

**NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOSTATISTICS**

**APPLICATION OF DATA MINING ALGORITHMS ON CORONARY
ARTERY DISEASE DATA FOR RULE DISCOVERY AND
EVALUATION**

Ph.D. THESIS

Meliz YUVALI

**Nicosia
June, 2022**

**NEAR EAST UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF BIOSTATISTICS**

**APPLICATION OF DATA MINING ALGORITHMS ON CORONARY
ARTERY DISEASE DATA FOR RULE DISCOVERY AND
EVALUATION**

Ph.D. THESIS

Meliz YUVALI

Supervisor

Assist. Prof. Dr. Özgür TOSUN

Nicosia

June, 2022

Approval

We certify that we have read the thesis submitted by Meliz YUVALI titled “**Application of Data Mining Algorithms on CAD Data for Rule Discovery and Evaluation**” and that in our combined opinion it is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy in Biostatistics.

Examining Committee	Name-Surname	Signature
Head of the Committee:	Prof. Dr. Şanda Çalı	
Committee Member:	Prof. Dr. İlker Etikan	
Committee Member:	Prof. Dr. Selim Yavuz Sanisoğlu	
Committee Member:	Assoc. Prof. Dr. Uğur Bilge	
Supervisor:	Assist. Prof. Dr. Özgür Tosun	

Approved by the Head of the Department

...../...../20...

Prof. Dr. İlker Etikan
Head of Department

Approved by the Institute of Graduate Studies

...../...../20...

Prof. Dr. Kemal Hüsnü Can Başer
Head of the Institute

Declaration

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Meliz Yuvalı

21/06/2022

Acknowledgments

First of all, I would like to thank my supervisor Assist. Prof. Dr. Özgür Tosun for his supervision, support, and sharing his knowledge with me during my thesis work.

I would like to thank the head of department Prof. Dr. İlker Etikan (Near East University), Prof. Dr. Yavuz Sanisoğlu (Yıldırım Beyazıt University), Prof. Dr. Şanda Çalı (Near East University), and Assoc. Prof. Dr. Uğur Bilge (Akdeniz University) for their support.

Meliz Yuvalı

To my dear family...

Abstract

Application of Data Mining Algorithms on Coronary Artery Disease Data for Rule Discovery and Evaluation

Yuvali, Meliz

PhD., Department of Biostatistics

June, 2022, 96 pages

Statistical methods and machine learning (ML) algorithms have been increasingly and efficiently used in medical decision-making for the last few decades. CAD (CAD) is a very common type of CVDs that causes many deaths each year. In this study, two CAD datasets have been obtained from TRNC and Iran. These datasets were used for understanding the classification efficiency of different supervised machine learning algorithms. Z-Alizadeh Sani dataset contained 303 individuals (216 patients, 87 control), while Near East University (NEU) Hospital dataset contained 475 individuals (305patients, 170 control). This research was conducted in 3 classification stages; Each of the two datasets and merged version were analyzed separately with ML algorithms. NEU Hospital dataset was assigned as the training data, while Z-Alizadeh Sani dataset was used as the test data; Z-Alizadeh Sani dataset was assigned as the training data, while NEU hospital dataset used as the test data. Among all ML algorithms, the Random Forest was shown to be successful in its classification performance at each stage. The least successful ML method was k-Nearest Neighbors, as it underperformed at all stages. Logistic regression was found to have successful classification performance.

Keywords: CAD (CAD); machine learning; logistic regression; validation; support vector machine (SVM)

Özet

Kural Keşfi ve Değerlendirmesi için Koroner Arter Hastalığı Verilerine Veri Madenciliği Algoritmalarının Uygulanması

Yuvalı, Meliz

PhD., Biyoistatistik Anabilim Dalı

Haziran, 2022, 96 sayfa

İstatistiksel yöntemler ve makine öğrenimi (ML) algoritmaları, son birkaç yılda tıbbi karar vermede giderek daha fazla ve verimli bir şekilde kullanılmaktadır. Koroner Arter Hastalığı (KAH), her yıl birçok ölüme neden olan çok yaygın bir kardiyovasküler hastalık türüdür. Bu çalışmada, farklı denetimli makine öğrenmesi algoritmalarının sınıflandırma verimliliğini anlamak için KKTC ve İran'dan elde edilen iki KAH veri seti kullanılmıştır. Z-Alizadeh Sani veri seti 303 bireyi (216 hasta, 87 kontrol), Yakın Doğu Üniversitesi (YDÜ) Hastanesi veri seti ise 475 bireyi (305 hasta, 170 kontrol) içeriyor. Bu çalışmada 3 sınıflandırma aşaması gerçekleştirildi; İki veri kümesinin her biri ve bunların birleştirilmiş versiyonu, eğitim-test alt kümelerinin elde edilmesi için uygulanan rastgele örnekleme yöntemiyle makine öğrenme algoritmaları ile ayrı ayrı analiz edilmiştir; YDÜ Hastanesi veri kümesi eğitim verisi olarak atanırken, test verisi olarak Z-Alizadeh Sani veri seti atandı; eğitim verisi olarak Z-Alizadeh Sani veri seti, test verisi olarak YDÜ hastane veri seti kullanıldı. Tüm makine öğrenme algoritmaları arasında Random Forest her aşamada sınıflandırma performansı açısından başarılı olduğu görülmüştür. En az başarılı makine öğrenme yöntemi, tüm aşamalarda düşük performans gösterdiği için kNN dir. Lojistik regresyonun başarılı sınıflandırma performansına sahip olduğu görülmüştür.

