**Heart disease Prediction**

**Abstract**

**Heart disease**is one of the most vital diseases that affect many people. Age, lack of exercise, obese, living style, lousy diet, diabetes, high blood pressure, etc., may cause **heart disease**. In addition, people with **heart disease** develop an increased risk of stroke, kidney disease, nerve damage, eye problem etc.

Existing practice procedures in the hospital are to gather needed data for **heart** disease identification across various tests, they offer a suitable cure centred on treatment. Big data analytics started playing a substantial part in healthcare organizations. Healthcare businesses have a vast amount of databases volume.

Applying big data analytics help in studying enormous datasets to discover hidden information, find unseen trends and patterns and locate information from the data to predict results appropriately.

 This paper proposed a heart disease prediction model for the classification of heart disease, which involves minor external causes liable for heart disease along with ordinary factors like BMI, age, glucose, insulin, etc.

**Introduction:**

The project idea is to develop a machine learning model to accurately predict whether or does not the patients in the dataset have heart disease or does not based on the dataset collected. The project will implement different algorithms to identify the algorithm with the best performance.

 Machine learning integrates several classifiers learning techniques used in predicting and finding the Accuracy of the provided dataset. Heart diseases have been common in recent days. World health organization estimated that 17.9 million people died from heart disease in 2019, demonstrating 32% of the global deaths. 85% of these deaths were due to heart attack and stroke (organization, 2022).

The study can assist in predicting people who likely have heart disease by looking at their health medical records. It will identify the people who have heart disease symptoms such as high blood pressure and chest pain, or it can improve in cure heart disease with a few medical tests and efficient treatments to be healed consequently.

### **Methods**

The paper analyses a dataset of 253,680 patients recorded with heart disease collected in 2015. The project applies a classifiers machine learning technique to predict if a patient has heart disease or does not.

The dataset contains 22 predicted attributes. The experiment used a subset of 10 features. The existence of heart disease in the patient considers a target feature. 0 value integer believes it has no heart disease, and integer value 1 has a condition of heart disease.

**Previous studies**

The project develops a heart disease detection model using three machine learning classification algorithms. This study predicts people with heart disease by isolating people's medical history, which leads to severe heart disease from a dataset that contains variables like sugar level, chest pain, blood pressure, etc.

 The study used Random Forest Classifier, KNN and Logistic regression. The accuracy of the model was 87.5%. The study claim using of additional training data makes sure the model gives a higher chance of accurately predicting whether a person has heart disease or does not. (Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath, 2021).

The paper suggested three techniques should be used to get better result for confusion matrix, F1 score, and precision specificity sensitivity can give a good outcome in the study. It found that machine learning algorithms performed better when using the better features selection technique in the dataset, and the Neighbour’s classifier worked better in the machine learning method when applied to data pre-processing.

Deploying the model reduces computational time. However, the researchers find that the dataset has to be normalized. If not, the training model will get overfitted, and the accuracy will not be achieved when the model is evaluated for the real world.

The study also finds that statistical analysis is essential in analysing a dataset, and it has to have a gaussian distribution. The outlier’s detection is vital, and the project used the Isolation Forest technique for handling it. (Rohit Bharti, Aditya Khamparia,Mohammad Shabaz,, 2021).

This paper designed an efficient machine learning model to detect heart abnormality. It focused on a model that learns by using the latest regularizer, which is entirely based on a standard deviation of the weight matrices.

The latest regularizer corrects the attributes coefficients from gaining high rates in the weight matrix outer space. As a result, the model achieved outstanding results and can assist medical practitioners when investigating heart abnormalities.

The holdout and 10-fold validation algorism are used, and the accuracy gained for detecting heart disease was 96.30%. Consequently, the integration of the planned regularizer with ANN exceeded other techniques in terms of accuracy. (Abdulaziz Albahr, Marwan Albahar, Mohammed Thanoon, Muhammad Binsawad , 2021).

This study builds an intelligent predictive model based on current machine learning algorithms for diagnosing and predicting heart disease. The established scheme used two heart disease datasets. In addition, the model was tested and trained on optimal features and full features as well.

The study used Ten classification algorithms KNN, RF, NB, DT, SVM, ET, GB, AB, ANN and LR. Four feature selection methods, FCBF, LASSO, mRMR, and Relief, are used to select the most meaningful features to reduce the classification errors and shrink the feature space.

