Myocardial Infraction Complications Predictions

Ijeoma Njoku injoku24@mmc.edu
Meharry Medical College
Biomedical Sciences Data Science
Nashville, Tennesee

***Abstract*— A myocardial infarction is also known as a heart attack. Heart attacks occur when blood flow to the heart is blocked thereby causing damage to the muscles of the heart. Unfortunately, there aren't typical myocardial infarctions, some may go undetected, or even lead to death. Myocardial Infarctions are an immediate medical emergency. Machine learning is a branch of AI that enables algorithms to discover or be taught patterns within a particular dataset which can make predictions on new or similar data. This dataset was collected in the Krasnoyarsk Interdistrict Clinical Hospital No.20, named after I.S Berzon(Russia) in 1992-1995.**

Keywords—Myocardial Infraction, Death, Hospitalization, Machine Learning.

I. Introduction

Around the world, half a billion people die from cardiovascular-type diseases that include myocardial infarction. It is estimated that 26% of women will die within a year of a heart attack, and 5 years after, 50% will die or develop complications. As a woman, this is an interesting and scary topic. This dataset will be used to predict Myocardial complications of patients by the end of the first day of stay. According to the World Health Federation, more than 80% of heart attacks can be prevented, and complications after are not always the cases. Most practitioners cannot predict the development of the complication to try and prevent it. Unfortunately, many complications can lead to death hopefully, with more guidance in the future, this will help aid practitioners around the world.

II. Data Description

A. Myocardial Infarction Complication Data: UCI Machine Learning Repository

The Myocardial Infarction dataset is publicly available and currently hosted by UCI Machine Learning Repository. This data set can be used to solve problems that challenge most in modern medicine today by using patient information to predict complications of Myocardial Infarction while admitted to the hospital.

B. Data Use

The dataset used in this project is publicly available from the UCI Machine Learning Repository and comprises data collected at the Krasnoyarsk Interdistrict Clinical Hospital No. 20 in Russia between 1992 and 1995. It includes 124 attributes related to patient health metrics and hospital admission details, with both categorical and integer data types. The dataset is suitable for classification tasks, aiming to predict whether a patient will develop complications within the first day of hospital admission. [6] The data is comprised of patient-related questions about the heart conditions that could increase the risk of a Myocardial infarction. For example, "coronary heart disease (CHD) in recent weeks, days before hospital admission" or "Cardiogenic shock at the time of admission to the intensive care unit." On top of heart conditions, the data also includes the patient's laboratory results, like white blood count, creatine, ALT, and sodium levels. Lastly, the target variables include myocardial rupture, chronic heart failure, and even death. The oldest patient in this data set is 92, while the youngest is 26. The average potassium level within this data set is 4.19, indicating a normal range. Only 6.47% of patients suffered from a cardiogenic shock (where the heart cannot pump enough blood), and lastly, 84% of the patients are still alive.

C. Dataset Characteristics:

- Type: Multivariate
- Number of Attributes: 124
- Attribute Type: Categorical & Integer
- Associated Task: Classification

D. Purpose and Use Cases

- Myocardial Infarctions can occur with or without complications that can lead to death.
- Four Possible time moments for complication prediction:
  - Time of admission
  - The end of the first day
  - The end of the second day
  - The end of the third day

E. Approach

Supervised learning models were implemented, including logistic regression, decision trees, and random forests, to predict the occurrence of complications. The steps in the approach include:

1. Data preprocessing: Handling missing values, standardizing numerical features, and encoding categorical variables using one-hot encoding.
2. Model training and evaluation**:** Dividing the dataset into training and test sets, followed by training the

models and evaluating their performance based on accuracy, precision, recall, and F1-score.

3. Hyperparameter tuning**:** Applying techniques such as grid search to optimize model parameters.
4. Model interpretability**:** Using feature importance metrics to identify key predictors.