Anahtar Kelimeler: koroner arter hastalığı (KAH); makine öğrenme; lojistik regresyon; validasyon; destek vektör makineleri (DKM)

Table of Contents

Approval	i
Declaration	ii
Acknowledgments	iii
Abstract	v
Özet	vi
List of Tables	ix
List of Figures	xi
List of Abbreviations	xii
CHAPTER I	1
Introduction	1
Coronary Artery Disease (CAD).....	1
Signs and symptoms.....	2
Causes.....	3
Statement of the Problem.....	4
Aim and Objectives.....	4
Significance and Justification of the Study.....	4
CHAPTER II	6
Literature Review	6
Statistical and Machine Learning Methods.....	7
Chi-Squared Test.....	7
Mann Whitney U test.....	7
Logistic Function and Logistic Regression.....	8
<i>Odds Ratio</i>	9
<i>Maximum Likelihood</i>	9
<i>ROC Curve</i>	10
Machine Learning (ML).....	11
<i>k-Nearest Neighbors (kNN)</i>	11
<i>Support Vector Machine (SVM)</i>	11
<i>The Random Forest (RF)</i>	12
<i>Artificial Neural Network (ANN)</i>	12

<i>Naïve Bayes (NB)</i>	13
Previous Studies	14
CHAPTER III	18
Materials and Methods	18
Research Design.....	18
Data Collection Tools/Materials	18
<i>Variables</i>	18
Software	20
Analysis Workflow	21
CHAPTER IV	27
Findings and Discussion	27
CHAPTER V	83
Discussion	83
<i>Limitation</i>	85
CHAPTER VI	86
Conclusion and Recommendation	86
REFERENCES	87
APPENDICES	91
Appendix A	91
Ethical Approval Document	91
Appendix B	93
Signed Similarity Report	93
Appendix C	94
Curriculum Vitae	94
PERSONAL INFORMATIONS	94

List of Tables

Table 1. The main characteristics of predictor variables	19
Table 2. Descriptive statistics for quantitative variables from NEU Hospital dataset.(n=475)	27
Table 3. Descriptive statistics for qualitative variables from NEU Hospital dataset (n=475)	29
Table 4. Comparison of quantitative variables between patients and controls in NEU Hospital dataset.(Mann Whitney U test).....	30
Table 5. Comparison of qualitative variable between patients and controls from NEU Hospital dataset.(Chi-Squared test)	32
Table 6. Bivariate Logistic Regression results of each variables in NEU Hospital dataset	34
Table 7. Multivariate Logistic Regression Equations Summary (NEU Hospital dataset)...	36
Table 8. Omnibus tests of Model Coefficients for the Multivariate Logistic Regression (NEU Hospital dataset).....	38
Table 9. Model Summary (Multivariate for the Multivariate Logistic Regression comprising of all Logistic Regression Models, NEU Hospital dataset)	39
Table 10. Hosmer and Lemeshow Test to Assesses the Model Fit (NEU Hospital dataset)	39
Table 11. Classification Table for the Multivariate Logistic Regression Model (NEU Hospital dataset)	40
Table 12. Area Under the Curve for the ROC for the quantitative variables (NEU Hospital dataset)	41
Table 13. Area Under the Curve for the ROC for the Multivariate Logistic Regression (NEU Hospital dataset).....	42
Table 14. Descriptive statistics for quantitative variables from Z-Alizadeh Sani dataset. (n=303).....	44
Table 15. Descriptive statistics for qualitative variables from Z-Alizadeh Sani dataset (n=303).....	45
Table 16. Comparison of quantitative variables between patients and controls in Z-Alizadeh Sani dataset.(Mann Whitney U test)	47
Table 17. Qualitative variable distributions between patients and controls from Z-Alizadeh Sani dataset (Chi-Squared test).....	49

Table 18. Bivariate Logistic Regression results of each variables in Z-Alizadeh Sani dataset.	51
Table 19. Multivariate Logistic Regression Equations Summary (Z-Alizadeh Sani dataset)	55
Table 20. Omnibus tests of Model Coefficients for the Multivariate Logistic Regression (Z-Alizadeh Sani dataset).....	58
Table 21. Model Summary (Multivariate for the Multivariate Logistic Regression comprising of all Logistic Regression Models, Z-Alizadeh Sani dataset).....	58
Table 22. Hosmer and Lemeshow Test to Assesses the Model Fit (Z-Alizadeh Sani dataset)	58
Table 23. Classification Table for the Multivariate Logistic Regression Model(Z-Alizadeh Sani dataset).....	59
Table 24. Area Under the Curve for the ROC for the quantitative variables (Z-Alizadeh Sani dataset).....	60
Table 25. Area Under the Curve for the ROC for the Multivariate Logistic Regression (Z-Alizadeh Sani dataset).....	61
Table 26. Multivariate Logistic Regression Equations Summary (Combined dataset).....	63
Table 27. Omnibus tests of Model Coefficients for the Multivariate Logistic Regression (Combined dataset).....	66
Table 28. Model Brief (The Multivariate Logistic Regression comprising of all Logistic Regression Models, Combined dataset).....	66
Table 29. Hosmer and Lemeshow Test to Assesses the Model Fit (Combined dataset).....	67
Table 30. Classification Table for the Multivariate Logistic Regression Model (Combined dataset).....	67
Table 31. Area Under the Curve for the ROC for the Multivariate Logistic Regression (Combined dataset).....	68
Table 32. Machine Learning Random Sampling Results for Step-1.....	70
Table 33. Machine Learning Classification Results for Step-2.....	77
Table 34. Machine Learning Classification Results for Step-3.....	80
Table 35. Classification success of the research.....	84

