



PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS IN THE DIAGNOSIS AND CLASSIFICATION OF HEART DISEASES

Olanloye Odunayo¹, Olawumi Olasunkanmi², Adetoye Adeyemo³, Adebayo Segun⁴

^{1,3,4} **Department of Computer Science, Bowen University, Iwo**

² **Department of Computer Science, University Of North Carolina, USA**

ABSTRACT

Clinical reports and research have established that heart diseases are a typical example of cardiovascular disease that has sent millions of people globally to an untimely grave. World Health Organization (WHO) also confirmed this assertion and as a result, there have been series of attempts by researchers in various fields to solve this problem. Certain researchers in computer and health informatics carried out predictive analytics to detect and classify the disease based on several biomarkers identified in the affected individual. Meanwhile, enough has not been done to determine the level of susceptibility of individuals to heart diseases with concerted effort on the key indicators such as age, sex, sugar level and some other related attributes before predictive analytics are made. This explores the attribute and it was finally established that sex, age, level of cholesterol etc. are strong markers to determining the level of susceptible of patient to heart disease. Moreover, Four ML models - KNN, NB, SVM and RF were implemented and evaluated in term of their performances in the classification of heart diseases using cross-validation and test dataset. At first, with every feature available in the dataset and later with only the correlated features identified in the descriptive analytics. It was established that accuracy improves across all models when only correlated features were used and SVM exhibits the highest accuracy and F1 Score (84%). Therefore, SVM performs better than KNN, RF and NB when all the models were evaluated on the 25% test set of the correlated features. It could be therefore concluded that in-depth understanding of features for identification of strong disease biomarkers will enhance more accurate diagnostics and this in turn will be of great assistance to the medical practitioners and other stake holders to track susceptibility of individuals with identified features to heart disease

Keywords: Heart, Cardiovascular, K-Nearest Neighbors, Support Vector Machine, Naive Bayes, Random Forest

1.0 INTRODUCTION

Cardiovascular diseases are described as the type of diseases that has to do with abnormality in the function of the heart and blood vessels. According to World Health Organization (2018), CVD includes Coronary heart diseases, Cerebrovascular diseases, arterial diseases, Rheumatic heart disease, Congenital heart diseases, Deep vein thrombosis and pulmonary embolism to mention few CVD are number one cause of death globally. More people die annually from CVD

than from any other cause. An estimated 17.9 million people died from CVD in 2016, 31% of a global death out of which 85% are due to heart diseases most especially, stroke.

Series of efforts are being made by individuals, parastatal, industries, government at all level to arrest this ugly situation. Researchers from various fields most especially in the area of science and technology are not left out in the struggle. Artificial Intelligence

experts are also contributing their quota using Machine Learning (ML) tools and technologies.

Machine Learning is an aspect of AI that deals with training computer machine with large amount of data to enable it acquire sufficient knowledge that can be used to solve some fundamental problems. It emerged due to the increase in amount of data made available by current technological development. Machine Learning makes it possible for machine to acquire knowledge from massive amount of data, which is very heavy and sometimes impossible for man to analyze (Youness and Mohamed, 2019).

A lot of researches are going on to determine various risk factors associated with heart diseases in patients with ML algorithms. Some used statistical approach and data mining approach. Researchers made use of several data mining technique to help the specialist or physicians and provide relevant diagnostic information (Avinash and Pavan, 2019). Some available ML algorithm used includes Decision Tree, K-Nearest Neighbour (KNN), Self-Organizing Map (SOM), Neural Network (NN), Support Vector Machine (SVM) etc. Machine Learning algorithms plays a vital role in managing huge amount of health care data and improving the quality of health care services offered to patients (Alarsan and Younes, 2019).

1.1 STATEMENT OF THE PROBLEM

Heart diseases is a typical example of cardiovascular disease that has sent millions of people globally to an untimely grave. According to WHO, there has been series of attempts by researchers in various fields to solve this problem. Computer scientists and in fact AI experts are not left out as they are putting all efforts to contribute their own quota in combating this deadly disease. What then is the way forward? Despite a lot of researches going on to predict this deadly disease, enough efforts have not been made to analyze the disease dataset to understand the level of susceptibility or vulnerability of humans to this deadly disease before the predictive analysis and also vast majority of the predictive models focus solely on accuracy score for evaluating the performance of ML models in the classification and prediction of heart disease.

1.2 OBJECTIVES/SUMMARY

- i. To download the heart diseases data-set and pre-process it.
- ii. To analyze the data in other to determine the level of susceptibility of humans to heart disease based on the available features.
- iii. To split into train and test set and randomly split the trainset into train and validation set using K-Fold Cross Validation:

- The entire disease dataset.
 - The aggregated dataset with features of importance identifiable in (ii)
- iv. To implement 4 ML (SVM, FT, KNN) models with the trainset, and make prediction/classification with the validation and test dataset using the dataset in (iii)a and (iii)b.
 - v. To compare the performance of the models implemented for binary classification of patients into (a) heart diseased or (b) healthy

2.0 LITERATURE REVIEW

Shashikant and Ashok (2012) presented a research work on diagnosis of heart disease using machine learning algorithm. India centric data was used for the diagnosis. The data was classified and the accuracy sensitivity and specificity valued were obtained. The researcher shows that SVM with sequential minimum optimization learning algorithm is better.

Alarsan and Younes (2019) published a research work titled Analysis and Classification of heart disease using heartbeat features. The research made use of ECG (Electrocardiogram) classification with machine learning approach. Multiple Classifiers were used for ECG classification with each of them influencing the final decision according to its performance on the training data. The data was divided into two groups – training and testing data. It was finally concluded that the proposed system is able to act as a helping tool to aid the cardiologist in reading the ECG heart signals and to know more about it.

Avinash and Pavan (2019) were able to use ML technique to predict heart disease. The authors were able to study various classification algorithms that can be used for classification of heart diseases. The algorithm used includes K-means clustering, Adaboost, K- nearness neighbour (KNN), decision trees. The research work presented the level of accuracy of each of the algorithm.

