MACHINE LEARNING MODEL ON HEART DISEASE PREDICTION

CIS4035-N-FJ1-2021

Abstract

The heart is the second most fundamental component of the human body. It circulates blood throughout the body, supplying it to all organs. Predictions are based on extra data which involve data analytics. The information gathered cn be used to predict the occurrence of future diseases. Linear Regression, Decision Tree Classifier, Random Forest and Support Vector Mechanism are some of the data mining and machine learning approaches we use to forecast heart disease.

EMMANUEL OKO-OSE

B1097931

INTRODUCTION

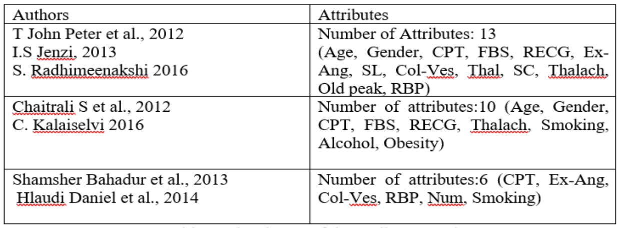
One of the most critical organs in a human being is the heart. One of the prominent cardiac disorders is heart attack. The body's circulatory system is pumped by the heart. As a result, if the heart does not operate properly, it can cause major health problems and even death.

Heart diseases claimed 1.7 million Indians in 2016, as per the 2016 Global Burden of Disease Report, unveiled on the 15th September 2017. Medical establishments across the world collect data from a wide range of health-related concerns. Using a variety of machine learning techniques, these datasets can just be exploited to gain insights. The volume of data, however, is vast, and it is usually inaccurate. These data, that are too huge for human brain to grasp, can be easily analyzed via ML algorithms. As a result, these models have already shown to be quite beneficial in reliably prediction and diagnosis or existence of heart-related problems.

Machine learning is a method of altering and acquiring implicit, formerly unknown/known, and crucially significant data." Machine Learning is a vastly complex area that is continuously developing in size and relevance. Machine Learning classifiers are used to evaluate and determine the precision of a sample and encompass supervised, unsupervised, and ensemble learning classifiers. We can utilize the knowledge to our HDPS program since it will be useful to many individuals.[2]

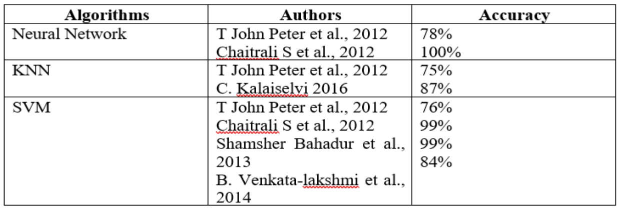
RELATED WORK

Heart disease diagnosis and prediction are major concerns. As a result, much research has been made in this subject. Existing research is divided into two groups:



Authors	Attributes
T John Peter et al., 2012 I.S Jenzi, 2013 S. Radhimeenakshi 2016	Number of Attributes: 13 (Age, Gender, CPT, FBS, RECG, Ex-Ang, SL, Col-Ves, Thal, SC, Thalach, Old peak, RBP)
Chaitrali S et al., 2012 C. Kalaiselvi 2016	Number of attributes:10 (Age, Gender, CPT, FBS, RECG, Thalach, Smoking, Alcohol, Obesity)
Shamsher Bahadur et al., 2013 Hlaudi Daniel et al., 2014	Number of attributes:6 (CPT, Ex-Ang, Col-Ves, RBP, Num, Smoking)

Table 1. The classes of the attributes used



Algorithms	Authors	Accuracy
Neural Network	T John Peter et al., 2012 Chaitrali S et al., 2012	78% 100%
KNN	T John Peter et al., 2012 C. Kalaiselvi 2016	75% 87%
SVM	T John Peter et al., 2012 Chaitrali S et al., 2012 Shamsher Bahadur et al., 2013 B. Venkata-lakshmi et al., 2014	76% 99% 99% 84%

Table 2. The most frequent cited algorithms in the papers

I. DATASET DESCRIPTION

The data originated with the CDC and is a key component of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect information on Americans' health.

The original dataset, which had nearly 350,000 variables, was reduced to little over 5,000. This dataset can be used to apply a variety of machine learning methods, including classifier models, in addition to traditional EDA (logistic regression, decision tree, random forest, etc.). The variable "heart disease" should be treated as a binary ("Yes" -

respondent had heart disease; "No" - respondent had no heart disease).

Column Descriptions

- Heart Disease:
- BMI: Body Mass Index (BMI).
- Smoking
- Alcohol Drinking:
- Stroke: (Ever told) (you had) a stroke?
- Physical Health
- Mental Health
- Different Walking
- Sex:
- Age Category
- Race
- Diabetic
- Physical Activity
- General Health
- Sleep Time: On average, how many hours of sleep do you get in a 24-hour period?
- Asthma
- Kidney Disease
- Skin Cancer

II. METHODS

a) Logistic Regression

Logistic Regression is a technique for analyzing a result from a set having any or much more self-reliant variables. The objective of Logistic Regression is to forecast the optimal link between the variables. [3]

b) Decision Tree Classifier

Decision Tree is used to create tree-like structures. While an associated decision tree is constructed sequentially, decision tree develops a smaller and decreasing subset of an issue. A decision tree can handle both qualitative and numerical values. A decision tree can learn to predict the target class by learning simple decision criteria from the dataset. We can quickly see how important a certain trait is based on the outcome of our decision tree. [3]

c) K-Nearest Neighbor Classifier (KNN)

K-Nearest Neighbor approach is a technique of data classification for calculating the possibility that a dataset will correspond with one of groups based on the attributes the data points closest to.

d) Random Forest

Random Forest is an algorithm which is supervised although it performs better in classification jobs in general. Random Forest ponder over numerous tree structure before getting a result. So, it's simply a collection of tree structures.