List of Figures

Figure 1. The Image of Coronary Artery Disease.....	2
Figure 2. <i>Coronary Arteries clogged up.</i>	3
Figure 3. <i>Classification workflow.</i>	22
Figure 4. <i>Step-1 Classification workflow for NEU Hospital Dataset</i>	23
Figure 5. <i>Step-1 Classification workflow for Z-Alizadeh Sani Dataset</i>	23
Figure 6. Step-1 Classification workflow for Combined (Neu Hospital & Z-Alizadeh Sani Dataset)	24
Figure 7. <i>Step-2 Classification workflow (NEU Hospital as the Training Dataset)</i>	25
Figure 8. <i>Step-3 Classification workflow (Z-Alizadeh Sani as the Training Dataset)</i> ...	26
Figure 9. <i>ROC Curve for the Quantitative Variables (NEU Hospital dataset)</i>	42
Figure 10. <i>ROC curve for the Final Multivariate Logistic Regression Model (NEU Hospital dataset)</i>	43
Figure 11. <i>ROC Curve for the Quantitative Variables (Z-Alizadeh Sani dataset)</i>	61
Figure 12. <i>ROC curve for the Final Multivariate Logistic Regression Model (Z-Alizadeh Sani dataset)</i>	62
Figure 13. <i>ROC curve for the Final Multivariate Logistic Regression Model (combined dataset)</i>	69
Figure 14 (a). <i>ROC for Table 32. (NEU Hospital dataset)</i>	71
Figure 15 (a). <i>ROC for Table 32. (NEU Hospital dataset)</i>	72
Figure 16 (b). <i>ROC for Table 32. (Z-Alizadeh Sani dataset)</i>	73
Figure 17 (b). <i>ROC for Table 32. (Z-Alizadeh Sani dataset)</i>	74
Figure 18 (c). <i>ROC for Table 32. (Combined Dataset)</i>	75
Figure 19 (c). <i>ROC for Table 32. (Combined Dataset)</i>	76
Figure 20 (a). <i>ROC graph of ML algorithm (SVM, kNN, RF) results of Step-2.</i>	78
Figure 21 (b). <i>ROC graph of ML algorithm (Logistic Regression, Naïve Bayes, ANN) results of Step-2.</i>	79
Figure 22 (a). <i>ROC graph of ML algorithm (kNN, RF, SVM) results of Step-3.</i>	81
Figure 23 (b). <i>ROC graph of ML algorithm (Logistic Regression, ANN, Naïve Bayes) results of Step-3.</i>	82

List of Abbreviations

CAD:	Coronary Artery Disease
ML:	Machine Learning
OR:	Odds Ratio
CI:	Confidence Interval
ROC:	Receiver Operating Characteristic
kNN:	k-Nearest Neighbors
SVM:	Support Vector Machine
RF:	Random Forest
ANN:	Artificial Neural Network
NB:	NaïveBayes
LR:	Logistic Regression
WHO:	World Health Organization
CVD:	Cardiovascular Disease
ECG:	Electrocardiography
DM:	Diabetes Mellitus
HT:	Hypertension
FH:	Family History
BP:	Blood Pressure
PR:	Pulse Rate
LVH:	Left Ventricular Hypertrophy
FBS:	Fasting Blood Sugar
CR:	Creatinine
TG:	Triglyceride
LDL:	Low-Density Lipoprotein
HDL:	High-Density Lipoprotein
BUN:	Blood Urea Nitrogen
Hb:	Hemoglobin

K: Potassium

Na: Sodium

WBC: White Blood Cell

PLT: Platelet

EF-TTE: Ejection-Fraction

Region-RWMA: Regional Wall Motion Abnormality

VHD: Valvular Heart Disease

CHAPTER I

Introduction

Human life is affected by fatal illnesses. These ailments are caused by genetic and environmental factors. Blood testing was crucial for diagnosis as blood tests, continual controls, treatments, and researchers hope to reduce the number of deaths caused by illnesses. In order to avoid this and decrease the hazards that may emerge, regular checkups and blood tests are performed.