Youness and Mohamed (2019) made use of swarm optimization and Ant Colony optimization in the prediction and classification of heart disease. A feature selection process as a pre-treatment step to machine learning was adopted for the purpose of size reduction, elimination of unresolved data, increase learning accuracy and improves understanding of result. Five algorithms were used for classification – SVM, K-nearest neighbour, Random Forest, ANN and later optimized using Ant Colony Optimization (ACO). The two experimental research were compared and it was established that the optimized classifier performed better than the one without classifier. The performance parameters among others includes: Mean absolute

error, Root mean squared error, Relative absolute error, Root Relative Squared error.

Beulah and Caroline (2018) published a research article titled improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. In this research work, a comparative analysis was done to determine how ensemble classifiers could be used to improve the prediction accuracy of heart disease. The data set consist of 14 attributes and 303 instances. There are 8 categorical attribute and 6 numerical attributes. The data set was divided into train and test group. The ensemble algorithm bagging, boosting, stacking and majority voting were employed in the experiment. Bagging produced an improvement of 6.92% in the accuracy. When boosting was used, there was an improvement of 5.94%; with majority voting, the accuracy improved by 7.26% and stacked brought an improvement of 6.93%.

Shamilulush and Aldabergen (2018), presented a performance analysis on four supervised machine learning algorithm K nearest neighbour, decision trees, Naive Bayes and logistic regression. This research work established the fact that logistic regression and Naïve Bayes method perform relatively better with over 80% level of accuracy. The data was obtained from UCI machine learning repository with 425 data sets, with 4 attributes and 300 instances. As part of the process 70% of the data was used as training data and the remaining 30% as testing data.

Rajesh et al (2018) were able to predicts heart disease using machine learning algorithms- Bayes algorithm and decision tree. It was finally concluded that if the data set is properly cleared, Naïve Bayes is more accurate when the data set is small but when the dataset is large, decision tree is more accurate.

Amin UI et al, (2018) were able to use an hybridized intelligent system with ML algorithm for predicting heart disease. In the proposed system, the researchers developed a ML based diagnosis system using heart disease dataset. Seven ML algorithms were used. The system can identify people with heart disease and those that are healthy.

3.0 METHODOLOGY

The heart disease data set used for this research was downloaded from UCI repository and the following operations were performed to achieve the stated objectives:

a. Data Preprocessing

Though there are no missing values across features of the dataset, the basic statistics revealed that the dataset contained over 700 duplicate entries.

b. Exploratory Analysis

For appropriate insights on the susceptibility of patients across various age range, gender and other important attribute features, to the disease, exploratory analysis was done and using Python Seaborn and Pandas library

c. Feature Selection

To measure the strength of the relationship between all variables, most importantly, the relationship of other features with the target variable, Pearson Correlation coefficient was used. The Correlation coefficient indicates the probability of patient's vulnerability to heart diseases as a result of some features. Features with high absolute coefficient contributes significantly to likelihood of a patient having the disease or not

d. Feature Scaling

Since majority of the Algorithms to be used are based on distance, features with bigger values might have a greater influence over others in the outcome. This necessitates feature scaling and minimax scaler was used. The real values such as Age, Maximum Heart Rate and Total resting BP were normalized as follows:

$$\begin{aligned} \text{Let } X &= \text{feature to be scaled} \\ &\text{for } i \text{ in range}(\text{length}((X))) \\ X_std &= (X - X.\text{min}) / (X.\text{max} - X.\text{min}) \\ X_scaled &= X_std * (\text{max} - \text{min}) + \text{min} \end{aligned}$$

e. Choice of Dataset Distribution

As part of the modality used in this research, attempt was made split the entire dataset of all features to 75% train and 25% test set in which the train set was validated using a k-fold cross validation. Afterwards, dataset of at least 1% correlated features were and cross validated for model training.

f. Model Implementation

Four models viz Support Vector Machine, K Nearest Neighbour, Random Forest and Naive Bayes Classifier were implemented on the dataset of all features and that of corelated features. The entire implementation was done using Python Programming Language and the overall workflow of the research is shown in Fig 1 below