F. Preliminary Results

Preliminary analysis has been performed using logistic regression and decision tree models. Initial results show that the logistic regression model achieved an accuracy of approximately 98.5%, while the decision tree model achieved 100% on the validation set. These initial findings demonstrate promising results, however, the perfect score for the decision tree suggests that there is room for improvement, particularly with model tuning and addressing the class imbalance in the dataset.

Future steps include implementing random forests, experimenting with hyperparameter tuning, and considering techniques such as synthetic minority oversampling to handle the imbalanced data distribution. One challenge encountered is the dataset's highly imbalanced nature, which can affect model performance and evaluation.

Please refer to notebook for more information.

III. Methods

A. Exploratory Data Analysis(EDA)

For the analysis, several preprocessing and exploratory data analysis (EDA) techniques were applied to prepare the dataset for modeling:

Handling Missing Values: The dataset contained over 1000 missing values across 153 features. These were handled by imputing with the mean to ensure no loss of critical information.

Encoding Categorical Variables: There were 105 categorical variables in the dataset, including age, sex, angina (chest pain), and presence of heart failure. These variables were encoded using one-hot encoding resulting in 40 new features for analysis.

Multicollinearity Analysis: Multicollinearity was analyzed using Variance Inflation Factor (VIF) and a correlation matrix]. A total of 20 highly correlated variables were identified, and features were removed based on their influence on other variables and their relevance to the target variable. The multicollinearity analysis graph is shown in Table 3 , highlighting the correlated features and justifying their removal.

Summary of EDA Findings: The EDA revealed key insights into the dataset, such as the older you are the higher chance that after one day you most likely have congestive heart failure, a younger person might have pulmonary edema and more males died of unknown causes. These findings informed the preprocessing decisions and model selection, ensuring the data was adequately prepared for robust analysis and accurate predictions.

B. Model Selection

We have implemented the following models to train and validate our data:

Logistic Regression: Logistic regression served as a baseline model due to its simplicity and interpretability. It models the probability of readmission as a linear combination of the features. The model's coefficients are straightforward to interpret, providing insights into the direction and strength of each feature's impact. However, it is limited in its ability to capture non-linear relationships in the data.

Random Forest: Random forest is an ensemble method that combines multiple decision trees, leveraging different subsets of data to enhance robustness and reduce overfitting. It performs well with non-linear data and provides feature importance metrics, making it useful for identifying key predictors. However, it can be computationally intensive and prone to overfitting if not carefully tuned.

Decision Tree: A decision tree is a simple, interpretable model that splits the data recursively based on feature thresholds. Each branch represents a decision rule, making it intuitive for users to understand. However, decision trees are prone to overfitting, especially in noisy datasets, and often require pruning or parameter tuning to improve generalizability.

XGBoost: XGBoost is an optimized version of gradient boosting designed for speed and efficiency. It incorporates regularization terms, making it less prone to overfitting and suitable for large datasets. While it offers high predictive power, it can be computationally intensive and requires careful parameter tuning.

C. Train Test Split

The dataset was divided into training and testing sets using a random or stratified approach. A total of 20% of the data was allocated for training, consisting of 340 instances, while the remaining 80% was reserved for testing, with 1360 instances. This split ensured that the models had sufficient data to learn patterns while maintaining a robust evaluation set. The class distribution was carefully maintained in both sets to reflect the original dataset composition.

Within the training set, we performed 5-fold cross-validation to identify the best hyperparameters for each model, ensuring optimal performance and reducing the risk of overfitting. This approach allowed us to fine-tune the models effectively while preserving the integrity of the test set for unbiased evaluation.

IV. Evaluation of Results

The performance of each model must be thoroughly assessed to predict complications of a myocardial infarction.

A. Model Evaluation

The performance of the conventional models was evaluated using metrics such as accuracy, sensitivity, specificity, precision, recall, and F1-score, as summarized in Table 2. This model also demonstrated strong performance across other metrics, making it the most effective conventional approach for the classification task.

The best-performing model Random Forest was optimized using the following hyperparameters listed in table 1 These hyperparameters were identified through cross-validation on the training set, ensuring a balance between generalization and performance. This configuration allowed the model to effectively capture patterns in the data while minimizing overfitting.

