Predicting Heart Disease

*F*airuz Diana bt Hamzah, Norsaza Amalia bt Mohd Saiful, AnisAmiera bt Zamrin Hasbullah, Jeya Jassvine a/p Jeyabala Sundram
Facultyof Informationand Communication Technology,
University Teknikal Malaysia Melaka,
76100 Melaka Malaysia.
{B031910307, B031910482}@student.utem.edu.my
{B031910491, B031910136}@student.utem.edu.my

Abstract - **Heart failure (HF) is one of the prevalent diseases that can lead to dangerous situations in the current period. Almost 26 million patients with this type of disease are affected every year. From the point of view of the heart consultant and the surgeon, predicting heart failure at the right time is difficult. Fortunately, there are models of classification and prediction that can support the medical field and can explain how to use medical data in an effective way. Using the Kaggle heart disease dataset, this paper aims to enhance the precision of HF prediction. In order to understand the data and predict the HF chances in a medical database, several machine learning methods were used for this. The findings and comparative studies showed that, in predicting heart disease, the current study increased the accuracy score. The integration of the machine learning model presented in this study would be useful for predicting HF or any other disease.**

Keywords—Kaggle,classification, database, accuracy, learning model

I. INTRODUCTION

Among various life-threatening diseases, heart disease has sought a great deal of attention in medical research. Approximately, there are almost 26 million people a round the world affecting with heart disease [1]. Hea rt diseases are one of the most common healthcare problems that occur in both young adults and adults [2]. This is somewhat of a preventable disease but is often diagnosed too late or inaccurately. The dia gnosis of hea rt disease is a challenging task. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There a re severa l factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, and family history of hea rt disease, obesity, high blood pressure, and lack of physical exercise.

A major challenge faced by the helthcarre such as hospitals and medical centers is the accuracy of diagnosis for example diagnosing a person if they have a heart disease or not. From here on, proper health advises, and treatments can be administered to the patients by the healthca re team. Using a heart disease prediction generator, it can offer an automated prediction about the heart condition of patient so that further treatment can be made effective before it is too late. This study therefore seeks to enhance the efficiency of the cla ssifiers [3] by carrying out experiments using various machine-learning models to allow better use of the dataset obtained from various medical databases.

This prediction will process the data of heart diseases we took from the Kaggle website and process, clean and filter the applicable attributes to determine the hidden knowledge and generate a prediction. Not only can this prediction model predict with a high level of accuracy, it is also expected to be able to be able to have knowledge on complex heart diseases too. Moreover, such a prediction system can come in very handy to a medical practitioner and aid in making decisions. Besides, a heart disease prediction system is important as it can predict and solve problems accurately even when there is no expert around.

II. PROBLEM BACKGROUND

Heart disease (HD) is one of the most common diseases nowadays, and an early diagnosis of such a disease is a crucial task for many health care providers to prevent their patients for such a disease and to save lives. A comparative analysis of different classifiers was performed for the classification of the Heart Disease dataset to correctly classify and or predict hea rt disease cases with minimal attributes. The model of algorithms used XGBoost, SVM, and KNN classifiers to show the performance of the selected algorithms to best predict the heart disease cases.

To predict heart disease, the test and training data is given as input to the algorithms and the accuracy is compared for analysis by using dataset from UCI (Kaggle). Next, data mining is used to discover the unknown knowledge from the known information and build predictive model. It is a step to discover knowledge from databases and summarizing into useful information. This project will be finding the model which fix better for the problem applying different metrics of performance to reduce the prediction error (error rate) and the best accuracy. Particularly, it will be looking at the fa lse negative error because this mean sends home a patient who in fact has heart disease (negative misclassification error).