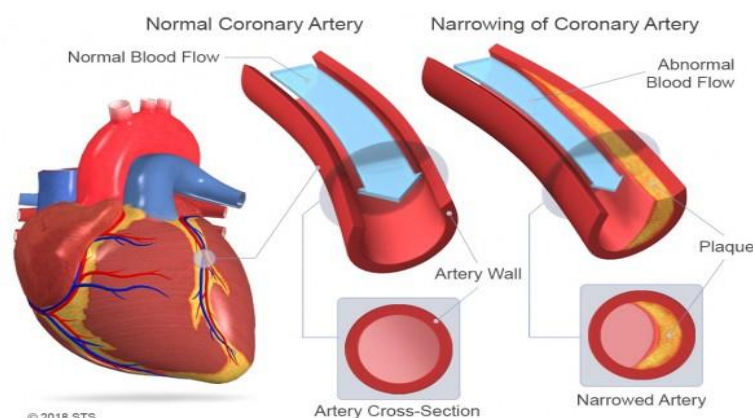
Much research for the prevention of fatal diseases is being held to contribute to the medical community by determining the most essential aspects for diagnosis of these diseases using artificial intelligence, physical examination records and blood tests. Coronary Artery Disease (CAD) forms a disabling condition that can strike women and men of all ages. In coronary artery and other heart disorders, it has been discovered that genetic predisposition and extrinsic influences are particularly efficient.

Coronary Artery Disease (CAD)

CAD is one of the leading causes of death. CAD is the blockage or narrowing of the vessels feeding the heart. They are located in the subepicardial connective tissue and on exterior part of the heart. CAD according to the degree of the disease; is treated with medication or surgery (Chen et al., 2020).

Figure 1.

The Image of Coronary Artery Disease



(The Society of Thoracic Surgeons, 2018)

The latest statistics from the American Heart Association CAD held responsible for 13% of deaths in the United States in 2018 (Akella and Akella, 2021). Angiography presented as the optimal diagnostic implement for CAD. Although Angiography is the optimal diagnostic method, it has been reported to be accountable for 0.1% to 0.14% mortality, 0.06% to 0.07% myocardial infarction, 0.07% to 0.14% cerebral palsy, 0.23% reaction to contrast agent, and 2% local vascular problems (Alizadehsani et al., 2018). It has been estimated that by 2025 the cause of 35 to 60% of worldwide deaths will be caused by CVDs (Ayatollahi et al., 2019).

Signs and symptoms

The most significant indication of CAD is chest pain and shortness of breath.

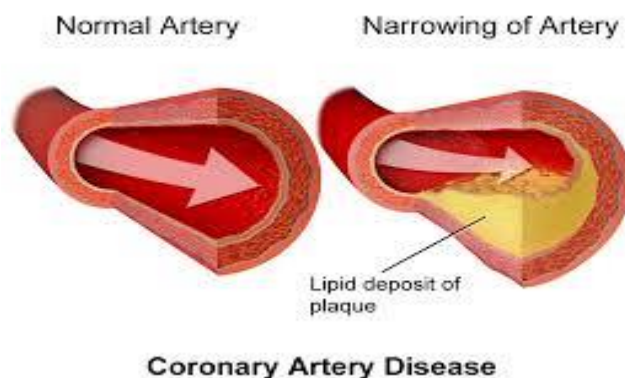
CAD usually seen in adults and people with heart disease.

Common complaints of patients suffering from simple CAD comprise:

- Chest Pain
- Unable to breathe
- Heart Attack
- Edema
- Weak Peripheral Pulse
- Weakness

Figure 2.

Coronary Arteries clogged up.



(Blaus, 2014)

In clinical trials, statistical analysis aids in demonstrating the effects of various statistical modelling tools. The development of multiple variable Logistic Regression (LR) models illustrates many risk variables that influence CAD and their effectiveness in diagnosing patients. Artificial intelligence algorithms multivariate LR assist in identifying sick persons and determining which variables are essential in defining the condition.

Causes

The emergence of the CAD is influenced by several factors. According to the scientific evidence, those with an irregular lifestyle and a family history of heart disease are more likely to develop CAD.

This disease may be caused by various factors, including:

- Family History
- Age
- High blood pressure
- High cholesterol
- Smoking
- Diabetes or insulin resistance

Statement of the Problem

At the beginning of the twentieth century, CVDs accounted for 10% of all deaths globally; this figure had risen to 25%. In order to forecast CAD and other disorders, researchers used clustering, classification, regression, and artificial intelligence algorithms. This research aspires to compare different methods to seek the most successful ML algorithm to contribute to science and utilize artificial intelligence algorithms with statistical methodologies (Ayatollahi et al.,2019).

Aim and Objectives

The objective is to explain the presence of CAD by discovering rules and verifying them by applying classical statistical methods as ML algorithms. Furthermore, the aim focuses on the methodologies to utilize the data as training and testing datasets, discover rules, and evaluate them to assess the performances of different statistical and data mining techniques. Validation techniques were applied to understand the validity of discovered rules across different data sets. Since data mining is still a developing technology, it is not possible to use it widely, but in the future, artificial intelligence has the capacity to be very successful in detecting diseases. From a statistical point of view, the development of data mining is of great importance in the early diagnosis and treatment of the disease in the medical field.

