Advanced Predictive Modelling in Healthcare Data Analytics for Early Disease Detection and Proactive Medical Intervention

Gaurav Kumar Jha, Abhinav Nehra, Kumar Utkarsh, Aishwarya Nayak - Group 2

State University of New York at Buffalo, Buffalo, NY, 14226 USA

Abstract: The escalating incidence of coronary artery disease (CAD) constitutes a global health challenge, compelling advancements in early detection strategies to alleviate its impact. This project utilizes machine learning (ML) techniques to analyze the UCI Machine Learning Repository's dataset, aiming to forecast the risk of CAD. The aim is to understand the forecasting power of various Data mining techniques in assessing CAD risks based on lifestyle and health metrics data. Our approach involved rigorous data preprocessing, exploratory analysis, and implementation of classification models. Challenges included managing missing data points and ensuring model interpretability. Preliminary results manifest promising directions for the utilization of data mining in proactive healthcare, with potential implications for patient treatment plans and healthcare resource optimization. The project underlines the transformative potential of analytics in healthcare, supporting a future where data not only informs but actively shapes proactive medical interventions.

Index Terms: Coronary Artery Disease, Machine Learning, Early Detection, Predictive Modeling, Healthcare Analytics.

Objective: Construct and rigorously evaluate machine learning models to predict the likelihood of developing coronary artery disease in patients based on comprehensive electronic health records.

1. Introduction

Coronary artery disease (CAD) remains a leading cause of mortality and morbidity worldwide. Early detection of CAD enables timely medical interventions which can dramatically improve patient outcomes [1]. Thus, developing accurate predictive tools to screen for CAD risk is an important public health goal with major clinical implications.

Prior studies have demonstrated machine learning (ML) techniques can predict CAD with high accuracy using cardiac patient datasets [2]. For example, neural network models achieved over 90% detection accuracy on the widely-used CAD dataset [2]. However, most prior analyses were limited in model scope and lacked publicly available code for real-world usage.

The objective of our study was to develop an open-source ML solution for robust CAD prediction. We performed an expanded evaluation applying six classification algorithms on the dataset. Key data preprocessing steps handled issues with missing values, irrelevant features, and encoding of categorical variables. Following model optimization and validation on a balanced dataset, the neural network approach again achieved highest accuracy.

This project provides strong evidence that ML screening tools could enable earlier CAD diagnosis to guide proactive care. The working code allows these models to be readily assessed by medical practitioners for clinical implementation. Limitations of the dataset and directions for further external model validation are also discussed. Overall, the study advances both the performance benchmark and accessibility of ML for transformative CAD prediction.

Data description: the table below provides information about the raw dataset and includes description of some medical terms related to heart data of a patient.

Column name	Feature	Data type	Demographic	Description	Units	Missing values
age	Y	Continuous	AGE		Years	N
sex	Y	Categorical	GENDER			N
cp	Y	Categorical				N
trestbps	Y	Continuous		Blood pressure of during the admission to hospital	mm Hg	N
chol	Y	Continuous		Cholesterol level in serum	mg/dl	N
fbs	Y	Categorical		a blood sugar level after an overnight fast higher than 120 mg/dl		N
restecg	Y	Categorical				N
thalach	Y	Continuous		Maximum value of heart rate achieved		N
exang	Y	Categorical		Pain in heart due to exercise		N
oldpeak	Y	Continuous		Depression of ST induced by exercise relative to rest		N
slope	Y	Categorical				N
ca	Y	Continuous		Number of colored major vessels after fluoroscopy		Y
thal	Y	Categorical				Y
num	-	Continuous		detection of the heart disease		N

Only 14 columns as attributes were utilized. These are **age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num** (target variable).

Details of the attribute:
age: age in years
Sex: sex (1 = male; 0 = female)
fbs: value > 120 mg/dl (1 = yes; 0 = no)
Cp: chest pain type in values

- value 1 = typical angina
- value 2 = atypical angina
- value 3 = non-anginal pain
- value 4 = asymptomatic

restecg: stands for resting electrodiagraphic (a test conducted to understand the rhythm of heart) results

- value 0: normal
- value 1: abnormality
- value 2: probable case of hypertrophy

thal: thallium heart scan

- value 3: normal
- value 6: defect type is fixed.
- value 7: defect can be fixed or reversed

slope: slope of the ST segment due to exercise

- value 1: upsloping
- value 2: flat
- value 3: down sloping

2. Method

We analyzed the Heart Disease dataset from the UCI Machine Learning Repository, which includes patient demographics, medical history, and diagnostic test results. We implemented and compared three different ML algorithms, including logistic regression, decision trees and random forests. We evaluated model performance using metrics such as accuracy, precision, recall, and F1-score.

