Heart Failure Prediciton

Trisha Jain
Department of CSE
PES University
Bengaluru, Karnantaka
jtrisha0403@gmail.com

Vandana Bhaskar
Department of CSE
PES University
Bengaluru, Karnantaka
beingvandana024@gmail.com

Veda N
Department of CSE
PES University
Bengaluru, Karnantaka
vedan1408@gmail.com

***Abstract*—Cardiovascular diseases are a very common cause of death globally. According to WHO an estimated 17.9 million lives each year are lost due to heart diseases, which accounts for 31% of all deaths worldwide. The existing state of the art techniques like Angiography are expensive and require high level of technical expertise. Therefore, this project aims to find an effective inexpensive machine learning based model to predict heart failure which can help in the early diagnosis of heart disease and determine the important features in predicting heart failure. Various data visualization techniques were used to understand the data well.**

***Index Terms*—heart failure, prediction, classification**

I. Introduction

Cardiovascular disease is a very common cause of mortality. They take an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs. People at high cardiovascular risk due to the presence of risk factors such as hypertension, diabetes or other diseases need early diagnosis. Heart failure (HF) is the failure of the heart to pump sufficient amounts of blood to meet the needs of the body. Narrowing or blockage of the coronary arteries is considered to be the main cause of HF. The common symptoms of HF include shortness of breath, swollen feet and weakness of the body. There are two main groups of risk factors. The first group includes patient's family history, sex and age. These risk factors cannot be changed. The second group includes risk factors that are related to the lifestyle of the patient. Hence, these factors can be changed e.g., high cholesterol level, smoking, physical inactivity. One of the best tools to diagnose heart failure is known to be Angiography. It is a type of diagnosis used to confirm heart disease and is regarded as a promising method for the diagnosis of HF. But Angiography has its own limitations and side effects. It is an expensive tool and requires a high level of technical expertise.

The positive side is that heart attacks are highly preventable and simple lifestyle modifications(such as reducing alcohol use; eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high risk patients because of the multi-factorial nature of several contributory risk factors. This is where machine learning and data mining come to the rescue.

Machine learning models can be used to predict heart failure. The use of data analytics allows for improvements to patient care, faster and more accurate diagnosis, preventive measures, and more informed decision-making. Additionally, it can lower costs, simplify internal operations and a lot more.

II. Review of Literature

In [1] the author discusses the performance of four popular machine learning models for predicting heart failure. They are Logistic Regression (LR), Naïve Bayes (NB), Random Forest (RF) and Support Vector Machine (SVM). These were selected by the author because of their good performance in medical related applications. The performance of the techniques are measured using various performance metrics such as accuracy, precision, recall, f1 score etc. From the experimental analysis conducted, the author listed the average performance score of all the four techniques by computing the average for all the evaluation criteria. It was noted that Random Forest produces the highest average performance score (0.88) when compared to other machine learning techniques. Also, an experiment conducted to determine the significant features to predict heart failure using RF indicated that all the features are important in making a good prediction for heart failure.

In [2], the author performs a comparative study of 18 machine learning models for heart failure prediction, with z-score or min-max normalization methods and Synthetic Minority Oversampling Technique (SMOTE) for the imbalance class problem and uses F1 score and accuracy as the evaluation metrics. The F1 score is the harmonic mean of Precision and Recall and gives a better measure of the incorrectly classified cases that reflects the robustness of the model, and the accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observations to the total observations.