Significance and Justification of the Study

Many studies have performed the statistical analysis and comprehensive evaluation of various health problems and diseases. Tougui et. al., (2020) used heart disease data which contains 13 features with 303 cases. In this study, 6 ML algorithms and data mining software were used. As a result of the study, they reported that the LR had a specificity of 88.12%. Kutrani et. al. (2019) used Benghazi Heart Center dataset with a sample size of 1,770. Among the ML algorithms used in this study, SVM and kNN showed 86% and 85% success in correct classification. Akella and Akella's (2021) article shows that CAD is a common disease worldwide, and its development is affected by various modifiable risk factors. The research estimated whether the patients included in the "Cleveland Dataset" had CAD by applying six different ML algorithms: LR, decision tree, SVM, kNN, and ANN.

In conclusion, the study indicated that all six machine learning (ML) algorithms have an accuracy of more than 80% and the "net neutral" algorithm has an accuracy of over 93%. Although significant work has been done, new methods are still being tested to understand the accuracy of different ML approaches for the classification of CAD patients.

CHAPTER II

Literature Review

Heart disease is a frequent and serious health concern that affects people all over the world. In developing countries, the incidence and mortality of cardiovascular illnesses have risen over the last few decades.

In CAD, fat layers accumulate in the coronary arteries; this damages the veins and induces narrowing of the blood flow, resulting in hypoxia of the heart muscle. In some cases, this circumstance can cause oxygen saturation in the heart and can be deadly (Shaima et al., 2016).

It's a serious health issue with a high mortality rate, especially among those in their middle and senior years. It builds up on the surface of the arteries that deliver blood to the heart. A blockage in the coronary artery due to a blood clot is the usual cause of a heart attack. According to the related source, CAD is one of the death leading diseases worldwide, with around 17.7 million deaths due to CAD in 2015 (Shaik Mohammad Naushad, et. al., 2018 & Amin et al., 2019).

Machine learning (ML) approaches have been used more in the field of medicine in recent years, and they have resulted in numerous advancements on various levels. Multiple ML algorithms have been proposed to detect and comprehend the progression of diseases (Cuvitoglu & Isik, 2018). Data mining is extremely beneficial for analyzing enormous data sets to reveal previously unknown and hidden patterns, correlations, and information that are difficult to detect using traditional statistical methods (Cuvitoglu & Isik, 2018).

To summarize, data mining is a significant advancement in knowledge discovery. Data mining in the healthcare industry is a rapidly growing field that has the potential to improve medical data interpretation.

Statistical and Machine Learning Methods

There have been numerous research using multivariate regression models and machine learning algorithms to classify CAD.

Chi-Squared Test

Chi-Squared test measures the parallelisation of the model to the examined data. It has two applications: to see if there is an association between the row and column variables to see if the two ratios are equal. The Chi-Squared test can be used to compare the rates of categorical outcomes against distinctive independent groups.

When comparison groups are independent and do not correlate, the Chi-Squared test can be used to determine independence between two variables to test an approach. The Chi-Squared test is a non-parametric test that expresses categorical data (Kim, 2017).

The formula for calculating a Chi-Squared statistic is:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

Where;

o_i = Actual frequency

e_i = Expected frequency

Mann Whitney U test

Mann-Whitney U test is used to determine whether there is a statistically significant in the dependent variable between two independent groups. It evaluates the allocation of the dependent variable of the two groups and hence from the same population. Mann-Whitney U test draws different outcomes about the data based on the distribution assumptions. These various conclusions are dependent on the form of data distributions (Zar,1999 & Zar, 2010).

Mann Whitney U test statistic is defined as:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} + R_1$$

Where;

n_1 And n_2 = First and second sample sizes respectively

R_1 = Rank sum

Logistic Function and Logistic Regression

LR analysis is a method that helps in classification and assignment. It is a regression method that obtains the expected values of the response variable as probabilities according to the explanatory variables. Discriminant analysis is a method that enables data to be classified and assigned to certain classes according to certain probabilities. It is possible to determine the effects of the variables in the data set on the classification. LR analysis is a method that provides the opportunity to make classification according to probability rules by calculating the estimated values of the dependent variable as possibilities.

For multivariate analysis, the selection of predictor variables to build a model depends on each one's statistical significance in the overall model (Zar,1999).

Binary Logistic Regression

Binary LR is an extrapolative model that fits in situations where the dependent variable is dichotomous or binary, for instance when the researcher is interested in whether a patient has CAD or not.

Most categories' codes are "0" and "1" since this allows for a direct interpretation. The case with the code "1" and the other category known as a "non-case", sometimes known as "0" is of particular relevance to the category (Zar,1999).

Odds Ratio

The Odds Ratio (OR) is the probability that an event (p) is divided by the probability that the event will not happen (1-p). The central mathematics concept that brings LR is the logit, which is the natural logarithm of an Odds Ratio (Peng et al., 2002). Usually, LR analysis is very convenient for portraying and testing hypotheses between a qualitative outcome variable and one or more qualitative or quantitative predictor variables (Zar, 1999).

$$Odds = \frac{p_i}{1-p_i}$$

The logit function calculates the natural logarithm of an outcome's chances.

