Analyze accuracy of machine learning algorithms for Heart Attack prediction

Abdullah Khondoker [1], Enam Ahmed Taufik [1], MD. Iftekhar Islam Tashik [1], Antara Fairuz Parsa [1], Prima Sarker [1]
1. School of Data Science, Brac University, 66 Mohakhali, Dhaka-1212

Abstract- The human heart is a complex organ that can continue to function even if it has suffered damage, although this may weaken the heart and reduce its ability to pump blood efficiently. To reduce the risk of further injury, it is important to promptly identify and treat potential problems, as well as make dietary adjustments after a heart attack. In this study, various machine learning algorithms, including Logistic regression, K-nearest neighbor classification, Random forest, Decision tree, Naive Bayes classification, Neural network classification, and Support vector classification, were used to predict the likelihood of a heart attack. The results showed that the Support vector classification algorithm had the highest accuracy of around 80.21%. This study also included an analysis of the correlation between different features and the visualization of mentioned features.

I. INTRODUCTION

Machine learning has the potential to revolutionize the way we predict and prevent heart attacks. By training algorithms on data that includes a variety of factors known to be associated with heart attacks, such as age, blood pressure, cholesterol levels and lifestyle factors like diet and exercise, machine learning can help identify patterns in the data that may indicate a high risk of a heart attack. These algorithms can then be used to make predictions about the likelihood of a heart attack for a new individual based on their data. While the accuracy of these predictions can vary, machine learning has the potential to significantly improve our ability to identify and intervene with individuals at high risk of a heart attack, potentially leading to better outcomes and fewer deaths from heart attacks. However, it is important to note that these predictions should always be interpreted and used in conjunction with a medical professional's advice. So, this paper aims to achieve better accuracy using various machine learning algorithms to make the system more efficient and predict the chance of heart attack with the highest verity.

.

II. METHODOLOGY

2.1. Data Collection & Preprocessing

The dataset used was heart attack analysis and prediction, which consists of 14 attributes. Therefore, we have used mostly processed datasets available on the Kaggle website for our analysis.[1] But there were some duplicate values which we changed. The complete description of the 14 attributes used is mentioned in Table 1 below.

Table 1. Features selected from dataset

Sl No.	Attribute Description	Distinct Values of Attribute
1.	age- Age of the patients'	Multiple values between 29 & 77
2.	sex- Sex of the patients' (0-Female, 1-Male)	0,1

3.	cp- chest pain type	0,1,2,3
4.	trtbps- resting blood pressure (in mm Hg)	Multiple values between 94 & 200
5.	chol- cholesterol in mg/dl fetched via BMI sensor	Multiple values between 126 & 564
6.	fbs- fasting blood sugar > 120mg/dl (1-True, 0-False)	0,1
7.	restecg- resting electrocardiographic results	0,1,2
8.	thalachh- maximum heart rate achieved	Multiple values between 71 & 202
9.	exng- Exercise induced angina (1-Yes, 0-No)	0,1
10.	oldpeak- previous peak	Multiple values between 0 & 6.2
11.	slp- slope	Multiple values between 0 & 2
12.	caa- number of major vessels	Multiple values between 0 & 4
13.	thall- Thal. rate	Multiple values between 0 & 3
14.	output- Target variable (0- less chance of heart attack, 1- more chance of heart attack)	0,1

Fig 01: Dataset Description Table

2.2. Dataset Splitting & Visualization

The dataset that we have used, has been split into two segments. Among them, 70% of the dataset has been used for training and 30% of the dataset has been used for testing.
The following chart portrays the route that has been proposed-

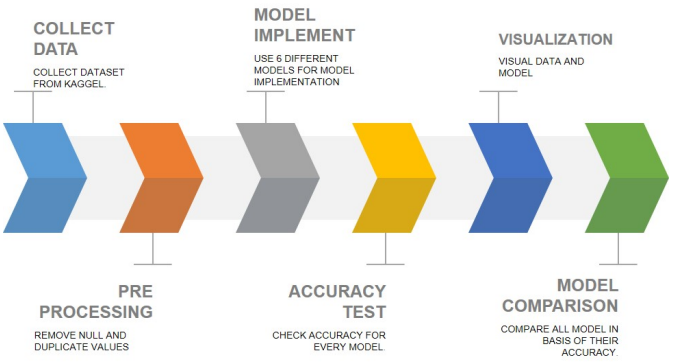

Fig 02: Diagram of Workflow

2.3. Logistic Regression

Logistic regression is a statistical method for predicting the probability of a binary outcome, such as whether or not someone will have a heart attack. Logistic regression might be used to predict the probability that a person will have a heart attack based on certain risk factors instead of fitting a straight line or hyperplane. The logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables in the dataset which make logistic regression good for classification. [2]

