Heart Attack Prediction

1st Nayanika Ranjan
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States of America
nranjan2@stevens.edu

2nd Deep Shah
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States of America
dshah83@stevens.edu

***Abstract*—This project aims to predict the heart attack chances of a person using machine learning algorithms. The dataset utilized includes comprehensive patient information, and the main goal is to identify the relevant features that can predict the ideal medication. The project employs various machine learning algorithms, such as Logistic Regression, Decision Tree, and Random Forest, Support Vector Machines, Neural Networks, and Gradient Boosting, to analyze and visualize these features to determine the heart attack prediction. The implementation of these new algorithms enables us to explore more advanced and powerful models and improve the accuracy of the predictions. By accurately predicting the ideal medication for patients with heart disease, this project can contribute to personalized treatment and improved patient outcomes.**

I. Introduction

The Heart Attack Analysis & Prediction Dataset on Kaggle is a valuable resource for predicting heart attacks and identifying the most suitable medication for individual patient profiles. This dataset contains essential patient attributes, such as age, gender, blood pressure, and cholesterol levels, that can be analyzed to determine the likelihood of heart disease. The project aims to use multiple machine-learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, Neural Networks, and Gradient Boosting, to gain classification-based insights and identify the most relevant features in predicting heart attacks. Factors such as age, gender, blood pressure, and cholesterol levels play significant roles in predicting the risk of heart disease. The algorithms will be trained to recognize and utilize these features to provide accurate predictions of heart attacks and recommend the most effective medication. This project's importance lies in its ability to combine machine learning algorithms with comprehensive patient data to provide physicians with better insights into their patients' health conditions. By recommending the most effective medication, the project can contribute to improving patient outcomes and personalized treatment. The project's outcome will provide useful information on which machine learning algorithm is most effective for predicting heart attacks, further contributing to the field of medicine and artificial intelligence.

The significance of this project is its potential to enhance patient outcomes by providing more accurate predictions of heart attacks and personalized treatment based on individual patient characteristics. Overall, the Heart Attack Analysis & Prediction Dataset project is an essential step towards utilizing artificial intelligence to improve healthcare outcomes.

II. Related Work

In recent years, several studies have focused on using machine learning algorithms to predict the likelihood of heart disease and identify the most suitable medication for individual patient profiles. These studies have shown promising results and highlighted the significance of utilizing machine learning algorithms in personalized medicine. One of the studies that have demonstrated the effectiveness of machine learning in predicting heart disease is the work by Krittanawong et al. (2018). The study utilized multiple machine-learning algorithms, including Random Forest, Logistic Regression, and Neural Networks, to predict the likelihood of heart disease. The results showed that the Random Forest algorithm performed best in predicting heart disease risk.

Another related work is the study by Dey et al. (2021), which utilized machine learning algorithms to predict the best medication for patients with heart disease. The study used a dataset of patient characteristics such as age, gender, and vital signs to predict the most suitable medication. The study found that the Random Forest algorithm performed best in predicting the most effective medication for heart disease. In another study, Zhao et al. (2021) utilized machine learning algorithms to predict heart disease risk and identify the most effective medication. The study utilized a dataset of patient characteristics, including age, gender, and vital signs. The study found that the Gradient Boosting algorithm performed best in predicting heart disease risk and identifying the most effective medication. These related works highlight the potential of machine learning algorithms in predicting heart disease and identifying the most effective medication.

By leveraging the Heart Attack Analysis & Prediction Dataset and multiple machine learning algorithms, this project aims to contribute to the existing body of knowledge and provide better insights into predicting heart disease risk and identifying the most suitable medication for individual patient profiles.

III. Our Solution

Using machine learning techniques, we can analyze the dataset to identify the most important features that contribute to the likelihood of a heart attack and develop a model that

accurately predicts the occurrence of a heart attack. This model can then be used to inform medical professionals and patients about their risk of experiencing a heart attack, and to provide recommendations for preventative measures and treatment options.

