Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Heart Attack Predication Using Machince Learning

[1]Divyanshu Tiwari,[2]Devpriyam Aggarwal,[3]Gopal Agarwal,[4]Gaurang Agarwal,[5]Md.Shahid

[12345]*Dept. of CSE,MIET,Meerut*

Abstract: The heart, often referred to as the body's pump, is crucial for circulating oxygenated and deoxygenated blood to all organs. Achieving precision and accuracy in diagnosing and analyzing heart conditions is imperative for timely intervention. This study employs Python programming language and machine learning techniques, utilizing libraries such as NumPy, Pandas, among others, to investigate the Multiple factors including blood pressure, blood sugar, and cholesterol levels contribute to the occurrence of heart attacks. The dataset comprises 14 elements, each representing individual values crucial for patient analysis. The primary objective of this research is to achieve maximum accuracy in heart attack detection through machine learning algorithms

Keywords:-NumPy, deoxygenated, oxygenated, performances matrix , svm, logistics regression .

1) Introduction-

The heart stands as the most vital organ in our bodies, necessitating close scrutiny to ensure its proper function. This project aims to contribute towards saving lives by addressing the prevalent issue of heart attacks, which have become increasingly common. Biologically termed as cardiovascular disease, it arises from the accumulation of fatty acids in blood vessel walls, impeding the heart's function and eventually leading to heart attacks. The primary motivation for undertaking this research stems from the staggering statistic of over 10 million lives lost due to undetected heart problems. Throughout our research journey, we encountered various challenges, notably the identification of key parameters leading to heart attacks. By delving into over 25 research papers on the subject, we gained insight into the major algorithms utilized in this field. Furthermore, we sought to apply our findings practically, leveraging past data to discern patterns and effectively treat future patients.

2) Related Work:-

Kohali et al. [4] conducted a comprehensive study focusing on the extrapolation of various medical conditions such as cardiovascular diseases, breast cancer, and diabetes prediction through the utilization of diverse machine learning algorithms, resulting in varying levels of accuracy. A. Mishra et al. [5] introduced a strategic framework for identifying and evaluating the most appropriate problems in medical image processing (MIP), aiming to enhance the assessment of MIP solutions. Sneha A, Mane et al. [6] applied two neural network algorithms, specifically as part of their implementation learning vector quantization and radial basis function, for diagnostic purposes, with a comparison conducted using MATLAB software to determine the optimal tool for medical analysis, thereby reducing analysis time and enhancing accuracy and output. Tijjani et al. [7] provided an overview of artificial neural network (ANN)-based approaches for kidney problem prediction, focusing on comparing patient mental behavior using MATLAB software. Gavhane et al. [8] proposed a technique utilizing a multilayer perceptron model for heart disease prediction, emphasizing precision through computer-aided design (CAD) technology. Kaur et al. [9] taken out a study on a large dataset of cardiovascular disease, comparing data mining approaches and various machine learning algorithms to ascertain the most precise method. Mohan et al. [10] explored heart disease prediction using diverse algorithms such as naïve Bayes, genetic algorithm, decision tree, and k-nearest neighbors (knn), introducing a hybrid algorithm that achieved an accuracy of 88%. Himanshu et al. [11] discussed the prediction of heart diseases based on large and small datasets, highlighting

that smaller datasets require less time for training and testing. They applied support vector machine (SVM) and knn algorithms for prediction, demonstrating that while some machine learning algorithms may not perform optimally for accurate prediction individually, hybridization can improve accuracy [12].

3) Proposed Framework:-

The data flow diagram below illustrates the process of predicting and analyzing heart attack risk using raw health record data stored in an Excel sheet. The data is initially cleaned of any null values, followed by univariate and bivariate analyses using tools such as histograms, pie charts, heatmaps, and box plots. Subsequently, they experimented with many type of machine learning (ML) algorithms such as logistic regresson, support vector machine (SVM), random forest, nave Bayes,

