Heart Failure Prediction and Analysis of the factors causing it

Atul Kumar Verma
Dept. of Mathematics
Indian Institute of
Technology, Bombay
19B090004
19B090004@iitb.ac.in

Abhishek Yadav
Dept. of Mathematics
Indian Institute of
Technology, Bombay
19B090001
19B090001@iitb.ac.in

***Abstract*—Many factors influence the conditions under which heart failure can occur. A person's lifestyle, along with their health conditions, can convey a lot about these factors. Using these factors, we can predict whether there is a chance that the person will have an unfortunate event leading to heart disease or not. Warning the person beforehand can lead to preventive measures taken on time to reduce the risk further. In this paper, we have done extensive Exploratory Data Analysis to find the factors responsible for heart failure. We have also employed various classification algorithms to predict heart failure from the attributes and compare their results based on different accuracy scores.**

***Index Terms*—Support Vector Machine, Decision Tree Classifier, Random Forest, Neural Network, Pytorch, PyCaret**

I. Introduction

Today, cardiovascular diseases are the leading cause of death worldwide, with 17.9 million deaths annually, as per the World Health Organization reports. Heart failure is a condition where the heart cannot pump the required amount of blood to the body. According to medical professionals, high cholesterol, increase in triglyceride levels, and hypertension are some of the reasons that lead to heart failure. In recent times, it has become easier to collect patients' medical data such as their ECG, glucose levels, and cholesterol levels and store it in an electronic form. Hence, a large amount of data is now becoming available for analysis, making it favourable to use machine learning or deep learning models to predict heart failure in individuals. While doing this project, one of our objectives was to verify whether conventional factors such as cholesterol, resting blood pressure, and heart rate affect the probability of heart failure in a statistically significant manner.

The data set that we used consisted of 11 features excluding the target variable, and hence it was not possible to visualize all of the data easily. For this purpose, we have performed Principal Component Analysis to reduce the number of dimensions to 2. We have also shown the decision boundaries generated by some commonly used classification algorithms.

For the actual prediction of heart failure, we have used classification algorithms such as SVM-C, Decision Tree, Random Forest and Multi-Layer Perceptron. Gridsearch has been used for finding the optimal hyperparameters. We have also used Auto ML technology (PyCaret) and compared the results with the models manually trained by us.

II. Background and Prior Work

Electronic health records are a valuable source of information to discover non-trivial correlations and relationships in the patients' data, and can be used for research and clinical practice, challenging traditional myths on risk factors. The use of machine learning to predict heart failure is an exciting prospect, and there has already been a lot of prior work regarding this topic. An interesting addition to the clinical data was the use of Natural Language Processing to incorporate clinical notes into the data. This was done by Zhang et al. [16] to achieve a 93.37 accuracy. The past studies are mainly based on a 13 feature dataset, including attributes such as Thalassemia.

Modelling survival for heart failure and cardiovascular diseases, in general, is still a problem, both in terms of achieving high prediction accuracy and identifying the driving factors. Most of the models developed for this purpose hit an upper cap on accuracy. More importantly, we cannot discern relationships of heart failure with the attributes from the models. More recent models have shown improvements, especially if the survival outcome is coupled with additional targets such as hospitalization. Although researchers have identified a broad set of predictors and indicators, there is no final agreement regarding their relative impact on survival prediction.

III. Data and Methodology

A. Dataset used

The data has been taken from Kaggle, where it has been made available by fedesoriano. This dataset was created by combining different datasets already available independently but not combined before. It contains a total of 918 observations taken from Cleveland, Switzerland, Hungary, and Long Beach.

B. Selected Attributes and their effects

The objective of our exploratory data analysis is to find the factors that majorly result in heart failures. We first start by finding these crucial factors which could result in heart failure. For this project, we have chosen the following attributes which can be used to predict heart diseases.

1) Age of the Patient
2) Gender of the Patient
3) Chest Pain Type experienced by the patient, the four chest pain types and their abbreviations are as follows

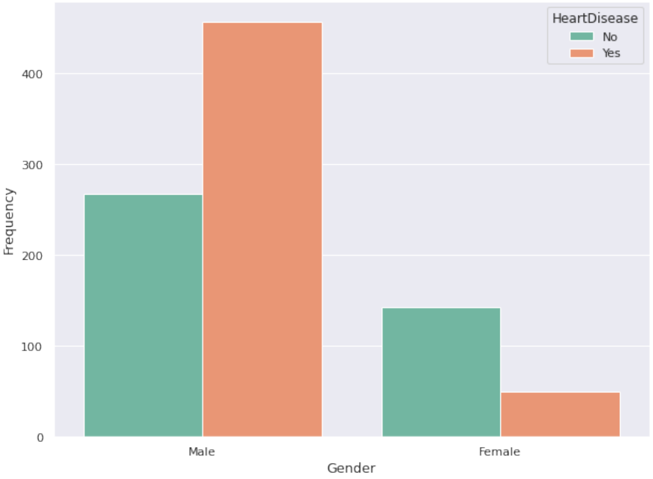

Fig. 1. Gender Count Plot

: Typical Angina - TA; Atypical Angina - ATA; Non-Anginal Pain - NAP; Asymptomatic - ASY

4) Resting Blood Pressure of the patient measured in mm/Hg
5) Cholesterol Level of the patient measured in mg/dl
6) Fasting Blood Sugar of the patient, considered high if it is above 120 mg/dl and low otherwise
7) Resting ECG of the patient which is classified into Normal, ST-T or Left Ventricular Hypertrophy (LVH)
8) Maximum Heart Rate of the patient measured in Beats Per Minute (BPM)
9) Exercise induced Angina experienced by the patient
10) Old Peak value of the patient which is the ST depression induced by exercise relative to rest
11) Slope of the peak exercise ST segment which can be classified into up sloping, flat and down sloping

Now that we have listed the attributes which affect heart failure, we can closely look at the effect of each of these attributes on heart failure. We can first analyse the effect of categorical variables such as Gender, Chest Pain Type, Fasting Blood Sugar, Resting ECG, Exercise Angina, and ST Slope.