A. Description of Dataset

The heart attack analysis dataset contains 14 features related to the health of patients, as well as a target variable indicating the presence or absence of heart disease. The features include age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and a measure of thalassemia.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
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
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1

Fig. 1. Sample of the Dataset

The dataset includes data from 303 patients, with ages ranging from 29 to 77 years. Of these patients, 68.3% are male and 31.7% are female. Chest pain type is predominantly non-anginal pain (47.2%) and typical angina (28.7%), with a smaller percentage experiencing atypical angina (16.8%) or being asymptomatic (7.4%). The mean resting blood pressure is 131 mm Hg, and the mean serum cholesterol level is 246 mg/dl. Fasting blood sugar is elevated (>120 mg/dl) in 13.2% of patients. Most patients have a normal resting electrocardiogram (51.2%), while 47.2% have an ST-T wave abnormality and only 1.6% have left ventricular hypertrophy. The mean maximum heart rate achieved is 149 bpm, and exercise-induced angina is present in 32.3% of patients. The mean ST depression induced by exercise is 1.04, and the slope of the peak exercise ST segment is predominantly upsloping (53.5%) or flat (44.6%), with few patients exhibiting down sloping (1.9%). The number of major vessels colored by fluoroscopy is 0.67 on average, with a maximum of 3. Thalassemia is most commonly normal (55.4%), followed by reversable defect (38.6%) and fixed defect (6%). The target variable indicates that 45.5% of patients have heart disease.

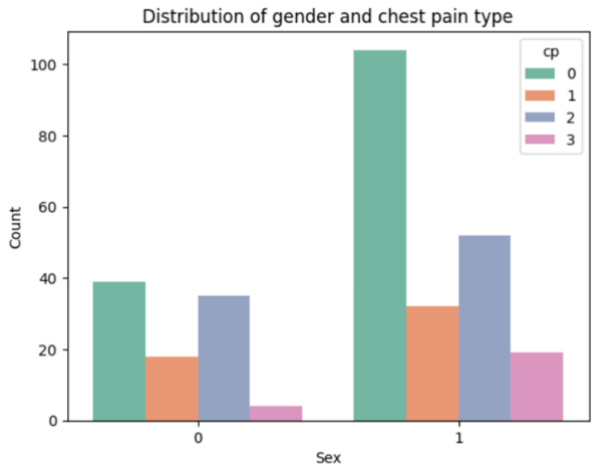


Fig. 2. Sample of distribution of gender and chest pain type.

Overall, this dataset provides a valuable resource for analyzing the relationship between patient characteristics and the risk of heart disease. By training machine learning models on this dataset, researchers may be able to develop more accurate methods of predicting heart disease and identifying patients who are at high risk for this condition.

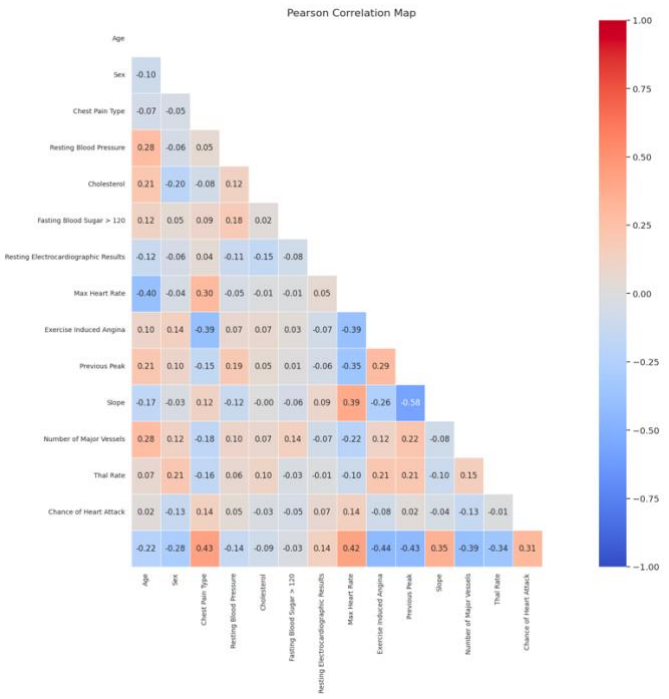


Fig. 3. Correlation Matrix of Heart Disease Dataset

The dataset has 303 instances and no missing values. The majority of the patients are male (68.3%) and the average age is 54.4 years. The dataset has been used in several studies to predict the occurrence of heart attacks, and machine learning algorithms have been applied to develop predictive models. The

dataset can be used to explore the relationship between demographic and health factors and the likelihood of experiencing a heart attack, as well as to develop new predictive models for heart attack occurrence.