2.4. K Neighbors Classification

K-nearest neighbor (KNN) classification is a machine learning method that can be used to predict the class of a sample based on the class of its nearest neighbors. KNN classification can be used to predict a person having a heart attack based on the characteristics of other people who have had heart attacks. The reason KNN might be used for heart attack prediction is that it can be effective at capturing non-linear relationships in the data. This can be especially important in the context of heart attacks, as there are likely many complex interactions

between different risk factors that can influence the likelihood of a heart attack. KNN is able to capture these interactions by considering the similarity of the new data point to its neighbors. [3]

2.5. Decision Tree Classification

Decision tree classification algorithm is a machine learning technique that builds a decision tree-like model based on data features. Beginning at the root node, the algorithm separates the data into different branches based on the feature that results in the greatest reduction in impurity. The algorithm keeps dividing into branches until it reaches a leaf node, at which point it predicts the class of the data point based on the leaf node's majority class. This algorithm can be effective for heart attack prediction because it is resistant to the effects of extreme values and missing values, which might exist in a medical dataset. [4]

2.6. Random Forest Classification

Random forest classification is a machine learning method that can be used to predict the class of a sample based on the class of a group of decision trees, where each tree is trained on a randomly selected subset of the data. A random forest classifier might be used to predict whether or not a person is likely to have a heart attack based on their risk factors. The algorithm's ability to handle categorical information makes it better at predicting heart attacks. This is important because our dataset contains a large number of categorical values, including sex, age, chest pain type etc. [5]

2.7. Naive Bayes Classification

Naive Bayes classification is a machine learning algorithm that is based on the idea of making predictions using Bayes' theorem. The algorithm in Naive Bayes classification predicts the likelihood of a specific class (e.g., heart attack or no heart attack) based on the presence or absence of certain features (e.g. age, sex, blood pressure, etc.) The reason Naive Bayes might be used for heart attack prediction is that it is known to perform well on datasets with a large number of features, which may be the case in a prediction task involving heart attacks. This is because Naive Bayes makes the assumption that all features are independent of one another, which can help to reduce the complexity of the model and improve its performance. [6]

2.8. Neural Network Classification

Neural networks are a type of machine learning algorithm that can be used for classification tasks, including predicting the likelihood of a heart attack. They are inspired by the structure and function of the human brain and are composed of interconnected "neurons" that can process and transmit information. The reason neural networks might be used for heart attack prediction is that they can handle large amounts of data and a large number of features, which may be the case in a prediction task involving heart attacks. This is because neural networks are able to automatically extract and learn important features from the data, which can help to improve their performance. But there are also some potential limitations to using neural networks for heart attack prediction. It can be more complex and harder to understand and interpret than some other machine learning

algorithms. Also require more data to train, which can have limitation in some cases. In addition, it can be sensitive to the quality and relevance of the data, and may require careful preprocessing and feature selection to achieve good performance. [7]

2.9. Support Vector Classification

Support vector classification (SVC) is a type of supervised machine learning algorithm that can be used for classification tasks, including predicting the likelihood of a heart attack. They work by finding the hyperplane in a high-dimensional feature space that maximally separates the different classes. The reason to use SVC for heart attack prediction is that they are able to learn complex, non-linear relationships in the data. This can be especially important in the context of heart attacks, as there are likely many complex interactions between different risk factors that can influence the likelihood of a heart attack. SVC are able to capture these interactions by finding the hyperplane that maximally separates the different classes in the data. Another reason SVCs might be used for heart attack prediction is that they are able to handle high-dimensional data, which may be the case in a prediction task involving heart attacks. This is because SVC is able to find the most important features in the data and use them to make predictions, which can help to improve their performance. [8]

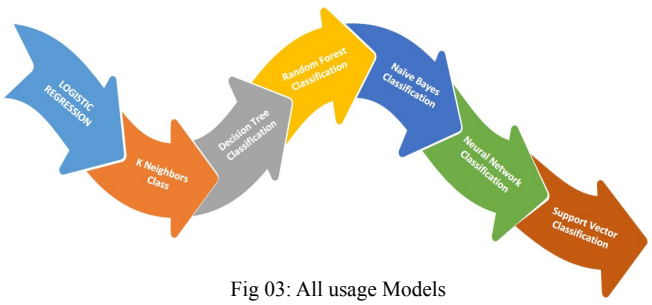

Fig 03: All usage Models

III. RESULT ANALYSIS

In this section, we describe the overall results of models accuracy.

a. Models and their accuracy:

From all of these models, the highest accuracy we got from Support Vector Classification. In this model, we got 80.21% accuracy.