This strategy was predicated on the idea that a larger number of trees will converge on the right answer. For classification, it adopts a polling mechanism and thereafter selects the group, whereas for regression, it considers the mean of all tree-based results. It works well with huge, multidimensional datasets. [2]

e) Gradient Boost Classification

Gradient Boosting machines are a collection of advanced machine-learning algorithms which have shown a huge success in a range of applications. They're very adaptable to the requirements of the uses, such as learning alternative loss functions. [4]

f) Artificial Neural Network

Artificial neural networks (ANNs) are useful tools for modeling complex ecosystems because they can predict how ecosystems respond to changes in environmental variables (e.g., nutrient inputs). ANNs can also be used to identify relations between variables, which helps with understanding ecosystem function.

g) Autoencoder (AE)

Unsupervised learning experts consider AE to be a compelling model. It's a unique neural network architecture in which the output and input are the same size. An AE is divided into two phases: an encoder that compresses the data for testing and a processor that reconstructs the presented content. Figure 1 was created by Shen [5]. AE is typically trained on normal data to provide the reconstruction error (RE) is the disparity seen between rebuilt and initial versions. The RE of normal data is believed to be lower since it is identical to learning data, but the RE of aberrant data should be greater. [6]. Autoencoder outperformed LR in credit card identification and performed well on unbalanced data [5][7]. Furthermore, with supervised learning, it proved the capacity to discern suspicious transactions requiring time-consuming feature engineering[8]. As a result of this research, an Autoencoder model was proposed to investigate its capacity to discern using different batch sizes and learning rates.

ReLu activation factors are used for the encoder and decoder layers, and they've lately gained favor for minimizing training loss and improving time for training [8]. Ridge regression, commonly referred as L2, is a regularize that helps to avoid fitting problem. Additionally, the model compilation includes Adam optimizer, a Stochastic Gradient Descent modification.

h) Sampling

Oversampling balances the data by increasing the number of minorities in comparison to the majority. The synthetic minority over-sampling technique (SMOTE) is commonly used to deal with skewness in heart detection difficulties [10] and has been thoroughly explored among the oversampling methods [9]. [11]. It uses k-NN to over-sample minorities and create synthetic samples. SMOTE enhances the model's accuracy, speed of convergence, and efficiency [12]. [13]. However, this method has significant drawbacks, like the risk of imbalanced datasets as well as the data obtained not matching the original.

The above oversampling difficulties appear to be mitigated by under sampling. For this technique, the class label instances get chosen at random and merged to the minority class in a 1:1 ratio. NearMiss [14] is a widely used under sampling approach for dealing with unbalanced data. However, like other under sampling strategies, Near Miss may exclude helpful data to creating rule classifiers thereby introducing prejudice to the dataset.

III. DATA STRATEGY

The strategy for the study was broken down into five categories: loading and preprocessing data, investigating data, data preparation for the model, training data, and investigating or evaluating findings. To select the optimal performance model, six experiments were analyzed using ROC-AUC and average precision score.

A. Data loading and Preprocessing

The Kaggle dataset for heart disease was obtained & processed into a data frame for further investigation. In this data, there were no missing values. There are 18 columns and 319,794 rows in the dataset.

B. Exploratory Data Analysis

Data comprised a total of 18 features, grouped into categorized and continuous data which distribution was visualized in Figure below.

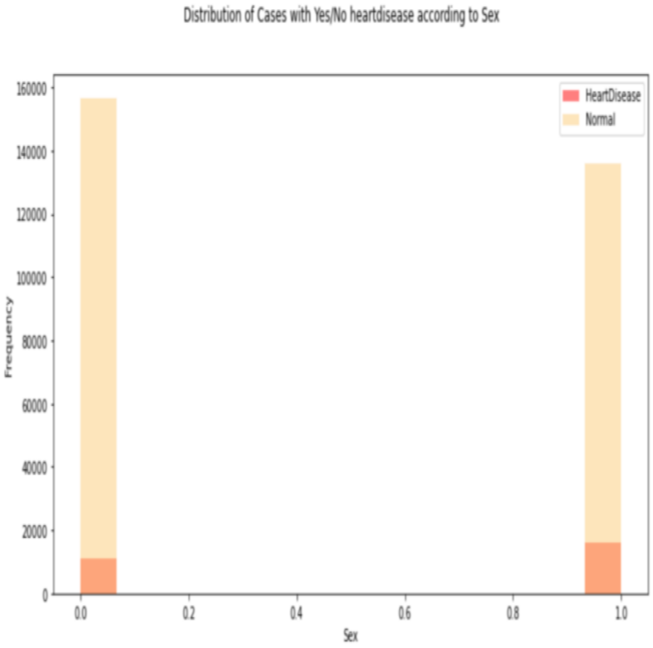


Fig 1

From the figure above, we can see that the males tend to have more heart disease than the females while the females have normal heart rate than the males.