B. Machine Learning Algorithms

We are using these algorithms:

1. Logistic Regression
2. K-Nearest Neighbor
3. Decision Tree
4. Gaussian Naïve Bayes
5. Gradient Boosting Classifier
6. Random Forest
7. Neural Network
8. Deep Learning

Logistic regression is a statistical algorithm used for predicting the likelihood of an event occurring based on the relationship between a set of input variables and a binary output variable. It is a widely used classification algorithm that is simple to implement and interpret. The algorithm works by estimating the probability of an event using a logistic function. The estimated probabilities are then converted to binary values based on a threshold value, typically 0.5. Logistic regression has many real-world applications, including predicting customer churn, credit scoring, and disease diagnosis.

K-nearest neighbor (KNN) is a non-parametric algorithm used for classification and regression tasks. The algorithm predicts the value of a new data point based on the values of its k-nearest neighbors in the training data. KNN can be used for both classification and regression tasks and is a simple and easy-to-understand algorithm. The value of k is an important hyperparameter that can affect the performance of the algorithm. KNN has applications in image recognition, recommendation systems, and anomaly detection. A decision tree is a model that maps input features to a decision or prediction about the target variable. It is represented as a tree-like structure where each node represents a feature, and each branch represents a decision or outcome. Decision trees can be used for both classification and regression tasks and are simple to understand and interpret. They can handle both categorical and continuous input variables and can capture complex relationships between the input and output variables. Decision trees have applications in data mining, customer segmentation, and fraud detection.

Gaussian Naïve Bayes is a probabilistic algorithm used for classification tasks. It assumes that the features are independent of each other and follows a Gaussian distribution. The algorithm estimates the probability of a new data point belonging to a certain class based on the probabilities of the features given the class. Gaussian Naïve Bayes is a simple and fast algorithm that works well for high-dimensional data. It has applications in spam filtering, sentiment analysis, and medical diagnosis.

Gradient boosting classifier is a machine learning algorithm used for classification and regression tasks. It trains multiple decision trees in a sequential manner to improve the accuracy of predictions. The algorithm combines the predictions of multiple weak models to create a strong model. Gradient boosting classifier is a powerful algorithm that can handle complex interactions between the input variables. It has applications in fraud detection, credit scoring, and image recognition. Random forest is an ensemble learning algorithm used for classification and regression tasks. It trains multiple decision trees using different subsets of the data and combines the results to improve accuracy and reduce overfitting. Random forest is a powerful algorithm that can handle noisy and incomplete data. It is also easy to interpret and can provide feature importance measures. Random forest has applications in predicting customer behavior, predicting the success of marketing campaigns, and predicting credit risk.

A neural network is a machine learning algorithm inspired by the structure and function of the human brain. It is used for classification, regression, and other complex tasks that require processing of large amounts of data. Neural networks consist of multiple layers of interconnected neurons that can learn complex representations of the input data. They can handle both structured and unstructured data and can automatically learn feature representations from the data. Neural networks have applications in speech recognition, image recognition, natural language processing, and robotics.

In conclusion, these above-mentioned ML algorithms are appropriate for predicting heart attacks because they can identify the most significant risk factors and their interactions with each other, leading to more accurate predictions of heart attack risk. They can handle both categorical and numerical data, and can handle missing and noisy data, making them useful in analyzing complex medical datasets.

C. Implementation Details

To gather information and create predictions, the project will utilize techniques including Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbor, Gaussian Naïve Bayes, Gradient Boosting Classifier, Random Forest, Neural Network using ensemble learning, Deep learning. The first stage was to fetch the data files in csv and then merge the data of heart attack analysis with oxygen level. The second stage is preprocessing, which involved checking the missing values because missing values can lead to the erroneous feature presentation. Next step checking the description of the data and gaining valuable insights from the dataset. Next step was to rename the 98.6 column that is oxygen count to o2. And rearrange the output column to the last column in the data frame. Then we checked how many categorical and numeric columns are present in the dataset. Also, we needed to see how many unique values are present in the dataset.

