Investigating the Effect of Information Theory in Machine Learning for Predicting Myocardial Infarction

Annajirao Challa, Omar Abueed, Rohan Hanamsagar, Upendhar Deshaboina, and Venkateswara Rao Thota

Department of Systems Science and Industrial Engineering, Binghamton University

May 2023

1 Abstract

Myocardial Infarction, also known as Heart attack is one of the most prevalent types of health problems worldwide. The study and prediction of myocardial infarction is a vital field of research in the healthcare sector that focuses on identifying and preventing cardiovascular diseases. This research examined the deployment of mutual information as a feature selection approach in machine learning models to predict the occurrence of heart attacks. Using a dataset that comprises electronic health records of patients. Four machine learning models were employed to predict the occurrence of heart attacks: Logistic Regression, Support Vector Machine Classifier, Decision Tree Classifier, and Gaussian Naive Bayes Classifier. These machine learning models' performance was validated and compared by relying on selected performance indicators including the confusion matrix and F1 score. The findings revealed that mutual information was efficient in the feature selection of heart disease parameters. Amongst the applied models, the SVM classifier and Logistic Regression outperformed the others. This research highlights the necessity of selecting a suitable classifier and exhibits the potential of mutual information as a feature selection strategy in machine learning for predicting heart attack probability.

2 Introduction

2.1 Myocardial Infarction

The heart is the vital organ that pumps blood throughout the body, oxygenating and nourishing tissues and maintaining the proper function of other organs in the body. Having a healthy heart is essential for healthy living.

Myocardial infarction (MI), another name for a heart attack, is one of the cardiovascular diseases that occur when the blood supply to a part of the heart is cut off, by a blood clot or plaque, composed of lipids, cholesterol, fat, and other chemicals, can build up in the artery walls over time and cause them to become harder and narrower. Heart attacks harm the myocardium(heart muscle). When the blood supply is restricted, the individual can expect serious and occasionally fatal repercussions.

Risk factors that can lead to a heart attack include smoking, diabetes, obesity, high blood pressure, and congenital heart diseases (heart problems occurs via genetics). Discomfort in the chest, shortness of breath, sweating, nausea, and lightheadedness are a few of the symptoms of a heart attack. Immediate medical attention is required to prevent permanent heart damage and reduce the severe impact.

To reduce heart-related problems and safeguard the heart from catastrophic risks, early detection of heart attack is essential. Based on medical history, expert symptom analysis reports, and physical laboratory results, invasive procedures diagnose cardiac disorders. Human intervention sometimes lengthens the diagnosis times, expensive, and time-consuming.

2.2 Machine learning

Machine learning is a subfield of Artificial Intelligence that aims at the development of computer systems or models that are capable of learning from historical data and making accurate predictions. The primary objective of machine learning is to develop algorithms and models that can identify patterns in data and use those patterns to predict and classify new data. To automate data-driven decision-making processes, machine learning enables computers to construct models from sample data. There are two primary categories of machine learning: supervised and unsupervised learning.

2.2.1 Supervised Learning

In supervised learning, the historical data fed to the model are labeled data. The idea behind this approach is to let the model learn the patterns and connections between independent variables and the target variable to forecast the new target variable based on new instances.

2.2.2 Unsupervised Learning

In unsupervised learning, the unlabelled data is given to the model. The main goal of unsupervised machine learning is to detect the patterns in unlabeled input data, which allows the model to classify the raw data. Huge amounts of data, or "big data," are produced by the healthcare sectors and can contain hidden knowledge or patterns which can be used for research. Machine learning techniques have become a potential tool for identifying and diagnosing heart attacks by analyzing vast amounts of patient data. Machine learning in healthcare helps with disease prognosis, better diagnosis, symptom assessment, delivery of appropriate medications, improvement of care quality, reduction of costs, prolongation of life, and reduction of cardiac patient mortality rates.

2.3 Information Theory

In Information Theory, Mutual information can be defined as the measure of statistical relationships between two variables. In addition, it demonstrates how much information can be gained from simultaneously observing two random variables. Mutual information is related to the concept of entropy. That is mutual information can be derived from entropy. A high value of mutual information signifies a substantial reduction in uncertainty, whereas a low value signifies a substantial increase in uncertainty. The two random variables are unrelated if they have less value for mutual information. The mutual information can aid in the selection of feature subsets from a sufficiently larger dataset in order to eliminate any unnecessary ones.