Confusion Matrix for Test Set: The confusion matrices for the top three models are illustrated in Table 4 providing a detailed breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These matrices offer insights into the classification performance of each model:

Model 1/RF: As shown in Table 2, Model 1 achieved the highest accuracy of 90% This performance demonstrates its strength in correctly identifying both positive and negative cases, making it the most robust among the evaluated models.

These visualizations highlight the comparative strengths and weaknesses of the top models, guiding the selection of the most appropriate model for the classification task.

B. Hypothesesis regarding Results compared to baselines:

I hypothesized that the more complex models (such as random forests, SVM, etc.) will outperform simpler models like logistic regression in terms of predictive accuracy and AUC-ROC.

V. Related Work

It is known that throughout many hospitals, there is a risk of complications after being admitted. Increased hospital stays can lead to hospital acquired complications and illnesses. Especially patients who are directly affected by health care disparities and other socioeconomic factors.

Most hospitals can be a mix of patients out of surgery, emergency visits turned admissions, same day appointments and even surgeries where the patient goes home same day. We are trying to predict a solution to a problem that can happen to any patient across the globe. Patients across the globe can not only suffer from a myocardial infarction but also from sepsis, thrombocytopenia, and death.[8]

Predictive modeling for cardiovascular complications has evolved from traditional statistical methods to more advanced machine learning approaches. Logistic regression, a widely used statistical technique, remains prevalent due to its interpretability and ease of implementation. However, it has limitations when dealing with complex, high-dimensional data. Decision trees and ensemble methods like random forests have been applied to improve predictive performance, benefiting from their ability to handle interactions between features.

A study by resident physicians at Standford hospital observed a mixture of medical and surgical patients who were septic in the first 24 hours or had to be transferred to the intensive care unit. To summarize, the study showed that the 3,862 patients observed had an increased 1-year mortality if they required escalation to the ICU within 24 hours of hospital admission. [7]

With more studies, clinicians can reduce the chances of patients acquiring complications after admission.

References

[1] E. L. Cahill, and D. F. Sharaf, "Modeling Risk of Complications in Cardiovascular Patients: A Machine Learning Perspective," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 998-1005, 2021.

[2] J. Thomas, "Facts and Statistics on Heart Disease," *Healthline*, Jul. 20, 2023.Available: https://www.healthline.com/health/heart-disease/statistics#at-risk

[3] G N. Ojha and A. S. Dhamoon, "Myocardial infarction," *National Library ofMedicine*,Aug.08,2023.Available: https://www.ncbi.nlm.nih.gov/books/NBK537076/

[4] K. Tipton et al., Interventions to decrease hospital length of stay. Agency for Healthcare Research and Quality (US), 2021. Available: https://www.ncbi.nlm.nih.gov/books/NBK574438/

[5] M. GAMIL and A. FANNING, "The first 24 hours after surgery.," Anaesthesia, vol. 46, no. 9, pp. 712–715, Sep. 1991, doi: https://doi.org/10.1111/j.1365-2044.1991.tb09761.x.

[6] S. Shah, F. Qureshi, S. Stanley, and E. Bennett-Guerrero, "Unplanned hospital admissions within 24 h after 53,185 surgical procedures at a U.S. ambulatory surgery center," Perioperative Medicine, vol. 13, no. 1, Aug. 2024, doi: https://doi.org/10.1186/s13741-024-00447-y.

[7] J. Leong, J. Madhok, and G. K. Lighthall, "Mortality of Patients Requiring Escalation to Intensive Care within 24 Hours of Admission in a Mixed Medical-Surgical Population," Clinical Medicine & Research, vol. 18, no. 2–3, pp. 68–74, Jan. 2020, doi: https://doi.org/10.3121/cmr.2019.1497.

[8] G. J. Duke et al., "Hospital-acquired complications in critically ill patients," Critical Care and Resuscitation, vol. 23, no. 3, pp. 285–291, Sep. 2021, doi: https://doi.org/10.51893/2021.3.oa5.