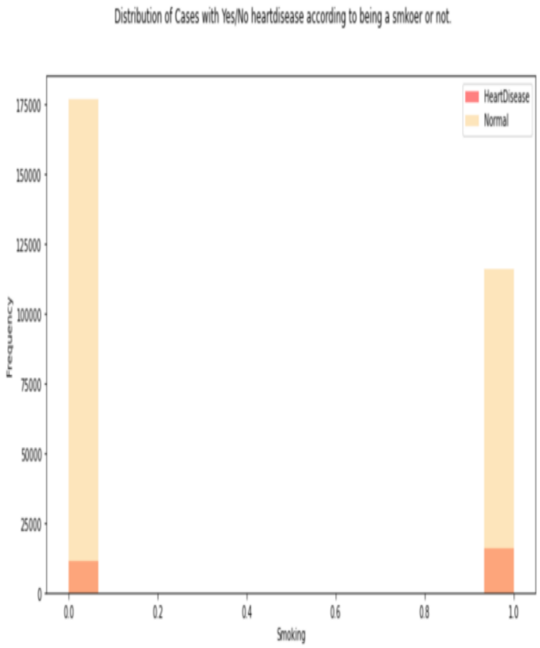


Fig 2

From the above figure, we have fewer male smokers but more with heart disease than female who have more smokers but have more person with normal heartbeat than those with heart disease.

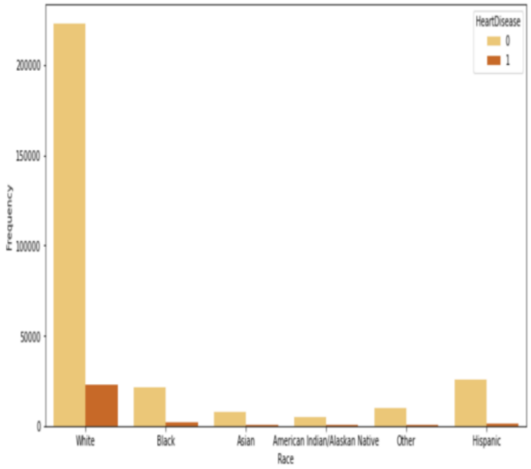


Fig 3

From the above figure, we have six (6) races with the whites the most populous and have more heart disease than other races.

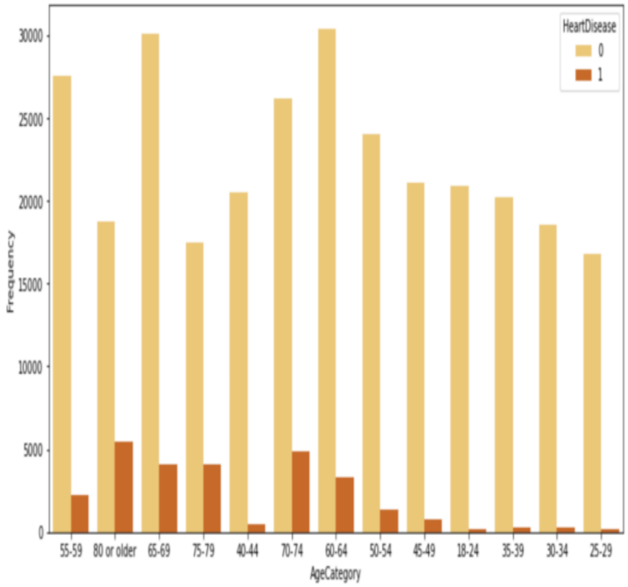


Fig 4

From the above figure, persons who are 80 or older have a higher risk of having heart diseases, seconded by 65-69 &75-79, 70-74, 60-64 and so on.

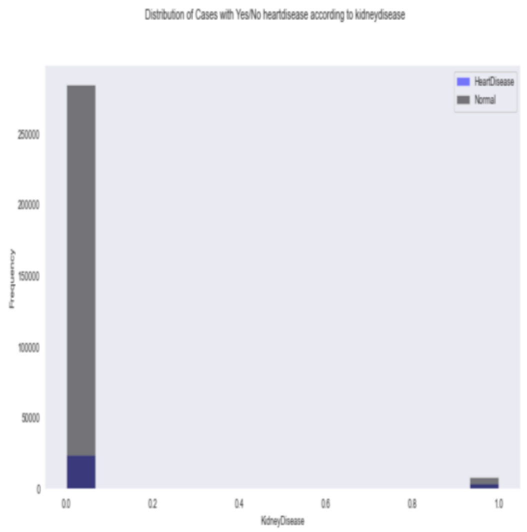


Fig 5

From the above figure, females have more kidney disease and also have more heart diseases than males with a lower population of skin disease and lower rate of heart disease (female=0, male=1).