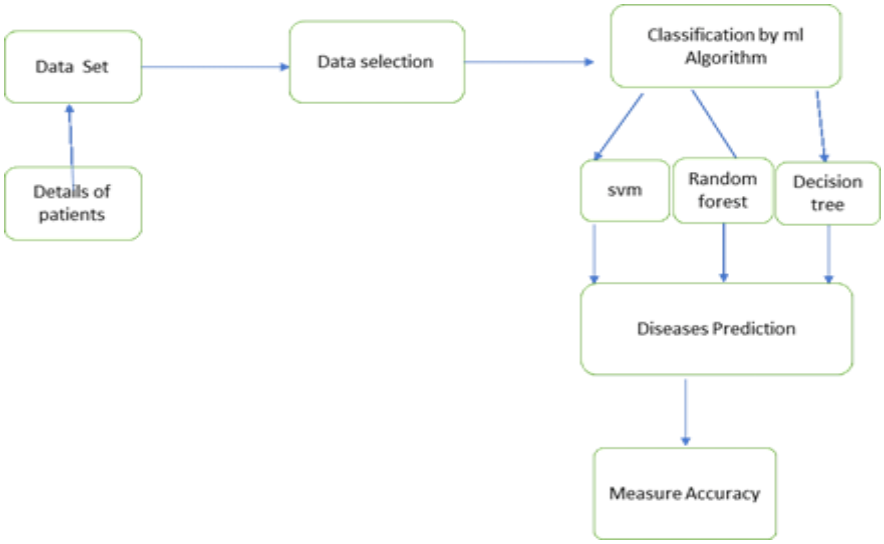


Fig.1 - Data Flow Diagram For Heart Attack Predication

3.1) Raw data is loaded.

3.2) Data from various individuals is collected into an Excel sheet.

3.3) Data selection is performed to remove abnormalities.

3.4) Selected data is processed using various algorithms such as SVM, random forest, and decision tree.

3.5) Heart attack risk for each individual is calculated based on parameters such as blood perssure and blood glucose level.

3.6) Model accuracy and algorithm performance are measured. Support Vector Machine:- This transformation makes the data more separable in that space. The selection of the kernel function is contingent upon both the dataset characteristics and the nature of the classification task. Subsequently, SVM identifies the hyperplane that optimally separates the classes by maximizing the margin. by solving an optimization problem. After identifying ,SVM can classify new data points by discerning their position relative to the hyperplane. SVM is a powerful algorithm that can work well in a variety of classification tasks, but it can be sensitive to the choice of kernel function and parameters, and it may not work well with very large datasets. Logistic Regrssion: Logistic Regrssion is a statistical technique which is employed for forecasting the binary outcome based on a set of independent variables or predictors. It works by using a mathematical function called the logistic functionor sigmoid function to estimate the probability of the binary outcome.

Random Forest :- This algorithm consist of various of factors which may included the Medical history lifestyle choices , and various health metric which make up the holistic Appreance of individual health profile .In essence

of random forest has a great key Advantage our capabilities to predict the and avoid the cardiovascular events through Innovations integration of ml algorithm .

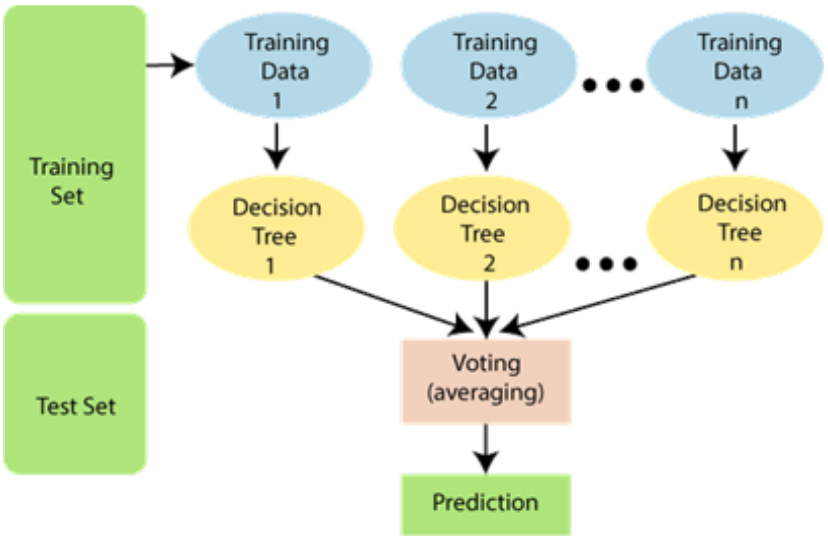


Fig-2 Diagram of model of random forestWorking of Random Forest Algorithm Working of Random Forest Algorithm:

Task1 we choose randomly sampls from the give datasetTask 2 For every samples a decision tree are made.

Task 3 Voting occurs by averaging decision treepredictions.

Task 4 The auspication with highest number of votes ischosen as the ultimate outcome.

This algorithm considers various factors including medical history and lifestyle choices to create a holistic health profile. By constructing decision trees for random samples and averaging their predictions, it selects the most voted result as the final prediction.