3 Literature Review

Cardiovascular diseases, particularly Myocardial Infarction is one of the main causes of global mortality. Early prediction and identification of this disease can benefit the reduction of negative outcomes and enhance patient care. Machine learning and Data Analytics have shown the promise in forecasting the likelihood of Myocardial Infarction based on electronic health record.

Several studies have looked into the use of machine learning algorithms to forecast the risk of myocardial infarction. Agrawal et al (2022) investigated the deployment of the machine learning algorithm for distinguishing healthy and infarcted patients using heart rate variability derived vector magnitude and direction. This current study contributes to establishing that VM-derived QT interval variability has a greater inherent value in identifying MI patients than RR variability—a measure that has received a lot of attention in previous studies. Furthermore, the study demonstrated that machine learning algorithms could correctly categorize healthy and MI patients based on interval data.

Indrakumari et al 2020, utilized Heart Disease Prediction Using Exploratory Data Analysis to demonstrate that a large volume of data may be used to create decisions that are more accurate than intuition. EDA identifies errors, locates relevant data, verifies assumptions, and discovers the relationship between explanatory factors. In this context, EDA is defined as data analysis that excludes conclusions and statistical modeling. Analytics is an important approach for every profession since it predicts the future and uncovers hidden patterns.

In recent years, data analytics has been regarded as a cost-effective tool that plays an important role in healthcare, including new research discoveries, emergency circumstances, and disease outbreaks. Analytics in healthcare enhances treatment by promoting preventative care, and EDA is a critical step in data analysis.

The risk variables that cause heart disease are evaluated and forecasted in this work using the K-means method, and the study is carried out using publically accessible data for heart disease. The dataset contains 209 records with 8 characteristics including age, kind of chest pain, blood pressure, blood glucose level, ECG at rest, heart rate, and four forms of chest pain. The k-means clustering method is used in conjunction with data analytics and visualization tool to forecast cardiac disease. The study goes through pre-processing techniques, classifier performance, and assessment criteria. The graphical data in the result section demonstrates that the forecast was correct.

Gupta et al 2021, presented a machine-learning approach for predicting heart attacks. The study utilized various supervised machine learning classifiers including Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression to develop a model for predicting myocardial infarction. Despite inconsistencies in the datasets, feature transformers were used to improve the consistency of the datasets. The resulting model achieved an average accuracy of 85.5% and a recall rate of 82% when trained on the Framingham dataset using the Gradient Boosting classifier. Additionally, the recall rate was further improved to 89.1% when training the Gradient Boosting classifier on the UCI Heart dataset. Following these approaches, a final model was deployed for predicting cardiac arrest using the Gradient Boosting Classifier.

This research could be expanded by integrating semi-supervised and deep learning techniques to further diversify the approach. The results indicate that the Gradient Boosting classifier achieved the highest accuracy score, with the model predicting the likelihood of a heart attack in binary form, where 1 represents a high chance of heart attack and 0 indicates no chance, and heart rate can also be beneficial in identifying individuals at risk of a heart attack. Among the most influential attributes, chest pain type is a significant predictor, with typical angina being the most influential and asymptotic chest pain being the least influential. High cholesterol levels, particularly those greater than 200mg/dL, are also strong predictors of heart attack risk, along with increased heart rate, and age. The study concludes that up to 80% of premature heart attacks can be prevented through a healthy diet and regular exercise, avoiding tobacco products, and drinking at least five glasses of water daily. Regular medical check-ups to monitor blood pressure, cholesterol levels, and heart rate can also aid in identifying individuals at risk of heart attack.

Overall, machine learning and data analytics have demonstrated potential in forecasting the likelihood of a heart attack. These approaches have the potential to enhance patient care and outcomes by identifying individuals who are at high risk of having a heart attack early on and allowing for prompt intervention and treatment. Further study, however, is required to verify these models in broader patient groups and in diverse healthcare settings.

4 Methodologies

4.1 Research Formulation

In machine learning, feature selection is a typical initial step followed before training the model, intended to improve predictive power, manageability, and interpretability. The performance of a machine learning model can be improved by eliminating undesirable features. In this research, Mutual Information is used as a feature selection technique for predicting the probability of getting a myocardial infarction. We are using four machine learning models and they are Support Vector Machine Classifier, Logistic Regression, Decision Tree Classifier, and Gaussian Naive Bayes Classifier.

4.1.1 Dataset Description

To predict the probability of occurrence of myocardial infarction, we will be considering the electronic health records of 303 patients as the dataset, which was taken from kaggle.com [1]. The variables are Age, Sex, Type of Chest Pain, Blood Pressure, Number of Major Vessels, Exercise-Induced Angina, Cholesterol, Blood Sugar, Electrocardiographic Results, Maximum Heart Rate and the Target variable which indicates the probability.