[9] P. McNair, T. Jackson, and D. Borovnicar, "Public hospital admissions for treating complications of clinical care: incidence, costs and funding strategy," Australian and New Zealand Journal of Public Health, vol. 34, no. 3, pp. 330–333, Jun. 2010, doi: https://doi.org/10.1111/j.1753-6405.2010.00536.x.

[10] P. M Paithane, "Heart Disease Prediction Using Multiple Machine Learning Algorithms," Advances in Robotic Technology, vol. 2, no. 1, pp. 1–5, Jan. 2024, doi: https://doi.org/10.23880/art-16000114.

Table 1

Models	Best Hyperparameters
Logistic	• 'C': 0.01, 0.1 • 'penalty': 'l2' • 'solver': 'liblinear', 'saga' • 'max_iter': 100, 500, 1000
Decision Tree	• 'max_depth': 8,20,12 • 'min_samples_split': 15,9 • 'min_samples_leaf': 3, 2 • 'max_features: sqrt, log2 • criterion': 'gini', 'entropy'
Random Forest	• 'n_estimators': 100,200 • 'max_depth': 10, 7 • 'min_samples_split': 10, 15 • 'min_samples_leaf': 3, 5 • Max_features: sqrt , log2 • Criterion : gini , entropy
XG Boost	• Min_child_weight: 1,8 • Gamma: 1,2 • Colsample_bytree: 0.3 , 0.5 • Max_depth: 3, 4

Table 2

	Logistics	**Decision Tr**	**Random Forest**	**XGBoost**
Accuracy	0.84	0.87	0.92	.90
F1 Score	0: 0.92 1: 0.68 2:0.17 3:1.00 4:0.33 5:0.00 6:0.00 7:0.15	0: 0.94 1: 0.36 2:0.00 3:0.67 4:0.00 5:0.00 6:0.00 7:0.00	0: 0.96 1: 0.68 2:0.00 3:1.00 4:0.00 5:0.00 6:0.00 7:0.00	0: 0.97 1: 0.69 2:0.00 3:0.00 4:0.00 5:0.00 6:0.00 7:0.00
Sensitivity (Recall)	0: 0.92 1: 0.68 2:0.17 3:1.00 4:0.33 5:0.00 6:0.00 7:0.15	0: 0.98 1: 0.30 2:0.00 3:0.64 4:0.00 5:0.00 6:0.00 7:0.00	0: 1.00 1: 0.61 2:0.00 3:1.00 4:0.00 5:0.00 6:0.00 7:0.00	0: 1.00 1: 0.92 2:0.00 3:0.00 4:0.00 5:0.00 6:0.00 7:0.00