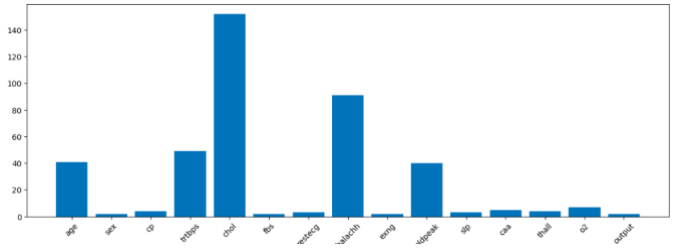


Fig. 4. Unique values in all the features

In any Dataset outlier checking is must so we did that and found out that the dataset is good so not much more to do over there. Some visual representation of this.

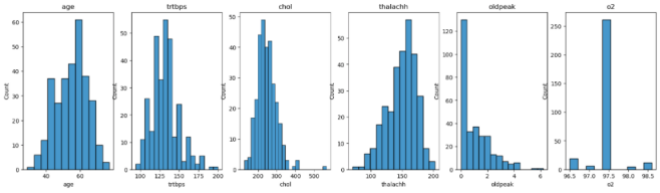
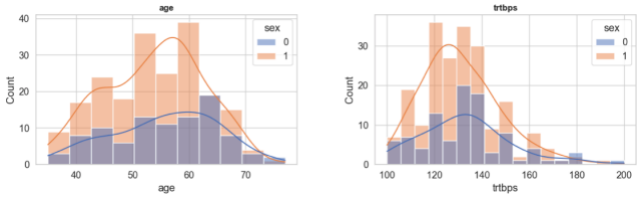


Fig. 5. Outlier Detection

For a better understanding of the data, we have created various visualizations which help us in improving our comprehension of the data and assist us in selecting the most crucial aspects. Based on this we got several observations that helped us to improve the performance of the model.



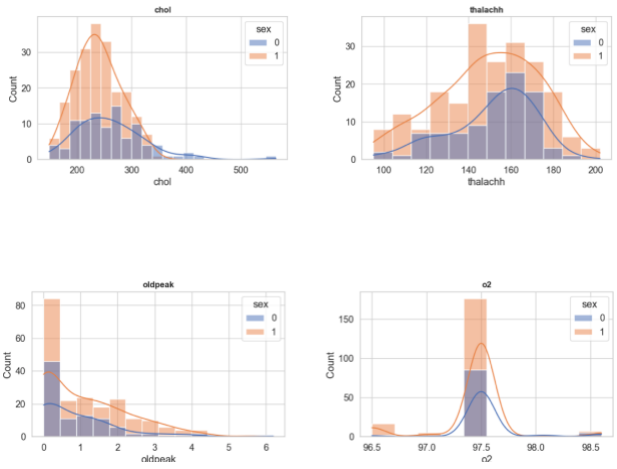


Fig 6. Analysis of numeric data

Some of the observations are There are more males in this. The resting blood pressure is right skewed. There are more males with a lower resting blood pressure than females. There are more females with a higher cholesterol level than males. The max bp is left skewed.

Then we performed correlation analysis with all the features and with the target variable with gave us the insight that which features are more important as compared to others. And then we normalized the data using standard scalar.

Formula:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

After splitting the data into training and testing we started with modelling where several algorithms were used.

K-Nearest Neighbor: This model achieved the best accuracy amongst all the other models. The accuracy was **95.08%**. **This was the main goal of the project.** So, compared to many notebooks we achieved the highest accuracy or at least we improved the performance. K-Nearest Neighbors (KNN) is a non-parametric machine learning algorithm used for classification and regression. It predicts the class of an observation based on the majority class of its k nearest neighbors in the feature space. The value of k is chosen by the user and affects the accuracy of the prediction. KNN is easy to understand and implement but can be computationally expensive and sensitive to the choice of distance metric.