1) Gender: We first start by checking if gender has any effect on heart failure. Do females have higher chances of having heart failure than men? To answer this, we can have a look at their count plot in Figure-1. We see that the percentage of men having heart failure is higher than the percentage of women having heart failure. Since our dataset contains more males than females, we can see that the peaks for males are higher than the peaks for females.

2) Chest Pain Type: We can ask the question, does the type of chest pain affect heart failure? Is any specific type of chest pain type indicative of risk of heart failure? Again we can look at the count plot to get a better insight into this. In figure-2 we can see that patients with chest pain of type-ASY have a very high probability of having heart failure, whereas patients experiencing Typical Anginal (TA) Chest Pain have a lower probability of having heart failure.

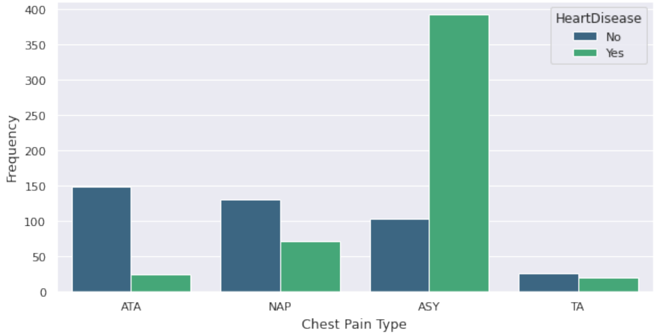

Fig. 2. Chest Pain Type Count Plot

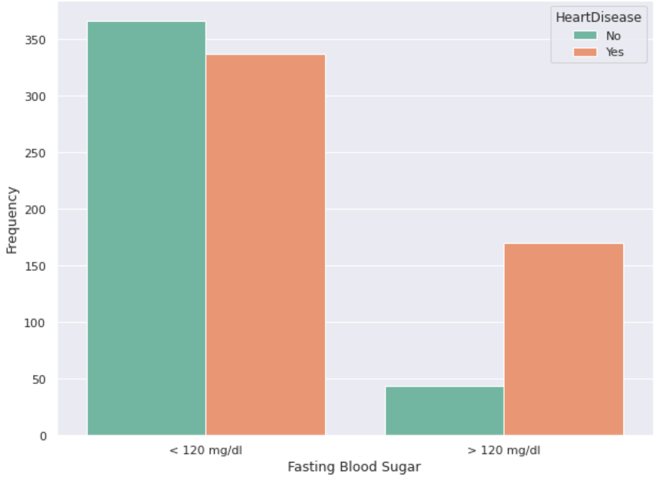

Fig. 3. Fasting Blood Sugar Count Plot

3) Fasting Blood Sugar: We might often think whether fasting blood sugar would have any effect if the patient has heart failure. To check if Fasting Blood Sugar Values have any effect on the patient having heart failure, we can have a look at their count plot in figure-3. We can see that patients having Fasting Blood Sugar levels above 120 mg/dl have a higher probability of having heart failure than patients having Fasting Blood Sugar levels below 120 mg/dl, which is intuitive.