Model	Accuracy (%)
Logistic Regression	79.12
K Neighbors Classification	75.82
Decision Tree Classification	71.42
Random Forest Classification	71.42
Naive Bayes Classification	79.12
Neural Network Classification	75.82
Support Vector Classification	80.21

Fig 04: Models' accuracy Table

From all of these models, we got the highest accuracy from Support Vector Classification. In this model, we got 80.21% accuracy.

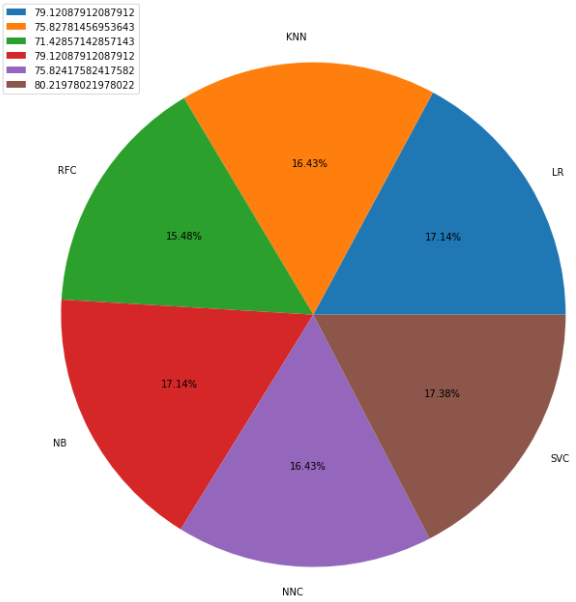

Fig 05: Models' accuracy in pie chart

b. Confusion matrix:

An N x N matrix called a "Confusion matrix," where N is the number of target classes, is used to assess the effectiveness of a classification model. The machine learning model's projected values are compared to the actual target values in the matrix.

MODEL	LABEL			
	TRUE POSITIVE	TRUE NEGATIVE	FALSE POSITIVE	FALSE NEGATIVE
Logistic Regression	29	43	9	10
K Neighbors Classification	29	42	9	11
Decision Tree Classification	28	37	10	16
Random Forest Classification	30	39	8	14
Naive Bayes Classification	31	41	7	12
Neural Network Classification	30	39	8	14
Support Vector Classification	29	44	9	9

Fig 06: Models' Confusion Matrix Table

In result analysis, we use heatmaps to visualize patterns and trends in data that may be related to the risk of a heart attack. By using colours to encode the values in a matrix, heatmaps allow researchers to quickly identify relationships and correlations in the data that might not be immediately apparent from looking at the raw data.

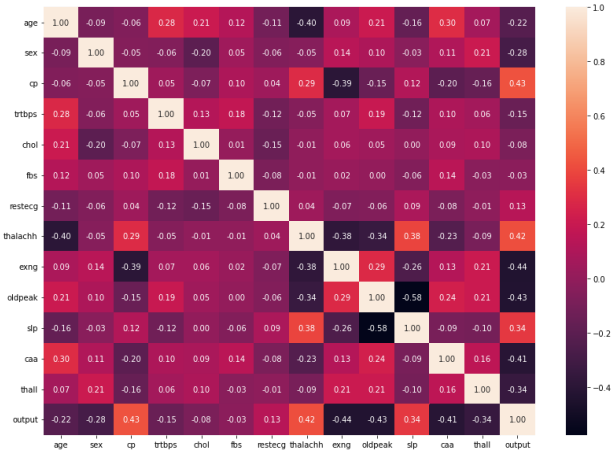

Fig 07: Heatmap

IV. CONCLUSION

In this research analysis the heart attack possibility performance is shown. The performance analysis has discovered a variety of categories, including f-measure and confusion metrics. Support Vector Classification performs the best overall, and it has an accuracy rate of 80.21%. To determine the best performance from our dataset, various machine learning techniques are used. In order to build a strong this study will analyze and

predict the probability of heart attacks using deep learning and artificial intelligence. Added more data as needed.

References

1. *Heart Attack Analysis & Prediction Dataset*. (n.d.). Retrieved December 21, 2022, from https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset
2. Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.
3. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175–185
4. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … others. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.
5. Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).
6. George H. John and Pat Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.
7. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.