IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

"HEART ATTACK PREDICTION USING DATA SCIENCE AND MACHINE LEARNING"

[1]Nikunj Aggarwal, [2]Raghav Gaur, [3] Yug Varshney [4] Divyansh Rastogi [5]Mayank Mahajan

[1]Computer Science Engineering,
[1]Chandigarh, NH-05, Ludhiana - Chandigarh State Hwy, Punjab , India

Abstract : This study has been undertaken to study Cardiovascular conditions, particularly heart attacks, remain a leading cause of global morbidity and mortality. Beforehand vaticination of individualities at threat can significantly ameliorate preventative interventions and patient issues. This exploration paper presents a comprehensive study on the integration of data wisdom and machine literacy ways for heart attack vaticination. The study employs a different dataset encompassing demographic information, life factors, and clinical pointers, collected from a large cohort of cases. Data preprocessing ways are applied to address missing values, and outliers, and ensure the quality of the dataset. point engineering is conducted to prize applicable information, and a relative analysis of colorful point selection styles is performed to identify the most influential predictors. Several machines learning algorithms, including but not limited to logistic retrogression, support vector machines, arbitrary timbers, and deep neural networks, are employed to develop prophetic models. The models are trained and validated using a robust cross-validation strategy to insure generalizability. Performance criteria similar as delicacy, perceptivity

Index Terms **- Heart attack · Decision Tree· Machine Learning · Random Forest**

1.INTRODUCTION

Heart attacks, continue to be a leading cause of mortality worldwide. The frequency of these life- changing events underscores the critical need for advanced styles of vaticination and forestallment. In recent times, the crossroad of data wisdom and machine literacy has surfaced as an important supporter in the healthcare sector, offering unknown openings to revise the way we approach heart attack vaticination. The traditional threat assessment models, while precious, frequently calculate on limited variables and may not capture the intricate interplay of multitudinous factors impacting heart health. The arrival of big data and the capability to reuse and dissect different datasets open new borders for refining prophetic models. Our exploration leverages these capabilities to develop a comprehensive and accurate heart attack vaticination system. This exploration paper navigates through the methodologies employed, including data preprocessing, point selection, and the perpetration of machine literacy algorithms. also, we explore ethical considerations girding the use of sensitive health data, emphasizing the significance of sequestration and informed concurrence in the period of data- driven healthcare. In conclusion, the emulsion of data wisdom and machine literacy in heart attack vaticination not only represents a technological vault but also signifies a vital moment in healthcare elaboration. By embracing the transformative power of these technologies, we aim to enhance the perfection of threat assessment, steering in a new period where visionary intervention becomes the foundation of cardiovascular health operation. As we navigate the intricate geography of heart attack vaticination, this exploration promises to contribute significantly to the ongoing global sweats to reduce the burden of cardiovascular conditions and ameliorate overall public health.

2. NEED OFF STUDY AND LITERATURE SURVEY:

In recent times, the healthcare assiduity has seen a significant advancement in the field of data mining and machine literacy. These ways have been extensively espoused and have demonstrated efficacity in colorful healthcare operations, particularly in the field of medical cardiology. The rapid-fire accumulation of medical data has presented experimenters with an unknown occasion to develop and test new algorithms in this field. Heart complaint remains a leading cause of mortality in developing nations and relating threat factors and early signs of the complaint has come an important area of exploration. The Algorithms 2023, 16, 88 3 of 14 application of data mining and machine literacy ways in this field can potentially prop in the early discovery and forestallment of heart complaint. The purpose of the study described by Narain teal.(2016) is to produce an innovative machine- literacy- grounded cardiovascular complaint(CVD) vaticination system in order to increase the perfection of the extensively used Framingham threat score(FRS). With the help of data from 689 individualities who had symptoms of CVD and a confirmation dataset from the Framingham exploration, the proposed system which uses an amount neural network to learn and fete patterns of CVD was experimentally validated and compared the accuracy of the proposed system in predicting the risk of CVD was found to be 98.57%, significantly higher than the accuracy of the FRS (19.22%) and other methods currently in use. The study's conclusions indicate that the recommended strategy may help physicians predict their patients' risk for CVD, develop more effective treatment regimens, and enable early diagnosis.

