**Coronary Artery Disease Risk Prediction Using Machine Learning**

A Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

Master of Computer Application

by

Shivam Sharma (2019CA58)

to the

**COMPUTER SCIENCE AND ENGINEERING DEPARTMENT**

MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY

ALLAHABAD

**June, 2021**

**UNDERTAKING**

We declare that the work presented in this report titled “Coronary Artery Disease Risk Prediction Using Machine Learning”, submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, for the award of the ***Master of Computer Application*** degree, is our original work. We have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, We accept that our degree may be unconditionally withdrawn.

June, 2021

Allahabad

Shivam Sharma (2019CA58)

**CERTIFICATE**

I Certified that the work contained in the report titled “Coronary Artery Disease Risk Prediction Using Machine Learning”, by Vaishali Jaiswal, Ayushi Barfa, Shivam Sharma, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

**Dr. Vibhav Prakash Singh**

Computer Science and Engineering Department

M.N.N.I.T, Allahabad

June, 2021

**Preface**

The development of coronary artery disease (CAD), a highly prevalent disease worldwide, Coronary artery disease is correlated with many preventable risk factors. Early diagnosis of CAD allows for prevention of worsening of CAD and its complications. Sometimes symptoms of CAD are so mild that it confused them from regular discomfort and not seek medical help at all, or they might reach the hospital but they have to go through various tests

Health Care today comes with a heavy price so in order to deal with this we decided to do this project, so that predictive models built using machine learning (ML) algorithms may assist clinicians in timely detection of CAD and may improve outcomes.

The key is that we can find the early stage of CAD when someone has 100% chance of survival.

Also we don’t need to go through unnecessary expensive and invasive medical treatments as it helps to reduce number of pathological tests to detect CAD.

**Acknowledgements**

The completion of this project could not have been possible without the participation and assistance of some individuals contributing to this project.

We would like to express our gratitude to **Dr. Vibhav Prakash Singh**, our project supervisor, for his patient guidance, enthusiastic encouragement and useful constructive recommendations on this project. Your insightful leadership and knowledge benefited us to complete this project successfully. Thank you so much for your continuous support and presence.

Last but not the least we would like to thank everyone who is involved in the project directly or indirectly.

**Contents**

Preface

Acknowledgements

1. Introduction………………………………………………………………….7
   * 1. Objective and Motivation……………………...……………...7
2. Dataset Analysis…………...………………………………………………...8
3. Methodology……………………………………………………………..9-12
   * 1. Feature Selection………………….………………………….10
     2. Data partitioning……………………….……………………..11
     3. Classification Models……………………….………………..12
4. Experimental Setup and Results Analysis………………...……………12-13
   * 1. Performance Measures………………………………………12
     2. Result ……………………………………………………….13
5. Conclusion & Future Work………………………………………………...14

References……………………………………………………………………15

**Introduction**

CAD is the major cause of morbidity and mortality. According to World Health Organization, coronary artery disease (CAD) taking 17.9 million lives each year an estimated 31% of all deaths globally.

Coronary Artery Disease often referred to as Atherosclerosis, affects the vessels that supply blood to the heart muscle. It is caused by deposits of calcium, fat, cholesterol, and fibrous tissues, which makes the lumen of arteries narrowed.

In such a case, an earlier diagnosis of CAD along with proper medication will drastically decrease the overall deaths in the country. Also instead of those expensive and time consuming approach a suitable health-care application for automatic CAD diagnosis using machine learning technique will assist a cardiologist in an early diagnosis of the disease.

**Objective and Motivation**

The objective of this project is to develop computational algorithm that can accurately detect coronary artery related diseases and to reduce the amount of expensive pathological tests done in order to diagnose whether patient has a CAD or not.

If we look at our health care today it is more like sick care than preventive medical and technology. We literally wait for the heart attack to occur and offer only post care treatment but by then irreversible damage has already been done.

We must establish a proactive health care technology. That’s why we come decided to work on this project.

**Dataset Analysis**

The dataset used in this project was obtained from the publically UCI repository heart disease dataset named Z- Alizadeh Sani medical dataset contains information about 303 patients as clinical records with 56 feature attributes freely available in the University of California, Irvine machine learning repository.

The dataset consists of 216 records of CAD patients and 87 records of healthy persons with 55 independent feature attributes or predictors and one output or response variable.

The medical dataset is the study based on angiography procedure conducted by Alizadeh Sani to measure the stenosis of each artery. The response variable has two values based on an angiographic disease status, namely (i) value 0 for the absence of CAD and (ii) value 1 for the presence of CAD.

The target class is set to value 0: When the narrowing diameter of an artery is less than 50% then the person is a non-CAD patient (a person is healthy or normal) and otherwise set to value 1: when the arteries have ≥ 50% diameter narrowing, the particular patient record is categorized as CAD patient (a person having CAD). The dataset is highly imbalanced with 71.29% patient records contributing to CAD patients, and the remaining 28.71% records are normal or healthy persons.