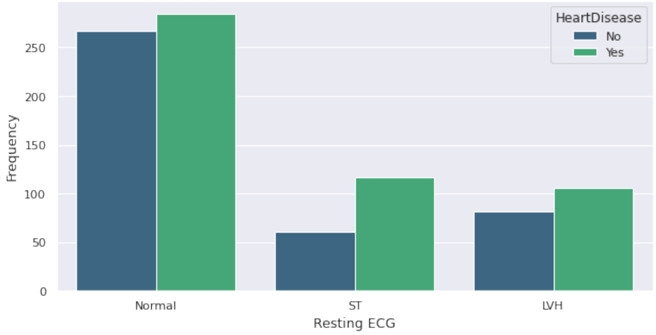

Fig. 4. Resting ECG Count Plot

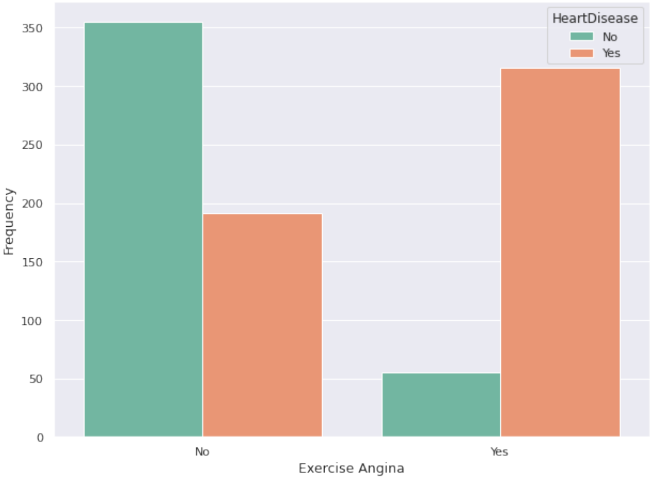

Fig. 5. Exercise induced Angina Count Plot

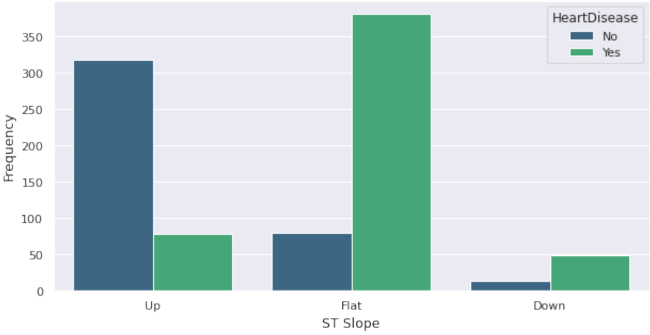

Fig. 6. Exercise induced Angina Count Plot

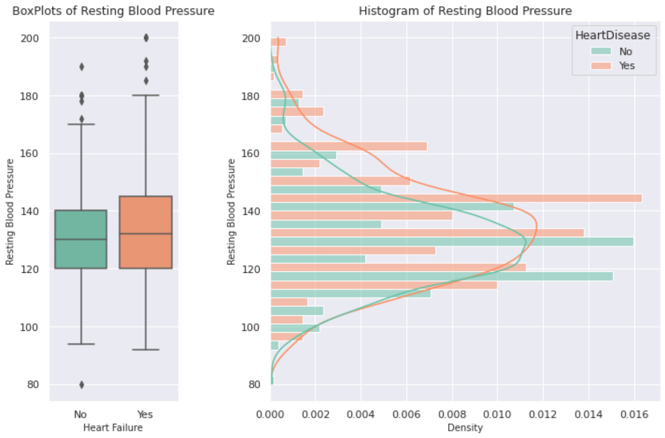

Fig. 7. Box and Histogram Plots of Resting Blood Pressure

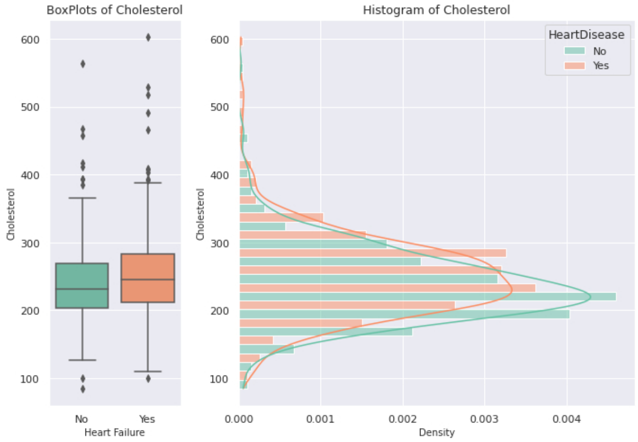

Fig. 8. Box and Histogram Plots of Cholesterol

4) Resting ECG: We can check if the resting ECG of the patient has any effect on them having heart failure. We can see in figure-4 that patients having ST-T and Left Ventricular Hypertrophy have a higher probability of having a heart failure than patients having normal resting ECG. We can see that even for Normal Resting ECG people have a high probability of having a heart failure (>0.5), i.e. a normal ECG might not necessarily mean a healthy heart.

5) Exercise induced Angina: Generally, one would think that exercise induced angina could possibly imply a potential heart failure. We can check if it is true by looking at its count plot in Figure-5. As one might have expected, patients having exercise induced angina stand at a higher risk of having heart failure than patients not having exercise induced angina.

6) ST Slope: We can ask whether the slope of the ST segment can show us any chances of the patient having heart failure. We see in Figure-24 patients having their ST Slope as upwards are at a lower risk of having heart failure than patients having their ST Slope as downwards or flat.

Now that we have covered the categorical variables, we can now see the distributions of continuous variables and how they affect heart failures. We can also check for correlation between any two continuous variables. For the attributes we have considered, the continuous variables are Resting Blood Pressure, Cholesterol, Maximum Heart Rate, Old Peak, and Age.

7) Resting Blood Pressure: We can start by asking the basic question, does resting blood pressure values of the patient affect the chances that the patient has heart disease? In figure-7 we see that there is a slight difference in the box plots of Resting Blood Pressure for patients with heart failure and patients without heart failure. We can see that the median value for Resting Blood Pressure is slightly higher for patients with heart failure; further, patients with heart failures have more outliers than patients without heart failure. This change can be further seen in their histogram. At higher blood pressures, the density of people having heart failure is higher than people not having heart failure.

8) Cholesterol: We have often heard that high cholesterol levels in a person often cause heart diseases. Let us see how well this statement really holds. In our dataset, there were around 170 missing cholesterol values which were marked as zero. For analysis, we drop these rows in our analysis. In figure-8 we see that similar to blood pressure, there is a slight difference in the box plots of Cholesterol for patients with