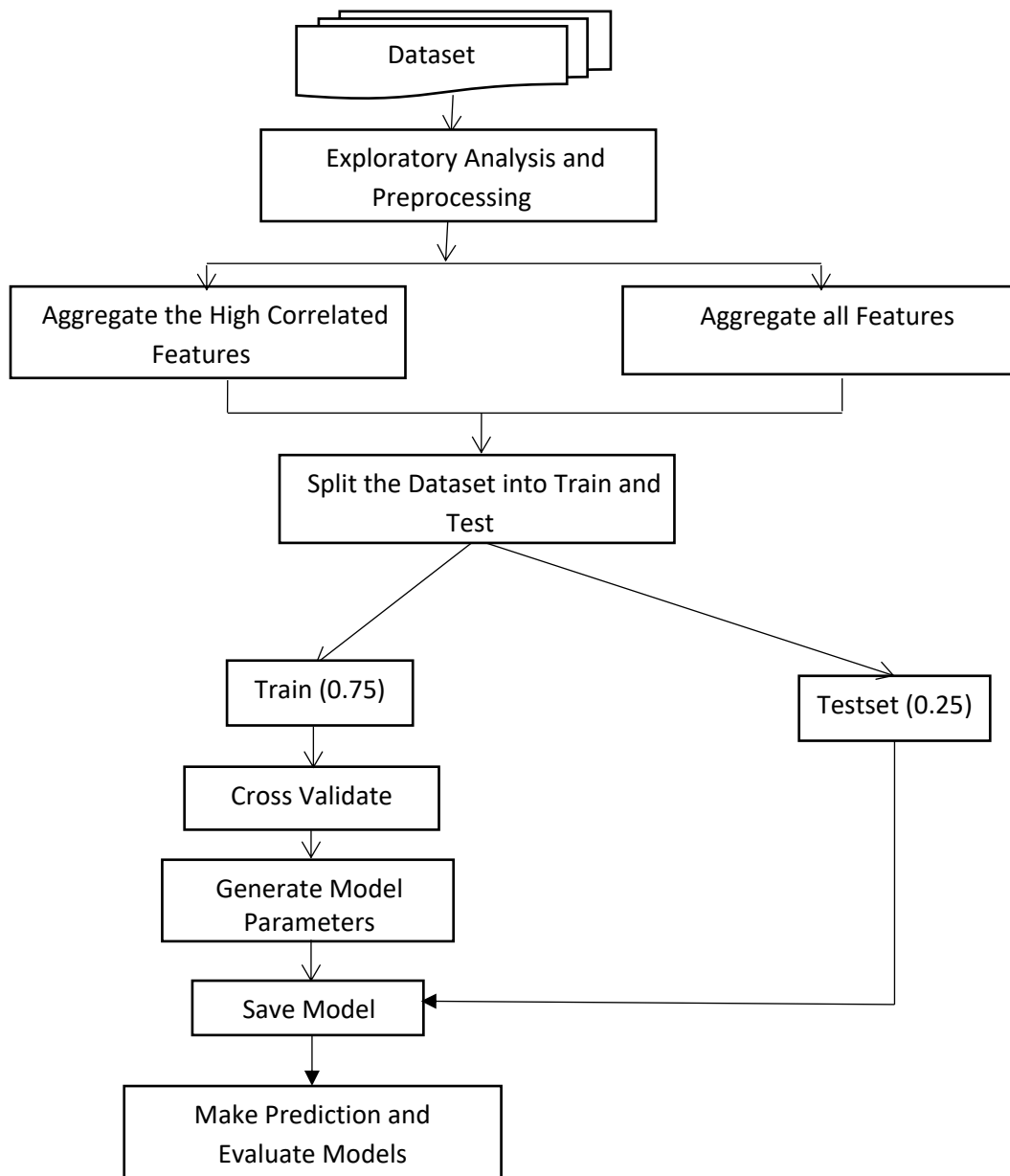


Fig 1: Research Workflow

4.0 ANALYSIS, RESULTS AND DISCUSSION

Exploratory Data Analysis

The dataset used in this study was carefully examined to detect anomalies, standardized data and accumulate insights about the dataset as follows:

Understanding the Features: The features in the heart disease dataset, their meaning and data types are

depicted in table 1 below. The Real values comprises of Integer/Float numbers; Boolean values evaluate to True or False. While Ordered: Values are arranged in order of Increasing importance, Nominal values are made up of categories where no weight/importance is attached to any of the categories;

Table 1: Dataset Features, Types and Meaning

Features	Types	Meaning
age	Real Values	Age of individual observations
sex	Binary	Gender of individuals Male - 1; Female - 0
cp	Nominal	Chest Pain (0 = typical; 1 = atypical; 2 = non-anginal pain; 3 = asymptomatic)
trestbps	Real Values	Total Resting Blood Pressure (in mm Hg on patients' admission to the hospital)
chol	Real values	Total Cholesterol Level in mg/dl
fbs	Boolean	Fasting Blood Sugar > 120 mg/dl (1 = True; 0 = False)
restecg	Nominal	Resting Electrocardiographic Results (0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophic)
thalach	Real values	maximum heart rate achieved
exang	Boolean	Angina is a type of chest pain caused by reduced blood flow to the heart. Exercise Induced Angina (1 = True; 0 = False)
oldpeak	Real values	ST depression induced by exercise relative to rest
slope	Ordered	The slope of the peak exercise ST segment (0 = upsloping; 1 = flat; 2 = down sloping)
ca	Real values	Number of major vessels (0-3) colored by fluoroscopy
thal	Nominal	(3 = normal; 6 = fixed defect; 7 = reversable defect)
target	Binary	An indicator for the heart disease (1 = Diseased; 0 = Not-diseased)

Removal of Duplicate Entries: The original data contains 1024 entries, 722 of which are duplicates. The duplicate entries were removed leaving 302 distinct observations for processing.

Understanding the Basic Statistics for all Features in the Dataset: The features were renamed for better

understanding, then the frequency, minimum, maximum, standard deviation and 25, 50 and 75 percentile quarter in each feature were estimated as shown in Table 2. The statistics shows that all values are in appropriate range and no outlier is present

Table 2: Basic Statistics for all Features in the Dataset

Features	Basic Statistics						
	mean	std	min	0.25 %	0.50 %	0.75%	max
Age	54.42	9.05	29	48	55.5	61	77
Gender	0.68	0.47	0	0	1	1	1
Chest_Pain	0.96	1.03	0	0	1	2	3
Resting_BP	131.60	17.56	94	120	130	140	200
Cholesterol_Level	246.50	51.75	126	211	240.5	274.75	564
FBlood_Sugar_Level	0.15	0.36	0	0	0	0	1
ECG Result	0.53	0.53	0	0	1	1	2
Max_Heart_rate	149.57	22.90	71	133.25	152.5	166	202
Excercise_ Induced_Chest Pain	0.33	0.47	0	0	0	1	1
Exercise_ Induced Depression	1.04	1.16	0	0	0.8	1.6	6.2
Peak_Exercise_Slope	1.40	0.62	0	1	1	2	2
Colored_Vessels	0.72	1.01	0	0	0	1	4
thal	2.31	0.61	0	2	2	3	3
Heart_Diseased	0.54	0.50	0	0	1	1	1

Understanding the value distribution in the target column

As shown in Fig 2., of the 302 patients tested in the dataset, 164(53.3%) have heart disease while

138(45.7%). This distribution is balanced enough to train the machine learning classifiers

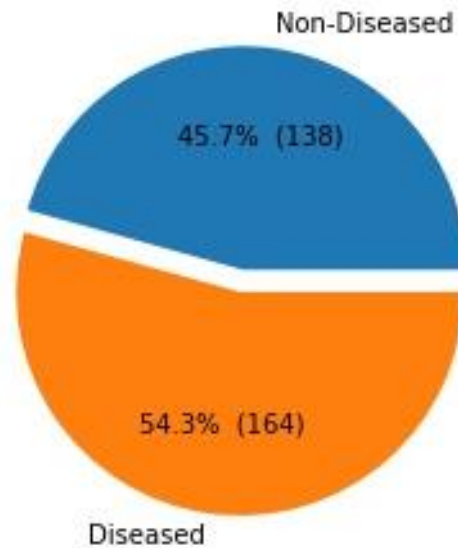


Fig 2: Distribution of Values within the Target Column

Exploratory Analysis and Visualization of Features based on patients who have heart disease and those who do not

Gender:

Fig 3 a) and Fig 3 b) shows that out of the 206 (68.2%) Male in the dataset, 90 of them have heart disease and

amongst the 96 (31.8%) females in the dataset, only 70 of them have the disease.