UCI machine learning repository also provided the data. Numerous supervised classification techniques, such as naive Bayes, decision trees, random forests, and nearest neighbor (KKN), were used by the authors. The study's findings showed that, at 90.8%, the KKN model had the highest degree of accuracy.
The study underlines the potential value of machine learning methods in the prediction of cardiovascular disease and stresses the significance of choosing the right models and methods to get the best outcomes.

Authors	Novel Approach	Best Accuracy	Dataset
Shorewall, 2021 [5]	Stacking of KNN, random forest, and SVM outputs with logistic regression as the metaclassifier	75.1% (stacked model)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Maiga et al., 2019 [7]	-Random forest -Naive Bayes -Logistic regression -KNN	70%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Waigi at el., 2020 [12]	Decision tree	72.77% (decision tree)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Our and ElSeddawy, 2021 [21]	Repeated random with random forest	89.01%(random forest classifier)	UCI cardiovascular dataset (303 patients, 14 attributes)
Khan and Mondal, 2020 [22]	Holdout cross-validation with the neural network for Kaggle dataset	71.82% (neural networks)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
	Cross-validation method with logistic regression (solver: lbfgs) where k = 30	72.72%	Kaggle cardiovascular disease dataset 1 (462 patients, 12 attributes)
	Cross-validation method with linear SVM where k = 10	72.22%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)

1.Related work on heart attack prediction

3. METHODOLOGY

Data Collection by compiling a thorough dataset with pertinent health metrics, demographic data, lifestyle choices, and medical histories of individuals. Both positive cases—those who have had a heart attack—and negative cases—those who have not—should be included in this dataset. Data Preprocessing to address missing values, outliers, and inconsistencies in the dataset, thoroughly clean the data. Normalize or standardize numerical features to guarantee scale consistency.
Use methods like one-hot encoding to encode categorical variables. Analyze exploratory data (EDA) to learn more about the relationships and distribution of the dataset. Feature Selection to determine the most important variables for heart attack prediction, apply feature selection techniques. To reduce the size of the feature set, apply statistical techniques, correlation analysis, or recursive feature elimination. Data Division to partition the dataset into training and testing sets in order to assess the effectiveness of the model.To make sure that the distribution of positive and negative cases in both sets is evenly distributed, think about utilising strategies like stratified sampling.

Model Building:
a. Random Forest:
Given its simplicity and interpretability, use logistic regression as a baseline model. .
b. Decision Tree:
determine which algorithm predicts heart attacks the best. To guarantee the robustness and generalizability of the models, take into consideration applying cross-validation techniques.

Model Interpretability: Use the models to interpret the significance of different features in heart attack prediction.
Create visualizations to share insights with stakeholders and healthcare professionals, such as feature importance plots.

3.1 Data Collection

This dataset is provided by KAGGLE and includes patient-specific details such as age, gender, cholesterol levels, blood pressure, heart rate, diabetes, family history, smoking habits, and other indicators, including heart health and lifestyle choices.
Provides comprehensive functionality related to.
, obesity and alcohol intake. Additionally, lifestyle factors such as exercise time, diet, stress levels, and sedentary time are taken into account. Medical aspects such as past heart disease, medication use, and triglyceride levels are taken into consideration.
These include socio-economic aspects such as income, and hemisphere. This dataset, consisting of 8,763 records from patients of classification around the world, was ultimately completed as a key binary classifier for indicating presence or absence of heart attack risk, and is useful for predictive analytics and research in the field of cardiovascular health.

3.2 Data Preprocessing:

It is clear that there are outliers within the dataset. These outliers may be due to data entry errors.
Removing these outliers can improve the performance of the prediction model. To resolve this issue, we removed all instances of ap_hi, ap_lo, weight, and height that were outside the range 2.5% to 97.5%.This process of identifying and eliminating outliers was performed manually.

3.3 Feature Selection

To improve the performance and interpretability of classification algorithms, the use of binning has been proposed as a way to transform continuous inputs, such as age, into categorical inputs.