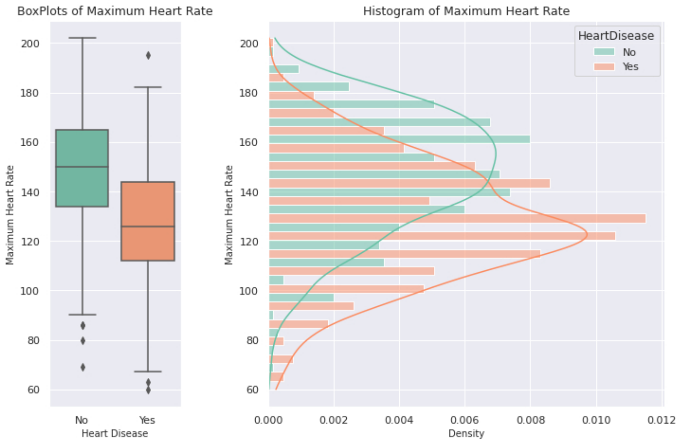

Fig. 9. Box and Histogram Plots of Maximum Heart Rate

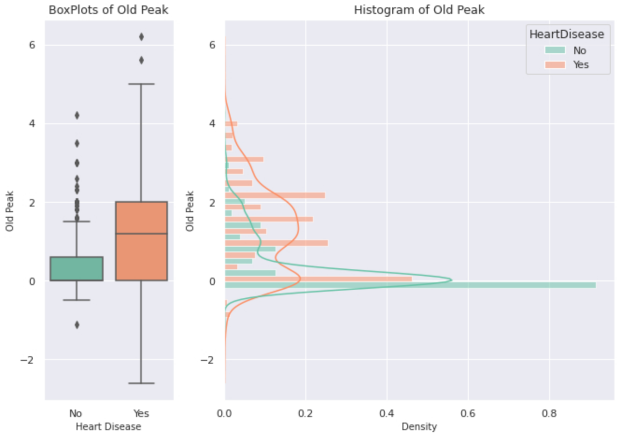

Fig. 10. Box and Histogram Plots of Old Peak

heart failure and patients without heart failure. The median value for Cholesterol is slightly higher for patients with heart failure. Further in their histogram, at higher cholesterol levels, the density of people having heart failure is slightly higher than people not having heart failure. However, broadly speaking, the cholesterol level distribution in both patients with and without heart failure is almost similar. Contrary to popular belief having high cholesterol levels does not significantly put you at a higher risk of heart failure.

9) Maximum Heart Rate: In figure-9 we see that patients with heart failures have a significantly lower maximum heart rate than patients with no heart failure. Even in the histogram, the distribution of maximum heart rate in heart failure patients is sharper and centered lower than the distribution of maximum heart rate in patients not having heart failure. The distribution for patients with heart failure is centered around 120 beats per minute, patients with a maximum heart rate lower than 120 beats per minute have a higher chance of having heart failure, while at the opposite end, patients having a maximum heart rate of above 150 have a lower chance of having a heart failure.

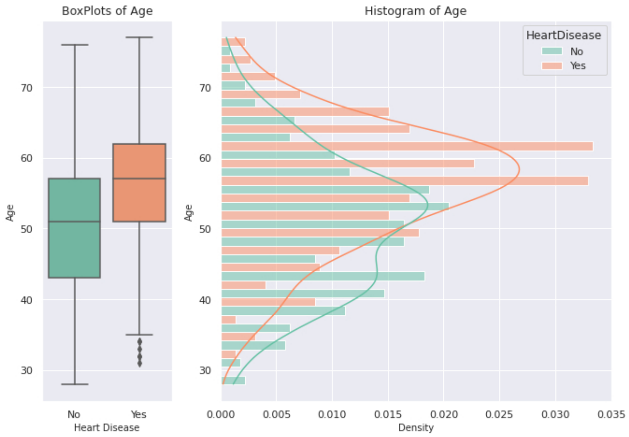

Fig. 11. Box and Histogram Plots of Age

10) Old Peak Values: Old Peak values generally measure the change caused due to exercise done by the patient. We can see that in figure-10 there is a contrast in the box plots of old peak values of patients with heart failures and the patients with no heart failure. The median old peak value is non-zero and higher for patients with heart failure, while the median Old peak value for patients without heart failure is zero. Even in the histogram, the distribution of old peak values in patients without heart failure is very sharp and centered at zero, while the distribution of old peak values in heart failure patients has two peaks, a sharper one near zero and another flat peak around 1.5. The distribution for patients with heart failure is skewed with a large interquartile range. We can state that patients having an old peak value near 0 have a low chance of having heart failure.

11) Age: Similar to cholesterol, we have often heard that aged patients are always at a higher risk of having heart failure. Is age an important factor in predicting heart disease? Is an aged person at a higher risk of getting heart failure? In figure-11 we see a contrast in the plots for patients with heart failure and patients without heart failure. The median age is quite higher for patients with heart failure. Further in their histogram, at higher ages, the density of people having heart failure is higher than people not having heart failure. The distribution of age in heart failure patients is sharper and centered at a higher age than the distribution of age in patients not having heart failure. Broadly speaking, age is an important factor in heart failure prediction and the finding is in accordance with the popular belief that aged people have a higher chance of having heart disease.

C. Relationship between continuous variables

We have seen how each individual attribute affects the chances of the patient having heart failure. We can now ask if there are underlying relationships among the continuous variables. Figure-12 shows the heat map of all the continuous variables. We can see that some pairs of continuous variables have almost zero correlation while others have a slightly higher correlation than others. The highest correlation is between