From Fig c), it could be observed that possibilities of both male and female developing a heart disease starts mostly at age 30 and above and comparatively, Female at extremely old age (60+) suffers from the heart disease than male. Although with slight increase in the male category, the highest potential risk in both distribution is found at age 40 and 50

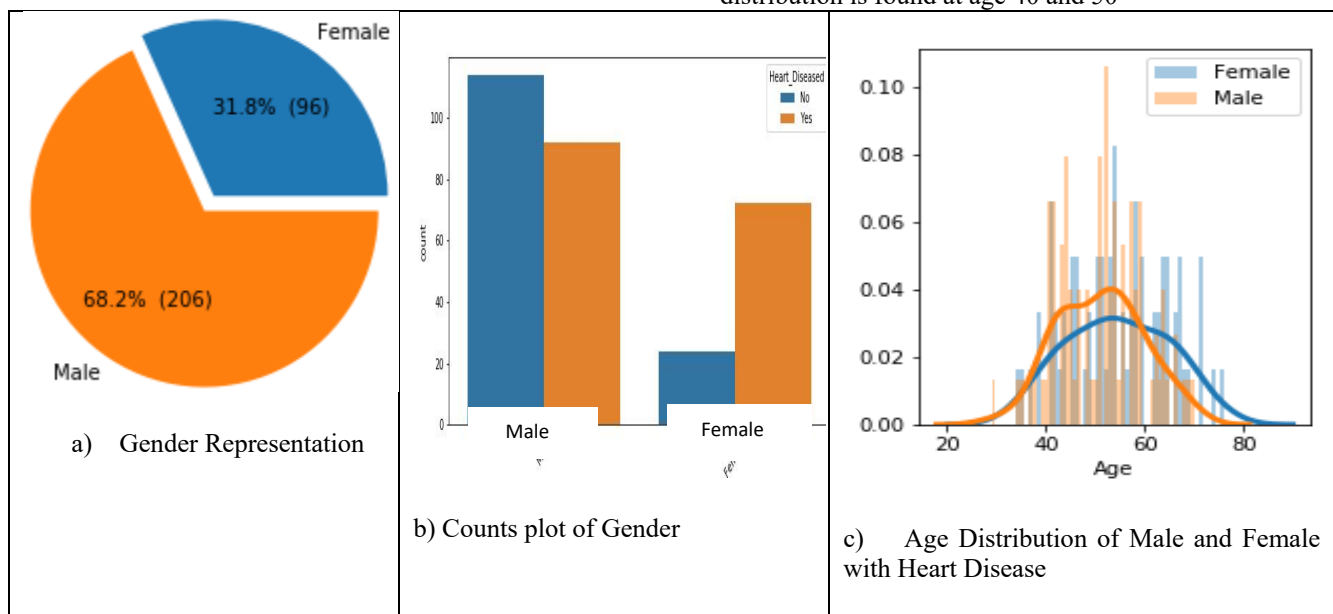


Fig 3: Gender Representation and Age Distribution

This might imply that men are most vulnerable to having the disease at early age 40-50 than women while the vulnerability shift to women once they pass the age bracket at old age.

Blood Sugar Level:

As depicted in Fig 4 a) and Fig b), majority 257(85.1%) of the 302 patients have normal blood sugar level, yet about 140 out of this majority has heart diseases. This

abnormality, as shown in Fig c), inherent the most in patients aged 51. Meanwhile, about 21 patients out of the minority who has high blood sugar suffers from heart diseases. This buttresses the low correlation (-0.03) recorded between blood sugar level and the target in Fig (Fig: Heatmap of Correlation Between Features), that, high blood sugar does not necessarily contribute to the vulnerability of patient to heart diseases