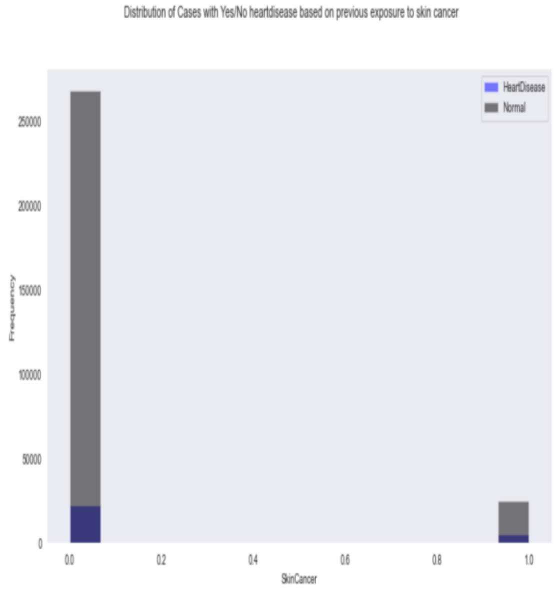


Fig 6

From the figure above, more females have exposure to skin cancer than males thereby making them have a higher rate of heart disease than males.

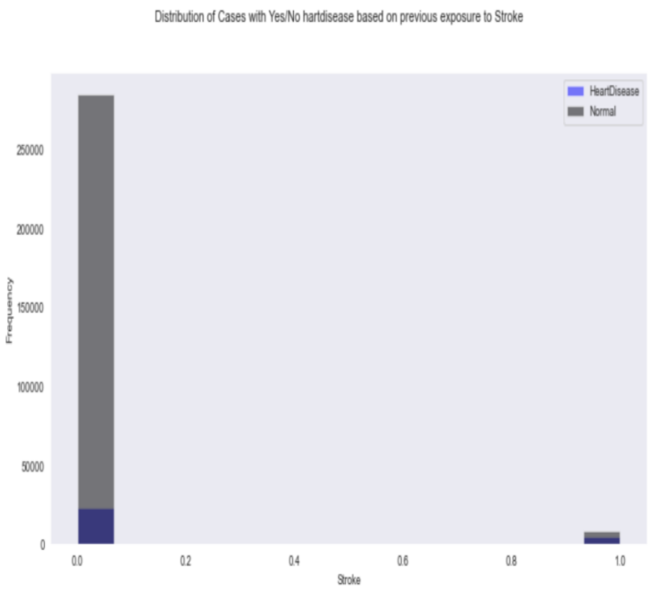


Fig 7

From the above figure, more females have previous exposure to stroke than makes and also have more rate of heart disease than males.

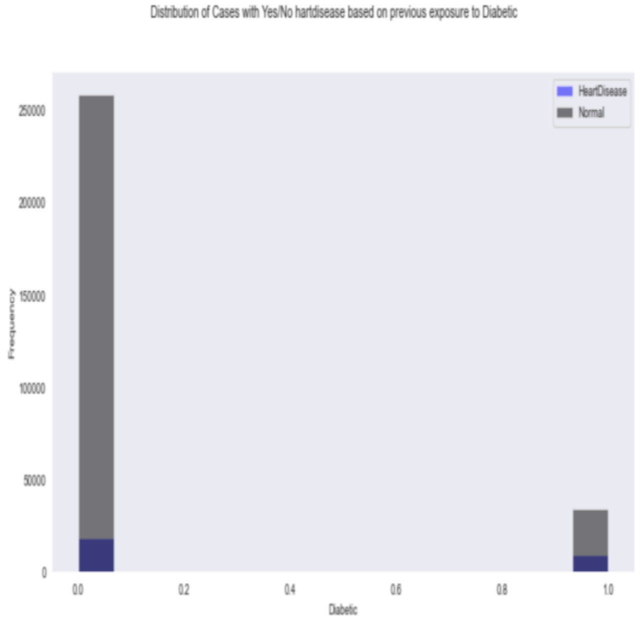


Fig 8

From the above figure, more females have high exposure to diabetes than males and therefore have more chances of heart disease than males.