5 Implementation

5.1 Importing and Loading Dataset

The dataset is imported using the function from the Pandas library.

5.2 Feature Selection

Feature Selection is a crucial machine learning technique for lowering the number of variables in a dataset. It is used to simplify the dataset and enhance the accuracy and effectiveness of machine learning models. This research exploits the Mutual Information concept as a technique for feature selection by calculating the mutual information between each input variable and the target variable. Based on the calculated mutual information values, the original dataset is divided into two datasets by a predetermined threshold value. A dataset titled "dataset_1" contains all variables with mutual information values exceeding the threshold. The remaining variables are organized in a separate dataset titled "dataset_2."

5.3 Test Train Split

Scikit-learn library "train_test_split" is used for splitting the datasets intp training dataset and testing the dataset in Python. This is a fundamental step prior to the training of the machine learning model. It aims to split the data into training dataset and testing dataset. The training dataset is used to train the model and the performance of the model is evaluated by comparing the target variable values predicted by the model using the testing dataset with the actual values of the target variable.

5.4 Model Selection

Models are selected depending on the nature of the issue under investigation. Since here we are performing classification, we have taken four machine learning models to test with the effect of mutual information - Support Vector Machine Classifier, Logistic Regression, Decision Tree Classifier, and Gaussian Naive Bayes Classifier.

5.5 Data Preprocessing and Model Training

First, we performed the data preprocessing stage, including data standardization. Then all the four models were trained on the dataset along with the mini datasets (dataset_1 and dataset_2). This involves fitting the models to the training data and optimizing their parameters.

5.6 Model Evaluation

The performance of the machine learning model is evaluated by using the testing dataset. This involves calculating metrics such as confusion matrix and F1 score. Also performed cross-validation and hyperparameter tuning for each and every model.

6 Result

In this research, we focused on predicting the risk of myocardial infarction through the effect of mutual information using machine learning models. The performance of each model was evaluated on three different datasets, including the original dataset, dataset_1, and dataset_2. The results showed that the performance of the models varied based on mutual information values. The confusion matrix, f1 score, and accuracy scores for these four models were displayed in figure i.

6.1 For Original Dataset

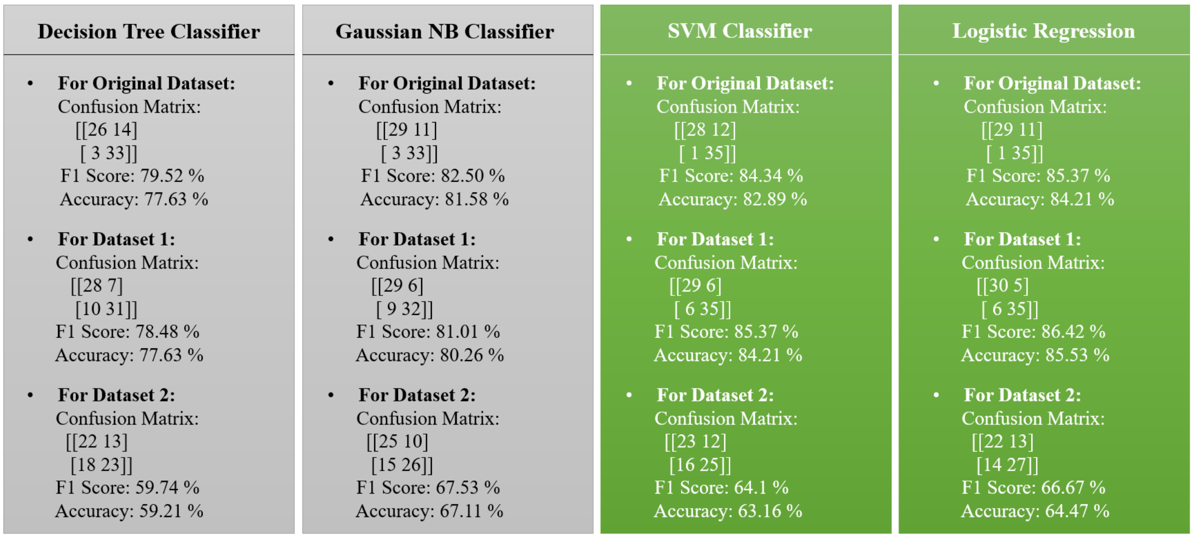
The accuracies of the four models are 77.63% (Decision Tree Classifier), 84.29% (Logistic Regression), 82.89% (Support Vector Machine Classifier), 81.58% (Gaussian Naive Bayes Classifier).