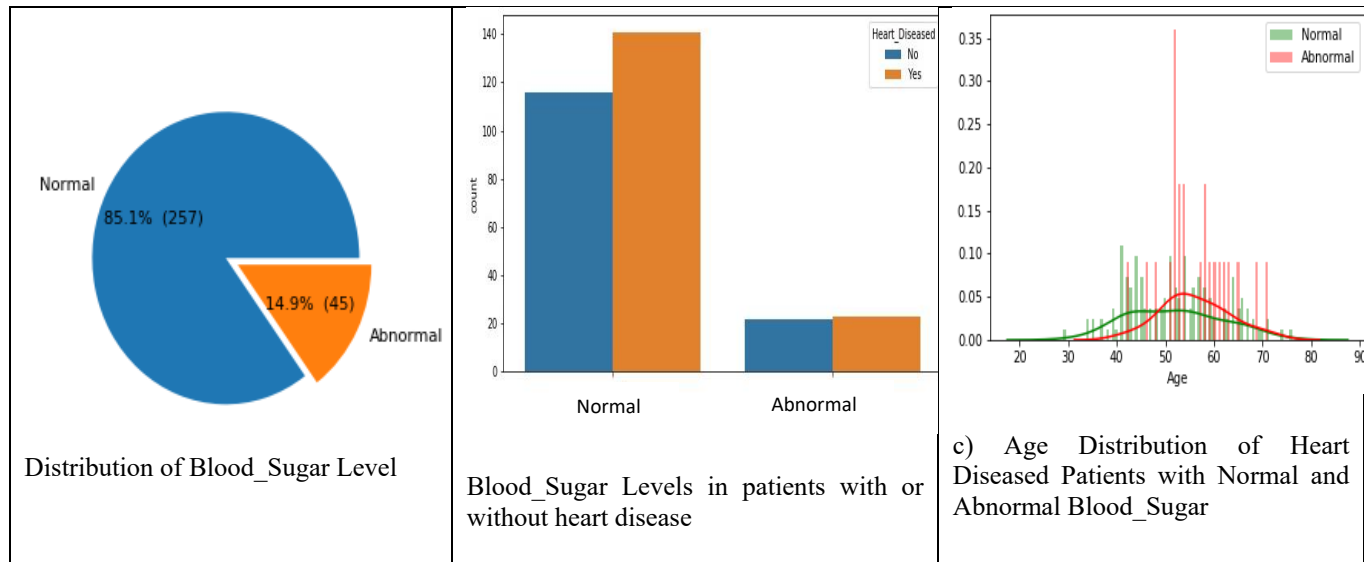
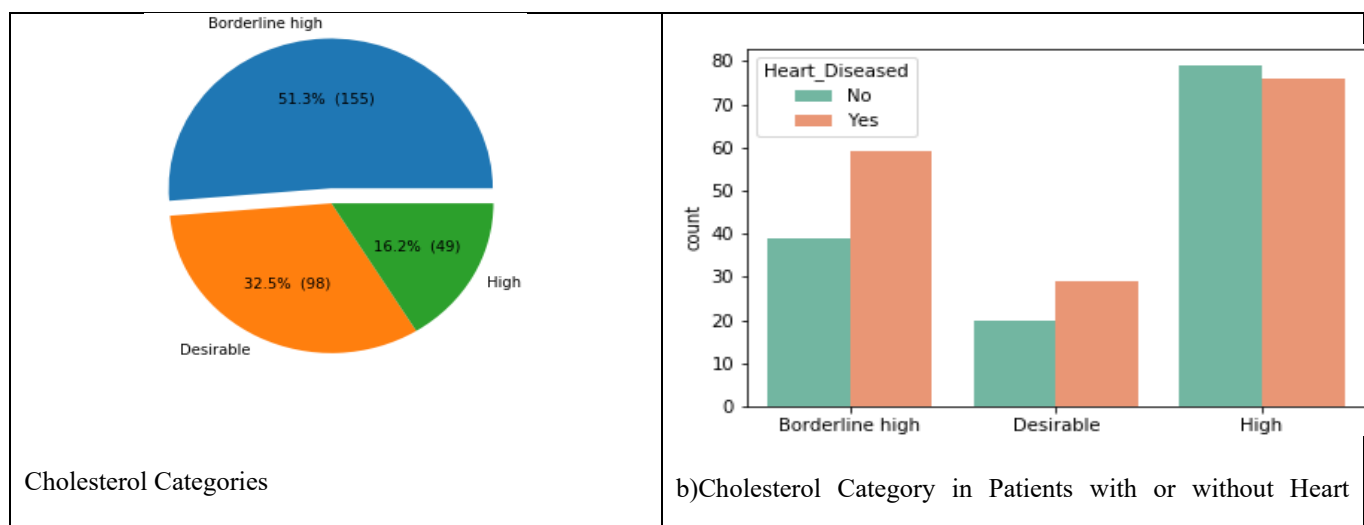


Fig 4: Analysis of Blood Sugar Level

Level of Cholesterol:

Though cholesterol level given as real values, are usually measured in milligrams (mg) of cholesterol per deciliter (dL) of blood, <https://medlineplus.gov/lab-tests/cholesterol-levels/> explained that the cholesterol level can be categorized as Desirable (cholesterol level

< 200), Borderline High (cholesterol level between 200 and 240) and High (cholesterol level = 240 above) these categories are shown in Fig5 a). in each division, Fig5 b) explained that most patients have Borderline to High Cholesterol levels which is mostly common in male than females as seen in Fig6 d)