The accuracy of the classification ET increases from 92.09 to 94.41%. The accuracy of GB rises from 91.34 to 93.36% with the FCBF feature selection algorithm. So, the ET classifier with the relief feature selection process performs brilliantly. (Yar Muhammad, 220).

The open-access conference (IOP) series for engineering and science materials applied Random Forest and K Nearest Neighbour (K-NN) algorithms to predict heart disease. The two-algorithm used in the practice study proved to work well.

The study said that machine learning could reduce the amount of heart disease damage, but a lot is needed to create new methods to tackle the problem more actively.(Apurv Garg, 2021)**.**

This paper proposes a novel technique whose goal is to find a noteworthy feature by implementing machine learning practices to improve the accuracy in predicting cardiovascular disease.

The technique produces enhanced performance with 88.7% accuracy by using a linear model (HRFLM) combined with the hybrid random frost. The study was directed to using the real-world data and finding a new and novel approach to challenge heart disease (SENTHILKUMAR MOHAN, 2019).

The study uses WEKA to compare various Decision Tree classification algorithms to seek better heart disease cure performance. The researchers tested several algorithms, J48 algorithm, Logistic regression, decision tree and Random Forest.

The study extracts the unseen patterns by implementing data mining practices, and it predicts heart disease in the patients. The J48 tree practice gives better results than others regarding the accuracy and decree amount of error tailed LMT and Random Forest. (Jaymin Patel, 2016).

This study analyses 299 patients in the heart failure dataset collected in 2015. It applies numerous classifiers techniques to predict the survival of the patients and rank the most critical factor corresponding to the risk factors of the features.

It also performs old-style biostatistics tests to compare the results with machine learning algorithms. The results show that ejection fraction and serum creatinine are adequate to predict heart failure survival from medical records and using the two-feature system gave better accurate prediction. (Davide Chicco, Giuseppe Jurman , 2020).

**Dataset**

The dataset is a part of the Behavioural Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the CDC. The dataset is for 2015 and was published on Kaggle in the same year.

The dataset has 253,680 survey responses for the binary classification of heart disease and 22 variables. In addition, it collects responses from over 400,000 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services (kaggle, 2015)**.**

**Data prepressing**

Building a machine learning models can help finding data-patterns when make predictions on unseen data. To predict accurate result, the project construct and transform the data correctly. The data were checked and examined to discover, understand, and compare data hypothesised value to develop the final dataset.

**Missing data and outliers:**

The project checked for missing data and find out the dataset is clean, and there is no missing data exists before progressing further. The dataset has some outlier, the outlier lead to a poor fit and a low down the model performance and productivity.

To avoid that the project used standard scaler technique to scale down the dataset to help in stopped raising the outliers added to that scaling dataset to general elements leads improved in accuracy (Abiyev, 2022).

**Feature Selection**

A attribute selection is one of the most essential steps in creating a machine learning model. Using a dataset directly from sources to create a model without checking the quality of the dataset and the inclusion of attributes may result in under or over fitting which can affect the model's ability to learn, so it is important to pre-process the data before to use it in a model.

The project implemented a correlation coefficient attribute selection first . The project removed the highly correlated variable, if they did not remove them, they will behave such as a duplicated feature.

The study used seaborn to perform person correlation to locate the correlations for all the independent features. The project visualizes the high correlation features that require to be eliminated.

Afterward, the project set a function for correlation and set the X train data inside it and created a loop function which ran through all independent variable and compared each column with each other and sees if they are greater than the threshold.

The project compared the independent variable with target variable where target variable heart disease (class) is highly correlated with an independent variable would keep and drop the other. Then the project check with other variable, same process is followed until last variable.

The project called the correlation function for X train and found out 11 high correlated features. The project dropped them from X train and X test data.

Graphical user interface, text, application, email

Description automatically generated

**Recursive Feature Elimination**

The idea is to select the suitable feature that provide a better result by reducing the number of the features. The project used Recursive Feature Elimination(RFE) method for the reduction of the dimensionality.

RFE is an optimization algorithm technique, the objective is to find the best performing feature subset. Recursive feature elimination repeatedly generates models and keeps aside the worst or the best performing feature at each iteration. It builds a model with the remain next features till entirely features are used, then ranks the features based on their elimination order. The methods used accuracy metric for ranking features based on their importance.

The RFE method which implemented used the quantity features required as input. After that, it gives ranks to entirely variables. It started with the most significant. True considered as relevant and false, consider as irrelevant features.