To identify the best model to fix the problem and classify correctly new data, it is necessary to reduce the bias of the estimation and reduce the variance of it. However, usually low bia s means high variance. Therefore, it is clear it needed the bias-variance trade-off in order to fix the model. The error when it does not fit well into the assumptions it makes when building the learning algorithm, and the high variance in the estimator indicates possible overfitting in the model. The objective-based on the re- sampling techniques is to achieve

III. OBJECTIVES

- To build model that predict the heart disease.
- Examine trends and correlations within the data.
- To analyze various Classification Models and yields greatest accuracy.

IV. DATA OVERVIEW

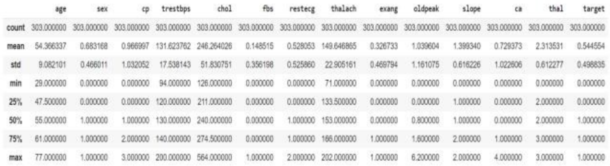
Data over viewing is a form of analysis where we identify the important attributes and contributing factors that will help make this system attain a good level of accuracy.

No.	Attribute Description	Distinct Values
1	***Age*** - The first attribute is defining the age of the person. [Minimum Age: 29, Maximum Age: 77]	Multiple values between 29 and 77
2	***Sex*** - The attribute number two describes the gender of a person. [-0‖ means Female and -1‖ means Male]	0, 1
3	***CP*** - The third attribute is defining the level of chest pain (CP) a patient suffering from, when reached to the hospital. There are four kind of distinct values defined for this attribute, where each value is describing a level of chest pain.	0, 1, 2, 3
4	***RestBP*** - The next attribute describes about the blood pressure (BP) figure for the patient while admitted to the hospital. [Minimum BP: 94, Maximum BP: 200]	Multiple values between 94 and 200
5	***Chol*** - This column is showing the cholesterol level recorded while admitting the patient in the hospital. [Minimum Chol: 126, Maximum Chol: 564]	Multiple values between 126 and 564
6	***FBS*** - The next attribute is describing the fasting blood sugar level in the patient. It has binary classified values. The values are depending on, if the patient has more than 120mg/dl sugar = 1, if not = 0.	0,1
7	***RestECG -*** This parameter is showing the result of ECG from 0 to 2. Where each value is showing the severity of the pain.	0, 1, 2
8	***HeartBeat*** - The maximum value of heartbeat counted at the time of admission [Minimum: 71, Maximum: 202]	Multiple values between 71 and 202
9	***Exang -*** This parameter was used to understand about, does exercise induce angina or not. If yes, the value will be -1‖, and -0‖ for not.	0, 1
10	***OldPeak -*** The next attribute is defining the patient's depression status. It is assigned as different real number values falls between 0 and 6.2.	Multiple real number values between 0 and 6.2.
11	***Slope -*** The condition of the patient during peak exercise. This value defined into three segments [Upsloping, Flat, Down sloping]	1, 2, 3
12	***CA***: This attribute is showing status of fluoroscopy. It is showing that how many vessels are colored.	0, 1, 2, 3
13	***Thal -*** This parameter is another kind of test required for the patient having chest pain or breathing difficulty. Four kind of values showing the result of Thallium test.	0, 1, 2, 3
14	***Target*** – This is the last column in the dataset. This Target column is also known as Class column or Label column. As this column describes the number of categories, (classes) defined in the data file. As per the dataset taken in this experiment. There are two different types of classes (0, 1), where ―0‖ means there is no chances of Heart Failure, where as ― 1 ‖ imply that there are strong chances of heart failure in a patient. The value ―0‖ and ―1‖ is based on the other 13 parameters described in this dataset above.	0, 1

V. PROCESSES

A. ExploratoryData Analysis

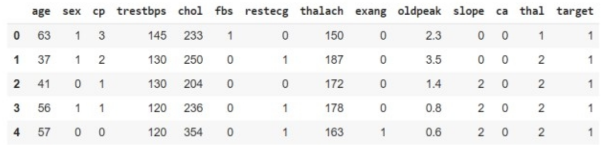
Fig.1 shows data exploration which display Count, mean, standard deviation, min, max,25%,50% and 70%.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Fig. 1 data exploration