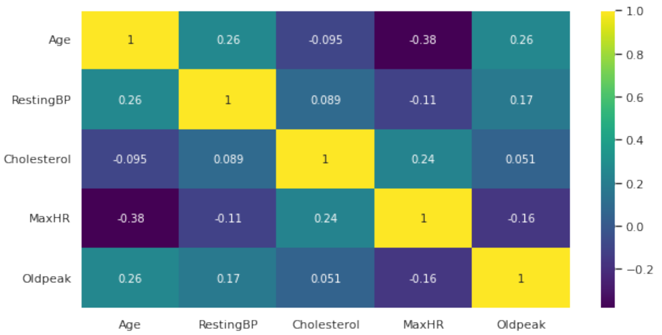

Fig. 12. Heat map Plot of Correlation

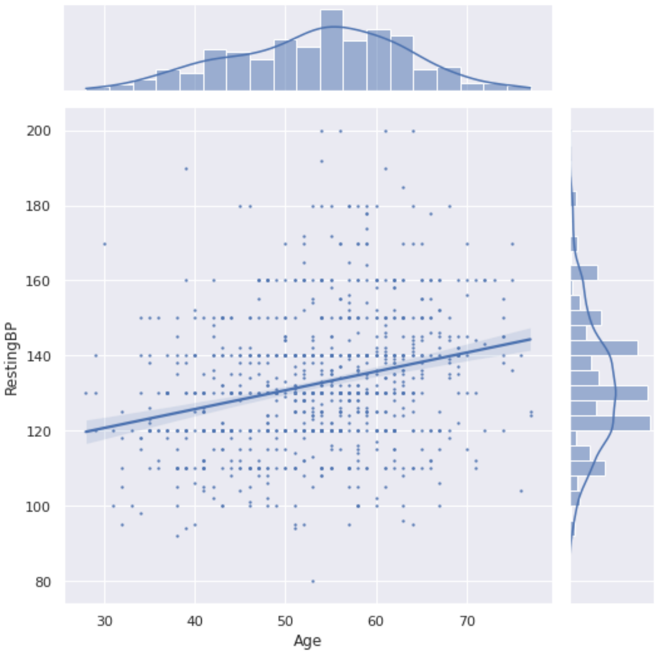

Fig. 13. Regression Plot of Age and Resting Blood Pressure

Age and Maximum Heart Rate variables, which is −0.38. Other variable pairs with higher correlation are Resting Blood Pressure and age, along with Cholesterol and Maximum Heart Rate.

We can see the relationship between Age and Resting Blood Pressure in Figure-13. We can see that aged patients have a higher resting blood pressure than young patients. Their combined effect can further be seen in figure-14, where we can see that patients with heart disease peak at a higher age and a higher resting blood pressure level.

We can also see the relationship between Age and Maximum Heart Rate in Figure-15. We can see a downward trend in the scatter plot; as the age of the patient increases, the maximum heart rate decreases. Further, in the scatter plot we see most of the patients with heart failure are clustered at higher ages and lower maximum heart rates, while patients without heart failure are clustered on the upper left with lower age and higher maximum heart rate. Their combined effect can

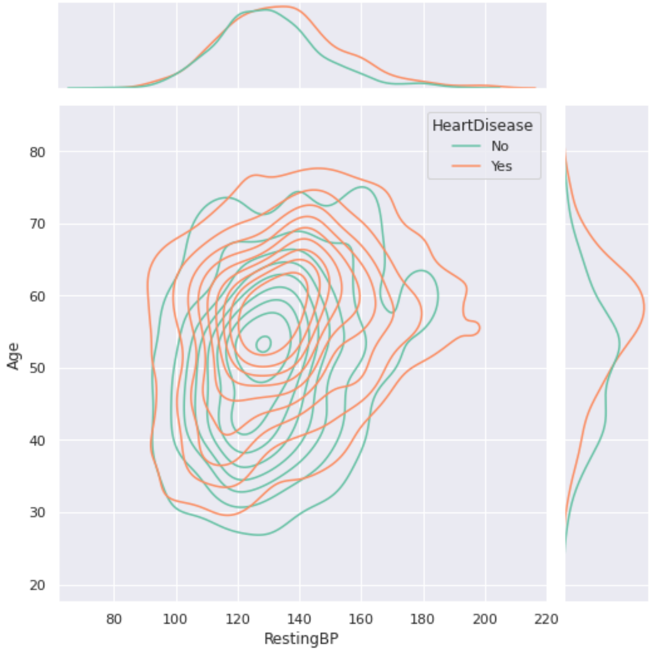

Fig. 14. KDE Plot of Age and Resting Blood Pressure

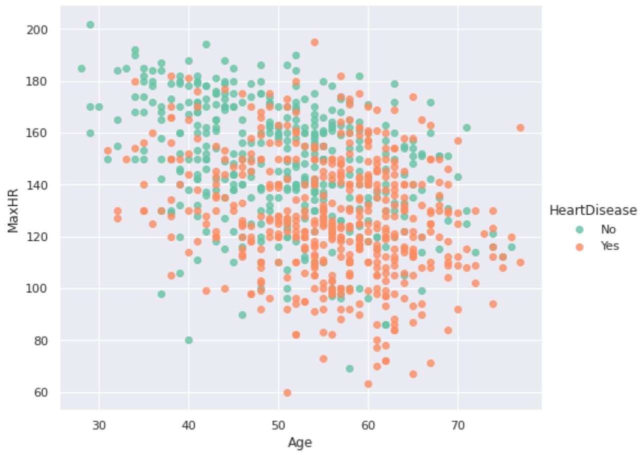

Fig. 15. Scatter plot illustrating Age and MaxHR effect on heart failure

further be seen in the KDE plot in figure-16, where we can see that patients with heart disease peak at a higher age and a lower maximum heart rate.

D. *PCA for visualization*

In order to get a visual sense of the data, we have used Principal Component Analysis to reduce the total number of dimensions to 2, to enable the plotting of data on a Cartesian plane. However, a significant amount of variance has been lost. The first principal component contains 21.66%, while the second principal component contains 10.40% of the variance, leading to capturing 32.06% of the original variance. Although PCA does not work well with categorical variables, the results can be considered a good approximation for visualization pur-