The dataset used is based on classification problem with the target variable having values ‘CAD’ and ‘Normal’ with no missing values. Which means it was already clean. During modelling phase, experiments were conducted to select significant features using feature selection methods, and risk prediction models were built using selected features and machine learning algorithms to exhibit the knowledge and classify the risk level of patients with CAD. Subsequently, the performance of prediction models was measured using different metrics in the evaluation phase. Finally the data mining results were reported, and the best performing risk prediction model will get deployed.

**Methodology**

**Dataset**

**Fig 1:** Framework for prediction of CAD

**Final prediction results**

**Compute the posterior Probability of prediction for each model**

**Apply classifier models on the test dataset**

Random Forest

Decision Tree

Naive Bayes

Logistic Regression

Support Vector Machine

K Nearest Neighbour

**Train the classification algorithm**

**10% of Testing Data**

**90% of Training Data**

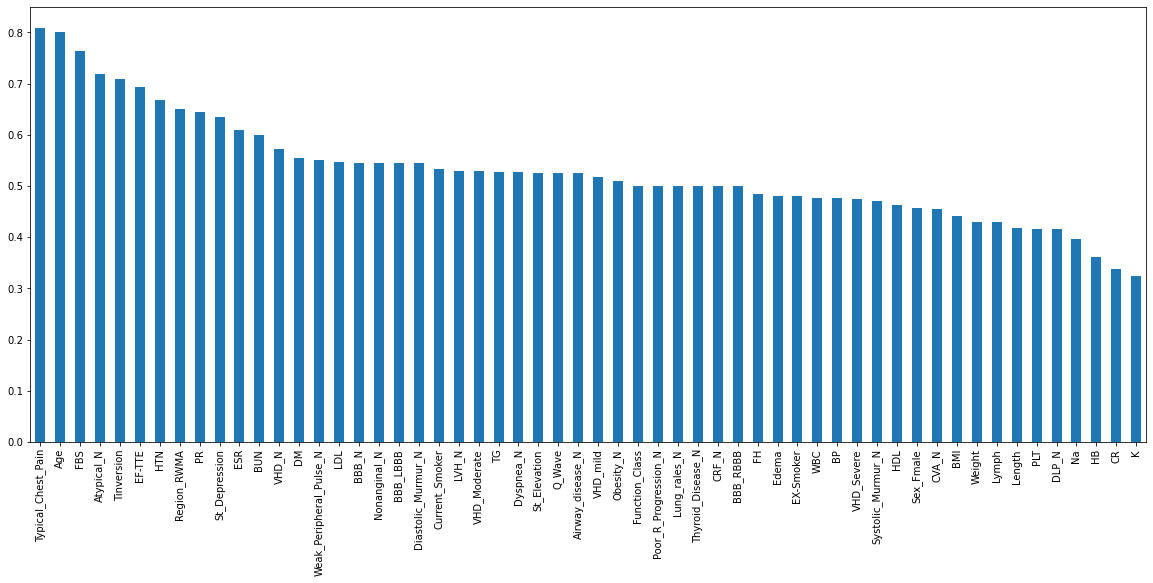
**Apply Train Test and split to split the dataset into two parts**

**Feature Selection**

**Feature Selection**

The feature selection method applied in this project are Remove constant features, Remove quasi-constant features, Remove correlated features, Univariate Selection, Univariate roc auc or mse.

Number of feature in this project was 56 but as we convert the categorical data into dummy or indicator variables it classified into sub features will become 79 features.



**Fig 2:** Feature selection using Univariate roc auc or mse

|  |  |
| --- | --- |
| **Feature Selection Method** | **Number of Features Selected** |
| Remove constant features | 78 |
| Remove quasi-constant features | 73 |
| Remove correlated features | 56 |
| Univariate Selection | -- |
| Univariate roc auc or mse | 30 |

**Table 1:** Selected Features using different Feature Selection Technique

Features got selected after applying all the above feature selection technique:

1. Age
2. DM
3. HTN
4. Current\_Smoker
5. PR
6. Typical\_Chest\_Pain
7. Q\_Wave
8. St\_Elevation
9. St\_Depression
10. Tinversion
11. FBS
12. TG
13. LDL
14. BUN
15. ESR
16. EF-TTE
17. Region\_RWMA
18. Obesity\_N
19. Airway\_disease\_N
20. Weak\_Peripheral\_Pulse\_N
21. Diastolic\_Murmur\_N
22. Dyspnea\_N
23. Atypical\_N
24. Nonanginal\_N
25. LVH\_N
26. BBB\_LBBB
27. BBB\_N
28. VHD\_Moderate
29. VHD\_N
30. VHD\_mild

**Data Partitioning**

The dataset is randomly divided into a training subset and testing subset with the selected features for the development of classifiers. The dataset is split into two parts with 90% of data for training and 10% for testing respectively. In the training process, 10-fold cross validation is utilized.

**Classification Models**

In applying machine learning models, it is generally understood that no single algorithm is superior to the others. There are six widely used supervised machine learning algorithm implemented for classification modelling in this project. These algorithms are:

K-Nearest Neighbour, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest.

Each algorithm is first trained (or fitted) with a fraction of the dataset, known as training set and then tested on a test set.