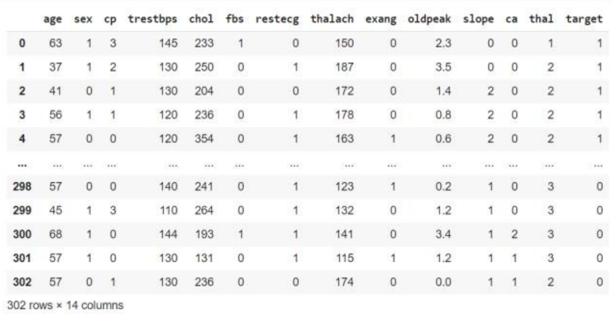
Fig.2 shows exploring dataset of Heart disease that consists of 303 data, 14 features.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 2 dataset of heart disease

Fig.3 shows about data that has been clean by drop all duplicates in heart diseases dataset.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

302 rows × 14 columns

Fig.3 the dataset after cleaning

B. Features Engineering

Fig.4 show data correlation which measures of a mutual relationship between two variables whether they a re causal or not. This degree of measurement could be measured on any kind of data type. So, in heart disease dataset, the most correlate data between target is cp, thalach and slope.

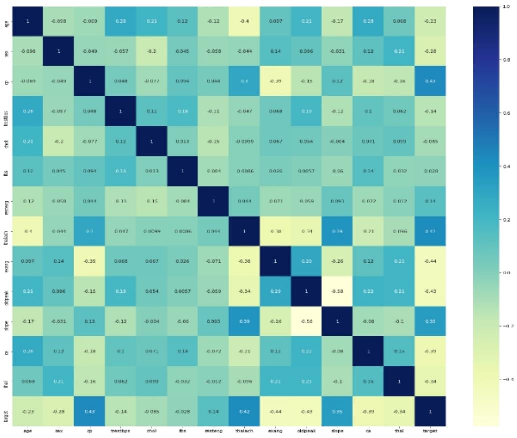
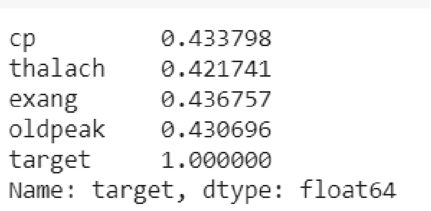


Fig.4 correlation

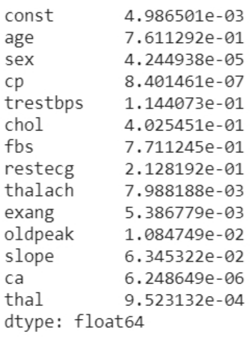
Fig. 5 shows the selecting of high correlated features which is cp, thalach, exang, oldpeak, target.



```
cp          0.433798
thalach     0.421741
exang       0.436757
oldpeak     0.430696
target      1.000000
Name: target, dtype: float64
```

Fig.5 selecting features

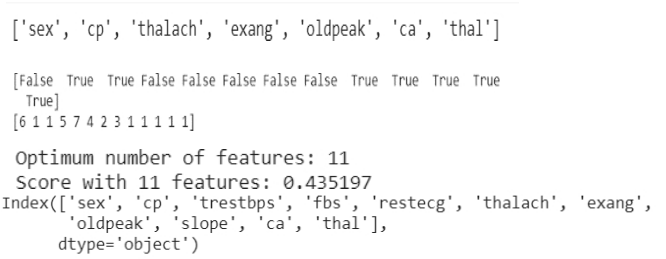
Fig.6 shows the linear model that estimates the intercept and regression coefficient.



```
const       4.986501e-03
age         7.611292e-01
sex         4.244938e-05
cp          8.401461e-07
trestbps    1.144073e-01
chol        4.025451e-01
fbs         7.711245e-01
restecg     2.128192e-01
thalach     7.988188e-03
exang       5.386779e-03
oldpeak     1.084749e-02
slope       6.345322e-02
ca          6.248649e-06
thal        9.523132e-04
dtype: float64
```