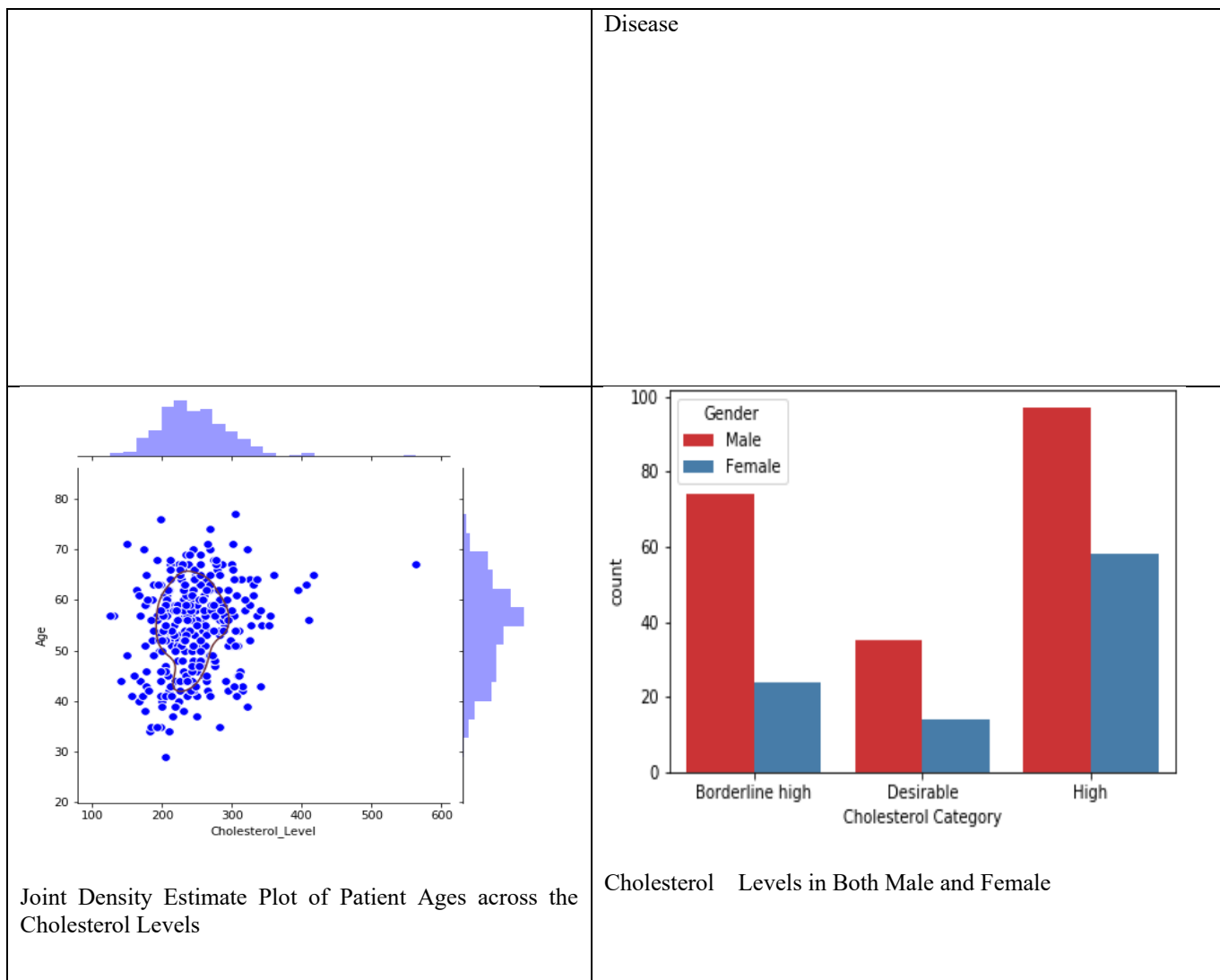


Fig 5: Analysis of Cholesterol Levels

As indicated in Fig 5 c), the cholesterol level of 200 - about 300 (Borderline to High) is common amongst age 41 -68. Meanwhile, an exceptional case of a-67-year-old patient whose cholesterol level is above 500.

This is referred to as an Outlier, since the single observation possess value aberrant to others. This might be due to an error during data capturing or an exceptional case in real sense.

Feature Selection

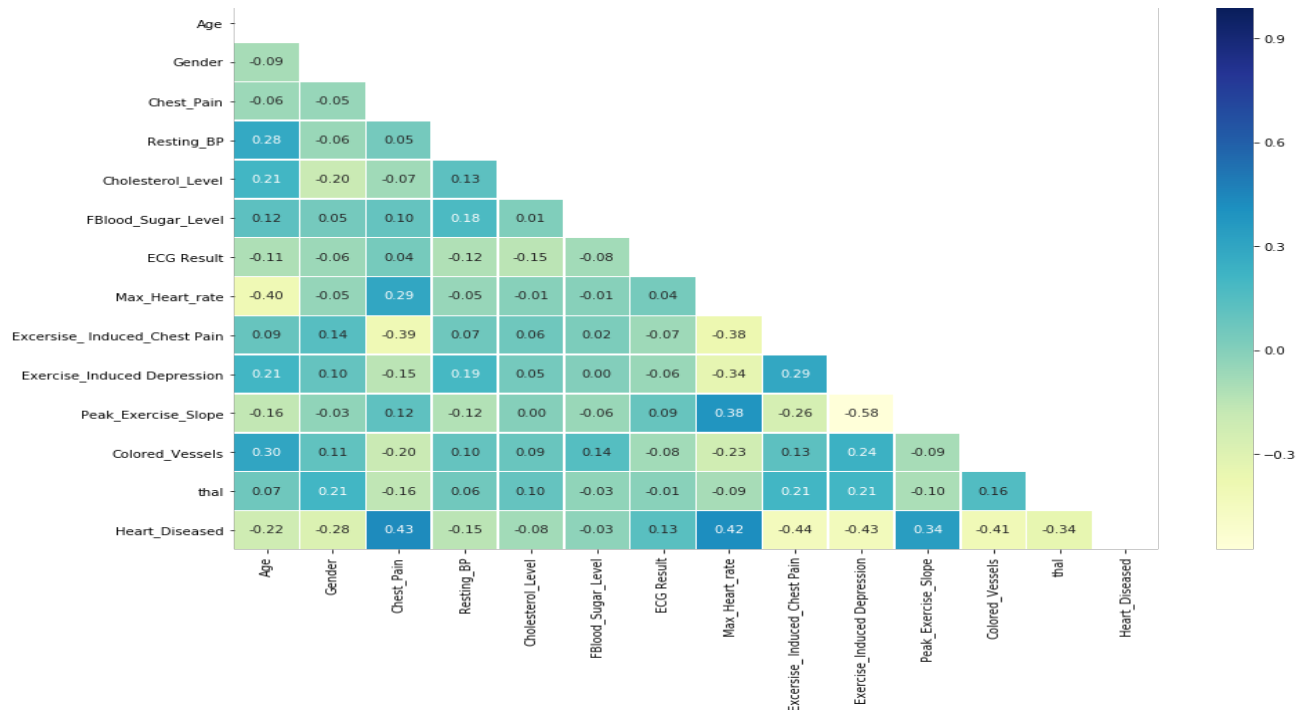


Fig 6: Heatmap of Correlation Between Features

Generally, in Fig 6 above, there is no strong correlation between any of the features and the target variable (since no coefficients approximately equal to 1). Pearson Correlation coefficient shows that there are relationships between Excercise_Induced_Chest Pain (-0.44), Exercise_Induced_Depression (-0.43), Colored_Vessels (-0.41), thal (-0.34), Max_Heart_rate (0.42), Chest_Pain (0.43) and the target. Features with absolute correlation coefficients less than 0.1 will be removed from the dataset. These are mainly Cholesterol_Level (-0.09) and FBlood_Sugar_Level (-0.02)

Machine Learning Model Implementation, Results and Evaluation

Before and after dropping the less correlated features, the models were trained and the following results depicted in Fig 7 and Fig 8 and Tables were obtained.

