Prevention of Heart Disease Through Early Detection

By: Omesha Prashanthika Samarakoon

Supervisor: Dr. Mykola Gordovskyy

Background

- The Heart is the one of most important organs for sustaining life since it pumps blood to all areas and vital organs in the body to supply oxygen and nutrients continuously whilst removing carbon dioxide via the lungs to keep the body functioning well. Therefore a normal and healthy life goes hand in hand with having a well functioning heart free of disease.
- Heart disease is one of the main causes of death in the world, with mortality rates reaching approximately 17.9 million annually across the globe (WHO,2023)
- According to a report compiled by the European Society of Cardiology (ESC), 26.5 million adults were already diagnosed as positive for heart disease, and annually identify 3.8 million were effected by heart disease. But unfortunately, 50-55% of heart disease patients lost their lives in the initial 1-3 years (Muhammad, Tahir, Hayat, and Chong, 2020).
- Early-stage detection of Heart disease significantly impacts decreasing the effect of this dangerous condition. Most existing methods of investigation such as angiography have been found to be complex, expensive, and time-consuming due to human error and the length of time taken during an assessment of the condition (Muhammad, Tahir, Hayat, and Chong, 2020).
- For this effort, machine learning is one of the suitable approaches, Because Machine learning is fine appropriate to handle complex data which comes through different data sources and a massive range of variables With permission from the limitations of the study, Machine learning is ingenious to catch and show the patterns hidden in the data.

Aims

- Investigate and identify the metrics with the highest level of correlation when it comes to predicting the early onset of heart disease
- Find machine learning algorithms best suited to predict heart disease with the highest possible accuracy utilizing the metrics the most appropriate metrics that have been deduced (above).
- Find the percentage of patients who are positive and negative according to the effect of each metric in the dataset.

Objectives

- Calculate the correlation between predictive variables and target variables and analyze the relationship between these in order to identify the variables that provide the strongest indicators in relation to predicting the increase in the risk of having heart disease.
- Implement a Machine Learning (ML) model using the most appropriate ML algorithms to help predict with a high level of accuracy, the possibility of contracting heart disease.
- Compare the relative performance of the machine learning model developed in this project with existing machine learning models used to achieve the same goal of predicting heart disease early.

Methodology

- Select a data set from Kaggle which is relevant for heart disease prediction. The one found to date which is most appropriate is the “Heart Failure Prediction Dataset”.
- Do the basic preprocessing checking null values, missing values, and duplicate values.
- Visualize the data according to each of the metrics to identify the distribution.
- Remove outliers from the data set to provide more accurate predictions when used with training algorithms.
- Calculate the correlation between variables and create violin plots to identify the relationship between predictive variables and target variables
- Visualize the categorical variables vs target variables to identify the relationship via bar plots
- To have a more accurate analysis use logistic regression to feature selection according to the Coefficient and odds ratio
- To apply logistic regression algorithms categorical variables, convert to numerical variables by using encoding
- Encode nominal variables, use the pandas dummy method to encode ordinal variables utilizing label encoding

Literature Review

Paper	Classifier	Dataset	Result
• **Early and accurate detection and diagnosis of heart disease using intelligent computational model (Muhammad, Tahir, M., Hayat, and Chong, 2020).**	• K-Nearest Neighbors • Decision Tree • Extra-Tree Classifier • Random Forest • Logistic Regression • Naïve Bayes • Artificial Neural Network • Support Vector Machine • Adaboost • Gradient Boosting	Cleveland and Hungarian heart disease datasets from UCI machine learning repository.	This research compares prediction accuracy using the full dataset with all features and dataset under feature selection so it gives different accuracy levels, high accuracy given by Extra-Tree Classifier(ET) and Gradient Boosting (GB) With all features, ET- 92.09% GB- 91.34% With feature selection, ET- 94.41% GB- 93.36%
• **Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison (Ali, Paul, Ahmed, Bui, Quinn and Moni 2021)**	• Multilayer perceptron(MP) • K-nearest neighbours (KNN), • Random forest(RF) • Decision tree(DT) • Logistic regression(LR) • AdaboostM1 (ABM1)	Heart disease dataset from Kaggle which has 14 attributes	High accuracy, sensitivity and specificity given by RF method and achieved 100%
• **Heart Disease Prediction Using Different Machine Learning Algorithms (Bhowmick, Mahato, Azad and Kumar, 2022)**	• Decision tree (DT) • Random forest (RF) • Logistic regression (LR)	Heart disease individuals data set from the Cleveland database of the UCI repository	Comparison of prediction accuracy of algorithms approximately, DT- 95% RF- 92% LR- 87% The DT algorithm gives the highest accuracy it exactly 94.7%