Fig.6 linear model

To find optimum features, backward elimination is used. It is to reverse process and all the independent variables a re entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation. Moreover, it also can remove those features that do not have a significant effect on the dependent variable or prediction of output. In addition, to find optimum feature, recursive feature elimination (RFE) also used in this prediction to compare selecting feature for best result.



```
['sex', 'cp', 'thalach', 'exang', 'oldpeak', 'ca', 'thal']


[False  True  True False False False False False  True  True  True  True
  True]
[6 1 1 5 7 4 2 3 1 1 1 1 1]

 Optimum number of features: 11
 Score with 11 features: 0.435197
Index(['sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang',
       'oldpeak', 'slope', 'ca', 'thal'],
      dtype='object')
```

Fig. 7 selecting optimum feature

Fig.7 shows the result of selecting feature by using backward elimination and recursive feature elimination (RFE).

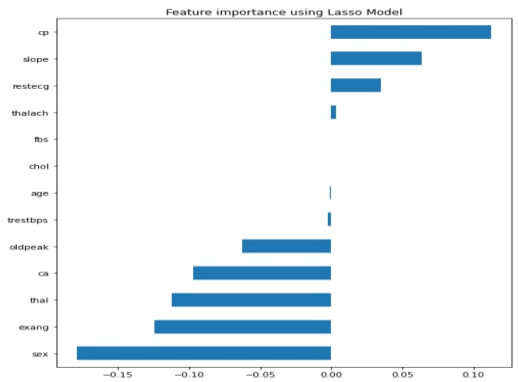
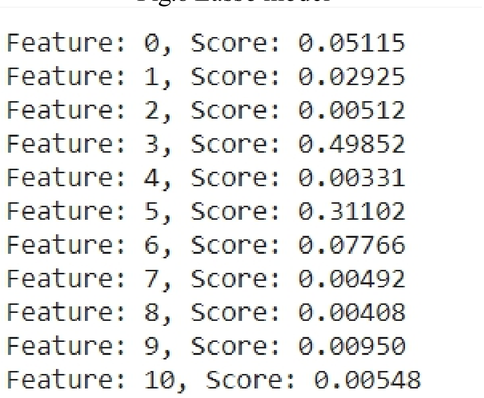


Fig.8 Lasso model

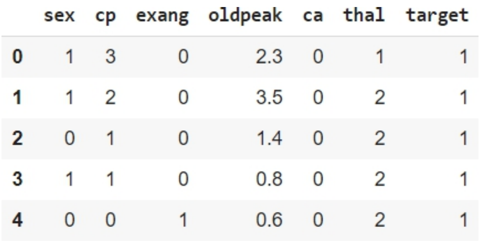


```
Feature: 0, Score: 0.05115
Feature: 1, Score: 0.02925
Feature: 2, Score: 0.00512
Feature: 3, Score: 0.49852
Feature: 4, Score: 0.00331
Feature: 5, Score: 0.31102
Feature: 6, Score: 0.07766
Feature: 7, Score: 0.00492
Feature: 8, Score: 0.00408
Feature: 9, Score: 0.00950
Feature: 10, Score: 0.00548
```

Fig.9 Decision tree

Based on Fig.8 and Fig.9 where Lasso model and decision tree has generated to compare the selected features with the model to finalize the optimum features. As the result, by comparing correlation, backward elimination, recursive feature elimination (RFE), lasso model and decision tree, the most optimum features are sex, cp, oldpeak, ca, exang and thalach.