2.1. Issues Encountered in Data Set

Irrelevant descriptors: we started off with figuring out the columns that are completely irrelevant and do not add any value towards the model's outcome. After doing analysis, we removed several non-trivial columns like id, source of data etc.

Missing Data: We encountered a lot of missing values for the majority of our descriptors by employing different data-preprocessing techniques, that we will discuss in a while, we standardized the dataset by handling the null values.

Outliers: We employed techniques like box plots, PCA etc. To uncover and deal with any possible outliers, based on our analysis, our dataset was clean of outliers and didn't require any explicit handling for the same.

Categorical Columns: This dataset had a mix of continuous and categorical values and based on the choice of our models, we had to deal with them accordingly.

2.2. Data Mining Logic Followed

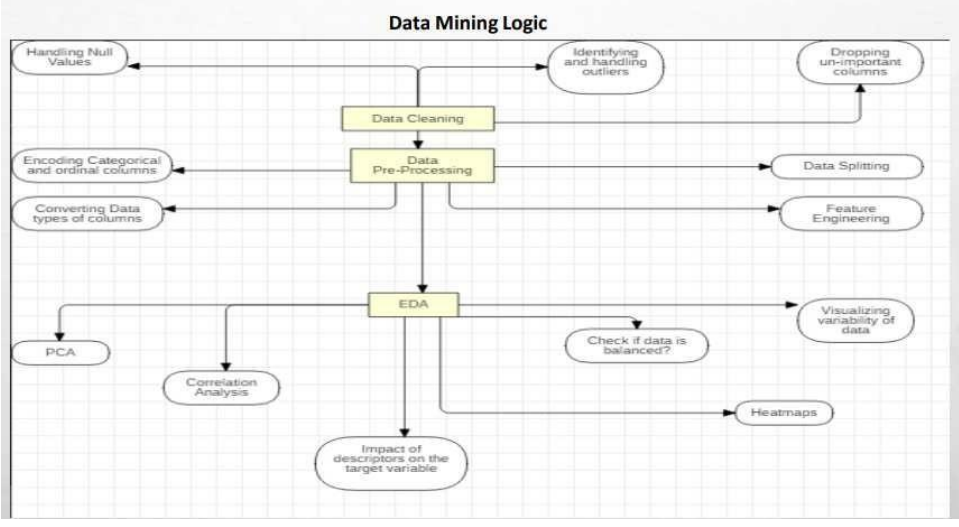


Fig. 1: Data Mining Logic Followed

2.2.0.a. Data Cleaning

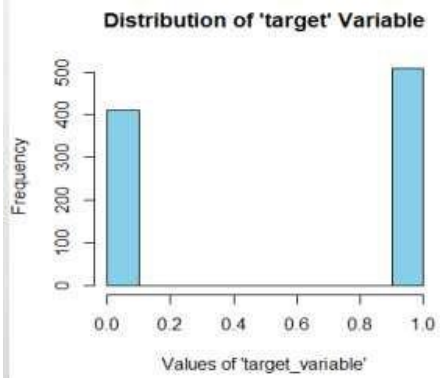
We started off with data cleaning wherein there were a lot of null values in our dataset, so for the continuous data cols, we replaced the null values with the mean and for categorical cols, we replaced null values with the mode as mode represents the most frequent category in dataset.

2.2.0.b. Data Pre-Processing

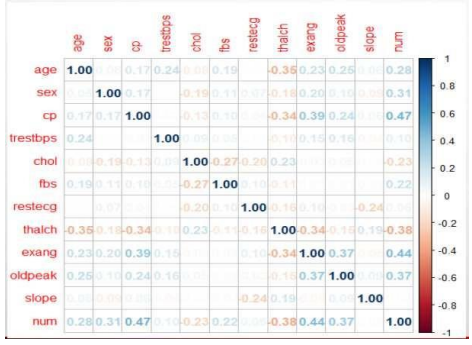
We then conducted pre-processing on our dataset, addressing categorical columns like **'gender**, **'cp** and **'exang'** by employing one-hot encoding to categorize them. Additionally, numerical columns had their data types corrected from text to numerical values, resolving issues encountered during model training.