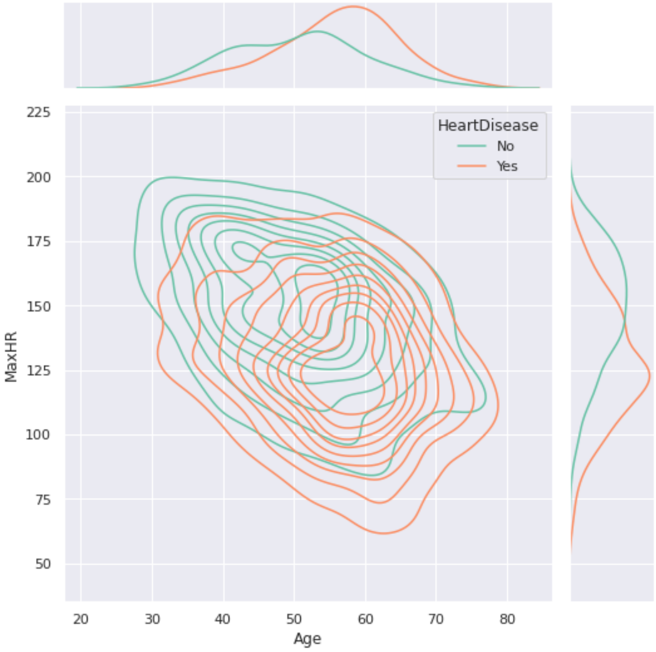

Fig. 16. KDE Plot of Age and Maximum Heart Rate

poses. 6/11 of the attributes used in the data were categorical. PCA cannot be used directly on categorical variables, hence One Hot Encoding was performed using a pandas built-in function. The resulting plot has been shown in figure- 17, where we can see a clear distinction between people who suffered heart failure and those who did not.

We have also visually demonstrated the decision boundaries of common classification algorithms such as Kernelized SVM, Logistic Regression, Naive Bayes and Random Forest. We have trained the models on the principal components obtained and plotted the decision boundary. The differences in the algorithms have been illustrated in figure- 18. Note that these models have not been tuned and have only been used for illustrative purposes.

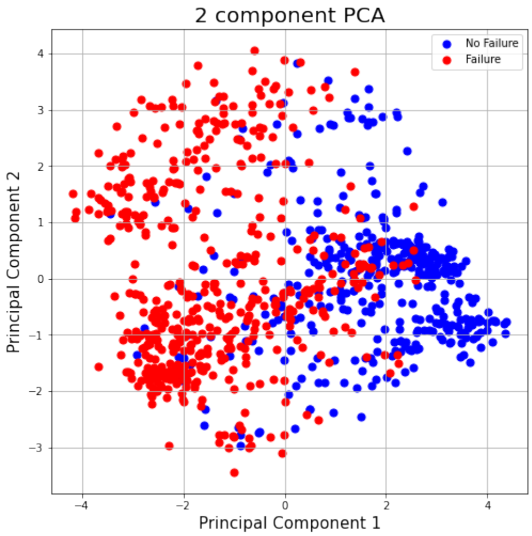

Fig. 17. Visualization using PCA

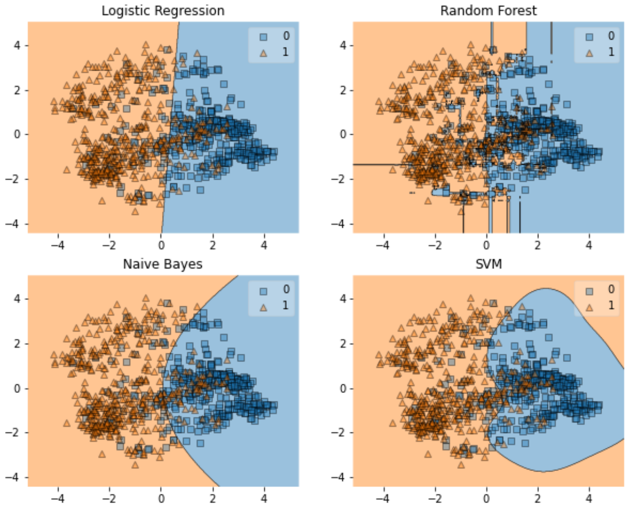

Fig. 18. Decision Boundaries Visualization

E. Hypothesis Testing

During our initial Exploratory Data Analysis, we observed that the maximum heart rate of people declined with age. This can be seen in figure- 15. We wished to verify this using formal hypothesis testing. Our objective was to show that people with age greater than 50 years have a different distribution of maximum heart rate compared to people with age less than 50 years. For this, we attempted to model the distribution of max HR of smaller age group as a normal distribution and a t distribution with 290 degrees of freedom (Sample size is 291). Both of these distributions have approximately the same density function and are fairly close to the actual distribution, as seen in figure- 20.

We have framed the null hypothesis and alternate hypotheses as follows. We have used a z test while assuming that population variance is the same as sample variance since the sample size is large. Let the population mean of HR of people with age less than 50 be $\mu_0 = 148.89$. Let the mean HR for people with age > 50 be μ

Null Hypothesis : $\mu_0 = \mu$

Alternate Hypothesis : $\mu_0 > \mu$

$\alpha = 0.05$ (threshold)

The p value obtained was 0.23, which is greater than threshold value chosen. Hence we fail to reject the null hypothesis.

Even though it seems apparent from the plot, **we have failed to conclusively show that the maximum heart rate reduces with age**, based on the data currently at our disposal.