SMOTE (Synthetic Minority Oversampling Technique) [7] is a statistical technique for increasing the number of cases in dataset in a balanced way. SMOTE is used when the dataset we are dealing with is imbalanced. The module generates new instances from existing minority cases that are supplied as input. This implementation of SMOTE does not change the number of majority cases. The output is original samples plus additional synthetically generated minority samples. It is important to note that increasing the number of cases using SMOTE does not guarantee to produce more accurate models. It is suggested to experiment with different feature sets, percentage and number of nearest neighbours and see

how it influences the model.
Author [2] compares the results of 18 models first, with the min-max normalization method and then with the z-score normalization method. From the comparisons it was noted that on average, accuracy and F1 score with SMOTE method is higher than that without SMOTE method and the z-score normalization method is better than the min-max normalization method. Further author evaluates feature importance graph for few of the models to determine the important input features in predicting heart failure. However different models may give different feature importance scores to different features. It is also noted by the author that SMOTE improved the accuracy and F1 score of a significant number of models out of 18 models that were used for comparative study. Thus, the SMOTE approach proved to be helpful.

A. Ishaq et al. [3], proposed ETC(Extra Tree Classifier) to be the most effective method for predicting heart patient's survival. Using this model they achieved an accuracy of 92.62%. A total of nine machine learning models were tried on the Heart-failure-clinical-records-dataset which is derived from the UCI machine learning repository. After using SMOTE for class imbalance and RF for feature selection, tree based classifiers were found to achieve the highest accuracy. Ensemble learning approaches could be applied to improve accuracy. SMOTE might have led to additional noise in the dataset which could have an impact on the accuracy.

L. Ali et al. [4], applied a stacked SVM approach for the prediction of heart failure. The first SVM model was L1 regularised and was used for feature selection and the second SVM model was L2 regularised and was used for the actual prediction. Using this approach they observed an improvement of 3.3% over the conventional SVM model. The proposed model is capable of showing better results with a few features which leads to a lower time complexity. Other ensemble learning approaches could be tried out for achieving better accuracy instead of stacked SVMs.

The purpose of the study [5] was to develop and validate a multivariate risk model to predict 1-, 2-, and 3-year survival in heart failure patients with the use of easily obtainable characteristics relating to clinical status.The Seattle Heart Failure Model was derived in a cohort of 1125 heart failure patients with the use of a multivariate Cox model. For medications and devices not available in the derivation database, hazard ratios were estimated from published literature. The model was prospectively validated in 5 additional cohorts totaling 9942 heart failure patients and 17 307 person-years of follow-up. The accuracy of the model was excellent, with predicted versus actual 1-year survival rates of 73.4% versus 74.3% in the derivation cohort and 90.5% versus 88.5%, 86.5% versus 86.5%, 83.8% versus 83.3%, 90.9% versus 91.0%, and 89.6% versus 86.7% in the 5 validation cohorts. For the lowest score, the 2-year survival was 92.8% compared with 88.7%, 77.8%, 58.1%, 29.5%, and 10.8% for scores of 0, 1, 2, 3, and 4, respectively. The overall receiver operating characteristic area under the curve was 0.729 (95% CI, 0.714 to 0.744). The model also allowed estimation of the benefit of adding medications or devices to an individual patient's therapeutic regimen.

The paper [6], talks about how heart diseases are complex and how it should be handled carefully by various classification techniques.The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB).In this work,a prediction model is produced using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods.The Hybrid Random Forest with Linear Model (HRFLM) method uses all features without any restrictions of feature selection.The highest accuracy is achieved by HRFLM classification method in comparison with existing methods such as Logistic Regression, Deep Learning, etc in terms of accuracy, classification error, precision, F-measure, sensitivity and specificity.

III. INITIAL INSIGHTS THROUGH EDA

A. Dataset

We utilized a common dataset in public: "https://www.kaggle.com/fedesoriano/heart-failure-prediction". This dataset [8] was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features. The five datasets used for its curation are Cleveland, Hungarian, Switzerland, Long Beach VA, Stalog (Heart) Data Set. All the dataset used is from [9] UCI Machine Learning Repository on: "https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/".
The dataset has a total of 1190 observations out of which 272 observations are duplicated. Hence the final dataset has 918 observations.
This data contains the following input characteristics and prediction targets. The target we want to predict is the heart failure.