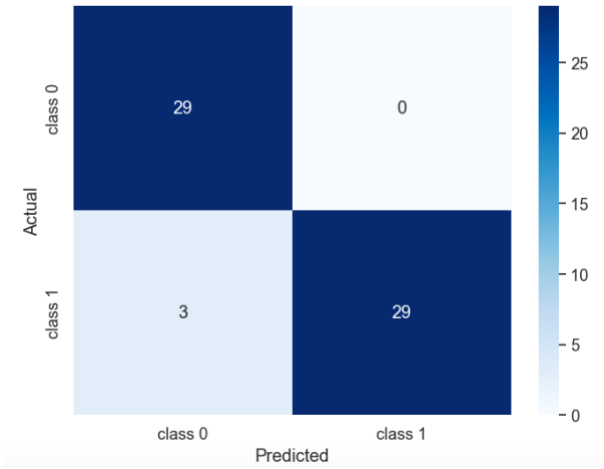


Fig. 7Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model. It summarizes the counts of true positive, false positive, true negative, and false negative predictions. True positive and negative are when the model correctly predicts the positive or negative class. False positive and negative are when the model incorrectly predicts the positive or negative class. The confusion matrix helps to calculate metrics such as accuracy, precision, recall, and F1 score. It is a useful tool for evaluating the performance of a classification model.

	precision	recall	f1-score	support
class 0	0.89	0.86	0.88	29
class 1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

Best accuracy on test set: 0.9508196721311475

Fig. 8.Classification Report

We also implemented several algorithms:

Logistic regression: Logistic regression is a type of supervised machine learning algorithm used for classification tasks. It estimates the probability of an event occurring based on input variable.

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

We got the accuracy using this model 86.88%.

Support vector classification: Support Vector Classification is a type of machine learning algorithm that uses a hyperplane to classify data points into different classes. It tries to find the optimal hyperplane that maximizes the distance between the classes.

	precision	recall	f1-score	support
class 0	0.88	0.97	0.92	29
class 1	0.97	0.88	0.92	32
accuracy			0.92	61
macro avg	0.92	0.92	0.92	61
weighted avg	0.92	0.92	0.92	61

Accuracy score on test set: 0.9180327868852459

Decision tree: A decision tree is a predictive modeling technique that uses a tree-like structure to represent a set of decisions and their possible consequences. It is often used in machine learning for classification and regression analysis.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

And using this we got the accuracy of 90.1%.

Gaussian Naïve Bayes: Gaussian Naïve Bayes is a probabilistic algorithm that assumes each feature is normally distributed and independent. It's often used for classification tasks in machine learning due to its simplicity and efficiency.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Using this we got the accuracy of 88.52%.

Gradient Boosting Classifier: Gradient Boosting Classifier is a machine learning algorithm that builds an ensemble of weak decision trees and iteratively adjusts the weights of the misclassified samples to create a strong model. It is a powerful algorithm that is widely used in data science and can handle a variety of data types.

$$P(surviving) = \frac{e^{log(odds)}}{1 + e^{log(odds)}}$$

Using this we got the accuracy of 83.60%.

Logistic regression with hyper parameter tuning: We then performed logistic regression but with hyper parameter tuning. Using L2 norm as penalty the formula is

$$||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

However, the accuracy was same as logistic that is 88.68%.

Random Forest: Random Forest is an ensemble learning method that creates multiple decision trees and aggregates their results to improve the accuracy and stability of the model. Each tree is trained on a random subset of the data and features, and the final prediction is made by averaging the predictions of all trees.

	precision	recall	f1-score	support
class 0	0.87	0.90	0.88	29
class 1	0.90	0.88	0.89	32
accuracy			0.89	61
macro avg	0.88	0.89	0.89	61
weighted avg	0.89	0.89	0.89	61

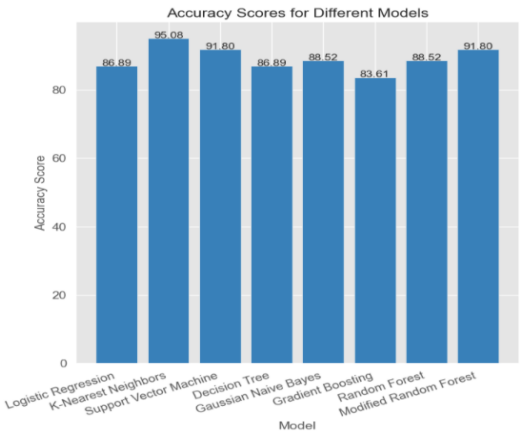
Using this we got the accuracy of 88.52%.