Decision Tree:-

All potential solutions to a problem or decision, considering given conditions, are represented by a decision tree. Named for its tree-like structure, it begins with a root node and branches out, resembling a tree. The tree construction is facilitated The CART algorithm, an abbreviation for Classification and Regression Tree algorithm, operates on the basis of decision trees. In essence, a decision tree asks a question and, based on the answer (Yes/No), partitions the tree into further subtrees..

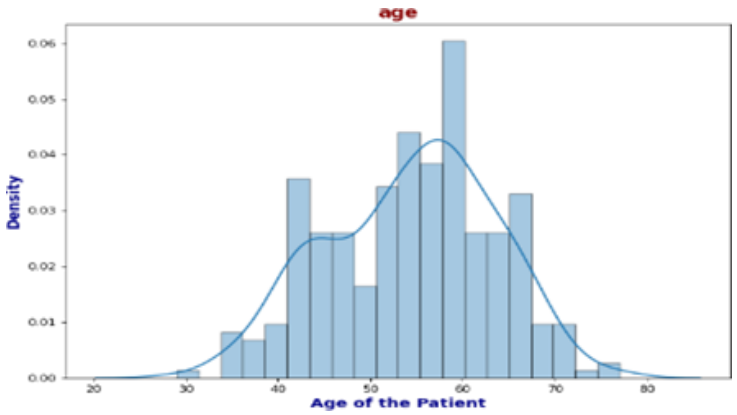
4) Experimental setup and datset:-

4.1) Data collection:-Here first we do for do is collection of the data for correct predication by system of chances ofheat attack after it is used for data cleaning and then usedfor training and testing of the dataset . In this project, 70%of the dataset is assigned for training, leaving the remaining 30% for testing purposeHere we have total 14 columns and 970 rows created using this data.

4.2) Data set description:-

1. Age: The individual's age expressed in years.
2. Sex: Gender of the individual, where 1 represents men and 0 represents women
3. Chest Pain Type (cp): Classification of chest pain, where values represent: 1 as typical angina, 2 as atypical angina, 3 as non-anginalpain, and 0 as asymptomatic.

4. Resting Blood Pressure (trestbps): Bloodpressure at rest, recorded in mm Hg uponhospital admission
5. Serum Cholesterol (Chol): Serum cholesterollevel expressed in mg/dl.
6. Fasting Blood glucose (fbs): Indicates if the fasting blood glucose level is over 120 mg/dl, with 1 denoting true and 0 denoting false.
7. Resting Electrocardiographic Results (restecg): Evaluation of relax electrocardiogram findings, categorized as 1 for normal, 2 for ST-T wave abnormality, and 0 for hypertrophy
8. Maximum Heart Rate Achieved (thalach): Highest heart rate attained during exercse.
9. Exercise Induced Angina (exang): Give alertthe presence of angina injected by exercise, with 1 denoting yes and 0 denoting no.

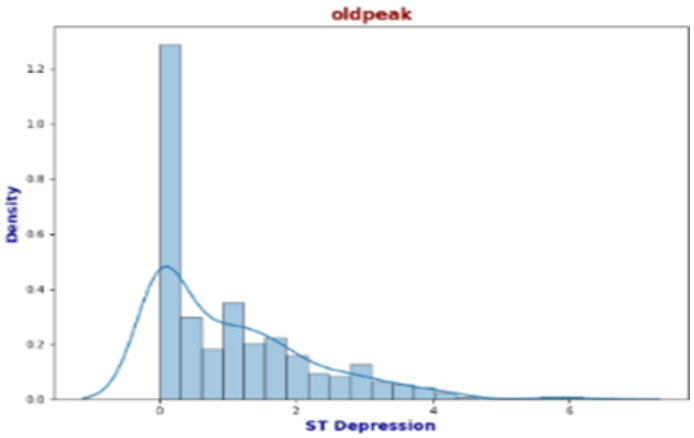


11. Slope of the Peak Exercise ST Segement (slope)Describes the slope of the peak exercise ST segment, where 2 represents upsloping, 1 represents flat, and 0 represents downsloping.
12. Number of Major Vessels (ca): show the number of major vessels (ranging from 0 to 3) that are colored by fluoroscopy.

13 . Thalassemia (thal): indicates the presence ofthalassemia

14. Diagnosis of Heart Disease (num): cardiac Cath disease status indicating the severity of heart disease (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing).
15. Target: Likelihood of experiencing a heartattack (0 = low probability; 1 = high probability

3.3) univariate analysis using histogram-



ST depression is relative to reduced induced to physical activities.