Accuracy Score Comparison

The accuracy score from all models in the two categories are generated as follows

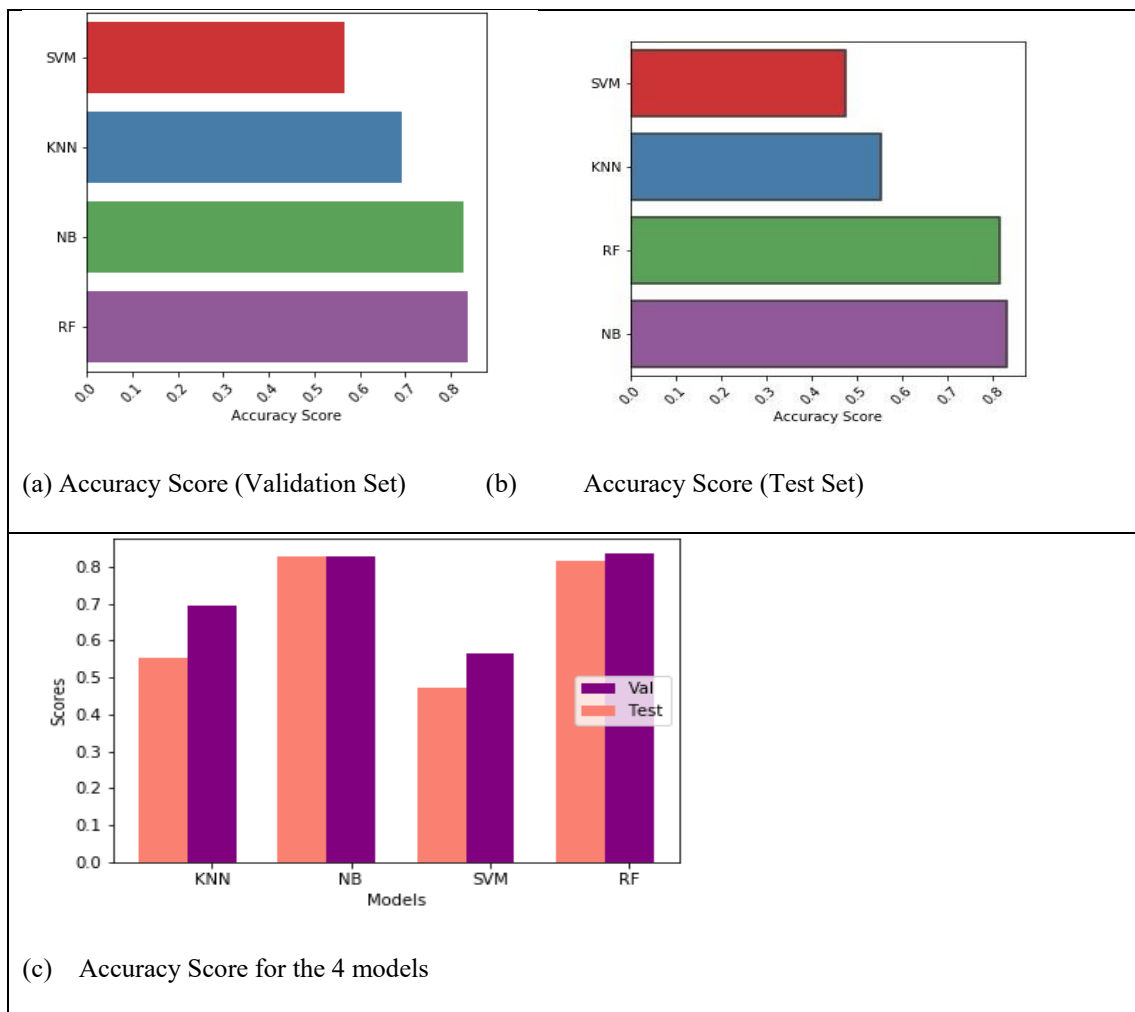


Fig 7: Classification Accuracy with Models Training and Testing on all Features

In addition to the low validation score from SVM and KNN, Fig 7 (a) shows that RF that has the highest validation accuracy compared with Fig (b) where NB produces the highest test score. This implies that using

all features for model training and testing, NB that has the highest test accuracy which is an improvement over the accuracy obtained in the validation set



Fig 8: Classification Accuracy with Models Training and Testing on Correlated Features

After removing features with $<1\%$ correlation, Fig8 (a, b, c) shows that though accuracy improves across all models and KNN, NB, and RF slightly overfits, albeit, SVM has the highest test and validation accuracy score.

Confusion Matrix

The confusion matrix in Fig 9 (a1) shows that, out of the 226 instances in the dataset, SVM correctly classify 72 non-diseased instances as well as 118

diseased instances. Meanwhile, 26 non diseased patients were wrongly classified as diseased and 10 diseased patients are categorized non-diseased by the model. The same SVM when implemented on the test data in Fig9 (a2), out of the 76 instances in the dataset, SVM correctly classify 31 non-diseased instances as well as 33 diseased instances. Meanwhile, 9 non diseased patients were wrongly classified as diseased and 3 diseased patients are categorized non-diseased by the model