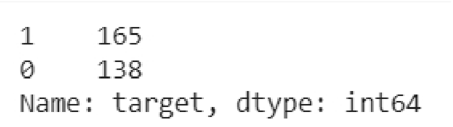
C. Split Train and Test Set



	sex	cp	exang	oldpeak	ca	thal	target
0	1	3	0	2.3	0	1	1
1	1	2	0	3.5	0	2	1
2	0	1	0	1.4	0	2	1
3	1	1	0	0.8	0	2	1
4	0	0	1	0.6	0	2	1

Fig. 10 Split features

Fig.10 shows about the dataset that has been split according to the optimum features.



```
1    165
0    138
Name: target, dtype: int64
```

Fig.11 Analyse feature target

Fig.11 shows to analyse the feature target and to determine the feature balance to train. The result show that the feature is not balance. Therefore, in train data phase, the feature needs random state and stratify to make it more ba lance. This can enhance the data more accurate. Random state is to ensure the number generated in the same order while stra tify is to test subsets that have the same proportion of the class labels as the input dataset.

D. Machine Learning

Firstly, XGBoost model is used. It is to train data and provide better solutions. XGBoost is boosting sequential technique which work on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy. At any instant t, the model outcomes a re weighed based on the outcomes of previous instant t-1. The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher.

Secondly, other than XGBoost, Support Vector Machine (SVM) model also has been applied in this prototype to compare the accuracy. In the SVM algorithm, each point is represented as a data item within the n-dimensiona l space where the value of each feature is the value of a specific coordinate.

Support Vector Machine algorithm is ma inly used to solve classification problems. Support vectors a re nothing but the coordinates of each data item. Support Vector Machine is a frontier that differentiates two classes using hyper-plane.

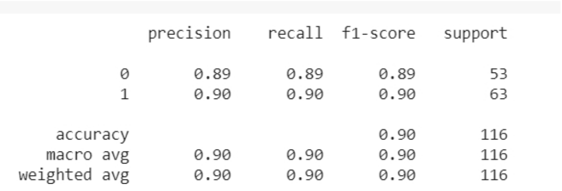
Third, KNN model is also used to train data. KNN algorithms use data and classify new data points based on similarity measures. Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nea rest neighbours. As it increases the number of nea rest neighbours, the va lue of k, accuracy might increase.

VI. RESULT

Table I presented the comparison of the experiment's results conducted in this study with different model. Overall, every classifier has shown good performance in this study. According to the table, the accuracy of the XGBoost model was the highest between all models, while the performance of the SVM and KNN has shown the lowest and constant accuracy in this study. Every model significantly enhanced the performances in this study and shown the satisfactory enhancement, which is greater than 80%.

MODEL	ACCURACY
XGboost	89.66%
SVM	83.62%
KNN	83.62%

Table I Accuracy comparison



```
              precision    recall  f1-score   support

           0       0.89      0.89      0.89        53
           1       0.90      0.90      0.90        63

    accuracy                           0.90       116
   macro avg       0.90      0.90      0.90       116
weighted avg       0.90      0.90      0.90       116
```

Fig. 1 Classification report

Based on Fig.1 the True values and predicted values, the class precision and class recall values computed and presented in the table. The class recall and class precision values a re helpful to identify the overall accuracy of the classifier. As per the displayed values in the Fig.1 the precision and recall a re 0.89 for true class (0) while true class (1) is 0.90.

VII. DISCUSSION

A. Correlation

Correlation can be an important tool for feature engineering in building machine learning models. Predictors which are uncorrelated with the objective variable are probably good candidates to trim from the model.In addition, if two predictors are strongly correlated to each other, then we only need to use one of them.

The correlation matrix is a (K × K) square and symmetrical matrix showing correlation coefficients between variables. The matrix is generally used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. Besides, it considers all variables involved on the same fundamental and bases where any analysis may be described as interdependence analysis. However, there are times where the variables do not have the same status as others. So, our interest may be in how some variables depend upon others and this case scenario arises specifically when there is a temporal ordering of the variables. This would intrigue us to find out how the later variables in a particular sequence depends on those which came in earlier enabling to make predictions about the values of the later variables.

Correlation evaluate the subset made of attribute vector which corelate with class label but independent to each other. Its to qualify the strength of dependence between variable x and y as x increase along with y. So we take the most correlated data that have nearest to 1 or -1.A coefficient of -1 is perfect negative linear correlation: a straight line trending downward. A +1 coefficient is, conversely, perfect positive linear correlation. A correlation of 0 is no linear correlation at all.