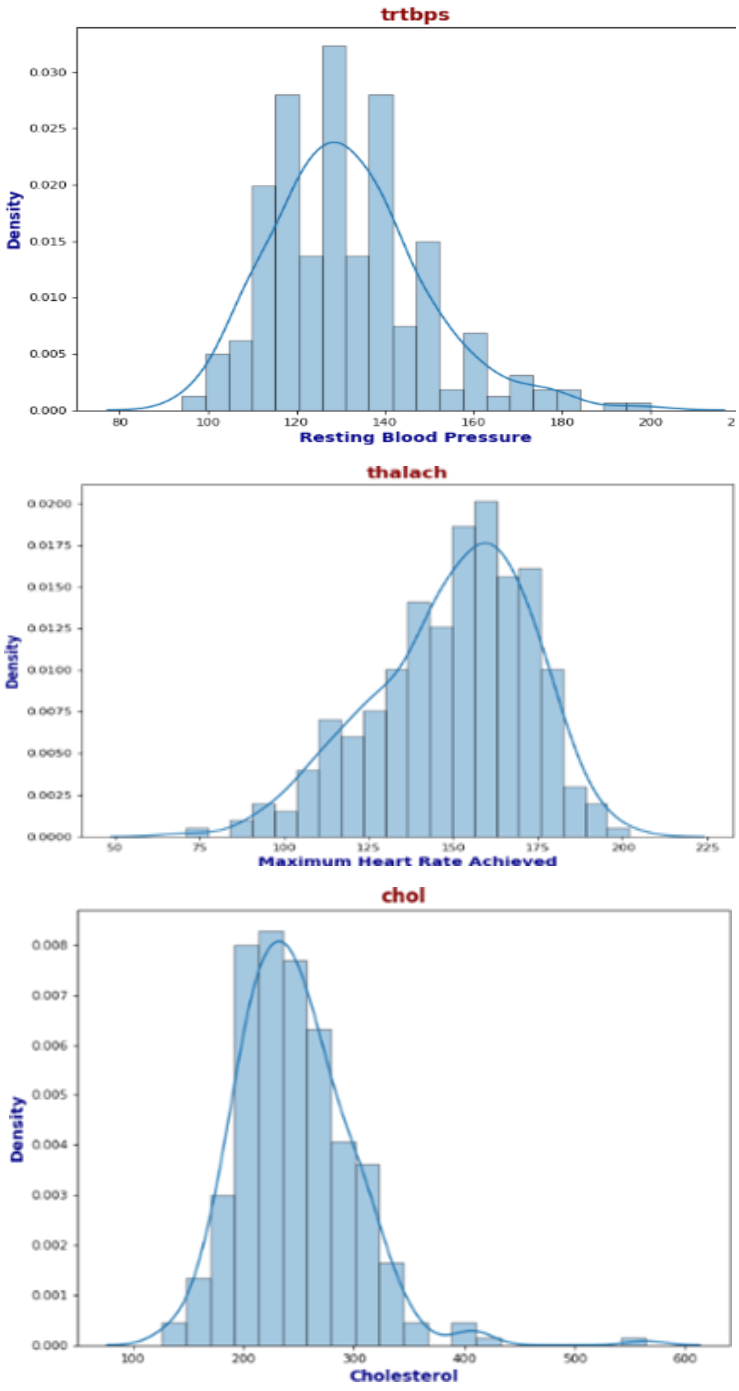


Fig 3 – Histogram of age, old peak, trtbps, chol ,thalach

4.3) Analysis Outputs(4)Age Variable:

- The majority of patients fall between the ages of50 and 60.
- A noticeable decline in patients is observedbetween the ages of 47 and 50.
- No outliers are detected in the age variable.Resting Blood Pressure (Trtbps) Variable:
- Most patients have resting blood pressurereadings ranging between 110 and 140 mm Hg.
- Values exceeding 180 mm Hg are consideredoutliers.

- Heavy patient distribution is observed between values of 115-120, 125-130, and 155-160 mm Hg.