	SEX	INF_ANAM	STENOK_AN	IBS_NASL	SIM_GIPERT	nr_11	nr_01	nr_02	nr_03	nr_04	nr_07	nr_08	np_01	np_04	np_05	np_07
	-0.116794	0.097365	0.184093	0.005925	0.014280	-0.016968	0.017423	0.020729	0.056942	0.059730	0.012424	0.023386	0.018132	-0.005574	-0.011377	-0.000481
	1.000000	0.040407	-0.090767	-0.013731	-0.065592	0.004970	-0.063023	-0.010727	-0.042578	-0.039491	0.018669	0.012273	-0.009027	-0.025545	0.016691	0.018695
	0.040407	1.000000	0.320776	-0.006183	-0.015833	0.049090	0.026047	0.063722	0.057939	-0.027206	0.071082	0.040578	0.038911	0.039284	0.060769	0.012989
	-0.090767	0.320776	1.000000	-0.008549	-0.020662	0.080423	0.034659	0.002452	0.025682	0.085629	-0.023762	-0.025385	0.009992	-0.003680	0.004863	0.037870
	-0.013731	-0.006183	-0.008549	1.000000	0.008200	0.033286	-0.045698	0.035104	0.025990	0.000000	-0.000000	-0.045698	0.000000	0.000000	-0.000000	-0.000000
	-0.065592	-0.015833	-0.020662	0.008200	1.000000	0.033637	-0.009010	-0.019725	-0.003884	-0.024443	-0.004501	-0.009010	-0.006356	-0.007786	-0.014945	-0.004493
	0.004970	0.049090	0.080423	0.033286	0.033637	1.000000	-0.007827	-0.017137	-0.023371	-0.021235	-0.003910	-0.007827	-0.005535	-0.006781	0.034257	-0.003913
	-0.063023	0.026047	0.034659	-0.045698	-0.009010	-0.007827	1.000000	-0.005228	-0.007130	-0.006479	-0.001193	-0.002388	-0.001689	-0.002069	-0.003971	-0.001194
	-0.010727	0.063722	0.002452	0.035104	-0.019725	-0.017137	-0.005228	1.000000	-0.015610	-0.014183	-0.002612	-0.005228	0.159547	-0.004529	0.061102	-0.002613
	-0.042578	0.057939	0.025682	0.025990	-0.003884	-0.023371	-0.007130	-0.015610	1.000000	-0.019344	-0.003562	-0.007130	-0.005042	-0.006177	0.039818	-0.003564
	-0.039491	-0.027206	0.085629	0.000000	-0.024443	-0.021235	-0.006479	-0.014183	-0.019344	1.000000	-0.003236	-0.006479	-0.004423	-0.005419	-0.010402	-0.003127
	0.018669	0.071082	-0.023762	-0.000000	-0.004501	-0.003910	-0.001193	-0.002612	-0.003562	-0.003236	1.000000	-0.001193	-0.000844	-0.001033	0.300609	-0.000596
	0.012273	0.040578	-0.025385	-0.045698	-0.009010	-0.007827	-0.002388	-0.005228	-0.007130	-0.006479	-0.001193	1.000000	-0.001689	0.287209	-0.003971	-0.001194
	-0.009027	0.038911	0.009992	0.000000	-0.006356	-0.005535	-0.001689	0.159547	-0.005042	-0.004423	-0.000844	-0.001689	1.000000	-0.001458	-0.002799	-0.000842
	-0.025545	0.039284	-0.003680	0.000000	-0.007786	-0.006781	-0.002069	-0.004529	-0.006177	-0.005419	-0.001033	0.287209	-0.001458	1.000000	-0.003430	-0.001031
	0.016691	0.060769	0.004863	-0.000000	-0.014945	0.034257	-0.003971	0.061102	0.039818	-0.010402	0.300609	-0.003971	-0.002799	-0.003430	1.000000	-0.001979
	0.018695	0.012989	0.037870	-0.000000	-0.004493	-0.003913	-0.001194	-0.002613	-0.003564	-0.003127	-0.000596	-0.001194	-0.000842	-0.001031	-0.001979	1.000000
	-0.036158	0.031863	-0.016261	0.000000	-0.011021	-0.009598	-0.002928	-0.006411	0.061119	-0.007671	-0.001463	-0.002928	-0.002064	-0.002529	-0.004855	-0.001459
	-0.009027	0.059448	-0.019062	-0.000000	-0.006356	-0.005535	-0.001689	-0.003697	-0.005042	-0.004423	-0.000844	-0.001689	-0.001190	-0.001458	-0.002799	-0.000842
	-0.025545	0.005737	0.065632	0.000000	-0.007786	-0.006781	-0.002069	-0.004529	-0.006177	0.102827	-0.001033	-0.002069	-0.001458	-0.001787	-0.003430	-0.001031
	-0.261425	0.102865	0.099284	0.002167	0.079793	-0.018939	0.087580	0.006909	0.003020	0.014177	-0.009670	-0.019358	-0.013655	-0.016729	0.010970	-0.009653
	-0.104925	-0.012514	0.051583	-0.004755	0.096414	-0.025135	-0.007668	0.019264	0.057180	0.008467	-0.003831	0.070554	-0.005409	0.083669	-0.012720	-0.003824

Confusion table for random forest

Table 4