B. Backward Elimination

The objective of the implementation of backward elimination method is to build a high quality multiple regression model that includes as few attributes as possible, without compromising the predictive ability of the model. This contains a full model that takes all the variables in a model into consideration. Then, to filter variables that would have some significant impact or contribution to the outcome, unwanted variables will be deleted one by one from the full model. A measure of the variable's contribution to the model will be needed to known in order to find the variable with the smallest test statistics. Any variables with less than the cut-off value or with the highest p value greater than the cut-off value or the least significant variable will be deleted first before the model is refitted without the deleted variable and the test statistics or p values are recomputed. This process will be repeatedly done until the remaining variables are all significant to the cut-off values which is also known to be associated with the p value.One of the advantages of backward elimination is the ability to evaluate the joint predictive ability of variables as the process starts with all variables being included in the model. It also eliminate the least important variables on an early stage leaving only the most important variables in the model. However, backward elimination does not allow eliminated variables to rejoin the model again

It's pretty easy to remove backwards, just pick a degree of importance or select the P-value. Typically, a 5 percent significance level is chosen in most situations.

This means 0.05 is going to be the P-value. Depending on the project, you can alter this value.The function that has the highest P-value has been defined. If this feature's P-value is greater than the meaning amount we selected, we will delete this feature from our dataset. If this feature's P-value, which is the highest in the package, is less than the level of significance. Delete the function with the highest P-value greater than the degree of significance.So we're going to delete the function from the dataset, and we're going to match the model with the new dataset again.This method continues until we reach a point where the highest P-value is less than the significance chosen from all the remaining features in the dataset. This implies that we iterate the code in our example and back until the highest P-value in the dataset is less than 0.05.

C. RFE

This approach conducts model training on a feature set that is increasingly smaller and smaller. The value or coefficients of the function are determined each time and the features with the lowest scores are removed. The optimal set of features is known at the end of this method.As this method involves repeatably training a model we need to instantiate an estimator first. RFE is an efficient approach for eliminating features from a training dataset for feature selection. RFE is a wrapper style function that also uses filter based feature selection internally. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model. This process is repeated until a specified number of features remains.

D. Lasso Model

The acronym "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. Lasso or L1 has the property of various forms of regularization that can shrink some of the coefficients to zero. The function may, \be removed from the model. After data pre-processing, feature selection should be performed. Number of features which coefficient was shrank to zero. So, Now number of coefficients with zero values is zero. As we can see, the logistic regression we used to extract non-important features from the dataset for the Lasso regularization.

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. Lasso regression can shrink the coefficient values of predictors to zero. some of the features are completely removed from the model and because of that it helps to reduce over-fitting in the model and also helps for feature selection. Which is why based on our system the elimination of one feature occurred as it was in the zero range.

E. XGBOOST

XGBoost stands for e**X**treme **G**radient **B**oosting. The reason why XG Boost is always widely used is because it is widely known for two factors. Firstly the speed and model execution.

As for XG Boost has always been bench marked as fast. Mostly its because of related to computer hardware. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. XG Boost follows the principal of gradient boosting.In Boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals. This is the opposite of bagging.

And now for gradient boosting, for the classification to produce a prediction model the model will rely on the intuition that is best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. Models of gradient boosting are a bit slower cause they have to follow sequence and lack scalability therefore being slower

And XGBoost is literally just an intense and faster version of gradient boosting. So XGBoost will provide parallelization in tree building, distributed computing, cache optimization and out of core computing. As an ensemble model, boosting is quite an easy to read and interprate algorithm and the prediction interpretations are easy to handle.

One difficult thing about this model is the fact that it is heavily dependant on outliers which are abnormal data and every classifier is obliged to fix the errors.