Cholesterol Variable:

- Cholesterol levels in most patients range between 200 and 280 mg/dl.
- Values above 380 mg/dl are considered outliers.Maximum Heart Rate Achieved (Thalach) Variable:
- The majority of patients achieve maximum heart rates between 145 and 170 beats per minute.
- Outliers are observed in values below 80 beats per minute. ST Depression Induced by Exercise Relative to Rest(Oldpeak) Variable:
- The majority of patients' values for STdepression range from 0 to 1.5.
- Values exceeding 2.5 are considered outliers.Categorical Variables (Analysis with Pie Chart):
- 69.3% of patients are male, while 31.7% arefemale.
- Half of the patients are asymptomatic, indicating they experience chest pain without symptoms.
- 29% of patients have atypical chest pain, and8% experience non-anginal pain.
- 85% of patients have fasting blood sugar levelsexceeding 120 mg/dl.
- More than half of the patients (50.2%) exhibitnormal resting electrocardiographic results.
- The majority of patients (57.6%) have 0 major vessels colored by fluoroscopy, indicating a lower risk.
- 54.5% of patients are at risk of heart attack,while 45.5% are not at risk.

3.4 Categorical Variables(Analysis with Pie Chart)

In our comprehensive analysis, it was revealed that a significant portion of the dataset, comprising 69.3% males and 31.7% females, provided valuable insights into various health parameters. Notably, half of the patients reported being asymptomatic, suggesting underlying health issues without apparent symptoms. Specifically, 29% exhibited atypical symptoms, while 8% reported non-anginal pain, indicating diverse manifestations of potential heart-related concerns. Furthermore, 85% of individuals displayed fasting blood sugar levels exceeding 120mg/dl, underscoring a prevalent risk factor for cardiovascular issues. Regarding angina, experienced by 50.2% of patients, and 48.5% classified as normal, the dataset highlighted a spectrum of cardiac health statuses. Moreover, a substantial proportion of patients showed no signs of exercise-induced angina, suggesting varying degrees of cardiovascular resilience. Delving deeper into the "ca" variable, which denotes the number of major vessels colored by fluoroscopy, 57.6% exhibited minimal vessel involvement (value 0), whereas 21.63% displayed larger vessel presentation (value 1).

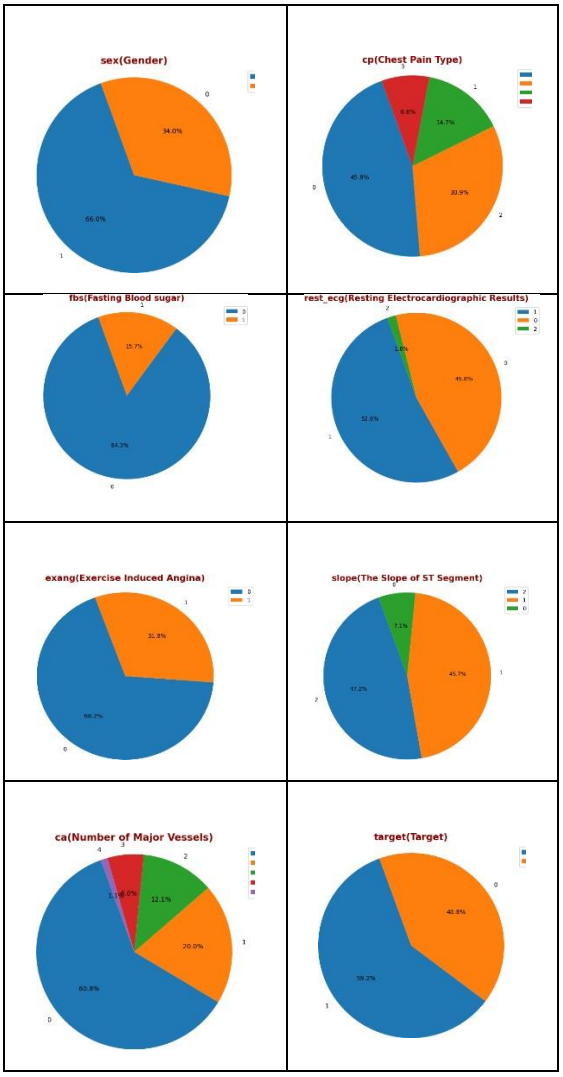


Figure-3 Analysis with pie chart

5) Result and Discussion:-5.1)Pair plot analysis-

Pair plot analysis visually examines the relationships between variables, which can be either continuous or categorical. This plot provides a comprehensive view of pairwise relationships within a dataset. Utilizing the Seaborn library, which offers an intuitive interface for creating informative statistical graphics, we conducted pair plot analysis.

Observations:

- The age variable exhibits the strongest relationship with the thalach variable, with a collection of points tending to occur in the southwest direction, indicating a negative correlation.
- Resting blood pressure (trtbhps) shows weak relationships with other variables, resulting in a cluttered graph, suggesting a generally positive correlation.
- Cholesterol (Chol) demonstrates the highest correlation with variables such as age and trtbps, indicating that cholesterol tends to increase with age.
- Thalach shows a negative relationship with the age variable, with scattered dots suggesting a weak correlation.