That is,

$$Logit = \ln \left(\frac{p}{1-p} \right)$$

The logarithm of odds is generally denoted by Logit in this logistic model, and it may be expressed as,

$$\log \left(\frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \dots + \beta_n \chi_n$$

In general, 1 and 0 which exemplify accomplishment and setback. LR uses a logit function to relate the probability of success and predictors and applies the maximum likelihood estimation method to estimate parameters (Usman 2014).

Maximum Likelihood

Maximum Likelihood estimation is the most popular general-purpose method of estimating a distribution from a finite sample (G.Lebanon, 2011). It employs the likelihood method to obtain the actual values of the parameters that are most likely to be the observed in the data. In practical concern, the performance of maximum likelihood estimators is well for the large database. It is one of the most flexible methods to fit the parametric statistical models in data (Ramachandran, 2009).

ROC Curve

It is a performance measure that identifies how well patients and healthy people differ from each other as a result of a diagnostic test.

ROC curve analysis examines continuous variables. The closer result is to 1, the more reliable the analysis curve is. The chart obtains sensitivity and specificity values, so it enables to understand that analysis is neither reliable or unreliable (Ramachandran, 2009).

Features:

It can be used to evaluate the performance of a diagnostic test.

It can be used to compare the performances of multiple diagnostic tests.

It can be used to determine the best cut-off point for diagnostic tests.

Hosmer Lemeshow's Test

Hosmer-Lemeshow (HL) test is a measure of accuracy analysis for LR, particularly risk prediction models. A goodness of fit test determines whether your data fits the model well. The HL test determines if observed event rates in subgroups match the projected event rates (Ramachandran, 2009).

Machine Learning (ML)

Data mining is critical for extracting valuable information from massive databases and generating outcomes for disease diagnosis and treatment (Kodati & Vivekanandam, 2018). Machine learning (ML) typically refers to changes in systems that conduct artificial intelligence-related tasks (AI). Detection, diagnosis, management, robot control, and prediction are examples of such activities. These could be either improvements to existing systems or the application of the new system from the ground up (Nilsson, 2005).

k-Nearest Neighbors (kNN)

kNN is the step towards addressing issues based on answers for similar previously solved problems. The parameter collection's required in every instance-based learning system:

- A distance function that determines how similar problems or data entries are.
- Several neighbors are taken into account when dealing with the new problem.
- A weighting function allows more precise quantification of discovered neighbors to improve prediction and learning quality.
- A method of evaluating outlines function how to use the discovered neighbors to solve a specific problem. Lazy learning methods include instance-based learning methods, which do not do any computation on the data before a query is presented to the system. These methods contrast with eager learning methods like Decision Trees, which attempt to shape data before processing inquiries (Kubat, 2021 & Sammut, 2010).

Support Vector Machine (SVM)

Support vector machines (SVMs) are a type of supervised learning method that is used for classification and regression. It is a prediction tool that employs machine learning theory to maximize forecast accuracy while avoiding over-fitting the data (Dwivedi, 2016). The link between dependent and independent variables is frequently explained using this approach (Ashraf et al., 2020).

The Random Forest (RF)

Random forests are a group of techniques that involve creating an ensemble (or forest) of decision trees using a randomized version of the tree induction procedure. Decision trees are excellent candidates for ensemble approaches because they typically have low bias and high variance, allowing them to benefit from the averaging process. RF's approach differs from one another in the manner. They add random perturbations into the induction phase (Breiman, 2001).

Artificial Neural Network (ANN)

The capacity of ANN to train quickly on sparse data sets is well-known. The ANN algorithm divides data into a set of output categories. ANNs are three-layer networks in which the input layer receives the training patterns, and the output layer has one neuron for each conceivable category. The Euclidean distance between data points finds neighbors between data points in this method. It's employed in studies with collected neighbors to solve classification and regression difficulties (Zeebaree, 2022). The ability to forecast patient success, like many other commercial data mining tools, would make decision-making easier. Various data mining techniques can help to increase prediction abilities. The data that gets examined by other models are fed into neural network models. The standard data mining procedure is to test all possible models while evaluating which one works best over time for a type of data. However, there are particular types of data where neural network models, such as regression or decision trees, frequently outperform the alternatives. ANN performs better complex interactions in the data, such as high degrees of nonlinearity. ANN can cope with both continuous and categorical data input, as they are versatile models that can be used in various data mining applications. The same can apply to regression models and decision trees, which help with the data mining modelling process (Kubat, 2021).

Naïve Bayes (NB)

Bayesian networks are made up of nodes, and direct interconnections represent dependence. They are called probabilistic directed acyclic graphical models. Each node reflects a characteristic relevant to the task, such as pollution levels in cities that calculates the likelihood of contracting lung cancer. The most basic Bayesian network is Naïve Bayes, which represents no relationships between attributes. It is usually never the case in real-world data mining activities. Hence this method tends to produce less-than-ideal outcomes compared to methods that are more sophisticated.

Normal Bayesian networks employ a known approach to predict the correlations between attributes and class labels and then use that information to compute the probabilities of various future event outcomes. It uses Bayes' theorem to learn about the condition of attributes and their relationships (Sammut, 2010).