F. Benefits of XGBOOST

Firstly, it carries our regularization which is a built in Lasso Regression (which we used in data selection earlier) that prevents the model from overfitting. Overfitting is when the model will train the data too well which includes noises and random data that is not supposed to be included in. This will then negatively impact the models ability to generalize.

Next, it makes use of parallel processing. This means it uses multiple CPU cores to execute the model which will then increase the speed of output production and hitting a good accuracy level.

Furthermore, XGBoost will allow to run cross validation at every iteration of the boosting process and thus it will be much easier to achieve optimum number of boosting iterations in a single run.

G. SVM

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. The objective of SVM is to find a hyperplane that distinctly classifies the data points.SVM differs from the other classification algorithms in the way that it chooses the decision boundary that maximizes the distance from the nearest data points of all the classes. An SVM doesn't merely find a decision boundary; it finds the most optimal decision boundary.

SVM is a new technique suitable for binary classification tasks. Like classical techniques, SVMs also classify a company as solvent or insolvent according to its score value, which is a function of selected financial ratios. But this function is neither linear nor parametric. The case of a linear SVM, where the score function is still linear and parametric, will first be introduced, in order to clarify the concept of margin maximization in a simplified context. Afterwards the SVM will be made non-linear and non-parametric by introducing a kernel. A kernel is a pathway to manipulate the data so basically the Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transformed to a linear equation in a higher number of dimension spaces.

H. BENEFITS OF SVM

The advantages are it is effective in high dimensional spaces and very versatile. It also uses subset of training points which are called support vectors and hence it is memory efficient. SVM also works great when there is a clear margin of separation between classes However what makes it incompatible is the fact that if the number of features are greater than the number of samples. In other words, it is incompatible with large datasets which will be inconvenient in our case. Next SVM also does not perform well when there is more noise such as target classes overlapping and so on.

I. KNN

KNN is a supervised algorithm that is widely used in classifications. It will manipulate the training data and classify the new test data based on matrices. For KNN we must choose an appropriate k value to achieve maximum accuracy. KNN stands for K Nearest Neighbor which pretty much sums up the concept of the model itself which is similar concepts existing in the same place.

K value indicates the count of the nearest neighbors. We have to be able to compute distances between test points. There are no predefined statistical methods to find the most favourable methods but we used the trial and error method and went with k=7 as there was no changes in accuracy.

The trial and error method can start with tweaking with the smaller values and if the model produces an unstable output we can increase the value of k which might bring about an accurate prediction. But then once there are errors, it means we have exceeded the limit of a decent value of k value.

J. BENEFITS OF KNN

The biggest benefit of KNN model is the fact that it is a lazy learner making it a very simple implementation process. Next, Robust with regard to the search space; for instance, classes don't have to be linearly separable. It also depends on very few parameters to tune and process.

However so, KNN model is extremely expensive despite its convenience and especially for larger datasets which in this case is extremely inconvenient. Next, it is also very sensitive to noisy and irrelevant data. However this is usually solved during the feature selection processes.

K. COMPARISONS

XGBoost is better than SVM and KNN.
Based on our project both KNN and SVM has the exact same accuracy levels. These models both have their pros and cons which have been discussed in this technical report.

XGBoost attained the highest level of accuracy and will be an appropriate model to use in the heart disease prediction system as it will produce a good level of diagnosis accuracy.

VIII. CONFUSION MATRICS

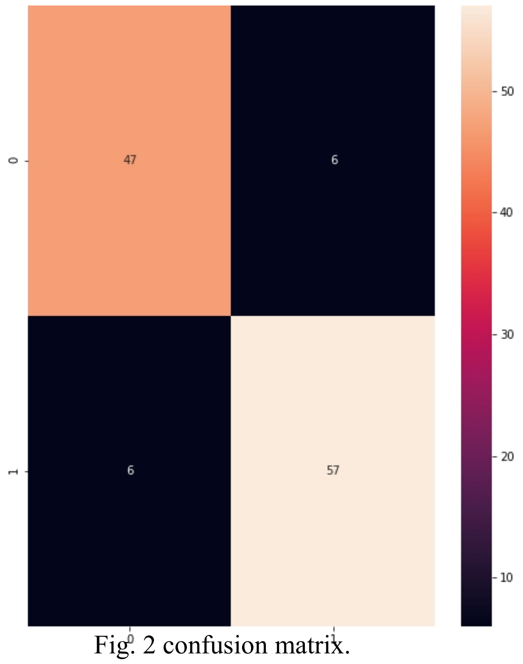


Fig. 2 confusion matrix.