- Old peak exhibits moderate relationships with all variables, indicating a closer relationship compared to others.

In summary, the pair plot analysis reveals varying degreesof correlation between different variables. While some variables show strong relationships, others exhibit weaker correlations, highlighting the complexity of interactions within the dataset.

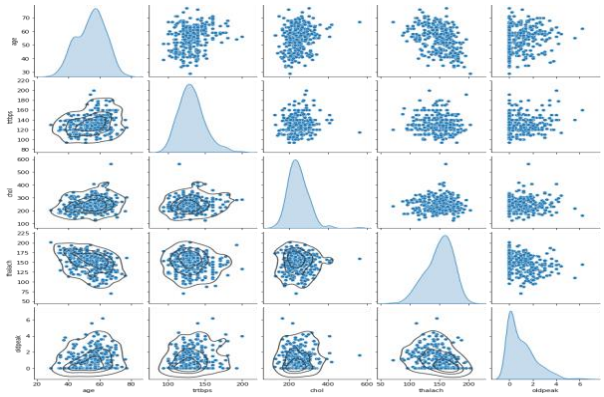


Fig -4 Pair plot for analysis for variousparameter

5.2) Heatmap to get the relationship between variable

In the heatmap analysis, we examine the correlations between variables. The highest negative correlation is observed between age and thalach, with a correlation coefficient of -0.40. This indicates an inverse relationship, suggesting that as age increases, heart rate decreases. Conversely, resting blood pressure (trtbps) shows a positive, low-intensity correlation with other variables.

Cholesterol (Chol) exhibits the highest correlation with age, with a coefficient of 0.21, implying that cholesterol levels tend to increase with age. Thalassemia (thal) has a correlation of 0.42 with the target variable, suggesting a potential trigger for heart attacks.

Old peak shows a significant negative correlation of -0.58 with changing variables, indicating a notable impact on variable alterations. Sex demonstrates a robust relationship with other variables. Chest pain type (cp) exhibits the highest correlation with the target variable.

Variables such as fasting blood sugar (fbs) and exercise- induced angina (exang) show low positive correlations with other variables. The slope variable demonstrates a moderate relationship with thalach and the target variable,while being fragile in nature with other variables.

Overall, the target variable correlates with more than one variable, indicating complex interactions within the dataset

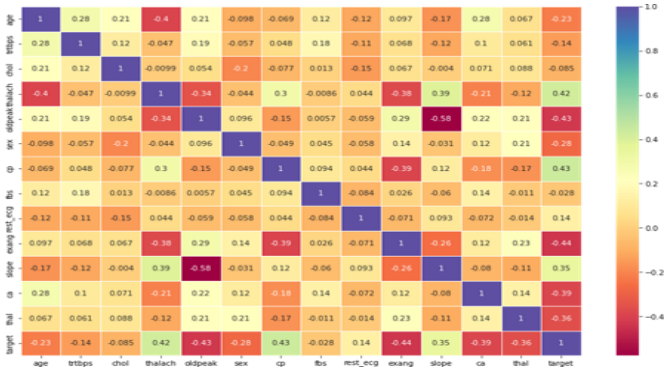
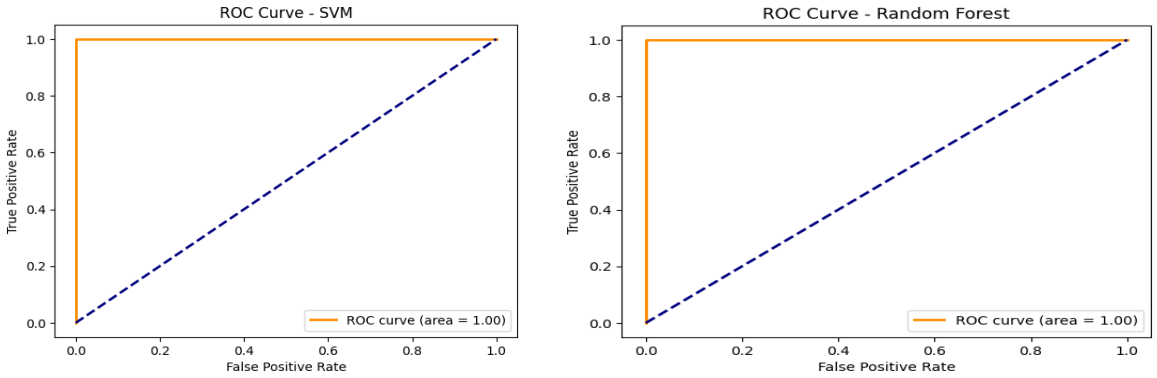