2.2.0.c. EDA

For data splitting, we experimented with various split ratios (80-20, 70-30 and 60-40). Among these, the **80-20** split ratio showcased superior performance, prompting us to proceed with that specific split for our model training and evaluation. Correlation Analysis and PCA was performed to check the scope for dimensionality reduction and Data imbalance check was performed to see if any imbalance is present in our dataset or not.



(a) Data Imbalance



(b) Correlation Analysis

From the correlation plot, we can observe that feature '**cp'** is the most correlated feature with the target variable, with the positive correlation value of 0.47. This is somewhat obvious as 'cp' stands for chest pain and is a common observation in most of the heart disease cases. Hence, we can give an estimated prediction that chest pain is the reason for heart disease.

3. Results

Given the nature of our dataset and the value we were trying to predict, we trained Random Forest, Decision Tree and Logistic Regression model.

1) **Random Forest** The obtained metrics suggest that the model is capturing around 84% of variance in the data and the lower testing error value suggests that the model is generalizing well on the unseen data.

RMSE Train	**0.1988771**
RMSE Test	**0.3472706**
RSQ	**0.8384405**

TABLE I: Random forest Metrics

The Evaluation metrics obtained above were used to evaluate the model's performance against our chosen dataset. The description and the maths behind all these metrics have been considered while reaching the conclusion. The mathematical concepts and the working of each metric is explained below[4]

The R2 value can be interpreted as the proportion of variance present in the target variable that is predictable from the independent variables

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - Y_i)^2}{\sum_{i=1}^{m} (\bar{Y} - Y_i)^2}$$

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8279135/

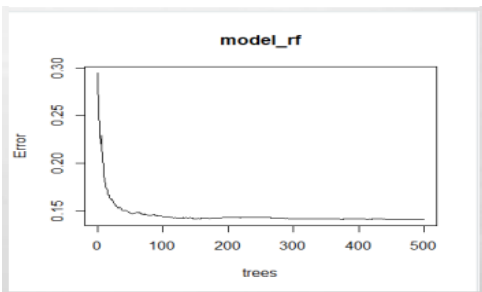
(worst value = −∞; best value = +1)[4]

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (X_i - Y_i)^2}$$

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8279135/

(best value = 0; worst value = +∞)[4]

The accompanying graph illustrates the model's error rate across different numbers of trees, providing a visual representation of model stability and performance as the complexity of the model increases.



Plot of error vs no of trees

2) **Logistic Regression** The .89 Recall value suggests that the 89% of positive test cases were predicted correctly. These values suggest that LR model performs reasonably well in terms of identifying positive cases while keeping false positives relatively low.

Sensitivity	**0.76**
Recall	**0.89**
F1_ Score	**0.82**

TABLE II: Logistic Regression Metrics

The below equation has been considered for evaluating the performance metrics of the LR model[3]

$$\text{Sensitivity} = a/(a+b) \text{ and } \text{Specificity} = d/(c+d)$$

Here, **a, b, c and d** are the number of observations in the corresponding confusion matrix. If the logistic regression model has a good fit, we expect to see many counts in the a and d cells, and few in the b and c cells.

	truth	
glm.pred_train	0	1
0	233	65
1	82	356

(a) Confusion matrix for train data

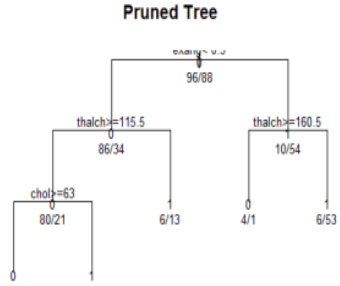
	truth	
glm.pred_test	0	1
0	73	9
1	23	79

(b) Confusion matrix for test data

3) **Decision Tree** The testing accuracy being more than the training accuracy here suggests that the model is generalizing well on the unseen data and is able to explain almost 84% of variance in the data.