Based on Fig.2 the model performance in the form of confusion matrix is displayed in table I. A confusion matrix is a table used for describing the performance of a classifier that executed on given test data where the “True” values are considered known data values. In this table, the True Class (1) means the known values for the class category (1), the patients having chances of heart failure. On the other side, True Class (0) denotes the known values for class category (0); the patients showing healthy sign. In the same way, the rows values illustrating the prediction computed for both classes.

IX. CONCLUSION

In a nutshell, heart diseases are as serious as a day-to-day disease and should not be overlooked. With fast evolving technologies and the ability to train and model a machine to a id in medical treatments should not be of wasted opportunity. A great tool like this can prevent deaths and maintain a healthy streak. As the rate of patients with heart diseases increase, we hope that technologies like these can help bring this situation to a healthy controlled level.

From the models that we have trained, XGBOOST has brought the highest level of accuracy. XGBOOST is scalable and accurate as it has been developed solely for the purpose of model performance and computational speed. In a nutshell XGBoost reduces overfitting and allows parallel processing which makes it faster. It also allows cross validation at each iteration of the boosting process which then leads to a optimum boosting level.

X. IMPROVEMENTS AND FUTURE WORK

As useful and convenient as our system has been, there is always room for improvements. For example, the data collected can be improved way better. There is a lot of room for human error to occur as the data is being tabulated by humans. A dataset is very crucial to a system like this because the system will only output what you feed it. Next, we also incorporate much more automation such as scanning and 3D printing of the diagnosis. Moreover, it can have a friendlier interface, as of now it is usable for the programmer but it has to be equally friendly to a third party user as well to sum up the entire objective of this system itself.

As for the future projects, this will be a good template experience for us to create a more evolved system with a better understanding and better experience.

XI. ACKNOWLEDGEMENTS

Firstly, we would like to wish a heartfelt thank you to *Professor Madya TS. DR. CHOO YUN HUOY*, our MACHINE LEARNING mentor who has been very professional and knowledgeable at this field of study. Besides being a good support, *Professor Madya TS. DR. CHOO YUN HUOY*, has also shun a great amount of knowledge and expose us to real life problems that can help us understand Machine Learning much better.

Next, we would like to be grateful for the knowledge learned and earned from this group project. As challenging as it was to pursue this completely online, it was equally fruitful to have worked with a team that is equally inquisitive and hardworking.

XII. REFERENCES

[1] G. Savarese and L. Lund, ―Global Public Health Burden of Heart Failure,‖ Card. Fail. Rev., vol. 3, no. 1, 2017.

[2] . S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and Bashir, ―Improving Heart Disease Prediction Using Feature Selection Approaches,‖ in 16th International Bhurban Conference on AppliedSciences and Technology (IBCAST), 2019, pp. 619–623.

[3] M. Gjoreski, A. Gradišek, M. Gams, M. Simjanoska, A. Peterlin, G. Poglajen, ―Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers,‖ in Proceedings - 2017 13th International Conference on Intelligent Environments, IE 2017, 2017, pp. 14–19.

[4] Centers for Disease Control and Prevention. 2021. *Heart Disease* | *cdc.gov*. [online] Available at: <https://www.cdc.gov/heartdisease/index.htm>

[5] .V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[6] S. S. Bashar, . S. Miah, A. H. M. Z. Karim, A. Al Mahmud, and Z. Hasan, ―A Machine Learning Approach for Heart Rate Estimation from PPG Signal