6.2 For Dataset 1 with High Mutual Information

The accuracies of the four models are 77.63% (Decision Tree Classifier), 85.53% (Logistic Regression), 84.21% (Support Vector Machine Classifier), 80.26% (Gaussian Naive Bayes Classifier).

6.3 For Dataset 2 with Low Mutual Information

The accuracies of the four models are 59.74% (Decision Tree Classifier), 64.47% (Logistic Regression), 63.16% (Support Vector Machine Classifier), 67.11% (Gaussian Naive Bayes Classifier).



Decision Tree Classifier	Gaussian NB Classifier	SVM Classifier	Logistic Regression
• **For Original Dataset:** Confusion Matrix: [[26 14] [3 33]] F1 Score: 79.52 % Accuracy: 77.63 %	• **For Original Dataset:** Confusion Matrix: [[29 11] [3 33]] F1 Score: 82.50 % Accuracy: 81.58 %	• **For Original Dataset:** Confusion Matrix: [[28 12] [1 35]] F1 Score: 84.34 % Accuracy: 82.89 %	• **For Original Dataset:** Confusion Matrix: [[29 11] [1 35]] F1 Score: 85.37 % Accuracy: 84.21 %
• **For Dataset 1:** Confusion Matrix: [[28 7] [10 31]] F1 Score: 78.48 % Accuracy: 77.63 %	• **For Dataset 1:** Confusion Matrix: [[29 6] [9 32]] F1 Score: 81.01 % Accuracy: 80.26 %	• **For Dataset 1:** Confusion Matrix: [[29 6] [6 35]] F1 Score: 85.37 % Accuracy: 84.21 %	• **For Dataset 1:** Confusion Matrix: [[30 5] [6 35]] F1 Score: 86.42 % Accuracy: 85.53 %
• **For Dataset 2:** Confusion Matrix: [[22 13] [18 23]] F1 Score: 59.74 % Accuracy: 59.21 %	• **For Dataset 2:** Confusion Matrix: [[25 10] [15 26]] F1 Score: 67.53 % Accuracy: 67.11 %	• **For Dataset 2:** Confusion Matrix: [[23 12] [16 25]] F1 Score: 64.1 % Accuracy: 63.16 %	• **For Dataset 2:** Confusion Matrix: [[22 13] [14 27]] F1 Score: 66.67 % Accuracy: 64.47 %

figure i: Result

7 Conclusion

To conclude, Mutual Information proved to be a useful feature selection technique in improving the performance of the models. The results also suggest that proper selection of threshold and machine learning model can significantly improve the accuracy. Therefore, the use of machine learning models in predicting myocardial infarction can be considered a promising approach, particularly when accompanied by appropriate feature selection and parameter optimization techniques. However, the research findings show that the performance of machine learning models varies depending on the dataset employed.

8 References

1. Rashik Rahman. (n.d.). Heart Attack Analysis Prediction Dataset. Www.kaggle.com. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

2. Gupta, S. K., Shrivastava, A., Upadhyay, S. P., Chaurasia, P. K. A Machine Learning Approach for Heart Attack Prediction.

3. Nahar, J., Imam, T., Tickle, K. S., Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert systems with applications, 40(1), 96-104.

4. Soudan, B., Dandachi, F. F., Nassif, A. B. (2022). Attempting cardiac arrest prediction using artificial intelligence on vital signs from Electronic Health Records. Smart Health, 25, 100294.

5. Vergara, J. R., Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural computing and applications, 24, 175-186.

6. Agrawal, R. K., Sewani, R. R., Delen, D., Benjamin, B. (2022). A machine learning approach for classifying healthy and infarcted patients using heart rate variabilities derived vector magnitude. Healthcare Analytics, 2, 100121.

7. Aggarwal, S., Pandey, K. (2023). Early identification of PCOS with commonly known

diseases: Obesity, Diabetes, High blood pressure and Heart disease using Machine Learning Techniques. Expert Systems with Applications, 119532..

8. Kwon, S. H., Dong, L. (2022). Flexible sensors and machine learning for heart monitoring. Nano Energy, 107632.

9. Ahsan, M. M., Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 102289.

10. Indrakumari, R., Poongodi, T., Jena, S. R. (2020). Heart disease prediction using exploratory data analysis. Procedia Computer Science, 173, 130-139.