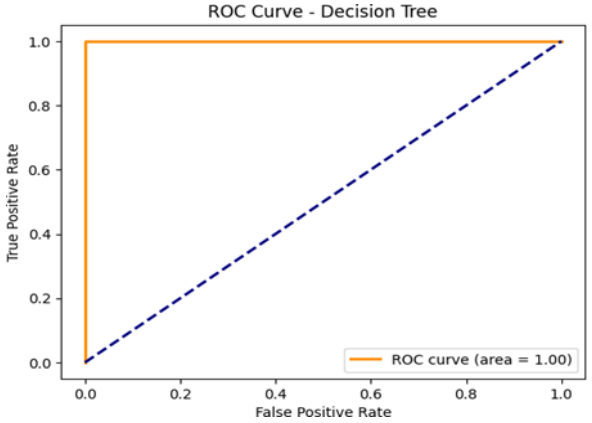
Fig-5 Heatmap for analysis of variable relationshipbetween them

5.3) Modeling of data using various algorithm and ROC and AUC: -

We utilized various machne learning algorithms such as Decision Tree, Support Vector Machine (SVM), Logistic Regression, and Random Forst to analyze the data. Subsequently, we evaluated the performance of these algorithms in predicting the risk of heart attacks using Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics. ROC, The ROC (Receiver Operating Characteristic) curve graphically illustrates the The effectiveness of a binary classification model is assessed through plotting the true positive rate (TPR) versus the false positive rate (FPR) at various classification thresholds. The AUC (Area Under the Curve) signifies the space enclosed by the ROC curve. It serves as a measure of the overall performance of the binary classification model. AUC values range from 0 to 1, with higher values indicating indicate better performance. Interpretation of AUC Values: AUC values between 0.9 and 1.0 are classified as excellent, while values falling between 0.8 and 0.9 are deemed very good. AUC values falling between 0.7 and 0.8 are considered as best. Values between from 0.6 to 0.7 indicate satisfactory performance. AUC values between 0.5 and 0.6 are deemed unsatisfactory. By analysing ROC and AUC curves, we can assess the predictive capabilities of each algorithm in determining the risk of getting the attack. We utilized various machine learning algorithms, including Decision Tree, Support Vector Machine (SVM), Logistic Regression, and Random Forest, to model the data and predict heart attack risk. Subsequently, we assessed their performance using Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics are used to evaluate classification models. The ROC curve illustrates the true positive rate (TPR) versus the false positive rate (FPR) across various thresholds., AUC quantifies the overall performance of the binary classification model. AUC values ranging from 0.9 to 1.0 indicate excellent performance, while those between 0.8 and 0.9 are considered very good, and values falling between 0.7 and 0.8 are classified as good. Values ranging from 0.6 and 0.7 indicate satisfactory performance, while those between 0.5 and 0.6 are considered unsatisfactory. These analyses enable us to evaluate the algorithms' predictive capabilities and select the most suitable model for determining heart attack risk.



.Fig 6 - ROC and AUC curve for logistic regression, decision tree, svm, random forest



Decision Tree	89	1	1	1	97
	accuracy	precision	recall	F1-measure	Support
Svm	88	1	1	1	97
Random forest	97	0.98	0.98	0.98	97

Table 1 : classification report of svm, random forest,decision tree

6) Conclusion : -

The heart is essential for the effective operation of the body. necessitating regular monitoring and check-ups for early detection and treatment. In our research , a many types of machine learning algorithms were utilized to prophesy the probability of an individual encountering the heart attack, relying on...14 parameters and a dataset of 971 values. Utilizing three major algorithms, we achieved accuracies of 88% and 83% with Decision Tree, 84.5% and 88% with SVM, and 91.4% and 97.6% with Random Forest under AUC curve. The results demonstrate Random Forest as the preferred algorithm, surpassing the accuracy of previous research papers. This highlights the improved accuracy and effectiveness of our model. Through the Hossain, Md. Imam. (2023). "Heart Disease Prediction Using Distinct Artificial Intelligence Techniques: Performance Analysis and Comparison." Iran Journal of Computer Science. Notably, Random Forest emerged as the superior model, outperforming previous research efforts. These findings signify a substantial enhancement in predictive capability,offering promising avenues for enhancing cardiovascular health outcomes.