- **age** of the patient (in years)
- **sex** of the patient
- **chest pain type**: Angina is discomfort caused when the heart muscles do not get enough oxygen rich blood. Angina feels like pressure, heaviness or pain in the chest. Typical angina, Atypical angina, Non-anginal pain, Asymptotic are type of chest pain.
- **serum cholesterol**: Cholesterol is a type of body fat and serum cholesterol level represents the amount of total cholesterol in blood.
- **fasting blood sugar**: It is a test to determine the blood glucose level after an overnight fast.
- **resting electrocardiogram results**: It is a simple, painless and quick test performed to detect certain heart condition. Some possible results from the test performed could be Normal, having ST-T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria.

- **maximum heart rate**: It has been shown that increase in heart rate is associated with increased risk of cardiac deaths and increased systolic blood pressure.
- **exercise-induced angina**: It is a common complaint of cardiac patients, particularly when exercising in the cold.
- **slope of the peak exercise**: The ST segment represents the interval between ventricular depolarization and repolarization. The important cause of ST segment elevation or depression is myocardial infraction. The ST segment shift relative to exercise-induced increments is more accurate ECG criterion for diagnosing CVD. (Up: upsloping, Flat: flat, Down: downsloping).
- **resting BP**: It's a tool doctors use to identify if a person is at risk for heart disease or stroke.
- **old peak**: ST (positions on the ECG plot) depression induced by exercise relative to rest.

There are a total of 12 attributes and 918 rows. There are no duplicate or missing rows and columns. The basic summary of the data can be seen in figure 1 by using the describe method of the Pandas library.

	count	mean	std	min	25%	50%	75%	max
Age	918.000000	53.510893	9.432617	28.000000	47.000000	54.000000	60.000000	77.000000
RestingBP	918.000000	132.396514	18.514154	0.000000	120.000000	130.000000	140.000000	200.000000
Cholesterol	918.000000	198.799564	109.384145	0.000000	173.250000	223.000000	267.000000	603.000000
FastingBS	918.000000	0.233115	0.423046	0.000000	0.000000	0.000000	0.000000	1.000000
MaxHR	918.000000	136.809368	25.460334	60.000000	120.000000	138.000000	156.000000	202.000000
Oldpeak	918.000000	0.887364	1.066570	-2.600000	0.000000	0.600000	1.500000	6.200000
HeartDisease	918.000000	0.553377	0.497414	0.000000	0.000000	1.000000	1.000000	1.000000

Fig. 1. Summary of dataset

The summary of the features is shown in the figure 2.

Numerical features	Age	in years
	RestingBP	in mm Hg
	Cholesterol	in mg/dl
	FastingBS	1: if more than 120mg/dl 0: if less than 120mg/dl
	MaxHR	maximum heart rate achieved (Range of values: 60 - 202)
	Oldpeak	ST (positions on the ECG plot) depression induced by exercise relative to rest (Range of values: -2.6 – 6.2)
Categorical features	Sex	2 unique variables 'M', 'F'
	ChestPainType	4 unique variables 'ATA', 'NAP', 'ASY', 'TA'
	RestingECG	3 unique variables 'Normal', 'ST', 'LVH'
	ExerciseAngina	2 unique variables 'N', 'Y'
	ST_Slope	3 unique variables 'Up', 'Flat', 'Down'

Fig. 2. Summary of features

The correlation map shown in figure 3 suggests that there is no strong correlation between any two features that don't belong to the target class.

The correlation of the numerical features to the target feature is shown by figure 4 This suggests that Oldpeak, MaxHR and Age are highly correlated to HeartDisease.

B. Categorical Data Analysis

Figure 5 to figure 8, indicate that if for an observation the feature Sex, has the value male then that observation is more likely to have the value of HeartDisease as 1. Similarly, if ChestPainType has the value ASY or ExerciseAgina has a

Fig. 3. Correlation between features

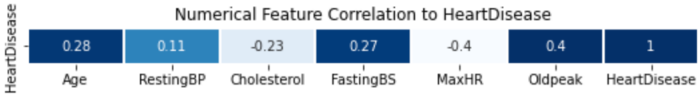

Fig. 4. Correlation of numerical features to HeartDisease

value Y or ST_slope has the value Flat then the chances of the value of HeartDisease being 1 are more.