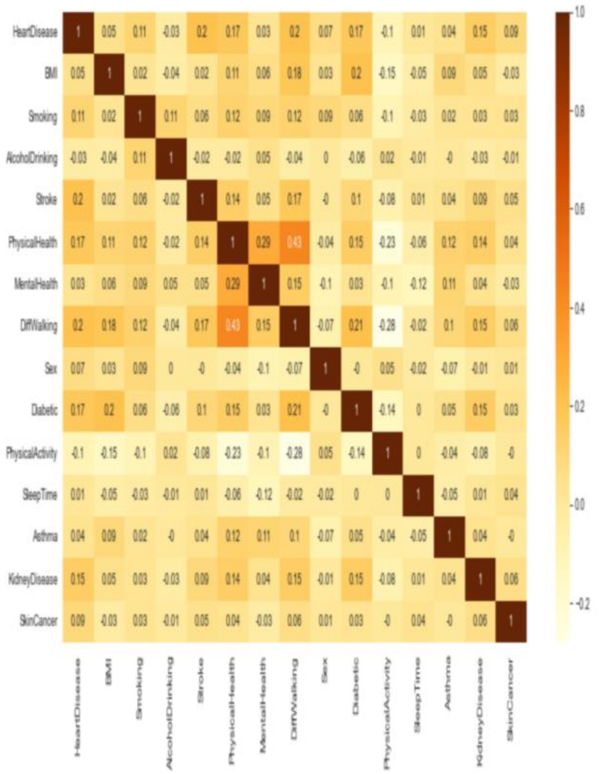


Fig 9: Correlation Matrix

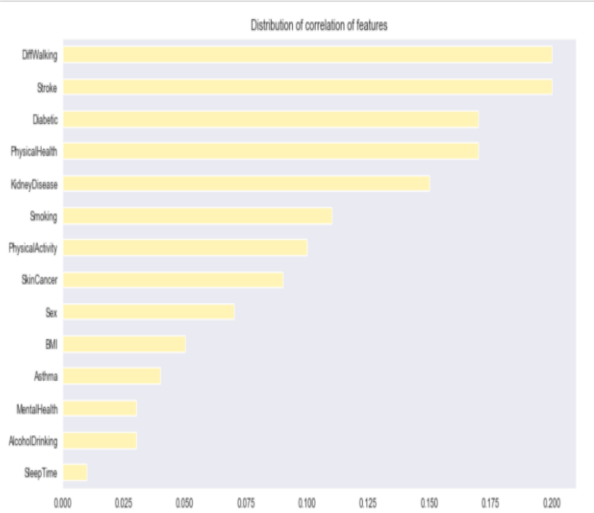


Fig 10

From the above correlation, it can be observed that Different Walking and Stroke both have a higher rate over other features.

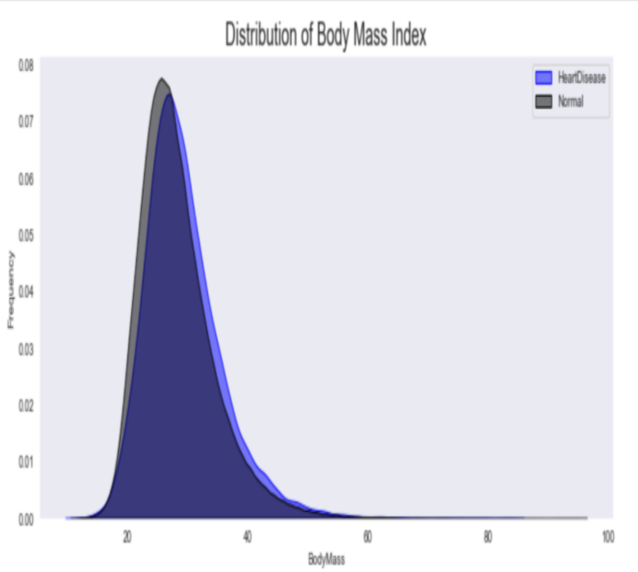


Fig 11a

Now for continuous data, we can see that with a BMI of above 20 have higher heart disease and reduces as BMI gets to 40.

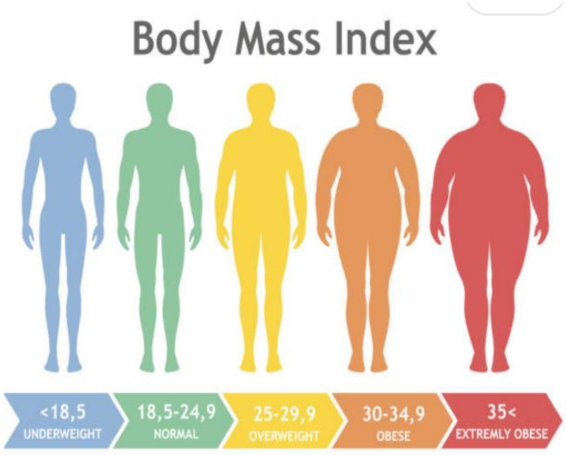


Fig 11b

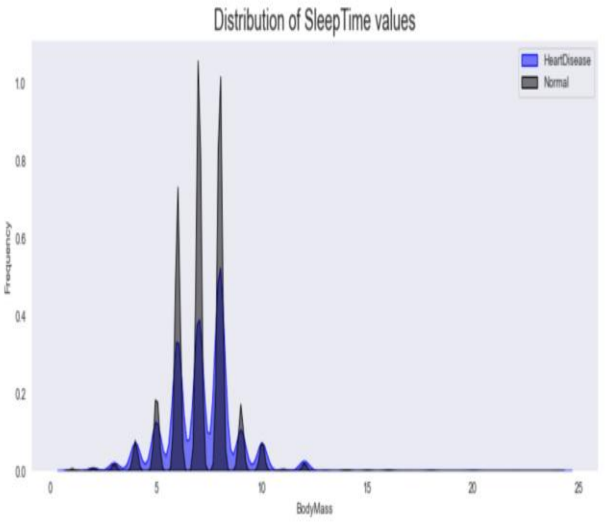


Fig 12

From the above figure, sleep time with 5hours usually have heart disease but reduces as it approaches 10.