The project took seven features. The RFE ranked the features; the choice of the amount of '22' was selected randomly, to find the optimum numbers of the feature with the highest accuracy. The project used a loop that began with one feature and built up to 13. Then it selected features with the highest accuracy

From the computation of the code, the features optimum number the project got is 10. Then it feeds 10 features as a number to the recursive feature elimination technique to get the final features set produced by the RFE technique.

Graphical user interface, text, application

Description automatically generated **Model performance**

The product calculated the accuracy of a classifier as a ratio of the total number of correctly predicted samples by the total number of samples. It is used to evaluate the classifier when it is a balanced data set. Accuracy metrics should not be used when the data set is imbalanced.

To value the projected model, the model used accuracy, confusion metric, Precision, recall, and F1 score.

**00000000**

**Model Accuracy**

**Fold cross validation**

The idea of the cross-validation is to use the same data twice for the training set and the other for the test set, then its turnover around. The data was previously used for the testing as training, and the data that used for training are used for testing and the model get the average error.

When we implemented 10-fold cross-validation, the model got 83.6 accuracy in Naïve Bayes and 90.5 accuracy in Logistic Regression

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Accuracy** | **Standard Deviation** |
| **Naive Bayes** | 83.6 | 0.89 |
| **Logistic Regression** | 90.5 | 0.47 |

**Hold-out method**

The project used the model mentioned above for splitting the data into several split up and operates one split for training and others for testing and validating the models. The project uses a 70-30 splitting percentage on its dataset, where 70% of the data used for training and 30% for testing the model. the model got 83.74 accuracy in Naïve Bayes and 90.51 accuracy in Logistic Regression

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| **Naive Bayes** | 83.74 |
| **Logistic Regression** | 90.51 |

**Evaluation**

Classification accuracy is the number of exact forecasts made as a ratio of the whole prediction. It is the most widespread assessment metric for classification challenges. Regrettably, it is also the most misused method.

It is appropriate when equal observations are in every class, which is sometimes the case. All predictions and estimate errors are similarly critical. Therefore, it is usually used to evaluate a classifier when the data set is imbalanced (Brownlee, 2020).

|  |  |  |
| --- | --- | --- |
|  | Predicted NO | Predicted YES |
| Actual NO | True Negative  (TN) | False Positive  (FP) |
| Actual YES | False Negative  (FN) | True Positive  (TP) |

True positives are the cases that predicted, yes, belonging to the class. YES, while true negative are cases that indicate no and belong to no class. False-positive are classes that expected yes, and they are NO. False-negative are classes where predictive NO and they’re in reality yes.

**10-fold Cross Validation**

The total accuracy of the classification is 83.6, Standard Deviation 0.89 and TN (202313), FN(14122), FP(27474) and TP are(9771).

|  |  |  |
| --- | --- | --- |
|  | Predicted NO | Predicted YES |
| Actual NO | 202313 (TN) | 27474 (FP) |
| Actual YES | 14122 (FN) | 9771(TP) |

Method 1 - Hold Out

The total accuracy of the classification is 83.35, Standard Deviation 0.89 and TN(66510), FN(4784), FP(9151) and TP are(93270).

|  |  |  |
| --- | --- | --- |
|  | Predicted NO | Predicted YES |
| Actual NO | 66510 (TN) | 9151 (FP) |
| Actual YES | 4784 (FN) | 3270 (TP) |

# Precision (Positive Predictive Value or positive)

Precision is a true positives (TP) ratio by the sum of true positives (TP) and false positives (FP). It tells how many totals predicted positive, actually positive. They used base on the use case as a metric to measure the classifier's quality (Kadamba, 2021).

**Recall** (True Positive Rate or Sensitivity or)

# The recall is a true positives (TP)ratio by the sum of true positives (TP) and false negatives (FN). Based on the use case, it tells how many of the classifiers predicted actual negative value is also used as a metric for quality measuring of the classifier (Xiaobo Tang, 2021).

**F1 Score**

F1 score is used when recall and precision are essential for the use cases. It is the harmonized mean of recall and precision, falling between (0,1).

**If False Positive (FP) is critical, then β lies between o and 1.**

# **If False Negative (FN) is essential then β > 1 (Kadamba, 2021).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| **0** | 0.93 | 0.88 | 0.91 | 229787 |
| **1** | 0.41 | 0.41 | 0.32 | 23893 |
| accuracy |  |  | 0.84 | 253680 |
| macro avg | 0.60 | 0.64 | 0.61 | 253680 |
| weighted avg | 0.87 | 0.84 | 0.85 | 253680 |

**Discussion**

The study result shows that the two-algorithm used gives high accuracy. When classification accuracy is used, it provides high accuracy, precision and recall, which could be helpful or possible to predict heart failure patients.