Future Work:

Moving forward, there are several avenues for enhancing our heart attack prediction model using machine learning. In addition to accurately identifying individuals at risk, we can offer personalized recommendations for lifestyle modifications, exercise regimens, and dietary changes to individuals not currently predisposed to heart attacks but susceptible in the future. Furthermore, implementing alerts to notify patients' family members about potential risks and collaborating with local hospitals can improve accuracy by incorporating more extensive datasets to capture diverse patterns and nuances. These advancements hold promise for refining preventive measures and promoting proactive heart health management.

References:-

[1] Goel, Rati. (2023). "Heart Disease Prediction Using Various Machine Learning Algorithms." Computer Science and Engineering, Inderprastha Engineering College. Problems Symptoms." Biomedical Engineering.

[2] G, Malavika et al. "Heart Disease Prediction Using Machine Learning Algorithms." Bioscience, Biotechnology Research Communications.

[3] Kohli, Pahulpreet Singh, and Shriya Arora. (2018). "Application of Machine Learning in Diseases Prediction." 4th International Conference on Computing,Communication, And Automation.

[4] Mishra, A., Rai, Abhishek, and Yadav, Akhilesh. (2014). "Medical Image Processing: A Challenging Analysis." International Journal of BioScience and BioTechnology.

[5] Mane, Sneha A., and Chougule, S R. (2016). "Neural Network of Kidney Stone Detection." International Journal of Science and Research.

[6] Adam, Tijjani et al. (2012). "Designing an Artificial Neural Network Model for the Prediction of Kidney utilization of machine learning algorithms, we have notably enhanced the precision of predicting heart attacks.

[7] Gavhane, Aditi et al. (2018). "Prediction of Heart Disease Using Machine Learning." Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology.

[8] Kaur, Amandeep, and Arora, Jyoti. (2019). "Heart Diseases Prediction Using Data Mining Techniques: A Survey." International Journal of Advanced Research inComputer Science.

[9] Kumar, M. Nikhil et al. (2019). "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools." International Journal of ScientificResearch in Computer Science, Engineering and Information Technology.

[10] harma, Himanshu, and Rizvi, M A. (2017). "Prediction of Heart Disease Using Machine Learning Algorithms: A Survey." International Journal on Recent and Innovation Trends in Computing and Communication.

[11] Hazra, A. et al. (2017). "A Heart Disease Diagnosis and Prediction Using Machine Learning and Data MiningTechniques: A Review." Advances in Computational Sciences and Technology.

[12] Shah, Devansh et al. (2020). "Heart Disease Prediction Using Machine Learning Techniques." Springer Nature Singapore Pte Ltd.

[13] Goel, R., and Jain, A. (2020). "Improved Detection of Kidney Stone in Ultrasound Images Using Segmentation Techniques." Advances in Data and Information Sciences. https://www.kaggle.com/datasets/divyanshu8956/hea rtattack-predication-using-ml.

[14] Folsom, A R et al. (1989). "Body Fat Distribution and Self-reported Prevalence of Hypertension, Heart Attack, and Other Heart Disease in Older Women." International Journal of Epidemiology.

[15] Ganna, A. et al. (2013). "Multilocus Genetic Risk Scores for Coronary Heart Disease Prediction." Arteriosclerosis, Thrombosis, and Vascular Biology.

[16] Jabbar, M A et al. (2013). "Heart Disease Prediction Using Lazy Associative Classification." International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing.

[17] Jee, S H et al. (2014). "A Coronary Heart Disease Prediction Model: The Korean Heart Study." BMJ Open.Kiyasu, J Y. (1982). U.S. Patent No. 4,338,396.

[18] Ordonez, C. (2006). "Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction." IEEE Transactions on Information Technology in Biomedicine.

[19] Parthiban, Latha, and Subramanian, R. (2008). "Intelligent Heart Disease Prediction System Using CANFIS and Genetic Algorithm." International Journalof Biological, Biomedical and Medical Sciences.

[20] Patel, S., and Chauhan, Y. (2014). "Heart Attack Detection and Medical Attention Using Motion Sensing Device - Kinect." International Journal of Scientific andResearch Publications.

[21] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.

[22] Worthen W J, Evans S M, Winter S C & Balding D (2002). U.S. Patent No. 6,432, 124. Washington, DC: U.S.Patent and Trademark Office.

[23] Zhang Y, Fogoros R, Thompson J, Ken knight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and TrademarAn overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[24] Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.