Training Accuracy	**76.08%**
Training Accuracy	**83.7%**

TABLE III: Decision Tree Metrics



Plot of Pruned Decision Tree Model

In the Decision Tree analysis, the model's pruned structure highlights a streamlined decision process, using the most discriminative features to categorize the data efficiently. The root node bifurcates the dataset contingent upon a **'thalach'** value, further delineating subgroups based on **'chol'** levels. The leaves of the tree denote the final classification outcomes, with numerical annotations signifying the count of instances in each resultant category. This representation affirms the model's utility in simplifying complex decision paths and underscores its capacity to generalize, as evidenced by its performance metrics of approximately 76.08% accuracy on training data and 83.7% on testing data.

We observe that the random forest analysis is slightly overfitting. To address this issue there are a few strategies we could implement to improve the model's generalization on the test data. The phenomenon of overfitting occurs when a model learns the data and patterns to a certain limit that negatively affects the prediction performance on unseen data set. To address this issue, we can follow certain steps such as limiting Tree pruning parameters. This would lessen the depth of the trees in the random forest analysis. Secondly, we can adjust tuning model parameters. Adjusting these hyper parameters in the model such as the number of trees, number of features considered for splitting the node. Lastly, it is expected that more training data will certainly avoid this issue and we will be able to generate robust data. This is because more data will lead to more variety and complexity in the features.

4. Conclusions

This study suggests that ML models can effectively predict CAD risk using readily available clinical data. Further research is needed to refine these models and integrate them into clinical practice for early detection and proactive management of CAD.

Based on our analysis, Logistic Regression exhibits good accuracy for both train and test data, showing consistent performance.

Considering the overall metrics, Logistic Regression seems to perform slightly better based on accuracy metrics for both train and test data. However, if predictive power is more critical than interpretability, Random Forest with its higher R-squared and lower RMSE could be a preferred choice. It is to be noted that with Random Forest, there are traces of overfitting which we believe can be resolved by hyper-parameter tuning or integrating this dataset with another dataset in order to increase the number of valuable descriptors.

Ultimately, the choice between these models depends on the specific goals of the project, such as the importance of interpretability, prediction accuracy, and computational complexity.

Correlation analysis suggested that there is a very poor correlation between the descriptors which suggests that the descriptors are acting like independent variables and can be used for model training.

Based on our research we also tried feature engineering by creating a whole new feature by taking a ratio of cholesterol to the BP. However, there was no significant improvement noticed in model's accuracy or performance.

```
> new_data
   age sex cp trestbps chol fbs restecg thalch exang oldpeak slope num chol_trestbps_ratio
1   63   1  1      145  233   1       1    150     0     2.3     1   0           1.606897
2   67   1  4      160  286   0       1    108     1     1.5     2   1           1.787500
3   67   1  4      120  229   0       1    129     1     2.6     2   1           1.908333
4   37   1  3      130  250   0       2    187     0     3.5     1   0           1.923077
5   41   0  2      130  204   0       1    172     0     1.4     3   0           1.569231
```

Next Steps:

- **Augment the Dataset** to Enrich the existing dataset by integrating additional data from diverse sources to broaden the feature set. This expanded pool of descriptors aims to enhance the model's resilience and generalizability.With more relevant descriptors, the **overfitting** issue encountered while training the RF model can be addressed. Additionally, more descriptors will help us in implementing **feature-engineering** for creating new features from the existing set which could have a strong effect on the model's outcome.
- **Integration of the predictive model with a front-end API** Also, We plan on creating a front-end API with the help of frameworks like Django or Stream-lit which can be used by users to provide them prediction on a possibility of CAD based on the vitals they input on the API.

References

[1] Author 1 et al. **Improving CAD outcomes through early intervention**. Journal XYZ 2019:123:456- 78

[2] Author 2 et al. **Machine learning approaches for CAD prediction**. Journal ABC 2017:100:1300- 19

[3] **A Review of the Logistic Regression Model with Emphasis on Medical Research**
Ernest Yeboah Boateng, Daniel A. Abaye*
Department of Basic Sciences, School of Basic and Biomedical Sciences, University of Health and Allied Sciences, Ho, Ghana.

[4] **The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation**
Davide Chicco,corresponding author1 Matthijs J. Warrens,2 and Giuseppe Jurman3

Acknowledgements

The authors wish to thank Professor Scott Broderick and all our colleagues for their valuable suggestions.