By classifying continuous input into different groups or sections, algorithms can distinguish between different classes of data based on specific values of input variables. For example, if the input variable is "age group" and the possible values are "young," "middle-aged," "," and "elderly," the classification algorithm uses this information to divide the data into different classes. Split or categorize based on the age range of people in your dataset. Additionally, converting continuous input to categorical input through binning also improves the interpretability of the results, as the relationship between and becomes easier to understand and interpret Input variables and output classes. On the other hand, using continuous inputs such as numbers in classification algorithms can be more difficult because the algorithm may have to make assumptions about where to draw the boundaries between different classes or categories. In this study, we applied a binning technique to the age attribute of the patient dataset.

Patient age was originally reported in days, but was converted to years by dividing by 365 for better analysis and prediction.

The age data was then divided into five-year interval bins ranging from 0-20 years old to 95-100 years old. The minimum age in the dataset is 30 years and the maximum age is 65 years.

So classes 30-35 will be marked with 0 and final class 60-65 will be marked with 6.Additionally, other attributes marked with continuous values were also converted to categorical values, such as height, weight, ap_hi, ap_lo. The results of this study show that converting continuous inputs into categorical inputs through binning can improve the performance and interpretability of classification algorithms.

3.4 Data Splitting

A training data set (80%) and a test data set (20%) are created from the data set. The model is trained using the training data set and its performance is evaluated using the test data set. Various classifiers such as decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost are applied on the clustered dataset and their performance is evaluated. The performance of each classifier is then evaluated using precision, precision, recall, and F-score. A training data set (80%) and a test data set (20%) are created from the data set. The model is trained using the training data set and its performance is evaluated using the test data set. Various classifiers such as decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost are applied on the clustered dataset and their performance is evaluated. The performance of each classifier is then evaluated using precision, precision, recall, and F-score.

3.5 Model Building

3.5.1 Random Forest

The Random Forest algorithm is a type of supervised classification method where several decision trees cooperate with one another. The class that receives the most votes is the one that our model predicts. The decision tree algorithm's drawbacks are removed as every tree in the random forest predicts a class. As a result, the dataset becomes less over fit and more accurate. If a sizable percentage of record values are missing, the random forest technique may still yield the same results when applied to huge datasets. The decision tree's samples may be stored and used to different kinds of data. In the study, using 500 estimators, random forest produced test accuracy of 73% and validation accuracy of 72%.

3.5.2 Decision Tree Classifier

Decision trees are treelike structures that are used to manage large datasets. They are often depicted as flowcharts, with outer branches representing the results and inner nodes representing the properties of the dataset. Decision trees are popular because they are efficient, reliable, and easy to understand. The projected class label for a decision tree originates from the tree's root. The following steps in the tree are decided by comparing the value of the root attribute with the information in the record. Following a jump on the next node, the matching branch is followed to the value shown by the comparison's outcome. When a decision tree node is used to split training instances into smaller groups, entropy changes. Information gain is the unit of measurement for this change in entropy.

Classification Report: DecisionTreeClassifier

	precision	recall	f1-score	support
0	0.59	0.59	0.59	1120
1	0.59	0.60	0.60	1130
accuracy			0.59	2250
macro avg	0.59	0.59	0.59	2250
weighted avg	0.59	0.59	0.59	2250

2.Decison Tree Classifer

3.5.3 Xgboost Classifier

A variation of gradient enhanced decision trees is called XGBoost. Using this approach, decision trees are created one after the other. Weights are assigned to each independent variable, and the decision tree uses these weights to generate predictions. The relevance of the pertinent factors is elevated and applied in the subsequent decision tree in the event that the tree predicts incorrectly.

After that, the output from each of these classifiers/predictors is combined to create a model that is more reliable and accurate. In a study, the XGBoost model with the following parameters: "learning_rate": 0.1, "max_depth": 4, "n_estimators": 100, and "cross-validation": 10 folds, on 70,000 CVD dataset with 49,000 training and 21,000 testing data instances, obtained 73% accuracy.