Previous Studies

Dahal and Gautam (2020) state that the current World Health Organization (WHO) data indicates that cardiovascular illnesses account for 31% of all fatalities worldwide, with 17.9 million people passing away each year. The death rate is predicted to rise dramatically in the next few years. According to their study, a heart attack is an initial symptom for some people. In 2017, 365,914 persons died in the United States of America because of CAD (Dahal and Gautam, 2020). They divided one dataset into two to they divided a single dataset into two and did a study on classification with training and testing methods; LR, RF, kNN, SVM, and Classification tree algorithms were used. The results indicate that the SVM model can predict the presence of CAD more effectively and accurately than other models with an accuracy of 0.8947, sensitivity of 0.9434, specificity of 0.7826, and AUC of 0.8868. The sensitivity rate of both LR and SVM is 0.9434, whilst AUC of LR is more successful than SVM with a performance rate of 90.32%.

According to Kolukisa et al. (2020), the World Health Organization (WHO), the global prevalence of CVD is increasing rapidly, with 30 million deaths expected by 2030. The process of uncovering legitimate, unique, potentially helpful, and eventually intelligible data patterns are knowledge discovery in databases (KDDs). Two different datasets were employed in their study towards the classification phase. K-fold cross-validation method was one of the training-test methods used for the data set that needed to be divided into training and test groups. In the end, LR came out as the most successful classification method in both datasets, with a success rate above 90% (Kolukisa et al., 2020).

In 2021, researchers hoped to meet the requirement to extract usable knowledge from clinical data, focusing on developing a Data Mining solution that can forecast the presence or absence of cardiovascular illnesses (Martins et al., 2021). Their goal was to emphasize the importance of detecting the danger of developing CAD early to avert deaths. The best model was the Optimized DT. The AUC of this performance is 78.8% (Martins et al., 2021).

The Malaysian database research carried out by Md Idris et al. (2020), which was registered in 2006, developed a typical Data Mining technique to assure the validity of experiment results. The methodology comprises six cyclic phases, with multiple iterations used to fine-tune study goals. Moreover, the purpose of the research set whilst finding key features and ML algorithms for the classification improvement of models to forecast CAD risk levels (Md Idris et al., 2020). In terms of AUC scores, all the top models have achieved more than 90%, whilst implicitly kNN is the best performing model with embedded DT features.

It has been reported every year, around 340,000 individuals in Turkey die as a result of CAD. Physicians' intuition and experience are frequently used in clinical decision-making (Nazlı, Yasemin, & Altural, 2020). They aimed to compare different Machine Learning techniques to find the most successful among them. Nazlı, Yasemin, and Altural (2020) used five different ML techniques and applied them to a single dataset. As a result of precision, RF was 100% successful whilst kNN showed the worst accuracy value (81.48%) among the others.

Muhammad et al. (2021) aimed to develop ML algorithms that are used in CAD classification. The dataset was used to create predictive models using machine learning algorithms such as SVM, kNN, random tree, NB, gradient boosting, and LR. The models were evaluated utilizing validity, clarity, responsiveness, and receiver operating curve (ROC) performance evaluation techniques (Muhammad et al., 2021). The sensitivity of the SVM-based ML model ranks high at 87.4 %, while the RF-based ML model emerged victorious with 92.20 %.

Besides, L. J. Muhammad (2019) demonstrates that Murtala Muhammad General Hospital and Abdullahi Wase General Hospitals in Kano State, Nigeria, provided the data utilized in the study to determine the quality of data mining. In 2017, the Ministry of Health in Kano, Nigeria, approved the data collection. Between 2003 and 2017, a total of 506 diagnostic cases of CAD were recorded in both hospitals. The algorithms' performance is assessed using the Weka machine learning program. Random Tree (87.35%) and Naïve Bayes (83.40%) made the lowest accuracy classifications.

As healthcare information systems hold a massive number of clinical data, information gathering, also known as data mining is very prevalent. The model was developed with patient datasets provided at the Mostar hospital's cardiovascular unit from 2011 to 2017. A total number of 507 patients with CVD were included in the study, with 123 dying and 384 surviving after 12 months (Imamovic, Babovic & Bijedic, 2020).

The CART algorithm forms a binary tree by branching syllables at each node according to the function specified for each input attribute based on the available input and output attributes. It is the most commonly used method for building decision trees, followed by Neural Networks and LR (Imamovic, Babovic & Bijedic, 2020).

According to the F- Measure result, Neural Networks achieved 83.12% success. F-Measure is controlled by Neural Network, indicating that precision and sensitivity are sufficiently high. (Imamovic, Babovic & Bijedic, 2020).

Presently, categorization is a chronic problem which influences various applications. The study begins with the data collection for categorization (Jinjri, Keikhosrokiani & Abdullah, 2021). The datasets were then separated into training and test sets after being pre-assessed. In 2021, around 77,000 clinical trial patient data records were collected by hospitals for cardiovascular disorders and included in the dataset (Jinjri, Keikhosrokiani & Abdullah, 2021). The dataset has three input functionalities: factual (practical information), analytical (medical research results), and subjective (previous anecdotes). The aim is to explore various ML algorithms and determine the most productive for CVD classification utilizing patient records. As a result of the study, the SVM emerges as the best-performing technique which can forecast the likelihood of CVD with much more accuracy (72.66%) for early diagnosis (Jinjri, Keikhosrokiani & Abdullah, 2021).