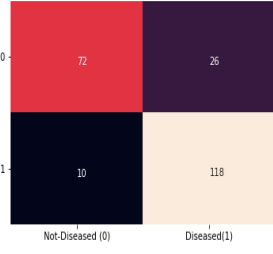
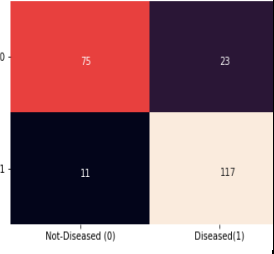
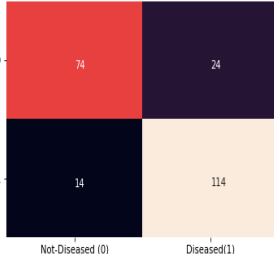
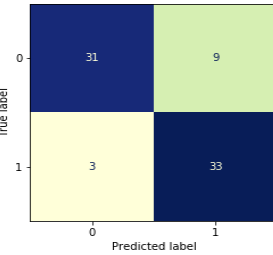
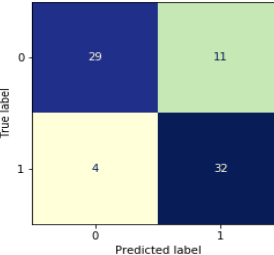
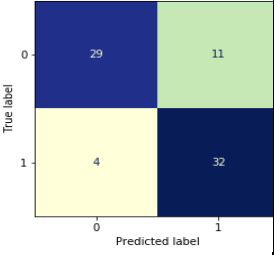
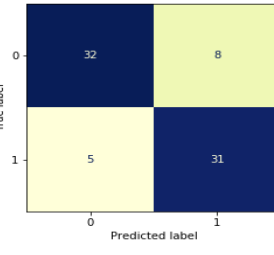
	SVM	KNN	RF	NB
K fold Val Set	 (a1)	 (b1)	 (c1)	 (d1)
	 (a2)	 (b2)	 (c2)	 (d2)

Fig 9(a-d): Confusion Matrix of the Classifiers with only Corelated Features

The implication of the misclassification is that diseased patient who need urgent medical attention are believed to be disease-free. As a result, other evaluation criteria are required to ascertain the best model with lesser misclassified positive class. These metrics are depicted in table 3 and 4 below.

Classification Reports

The classification report gives detailed evaluation report for the trained classifiers for selection of suitable reports for the target distribution and the problem domain. Since F1 score evaluates that balance between Precision and Recall, F1 score will be used as the overall metrics.

Table 3: Model Validation with Correlated Features (Val Set)

EVALUATION PARAMETERS	MODEL Evaluation with Validation Set											
	SVM			KNN			RF			NB		
	0	1	Avg	0	1	Avg	0	1	Avg	0	1	Avg
Precision	.88	.82	.84	.85	.81	.83	.87	.84	.85	.84	.83	.83
Recall	.73	.92	.84	.71	.91	.82	.77	.91	.85	.76	.89	.83
F1 Score	.80	.87	.84*	.78	.85	.82	.82	.84	.83	.80	.86	.83
Support	226(98/128)											

Table 4: Model Testing with Correlated Features

EVALUATION PARAMETERS	Test Evaluation of Model											
	SVM			KNN			RF			NB		
	0	1	Avg	0	1	Avg	0	1	Avg	0	1	Avg
Precision	.91	.79	.82	.88	.74	.82	.88	.74	.82	.86	.79	.83
Recall	.78	.92	.80	.72	.89	.80	.72	.89	.80	.80	.86	.83
F1 Score	.84	.85	.84*	.79	.81	.80	.79	.81	.80	.83	.83	.83
Support	76(40/36)											

The highest F1 Score (84%) was recorded from SVM, therefore, SVM performs better than KNN, RF and NB all the models were evaluated on the 25% test set.

Conclusion

It was concluded that age, gender, level of cholesterol and some other related factors will contribute to the level of susceptibility of man to heart disease. Both females and males between the age of 30 and 40 years are equally susceptible to the disease. However, men are much more susceptible than female between the age of 40 and 50 years, But, females above 50 years are more susceptible to heart disease than males. Meanwhile, high level of cholesterol increases the chance of having heart disease.

Again, for the purpose of classification, four ML models - KNN, SVM, RF and NB were implemented with features of at least 1 % correlation with the target. The performance of the models was properly evaluated with validation and test dataset. The performances were compared and it was concluded that NB performs better than SVM, RF and NB for the model validation while SVM outperforms others the model evaluated on the test set for the heart disease classification task analyzed in this study.

References

1. Ramya, M. and Radha, N. (2016). Diagnosis of CKD using ML algorithm. International Journal of Innovative Research in Computer and Communication Engineering. Vol 4, Issue 1.
2. Baylor, Kumukseven and Kalu (2008) Central of Differentially Driven Mobile Robot using Radia Basis Function Based NN World Scientific and Eng Academic and Society transformation on system and central, Vol 2, No 12, pp 1002-1013
3. Snethil, P. Aritha (2019). Comparison of feature selection method for CK data set using data mining classification Analytical Model. International Research Journal of Engineering and Technology
4. World Health Organization (2018). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
5. Youness K., Mohamed B.(2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. International Journal of Intelligent Engineering and Systems, Vol.12, No.1, DOI: 10.22266/ijies2019.0228.24.
6. Avinash Golande, Pavan Kumar T.(2019). Heart Disease Prediction Using Effective Machine Learning Techniques. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4.
7. Alarsan, F.I., Younes, M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *J Big Data* 6, 81 (2019). <https://doi.org/10.1186/s40537-019-0244-x>
8. Shashikant U. G. and Ashok G.(2012). Heart Disease Diagnosis Using Machine Learning Algorithm. In book: Proceedings of the International Conference on Information Systems Design and Intelligent Applications (INDIA 2012) held in Visakhapatnam, India, 10.1007/978-3-642-27443-5_25
9. C. Beulah ChristalinLathaS. CarolinJeeva (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked. Volume 16, 100203. <https://doi.org/10.1016/j.imu.2019.100203>
10. Amin Ul H., Jian P., Muhammad H. M., Shah N., and Ruinan S.(2018). A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Wearable Technology and Mobile Applications for Healthcare. Vol. 18, pp 1-21. <https://doi.org/10.1155/2018/3860146>.
11. Rajesh N., T. Maneesha, Shaik H. and Hari K. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering & Technology 7(2):363-366. [10.14419/ijet.v7i2.32.15714](https://doi.org/10.14419/ijet.v7i2.32.15714).
12. Shahid, R., Bertazzon, S., Knudtson, M.L. (2009). Comparison of distance measures in spatial analytical modeling for health service planning. *BMC Health Serv Res* 9, 200. <https://doi.org/10.1186/1472-6963-9-200>
13. Taiwo O. A. (2010). types of Machine Learning Algorithms. University of Portsmouth, United Kingdom. <https://www.researchgatepublications/221907660>