**Performance Measures**

**Table 2:** Performance metrics of the ML models applied

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **ACC%** | **Precision** | **Recall (sensitivity)** | **F-score** |
| **K-NN** | 70.97 | 0.68 | 1.0 | 0.81 |
| **SVM** | 87.1 | 0.83 | 1.0 | 0.90 |
| **Logistic Regression** | 83.87 | 0.82 | 0.95 | 0.88 |
| **Naive Bayes** | 70.97 | 1.0 | 0.55 | 0.70 |
| **Decision Tree** | **93.55** | **0.90** | **1.0** | **0.95** |
| **Random Forest** | **93.55** | **0.90** | **1.0** | **0.95** |

The following four metrics were used to evaluate the performance of the predicted models and compare them with one another:

1. Accuracy: the proportion of total dataset instances that were correctly predicted out of the total instances

Accuracy = (true positives + true negatives)/total

1. Precision: Precision is the ratio of correctly predicted high risk to the total of high-risk cases

Precision = true positives / (true positives + false positives)

1. Recall (sensitivity): the proportion of the predicted positive dataset instances out of the actual positive instances

Sensitivity = true positives/ (true positives + false negatives)

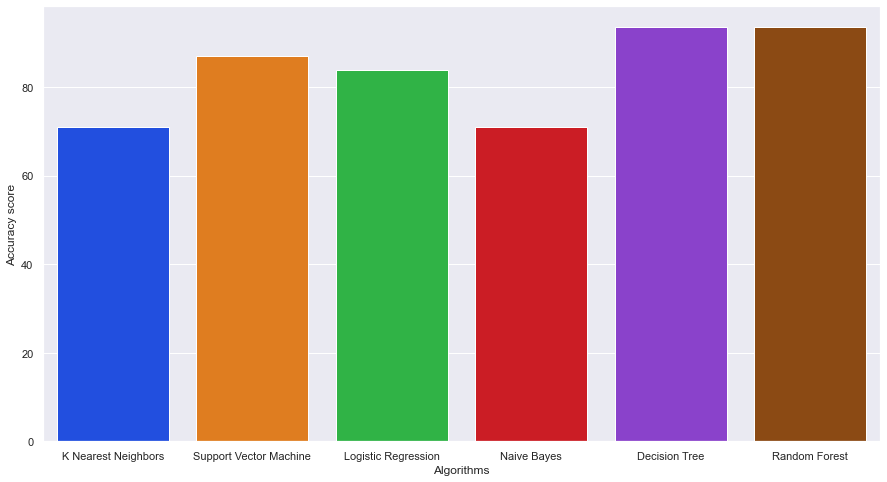
1. F1 score: a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances

Precision = true positives/(true positives + false positives)

The F1 score is estimated as

F 1=2 × (precision × recall)/(precision + recall)

**Result**



**Fig 3:** Classifier performance comparison with accuracy obtained in applying machine learning algorithms to the Z-Alizadeh Sani medical dataset

**Conclusion & Future Work**

* In this project, we demonstrated those ML algorithms were evaluated based on high accuracy and recall to detect the presence of CAD using a publicly available dataset.
* The six ML models performed well, with accuracies found to be greater than 70.9 % and among all of them **Random Forest** and **Decision Tree** found with highest accuracy of **93.55 %.**

* Although CAD is both widely prevalent and may lead to fatal consequences, timely detection of CAD would empower clinicians to treat modifiable risk factors associated with the progression of CAD.
* Using an ML approach provides the ability to predict the presence of CAD with high accuracy and recall, and thus allows practitioners to practice preventive medicine in patients with CAD in a timelier manner. However at such initial stages, it should be noted that ML serves solely as a predictor of CAD rather than a diagnostic tool.

* We hope that as more datasets are available for training and algorithm, we may be able to label ML algorithms as diagnostic steps in CAD management. Because machine learning utilizes the dataset of patients who have already been diagnosed, the predictive ability of an ML algorithm for CAD would improve as more data are supplied to the algorithm.
* We look forward to gain access to larger datasets for further validation and refinement with the eventual goal of providing an open source solution to aid healthcare practitioners in the detection and treatment of CAD.

**References**

**[1]** Z. A. Sani, R. Alizadehsani, M. Roshanzamir, Z-Alizadeh Sani data set, 2020, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani, Accessed January.

**[2]** https://www.who.int/health-topics/cardiovascular-diseases#tab=tab\_1

**[3]** Organizing machine learning projects: project management guidelines. https://www.jeremyjordan.me/ml-projects-guide/

**[4]** Alizadehsani R, Abdar M, Roshanzamir M et al. Machine learning-based coronary artery disease diagnosis: a comprehensive review. Compute. Biol. Med. 111, 103346 (2019).

**[5]** Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12(4), e0174944 (2017).

**[6]** Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine learning outperforms ACC/AHA CVD risk calculator in MESA. J. Am. Heart Assoc. 7(22), e009476 (2018).

**[7]** Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica 160, 3–24 (2007).

**[8]** James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer, New York, 112, 3–7 (2013).

**[9]** Fawcett T. An introduction to ROC analysis. Pattern Recognition Lett. 27(8), 861–874 (2006).