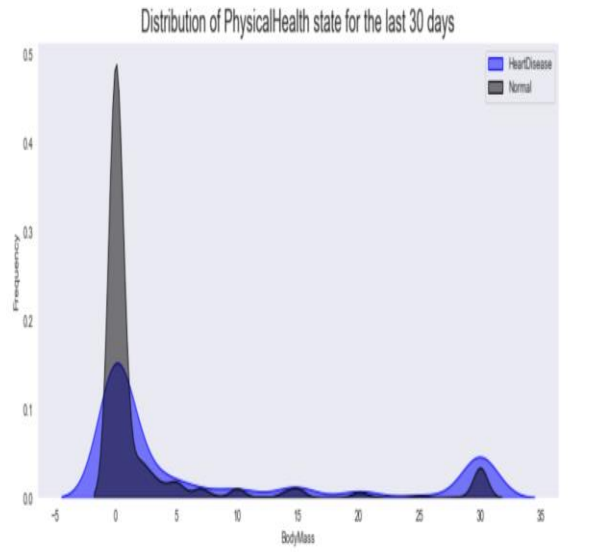


Fig 13

From the above figure, for those whose state of health in the last 30days was 0 have heart disease over others.

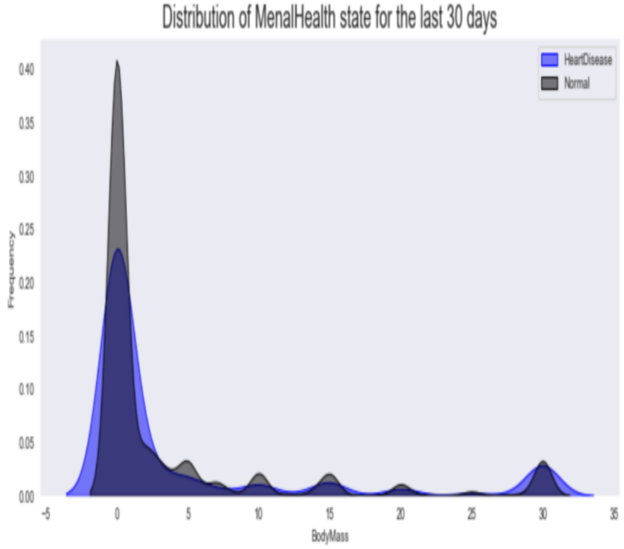


Fig 14

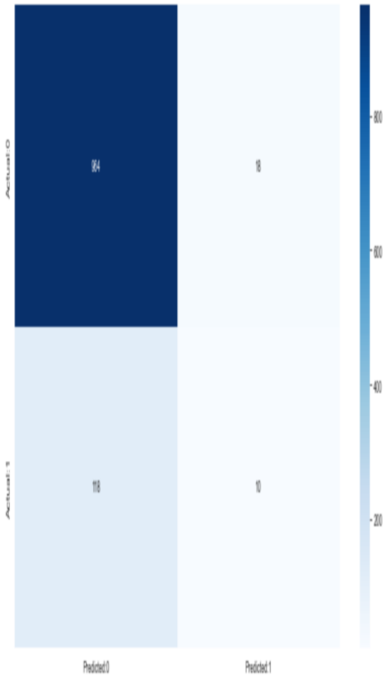
III. DATA PREPARATION FOR ML

To avoid outliers and improve model generalization, 5,550 random samples were picked, and "Standard Scaler" normalized the "Amount" feature. The columns 'Age Category,' 'Race,' and 'General Health' were eliminated due to their lack of significance. The new sample data was then visualized to ensure that the connection and dispersal were correct. The sample data was then divided into train and test sets in an 80:20 ratio

IV. RESULT

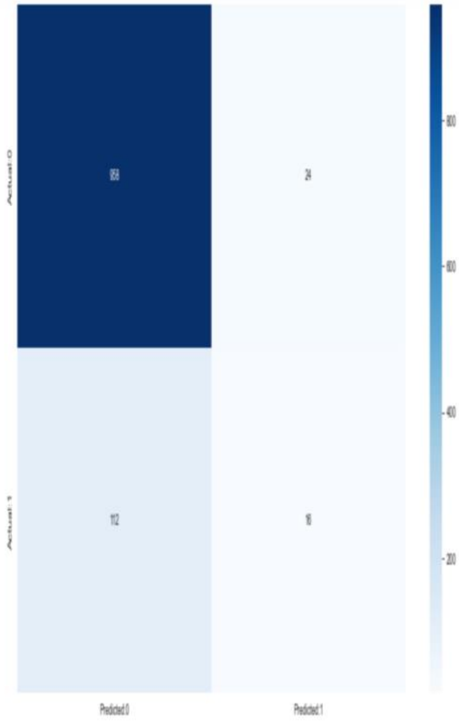
a) Logistic Regression

The train's precision is **35.71%** and its accuracy is **87.74%**. It works well, although not optimally for us. Logistic Regression has the merit of making smaller computation power and being exceptionally interpretable. Therefore, using Logistic Regression is helpful, simple and sufficient. Nevertheless, Logistic Regression has a drawback of assuming linearity between the dataset's features.