IV. Experiments and Results

With the technological advancements day by day it is easier for us to predict using the dataset whether a particular

TABLE 1

Summary of Algorithms used

Algorithm	Hyperparameters	Accuracy (on Test Data)
Support Vector Classifier	C = 2000 and kernel = rbf	0.818
Decision Tree Classifier	ccpalpha = 0, Criterion = gini and maxfeatures = $log2$	0.79
Random Forest Classifier	ccpalpha = 0, Criterion = gini and maxfeatures = $log2$	0.82
MLP Classifier	activation = tanh, hidden layers = (9,5,2), learning rate initial= 0.01,and solver = Adam	0.76

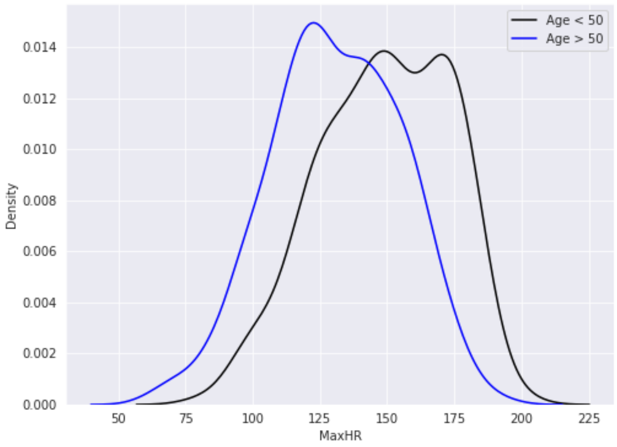

Fig. 19. Maximum Heart Rate for 2 age groups

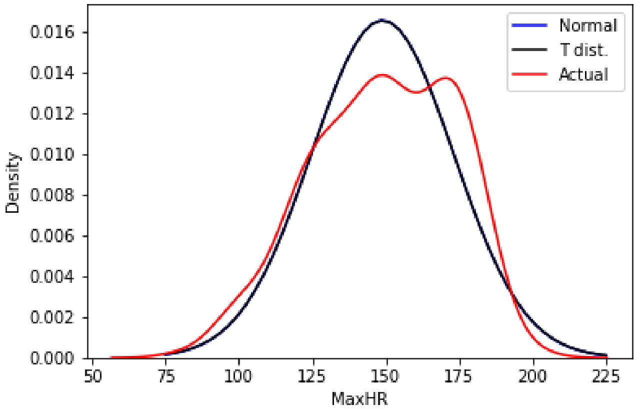

Fig. 20. Modelling distribution of heart rate

person has a risk of heart failure or not.We used different classifiers like Support Vector Machine Classifier or Decision Tree Classifier and then compared our results to get the best model.

A. *Choice of Machine Learning Frameworks*

After careful deliberation, we chose 4 classification Algorithms for the task at hand:

1) Support Vector Machine
2) Decision Tree
3) Random Forest
4) Multi Layer Perceptron

Support vector Machine: A supervised machine learning algorithm that can be used for both classification and regression challenges. In this SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features or attributes you have). Then, we perform classification to find the exact hyper-plane that differentiates the two classes very well and maximizing the distances between the nearest data point (either class), and the hyper-plane. This optimization will help us to decide the right hyper-plane. The two main hyperparameters are C and the kernel type.

Decision Tree Algorithm: A flowchart-like structure in which each internal node represents a test on a feature, each leaf node represents a class label, and branches represent conjunctions of features that lead to those class labels. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. The three main hyperparameters are ccpalpha, criterion, and max features.

Random Forest Algorithm: A classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the accuracy. Instead of relying on just one decision tree, the random forest takes the prediction from each tree and, based on the majority votes of predictions, predicts the final output. Greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting with the training data. The three main hyperparameters are ccpalpha, criterion, and max features.

Multi Layer Perceptron: A supervised learning algorithm that learns a function by training on a dataset. Given a set of features and a target, it can learn a non-linear function approximator for classification. It is different from logistic regression, as between the input and the output layer, and there can be multiple non-linear layers, called hidden layers. The four main hyperparameters are activation, learning rate, the number of hidden layers, and the number of nodes in each hidden layer.

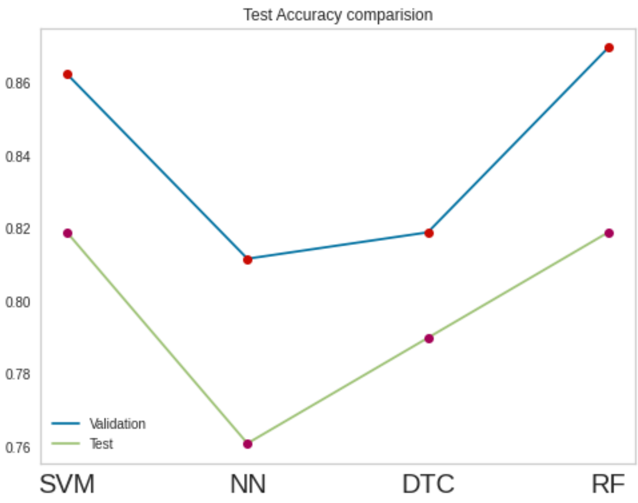

Fig. 21. Accuracies obtained by models using sklearn

were selected using grid search techniques are also stated.

Table 2 summarizes the top 5 models available on PyCaret based on their F1 scores along with the scores obtained on the dataset related to the current task on hand. We created a blended model using these top 5 models to get better performance on the dataset.

We also trained a simple neural network using **Pytorch** with three different linear layers and four different activations after each layer, like PReLU, Sigmoid, LogSoftmax, and ReLU for 80, 100, 120, 140 epochs. We tested with two different types of losses: NLL loss, which has its final activation function as LogSoftMax, and Cross Entropy Loss (BCE Loss), which has its final activation function as Sigmoid. Further evaluation in the fall of the losses is done as the model is trained for many epochs. The training and validation accuracy and loss curves can be seen in Figures 22 and 23.

