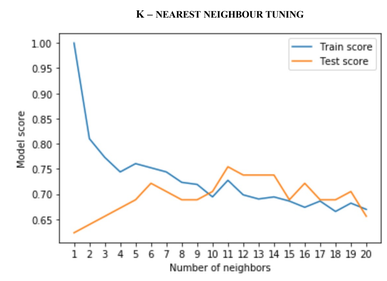

FIG.9 KNN HYPER PARAMETER TUNING

EVALUATING THE BEST MODEL

BEST MODEL: LOGISTIC REGRESSION

ROC curve and AUC score:

- [5]The ROC curve answers the question: How often will a randomly chosen 1 outcome have a higher probability of being predicted to be a 1 outcome than a randomly chosen true 0?
- The larger the area under the ROC curve AUC-ROC, the better is the discrimination.
- The Logistic Regression model has an AUC value of 0.93, indicating that it performs better in distinguishing patients with Heart Disease from those who do not.

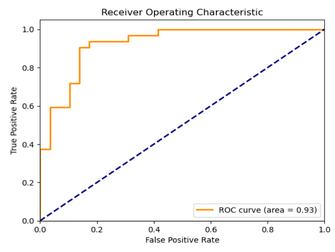

FIG.10 ROC CURVE FOR LOGISTIC REGRESSION

CONFUSION MATRIX:

- TP: 25 instances were correctly predicted as positive.
- FP: 4 instances were wrongly predicted as positive when they were actually negative.
- FN: 3 instances were wrongly predicted as negative when they were actually positive.
- TN: 29 instances were correctly predicted as negative.

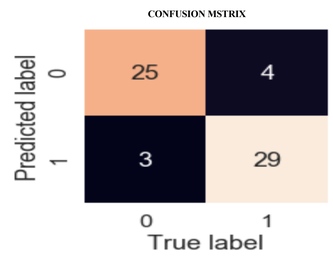

FIG.11 CONFUSION MATRIX

CROSS VALIDATION SCORES:

- [6]**Cross-validated Precision**: The precision is a metric that assesses the proportion of positive identifications (true positives) that are actually correct. In your case, the cross-validated precision has been calculated using a 5-fold cross-validation. The average precision across the folds is approximately 0.82.
- **Cross-validated Recall (Sensitivity)**: The recall measures the proportion of actual positive instances that were correctly identified by the model (true positives). The cross-validated recall has been calculated using a 5-fold cross-validation. The average recall across the folds is approximately 0.92.
- **Cross-validated F1-Score**: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of a model's accuracy. The cross-validated F1-score has been calculated using a 5-fold cross-validation. The average F1-score across the folds is approximately 0.87.
- **Cross-validated Accuracy**: The cross-validated accuracy results indicate how well the classifier is performing on different subsets of the data. The average accuracy score (approximately 0.844) gives you an estimate of the model's general performance on the entire dataset. This cross-validation approach helps provide a more robust evaluation of the model's accuracy, as it considers multiple splits of the data for training and testing.

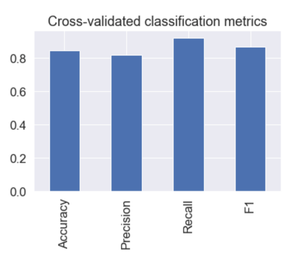

FIG.12 ACCURACY, PRECISION, RECALL, F1

IDENTIFYING THE POTENTIAL IMPORTANT FEATURE :

- [7]The magnitude and direction of the coefficients can provide insights into how each feature impacts the model's predictions. Positive coefficients suggest a positive impact on the target event, while negative coefficients suggest a negative impact.

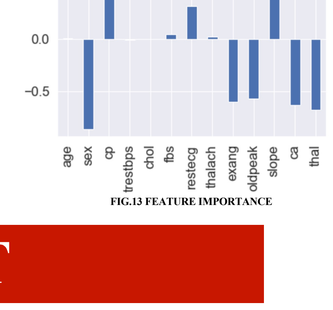

FIG.13 FEATURE IMPORTANCE

RESULT

- We have considered 4 models namely Logistic, Random forest, KNN, DNN.Based on the evaluation of four models (Logistic Regression, Random Forest, K-Nearest Neighbour's, Deep Neural Network), after hyper-parameter tuning and fitting the optimised models on balanced data, Logistic Regression emerged as the best-performing model based on accuracy.
- Subsequently, the evaluation of the Logistic Regression model included the following metrics to gauge its predictive capacity:

1. Area Under the ROC Curve (AUC):
2. Sensitivity (True Positive Rate):
3. Specificity (True Negative Rate):
4. F1 Score:

- These metrics collectively offer a comprehensive understanding of the model's performance across various aspects. While accuracy is a valuable metric, these additional measures give deeper insights into how well the model performs under different scenarios and classes. Based on the evaluated metrics, the Logistic Regression model's AUC, Sensitivity, Specificity, and F1 Score provide a more nuanced assessment of its predictive capacity.

IMPROVEMENTS

- We can collect more data which trains the model in more diverse situations.
- We can try better model like CatBoost or XGBoost.

REFERENCES

1. Machine Learning Technology-Based Heart Disease Detection Models, Authors:Umarani Nagavelli, Debabrata Samanta, and Partha Chakraborty
2. Effective Heart Disease Prediction Using Machine Learning Techniques, Authors:Chintan M. Bhatt ,Parth Patel,Tarang Ghetia and Pier Luigi Mazzeo
3. Heart disease prediction using machine learning algorithms, Authors: Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath Published under licence by IOP Publishing Ltd
4. Diagnosis And Prediction Of Heart Disease Using Machine Learning Techniques, Author:J.Jeyaganesan, A.Sathiya , S.Keerthana, Aaradhyanidhi Aiyer
5. A. L. Bui, T. B. Horwich, and G. C. Fonarow, ''Epidemiology and risk profile of heart failure,'' *Nature Rev. Cardiol.*, vol. 8, no. 1, p. 30, 2011.
6. M. Durairaj and N. Ramasamy, ''A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,'' *Int. J. Control Theory Appl.*, vol. 9, no. 27, pp. 255–260, 2016.
7. S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ''Innovative artificial neural networks-based decision support system for heart diseases diagno- sis,'' *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, 2013, Art. no. 35396.