Neural Network

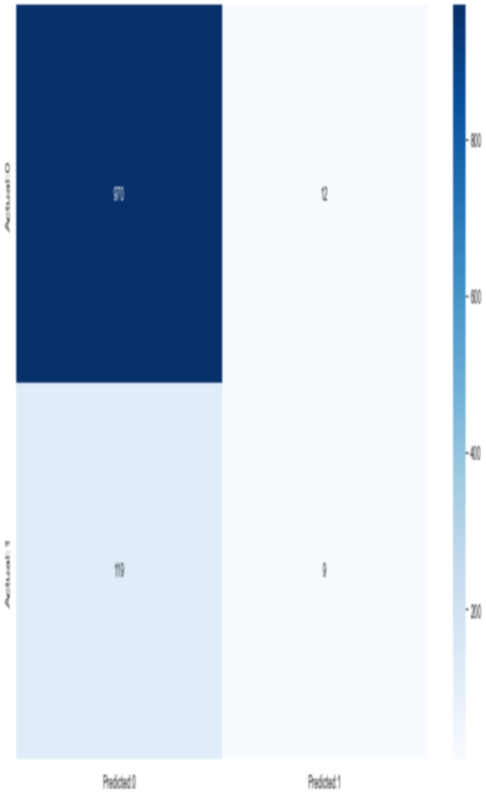
I added 6 units to the first layer and 1 unit to the second layer at the start of the operation. The dataset was then evaluated. We utilize Adam instead of SGD for the optimization because Adam is a combination of RMSprop and SGD with momentum, and it takes advantage of momentum by altering the gradient's moving average. Because our dataset is large, we set the batch size to 32, which is sufficient for training. To avoid overfitting, we run 30 epochs. The train has an accuracy of **88.01%** and a precision of **41.94%**.



The loss for test data is reaching **0.249** and test accuracy is nearing **90.4%** as the epochs grow. A neural network has the advantage of being able to handle large datasets with high dimensional features (for example, photos) and make accurate predictions by constructing numerous hidden layers. The neural network, on the other hand, does not do well with limited datasets since it becomes convoluted.

Gradient Boosting Classifier

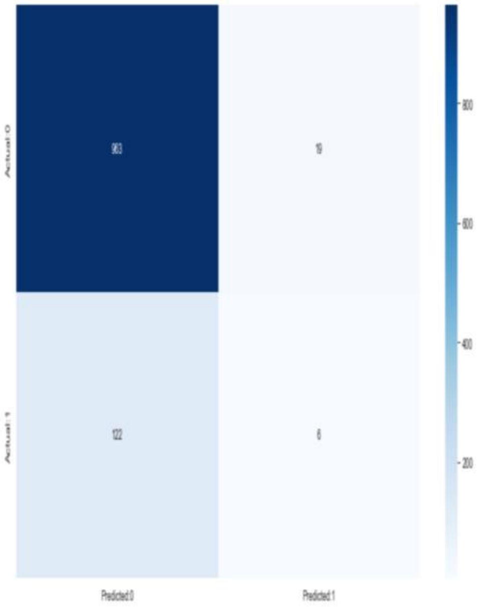
Rf.fit was used to train the Gradient Boosting Classifier model, while rf.predict was used to test it. The outcome is displayed in the table below, with P denoting predicted and A denoting actual.



The accuracy of the Gradient Boost Classifier was **88.189%**. The Random Forest model testing predicted **970** Nos and **9** Yes out of a total of 1,110. GBM's key advantage is that it frequently achieves unrivalled forecasting accuracy.

Random Forest

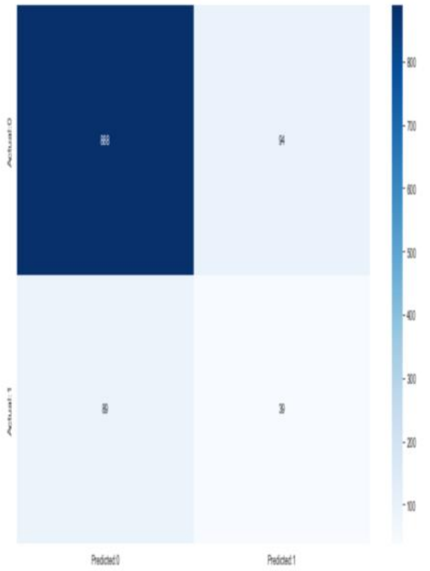
Rf.fit was used to train the random forest model, while rf. predict was used to test it. The outcome is displayed in the table below, with P denoting predicted and A denoting actual.



The accuracy of random forest was **87.29%**. The Random Forest model testing predicted **963** Nos and **6** Yes out of a total of 1,110. Random Forests have a significant edge over CART in terms of classification accuracy.

Decision Tree

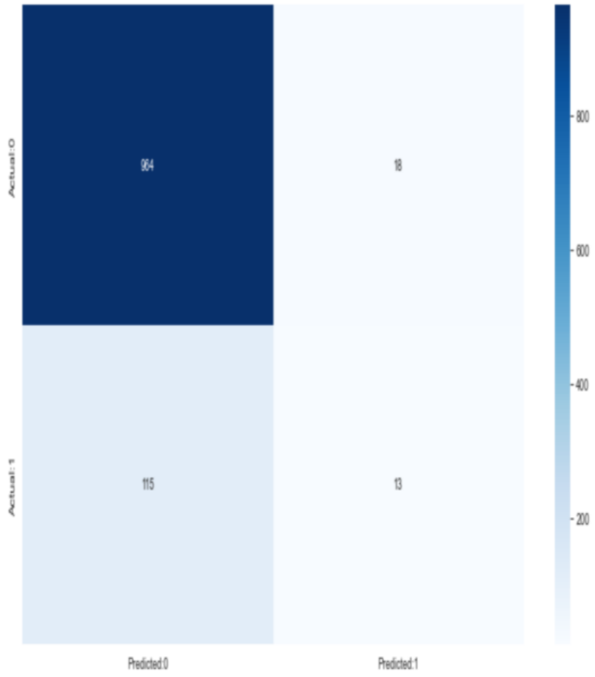
Rf.fit was used to train the random forest model, while rf. predict was used to test it. The outcome is displayed in the table below, with P denoting predicted and A denoting actual.