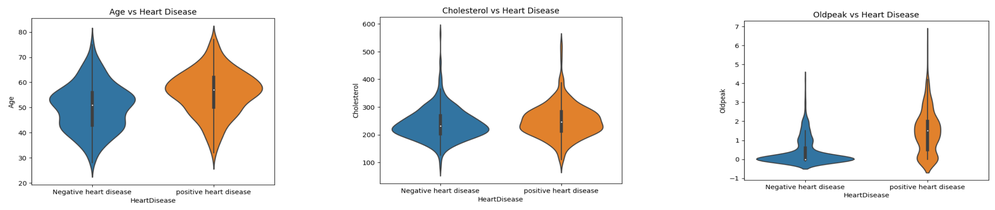
Analysis of Numerical Data

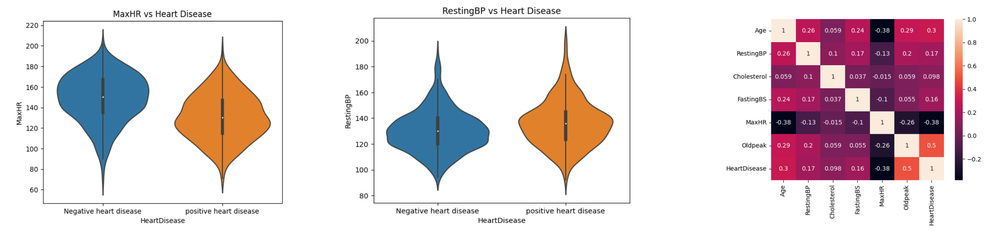
•Age: Having heart disease risk worsens with age and shows positive correlation and positive mean around 55-60 and Negative mean around 45-50

• Cholesterol: Cannot identify clear correlation on violin plot

• Max Heart Rate: It shows a negative correlation with heart disease when maximum heart rate is low and below 120, the risk of being positive in heart diseases getting high

•Old Peak: Old Peak is around ‘0’ there is no risk of having heart disease and it is near to ‘2’ the positivity of getting heart disease is the high, old peak, and the presence of heart disease have a positive correlation





Analysis of Categorical Data

•Chest Pain type: we can identify the ‘ASY’- Asymptomatic chest pain type is more affected type in positive heart disease if the patient has Asymptomatic chest pain he has high risk than others

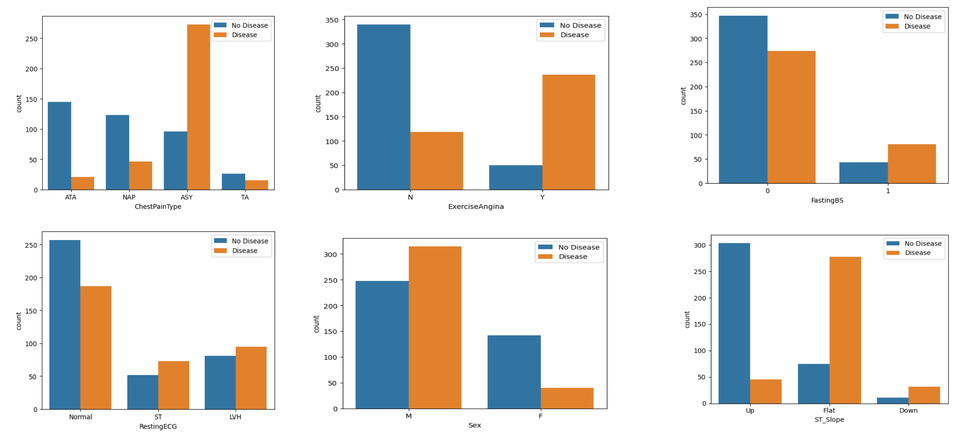
•Exercise Angina: If the patient has exercise angina the probability of having heart disease is high

•Blood Sugar: If patient has high blood sugar the probability of having heart disease is higher than non-diabetics patient

• Resting ECG: If the patient has ST or LHV, ECG report he has a high probability of having heart diseases than a patient who has a Normal ECG report.

•Sex: Males have a high probability of having heart diseases than female

•ST Slope: If the patient has a Flat ST Slope the chance of positive heart Disease patient is higher than Up and Down



Logistic Regression to identify the relationship using coefficient

- According to the odds ratio, Age, Resting BP, Cholesterol, fasting blood sugar, old peak, sex, Exercise angina and Resting ECG/LVH have a positive relationship with target variable heart diseases.
- Age, RestingBP, Cholesterol, MaxHR, old peak, ST Slope, Sex, Exercise Angina and chest pain type shows a strong relationship with the target variable which is positive or negative heart diseases. Some of them show positive relationships and some of them show negative relationships according to the coefficient which calculates by logistic regression

Variable Name	Coefficient	Odds Ratio
Age	1.0826	2.9523
Chest Pain Type	-0.5911	0.5537
Resting BP	0.6808	1.9755
Cholesterol	0.5036	1.6547
Fasting BS	0.1918	1.2115
MaxHR	-0.8417	0.4309
Oldpeak	1.4095	4.0939
ST_Slope	-1.8888	0.1512
Sex	1.4589	4.3012
ExerciseAngina	1.1896	3.2857
LVH	0.2928	1.3401
RestingECG/Normal	-0.2252	0.7984
RestingECG/ST	-0.0671	0.9351

Reference

- Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M. and Moni, M.A., 2021. Heart disease prediction using
- supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine, 136,* p.104672.
- Bhowmick, A., Mahato, K.D., Azad, C. and Kumar, U., 2022, June. Heart Disease Prediction Using Different Machine Learning Algorithms. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 60-65). IEEE.
- Muhammad, Y., Tahir, M., Hayat, M. and Chong, K.T., 2020. Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific reports, 10*(1), p.19747.
- World health organization, 2023, Cardiovascular diseases, Available at:

https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1