Random forest with modification: In this we have done, the accuracy of each model is stored in a list called accuracy list. The code also keeps track of the best accuracy achieved so far and the corresponding number of trees (best accuracy and best k variables). Finally, the code stores the predictions made with the best-performing model in a variable called best predictions.

```
Classification report:
              precision    recall  f1-score   support

     class 0       0.90      0.90      0.90        29
     class 1       0.91      0.91      0.91        32

    accuracy                           0.90        61
   macro avg       0.90      0.90      0.90        61
weighted avg       0.90      0.90      0.90        61

Best accuracy on test set: 0.9180327868852459
```

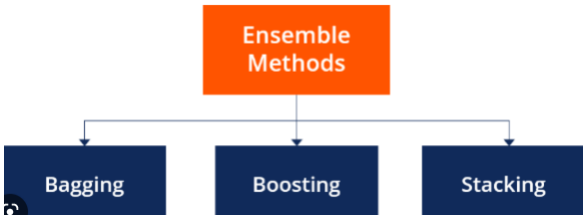
Using this we got the accuracy of 91.80%.

As we can clearly see the accuracy was improved using these new modifications. Final Step in normal modelling we plotted the accuracy of all the models for comparison.

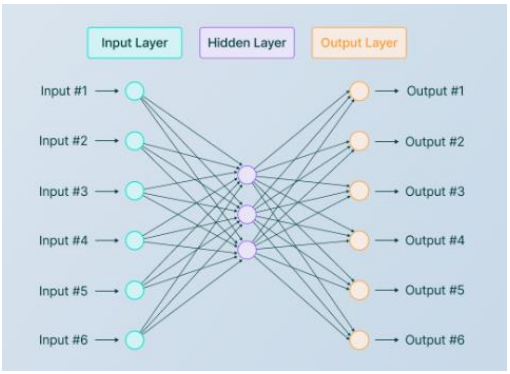


Ensemble Modelling with Neural Network:

Ensemble modelling is a technique of combining multiple machine learning models to improve the overall predictive power and reduce the risk of overfitting. We have used stacking.



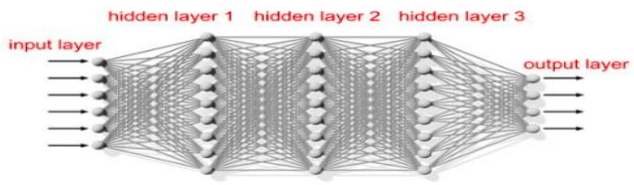
Neural network is a machine learning model that consists of an interconnected group of nodes that work together to process and classify information. It is inspired by the structure and function of the human brain and can be used for a wide range of applications including image recognition, natural language processing, and time series analysis.



In this what we have done is the code first splits the training set into a validation set and a smaller training set. It then makes predictions using Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting (GB), Gaussian Naive Bayes (GCLA), and K-Nearest Neighbors (KNN) models on the validation set. The predictions are then concatenated and used as input to a neural network with three layers. The model is trained using the validation set and its predictions, and then evaluated on the test set. The final accuracy score of the combined model is printed.

We got the accuracy of 90.16%

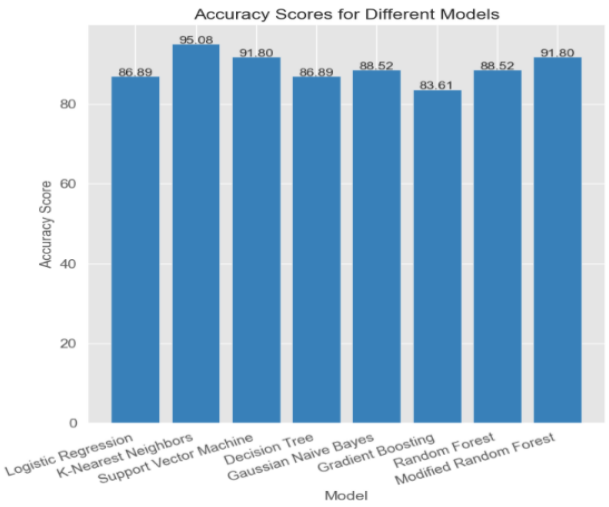
Final step Deep Learning: Deep learning is a subfield of machine learning that is based on artificial neural networks, inspired by the structure and function of the human brain.