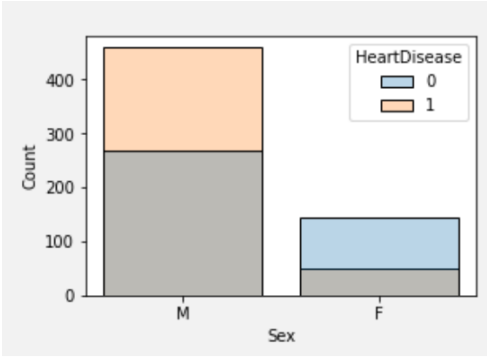

Fig. 5. Impact of Gender on Heart Disease

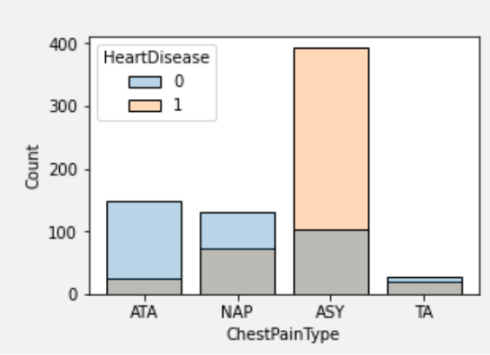

Fig. 6. Impact of Chest Pain Type on Heart Disease

C. Numerical Data Analysis

Figure 9 to figure 12, indicate that, there is a higher chance of getting a heart disease if for an observation the value of Age

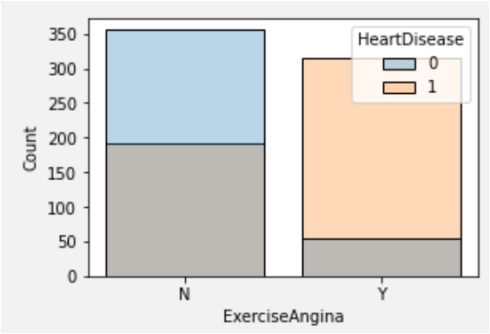

Fig. 7. Impact of Exercise Agina on Heart Disease

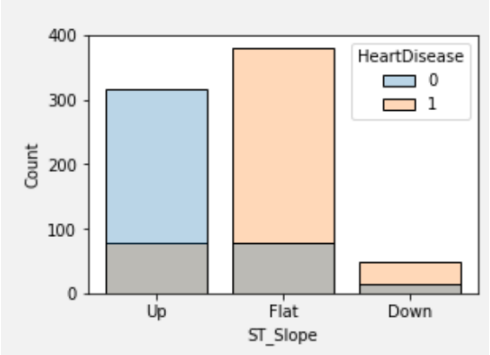

Fig. 8. Impact of STslope on Heart Disease

is greater than 55, or the value of FastingBS is 1, or MaxHR is low or OldPeak is high.

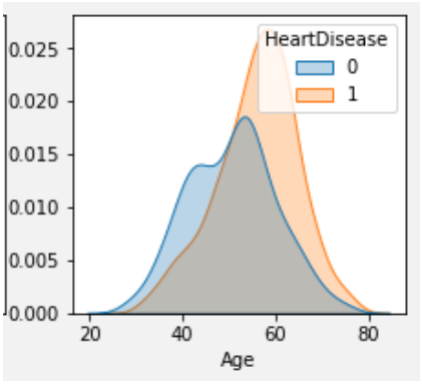

Fig. 9. Impact of Age on Heart Disease

D. Data Preprocessing

- Removing outliers : RestingBP had an extreme outlier which was removed.
- Imputing incorrect values : Cholesterol had a lot of zero values which were imputed with the median
- Normalizing the numerical values : To adjust the difference in the order of magnitude between variables they were normalized.
- Encoding categorical values : Using LabelEncoder categorical variables were converted to numerical variables.