Classification Report: XGBClassifier

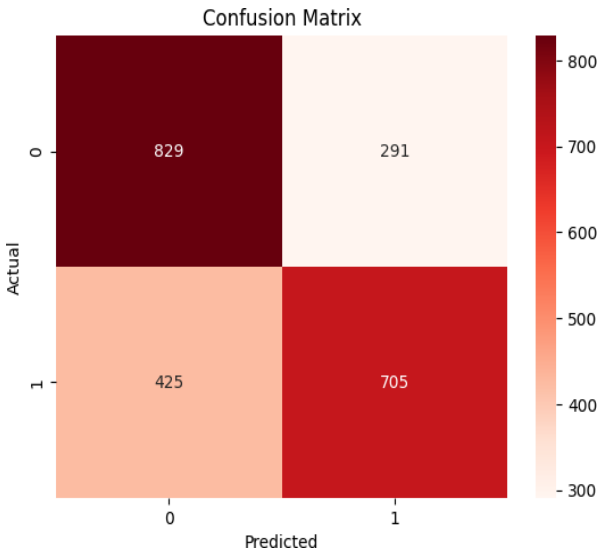
	precision	recall	f1-score	support
0	0.63	0.65	0.64	1120
1	0.64	0.62	0.63	1130
accuracy			0.63	2250
macro avg	0.63	0.63	0.63	2250
weighted avg	0.63	0.63	0.63	2250

ROC_AUC_SCORE is 0.6325023704171935

3.Report XGBoost

4. RESULTS AND DISCUSSION

The random forest, decision tree, multilayer perception, and XGBoost classifier methods were employed in this investigation. This study employed a number of performance metrics, including area under the ROC curve, F1 score, accuracy, precision, and recall. Eighty percent of the dataset was utilised to train the model, and the remaining twenty percent was used for testing.We used the GridSearchCV method as part of an automated hyperparameter tuning process. GridSearchCV produces the optimal collection of hyperparameters that maximises the scoring method given an estimator, a set of hyperparameters to be searched over, and a scoring method. This technique, which is included in the scikit-learn toolkit, assesses the effectiveness of various sets of hyperparameters using k-fold cross-validation.



4.Confusion Matrix

5. CONCLUSION

This study's main goal was to categorize cardiac disease using several models and an actual dataset. To predict the existence of heart disease in a dataset of patients, the k-modes clustering technique was used. The age attribute was converted to years and split into bins of five-year intervals. The diastolic and systolic blood pressure data were split into bins of ten intervals as part of the preprocessing of the dataset. To account for the distinct traits and course of heart disease in men and women, the dataset was further divided based on gender.For both the male and female datasets, the ideal number of clusters was ascertained using the elbow curve approach. The MLP model has the maximum accuracy of 87.23%, according to the data. These results show the possibility of clustering using k-modes. : to directly predict heart complaint and suggest that the algorithm could be a precious tool in the development of targeted individual and treatment strategies for the complaint. The study employed the Kaggle cardiovascular complaint dataset with 70,000 cases, and all algorithms were executed on Google Colab. The rigor of all algorithms was above 86 with the lowest delicacy of86.37 given by decision trees and the topmost delicacy given by multilayer perceptron, as previously mentioned. Limitations. First, the study was predicated on a single dataset and may not be generalizable to other populations or patient groups. likewise, the study only considered a limited set of demographic and clinical variables and did not take into account other implicit trouble factors for heart complaint, analogous as life factors or heritable tendencies. also, the performance of the model on a held- out test dataset was not estimated, which would have handed insight on how well the model generalizes to new, unseen data. Initially, the interpretability of the results and the capability to explain the clusters formed by the algorithm was not estimated.

In light of these limitations, it's recommended to conduct further disquisition to address these issues and to more understand the eventuality of k- modes clustering for heart complaint prophecy .

.

REFERENCES

1. Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.;
et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904. [CrossRef] [PubMed]
2. Droz˙dz˙ , K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman,W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H.
Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240. [CrossRef] [PubMed]
3. Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332. [CrossRef]
4. Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng,
S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. Circulation 2019, 139, e56–e528. [CrossRef] [PubMed]
5. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655. [CrossRef]
6. Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.;
Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322. [CrossRef]
7. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48. [CrossRef]
8. Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P.Work stress and cardiovascular disease: A life course perspective. J. Occup. Health 2016, 58, 216–219. [CrossRef]
9. Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. Procedia Comput. Sci. 2016, 85, 962–969. [CrossRef]
10. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.
Int. J. Comput. Appl. 2011, 17, 43–48. [CrossRef]
11. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.