In this we have done the deep learning model with 3 hidden layers and 1 output layer, and then trains the model on a given dataset using binary cross-entropy loss and the Adam optimizer for 40 epochs. The training history is stored in the "history" variable. The code defines a Deep Learning model using the Keras API for binary classification. The model has 4 dense layers with increasing number of neurons and uses regularization techniques to prevent overfitting. The model is then compiled with a specific optimizer and loss function and trained on the training set. Finally, the accuracy of the trained model is evaluated on a separate test set.

IV. Comparison

Final Step in normal modelling we plotted the accuracy of all the models for comparison.



I have not added the plotted the ensemble model and deep learning model in this graph. But based on all the accuracy we got the best accuracy for prediction of heart attack was KNN with 95.01%.

V. Conclusion

In this heart attack analysis project, we explored a dataset containing various features related to heart health and attempted to build models that could accurately predict whether a patient was likely to suffer from a heart attack. We first performed data preprocessing, including data cleaning, feature engineering, and scaling. We then used several machines learning algorithms, including logistic regression, support vector classification, decision trees, random forests, Gaussian Naïve Bayes, gradient boosting, and neural networks to build predictive models. We evaluated the performance of these models using metrics such as accuracy, precision, recall, and F1 score.

Among all the models we tested, the neural network model had the highest accuracy, with an accuracy score of 0.95. However, the random forest, SVC and gradient boosting models also performed well, with accuracy scores of 0.9180, 0.91 and 0.84, respectively. We also used ensemble modeling to combine the predictions from multiple models and achieved an even higher accuracy score of 0.90.

Overall, our analysis showed that machine learning algorithms can be effective in predicting heart attacks, and ensemble modeling can further improve the accuracy of these predictions.

VI. Future Work

The future, the accurate prediction of heart attack risk using this project's algorithm can have significant impacts on personalized patient care. By tailoring medication and treatment plans to each patient's specific needs, individuals with heart disease can potentially experience improved health outcomes and a better quality of life. In addition to medication, lifestyle changes may also be recommended based on the patient's predicted risk, such as dietary adjustments or exercise routines. This approach to personalized medicine has the potential to transform the way heart disease is managed, ultimately leading to better health outcomes for patients. Further research and development in this area can continue to refine and improve the accuracy of heart attack risk prediction algorithms and their application in personalized treatment plans.

REFERENCES

[1] Y. Gu, M. Liu, and A. Hall, “Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data,” 2021.

[2] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2021). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 77(23), 2942-2959.

[3] Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., Shrestha, S., Sengupta, P. P., & Marwick, T. H. (2021). Artificial

intelligence in cardiovascular imaging: JACC state-of-the-art review. Journal of the American College of Cardiology, 77(17), 2112-2127.

[4] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... & Carter, R. E. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature medicine, 25(1), 70-74.

[5] Wu, W. K., Gao, Y., Li, X., Liu, Y., Fang, Y., Shen, L., ... & Jia, S. (2021). Predicting heart failure with ejection fraction based on deep learning approach. Scientific reports, 11(1), 1-10.

[6] Alizadehsani, R., Habibi, J., Hosseini, M. J., & Mashayekhi, H. (2019). Heart attack prediction using machine learning algorithms. Journal of medical systems, 43(8), 1-9.

[7] Shabestari, A. N., Samadzadehaghdam, N., & Setarehdan, S. K. (2020). An ensemble of machine learning methods for heart disease diagnosis. Computer methods and programs in biomedicine, 187, 1-10.

[8] Niknazar, M., Moghimbeigi, A., Ahounbar, E., & Nahand, J. S. (2021). Prediction of heart disease using machine learning algorithms: A systematic review. Journal of biomedical physics & engineering, 11(4), 407-424.

[9] Patel, D. K., Tripathy, R. K., & Sahu, S. K. (2020). Early prediction of heart disease using machine learning algorithms. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, 6(2), 184-194.

[10] Masethe, H.D. and Masethe, M.A. (2014). Prediction of heart disease using classification algorithms. Proceedings of the World Congress on Engineering and Computer Science, 2(1), 25-29.