B. Results

The overall pipeline for the classification task consists of a train-validation-test split (70:15:15) where the validation set was used to check for overfitting while training. We were getting an average F1 score of **0.877** which was more tha n what was obtained by the **Neural Network** built using Pytorch (**0.833**)

V. Conclusions and Final Result

In a medical application such as this, false negatives are much more problematic than false positives. If the model falsely predicts that a person is at risk of heart failure, it would not harm the person to start taking healthy and preventive measures. However, a person who is actually at risk of heart failure but whose risk is undetected will not be made aware

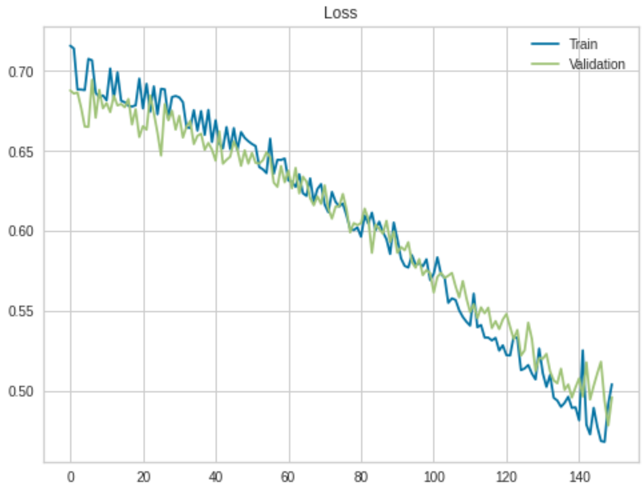

Fig. 22. Loss obtained while training torch model

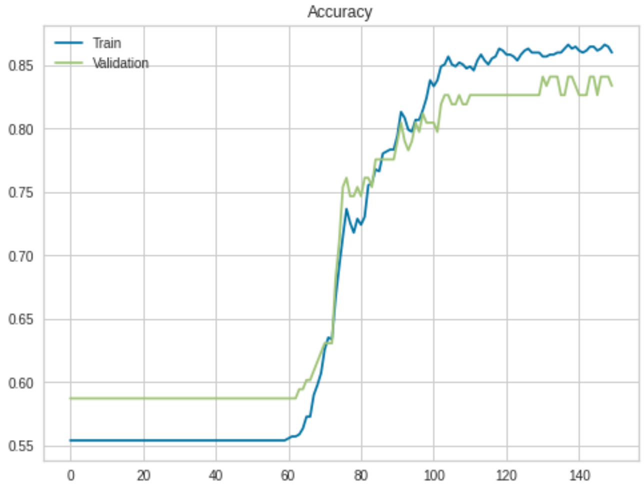

Fig. 23. Accuracy obtained while training torch model

of the danger. This should be kept in mind while preparing future models.

We came across an interesting case of possible misinterpretation of data. While doing Exploratory Data Analysis, we remarked that the people having asymptomatic, i.e. ASY type of chest pain have highest fraction of patients suffering heart failure. This might imply that a person with no chest pain is at a higher risk of heart failure than a person having chest pain. However, after some scrutinizing, we came to the following conclusion : A person who suffers chest pain will get himself diagnosed, and start taking measures to improve his health. On the other hand, a person with no chest pain will not discover whether he is at risk until he gets a general checkup. Hence he is less likely to improve his health and hence more likely to suffer heart failure.

Finally, after evaluating the various models listed above, When run on test data,we get the final results, an accuracy of **0.8406** and an F1 score of **0.8608**

References

[1] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [1 November 2021] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[2] Galarnyk, Michael. 'PCA Using Python (Scikit-Learn)'. Medium, 17 Nov. 2021, https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60.

[3] Raschka, Sebastian. 'MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack'. Journal of Open Source Software, vol. 3, no. 24, Apr. 2018, p. 638. DOI.org (Crossref), https://doi.org/10.21105/joss.00638.

[4] 'Hypothesis Testing'. Statistics How To, https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/. Accessed 22 Nov. 2021.

[5] '1.17. Neural Network Models (Supervised)'. Scikit-Learn, https://scikit-learn/stable/modules/neural_networks_supervised.html. Accessed 24 Nov. 2021.

[6] Harris, C.R. et al., 2020. Array programming with NumPy. Nature, 585, pp.357–362.

[7] Virtanen, P. et al., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, pp.261–272.

[8] McKinney, W. others, 2010. Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference. pp. 51–56.

[9] Van Rossum, G., 2020. The Python Library Reference, release 3.8.2, Python Software Foundation.

[10] Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in science amp; engineering, 9(3), pp.90–95.

[11] Waskom, M. et al., 2017. mwaskom/seaborn: v0.8.1 (September 2017), Zenodo. Available at: https://doi.org/10.5281/zenodo.883859.

[12] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.

[13] Seabold, S. Perktold, J., 2010. statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference.

[14] Bharti, Rohit, et al. "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning." Computational Intelligence and Neuroscience, vol. 2021, July 2021, p. e8387680. www.hindawi.com, https://doi.org/10.1155/2021/8387680.

[15] Chicco, Davide, and Giuseppe Jurman. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone." BMC Medical Informatics and Decision Making, vol. 20, no. 1, Feb. 2020, p. 16. BioMed Central, https://doi.org/10.1186/s12911-020-1023-5.

[16] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York heart association classification from clinical notes," in Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.

[17] Role-Type Classification. https://bolt.mph.ufl.edu/6050-6052/unit-1/role-type-classification/. Accessed 25 Nov. 2021.

[18] Finkelhor, R. S., et al. 'The ST Segment/Heart Rate Slope as a Predictor of Coronary Artery Disease: Comparison with Quantitative Thallium Imaging and Conventional ST Segment Criteria'. American Heart Journal, vol. 112, no. 2, Aug. 1986, pp. 296–304. PubMed, https://doi.org/10.1016/0002-8703(86)90265-6.