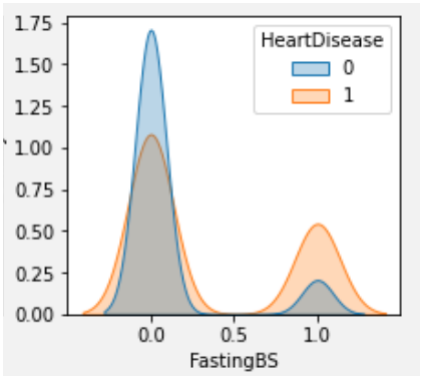

Fig. 10. Impact of FastingBS on Heart Disease

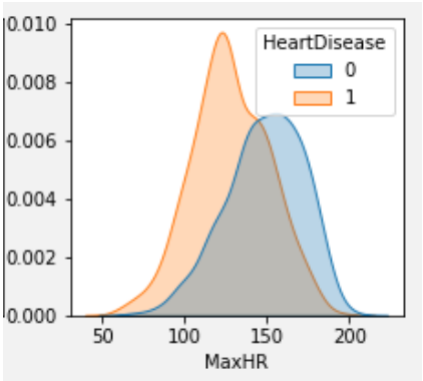

Fig. 11. Impact of MaxHR on Heart Disease

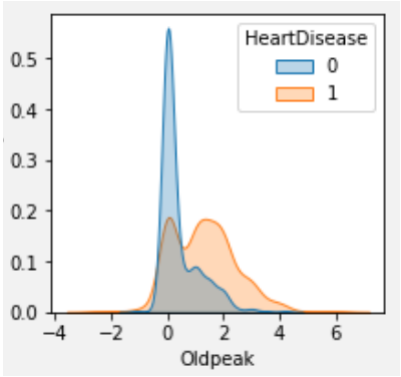

Fig. 12. Impact of Old Peak on Heart Disease

IV. Conclusion

Exploratory data analysis (EDA) is used to analyse and investigate data sets and summarize the main characteristics. It helps determine how best to manipulate data sources to get the answers that are needed. The initial insights obtained after performing exploratory data analysis/ data visualization is noted. Data pre-processing, a data mining technique like removing outliers, imputing incorrect values, normalising the incorrect values and encoding categorical values are performed to transform the raw data to a useful and efficient format. We plan to experiment with various models in order to choose the

model which outperforms the other in predicting heart failure effectively and accurately.

V. Acknowledgment

We would like to express our gratitude to our Data Analytics Course Professor Dr. Gowri Srinivasa for providing constant guidance and leading us in the right path during each phase of our project. We would also like to acknowledge our assistant professors who have meticulously prepared the course content and also the teaching assistants who have been consistently supportive and helped us with resources to practice the learnt concepts.

References

[1] Huang, Nur & Ibrahim, Zaidah & Diah, Norizan. (2021). Machine Learning Techniques for Heart Failure Prediction. MALAYSIAN JOURNAL OF COMPUTING. 6. 872. 10.24191/mjoc.v6i2.13708.

[2] Wang, Jing. (2021). Heart Failure Prediction with Machine Learning: A Comparative Study. Journal of Physics: Conference Series. 2031. 012068. 10.1088/1742-6596/2031/1/012068.

[3] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," in IEEE Access, vol. 9, pp. 39707-39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[4] L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure," in IEEE Access, vol. 7, pp. 54007-54014, 2019, doi: 10.1109/ACCESS.2019.2909969.

[5] Wayne C. Levy et al., "The Seattle Heart Failure Model," in Circulation 2006;113:1424–1433, doi: 10.1161/CIRCULATIONAHA.105.584102.

[6] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[7] "ML Studio (classic): SMOTE – Azure." ML Studio (classic): SMOTE - Azure — Microsoft Docs. Accessed October 17, 2021. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote.

[8] "Heart Failure Prediction Dataset." Kaggle. Publication September 10, 2021. https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.