The accuracy of Decision Tree was **83.51%**. Random Forest model testing predicted **888** No and **39** Yes from N = 1,110. The advantage of a decision tree classifier is that it can be used for both classification and regression issues, and it's straightforward to interpret, grasp, and visualize. A decision tree's output is also simple to comprehend.

KNeighborsClassifier

The KNeighborsClassifier model was trained and tested using rf.fit and rf.predict. The outcome is displayed in the table below, with P denoting predicted and A denoting actual.



The accuracy of the KNN forest was **88.01%**. Random Forest model testing predicted **964** Nos and **13** Yes out of a total of 1,110.

This algorithm has the advantage of being adaptable to various proximity estimations, being relatively intuitive, and being memory-based.

V. CONCLUSION

The idea behind this project is to predict heart disease and predict people who have high risk of having heart disease. I used six algorithms with gradient boost classifier (GBV)having the highest accuracy of 88.189% and Decision Tree Classifier (DT) being the lowest with 83.01%.

VI. ETHICAL ISSUES

The major ethical issue is oversampling as it ends up adding multiple observations of several types, thus leading to overfitting

VII. FUTURE WORK

Looking at the focus subject, stroke is a very common disease that continues to be important, which allows for more research into forecasting, early detection, and Artificial intelligence is a very wide area that can still come up with potential solutions and algorithms that can aid this out comes in the health sector.

REFERENCE

1. Baban.U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade. Heart Disease Prediction Using Machine Learning, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 5, Issue 1, May 2021, PP. 267-273.
2. Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain & Preeti Nagrath. Heart disease prediction using machine learning algorithms, Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1, pp.1-11.
3. S.Nandhini, Monojit Debnath, Anurag Sharma, Pushkar. Heart Disease Prediction using Machine Learning, International Journal of Recent Engineering Research and Development (IJRERD) ISSN: 2455-8761 www.ijrerd.com || Volume 03 – Issue 10 || October 2018 || PP. 39-46 39.
4. Alexey Natekin & Alois Knoll. Gradient boosting machines, a tutorial, Frontiers in Neurorobotics www.frontiersin.org December 2013 | Volume 7 | Article 21, pp.1-21.

5. [14] J. Shen, “Fraud Detection Using Autoencoder-Based Deep Neural Networks,” Apr. 2021, pp. 673–677, doi: 10.1109/icbaie52039.2021.9389940.

6. [15] R. C. Aygun and A. G. Yavuz, “Network Anomaly Detection with Stochastically Improved Autoencoder Based Models,” in *Proceedings - 4th IEEE International Conference on Cyber Security and Cloud Computing, CSCloud 2017 and 3rd IEEE International Conference of Scalable and Smart Cloud, SSC 2017*, Jul. 2017, pp. 193–198, doi: 10.1109/CSCloud.2017.39.
7. [16] M. A. Al-Shabi, “Credit Card Detection Using Autoencoder Model in Unbalanced Datasets,” *J. Adv. Math. Comput. Sci.*, vol. 33, no. 5, pp. 1–16, Aug. 2019, doi: 10.9734/jamcs/2019/v33i530192.
8. [17] A. Alazizi, A. Habrard, F. Jacquenet, L. He-Guelton, and F. Oblé, “Dual Sequential Variational Autoencoders for Fraud Detection,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Apr. 2020, vol. 12080 LNCS, pp. 14–26, doi: 10.1007/978-3-030-44584-3_2.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
10. A. Mishra and C. Ghorpade, “Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques,” in *2018 IEEE International Students’ Conference on Electrical, Electronics and Computer Science, SCEECS 2018*, Nov. 2018, doi: 10.1109/SCEECS.2018.8546939.
11. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, Feb. 2011, doi: 10.1016/j.dss.2010.08.008.
12. R. Sailusha, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao, “Fraud Detection Using Machine Learning,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, May 2020, pp. 1264–1270, doi: 10.1109/ICICCS48265.2020.9121114.
13. D. Almhaithawi, A. Jafar, and M. Aljnidi, “Example-dependent cost-sensitive credit cards detection using SMOTE and Bayes minimum risk,” *SN Appl. Sci.*, vol. 2, no. 9, p. 1574, Sep. 2020, doi: 10.1007/s42452-020-03375-w.
14. C. V. Priscilla and D. P. Prabha, “Influence of optimizing xgboost to handle class imbalance in credit card detection,” in *Proceedings of the 3rd*

International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Aug. 2020, pp. 1309–1315, doi: 10.1109/ICSSIT48917.2020.9214206.

15. https://towardsdatascience.com/k-nearest-neighbours-explained-7c49853633b6