The project implemented a 10-Fold cross-validation and Hold-out method with Naive Bayes and Logistic Regression, in model performance. Both algorithms give kind of the same result. Moreover, the accuracy classification also showed a similar outcome.

Although the resulting outcome is promising, the study needs a further experiment by testing more algorithms and using the dataset from different sources such as the internet ‘of Things (IoT) and seeing the result before deciding on the effeteness of the study.

**Conclusion**

The project builds a classification machine learning model that predict if a patient has a heart disease or not based on a historical dataset, the Behavioural Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the CDC in 2015.

The project builds a classification machine learning model that predict if a patient has a heart disease or not based on a historical dataset, the Behavioural Risk Factor Surveillance System (BRFSS), a health-related telephone survey collected annually by the CDC in 2015.

The outcomes of the model show that based on the Fold cross validation algorithm Naive Bayes got 83.6 and Logistic Regression got 90.5, while when used Hold-out method Naive Bayes got 83.74 and Logistic Regression got 90.51.

The evaluate the experiment further the project implemented confusion metric, **10-fold cross validation accuracy was** 83.6and TN (202313), FN (14122), FP(27474) and TP are(9771).**Method 1 - Hold Out accuracy is**  83.35and TN(66510), FN(4784), FP(9151) and TP are(93270).

This project is limited on the historical dataset collected by CDC in 2015. In the future the study can be expanded by using other data sources and more algorithms this will help to compare the result and it may give better outcome.

# References

Abdulaziz Albahr, Marwan Albahar, Mohammed Thanoon, Muhammad Binsawad , 2021. Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer. *Comput Intell Neurosci,* 2021, Article ID 86283350(Computational Intelligence and Neuroscience), p. 10.

Abiyev, A., 2022. *towardsdatascience.* [Online]   
Available at: https://pub.towardsai.net/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff  
[Accessed 02.04 April 2022].

Apurv Garg, B. S. R. K., 2021. *Heart disease prediction using machine learning techniques.* Uttar Pradesh, IOP Conf. Series: Materials Science and Engineering.

Brownlee, J., 2020. *machinelearningmastery.* [Online]   
Available at: https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/  
[Accessed 03 April 2022].

uuuu

Davide Chicco, Giuseppe Jurman , 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 03 02, p. 80.

uuuu

Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath, 2021. *Heart disease prediction using machine learning algorithms.* New Delhi, IOP Conf. Series: Materials Science and Engineering .

yyyy

Jaymin Patel, T. P., 2016. Heart Disease Prediction Using Machine learning and Data Mining Technique. *IJCSC,* 7(IJCSC), pp. 129-136.

jtytyt

Kadamba, V. G., 2021. *analytics-vidhya.* [Online]   
Available at: https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-problems-with-implementation-in-python-a20193b4f2c3  
[Accessed 03 April 2022].

Kadamba, V. G., 2021. *Analytics-Vidhya.* [Online]   
Available at: https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-problems-with-implementation-in-python-a20193b4f2c3  
[Accessed 02 April 2022].

hjjtyjt

kaggle, 2015. *kaggle.* [Online]   
Available at: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset  
[Accessed 12 march 2022].

hjjyuk

organization, W. h., 2022. *World Health Organization.* [Online]   
Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Cardiovascular%20diseases%20(CVDs)%20are%20the,-%20and%20middle-income%20countries.  
[Accessed 19 March 2022].

jjktyukt

Rohit Bharti, Aditya Khamparia,Mohammad Shabaz,, 2021. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,Gaurav Dhiman,Sagar Pande,Parneet Singh. *Computational Intelligence and Neuroscience,* 16 May, 2021, Article ID 8387680, 11 pages(All India Institute of Medical Science, Rishikesh, India), p. 10.

hj

SENTHILKUMAR MOHAN, C. T. S., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Acess,* 7 -2019(IEEE), pp. 81542-81552.

hjjr

Xiaobo Tang, H. M. L. a. X. D., 2021. Research on automatic labeling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching. *Scientific Reports,* 11(Springer Nature Limited), p. 17.

Yar Muhammad, M. H. C., 220. Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific Reports,* 10(https://www.nature.com/srep/), p. 22.