According to research conducted by Cuvitoglu & Isik (2018), multiple machine-learning algorithms are applied. Such as; NB, RF, SVM, ANN, and kNN . The study aims to reveal how successful artificial intelligence can be in classification with the application of different methods. The use of a Cross-Validation (CV) scheme has substantial impact on testing of a machine learning approach, with an accuracy level

of more than 80%. As a result of their study, The ANN outcomes on AUC results were quite promising, with a success rate of 93% (Cuvitoglu & Isik, 2018).

Again, Dwivedi (2016) used six machine learning approaches to CAD data: ANN, SVM, LR, kNN, Classification tree, and NB. Receiver Operating Characteristic (ROC) curves and turning plots were used to double-check the results of the approaches. As a result, kNN gave the highest negative precision value rate for misclassification and F1(83%) measurements (Dwivedi, 2016).

Moreover, between 2014 and 2017, the Department of Advanced Biomedical Sciences at the University Hospital Federico II of Naples assessed 10,265 patients with suspected or known CAD for myocardial perfusion deficit. Clinicians gathered data on traditional cardiovascular risk variables such as age, gender, blood pressure, smoking history, serum cholesterol, family history of CAD, resting ECG features, diabetes and associated consequences, and ECG stress testing as part of their initial check-up. In R programming, the MASS package was used (Ricciardi et al., 2020). The data was split into two halves. A training set was used to validate the data, and the outcomes were collected on a test set. The accuracy of the classification was 84.5%, only utilizing LDA (Ricciardi et al., 2020).

Tasnim & Habiba (2021) report that they utilized the Cleveland, statlog Cleveland and Hungarian datasets from the UCI machine learning repository. This dataset has 303 samples with 14 attributes. NB, RF, kNN, SVM, LR, Xgboost, ANN, and Decision Tree were used to analyse observe the raw data (Tasnim & Habiba, 2021). The biomarker values of 104 people are included in the data collection of Saharan (2021). In addition to the ultimately aimed characteristic that allocates individuals to the CAD (39 individuals) or Control (65 individuals) groups, 35 cytokine biomarkers were tested (Tasnim & Habiba, 2021). The model's feature space includes 35 cytokine biomarkers to express resemblance and, lastly, to classify CAD or Control. The ROSE Package from the R programming language was used. The final balanced data, which consists of 52% CAD and 48% Control, is suitable for kNN and RF implementation. This research has attained the maximum classification accuracy of 92.85% by employing an RF classifier and Principle Component Analysis (Tasnim & Habiba, 2021).

CHAPTER III

Materials and Methods

In this section, the research model, research study group, the collection and analysis of data will be reported.

Research Design

The current dataset used in this study is two independent CAD data that were gathered from NEU Hospital Department of Cardiology and UCI Z-Alizadeh Sani. A total of 778 patient data were collected. Firstly, descriptive statistical analysis was performed on both data sets, and then LR analysis was applied to determine the statistically significant variables. Secondly, to find the classification success of the ML method, all variables were included, and the results of the algorithm were evaluated for each data set separately and combined.

Data Collection Tools/Materials

This research used two independent CAD datasets. The first dataset is collected from NEU Hospital. It was obtained from the computer information system of the cardiology department with the ethics committee permission. The second dataset is obtained from an open-source titled; Z-Alizadeh Sani dataset of UCI ("UCI Machine Learning Repository: Z-Alizadeh Sani Data Set", 2020). Near East University Hospital dataset (protocol code NEU/2019/74/931 and date 21 November 2019) consists of 475 patients (305 CAD patients, 170 control), and Z-Alizadeh Sani dataset consists of 303 patients (216 CAD patients, 87 control). Whole computation and analysis were performed on a laptop with Intel(R) Core (TM) i5-7200U CPU@2.50 GHz, installed RAM 4.00 GB, Windows 10 and a 64-bit operating system.

Variables

The variables those were common in both datasets were filtered and used for the current study. These datasets have 30 variables (13 qualitative, 17 quantitative), one dependent, and twenty-nine independent variables. The dependent variable is binary and signifies CAD.

The dependent variable has two categories:

- CAD: The patient has Coronary Artery Disease
- Normal: The patient has no CAD

Independent variables include; Age, Gender, DM, HT, Smoking Status, FH, BP, PR, Edema, Systolic Murmur, Chest Pain, Dyspnea, LVH, FBS, CR, TG, LDL, HDL, BUN, Hb, K, Na, WBC, Lymph, Neut, PLT, EF-TTE, Region RWMA, VHD (Table 1).

Table 1.

The main characteristics of predictor variables

Variables	Explanation	Variable Type
Age	Patient's Age	Quantitative
Gender	Patient's Gender	Qualitative
DM	Diabetes Mellitus	Qualitative
HT	Hypertension	Qualitative
Smoking Status	Active Smoker	Qualitative
FH	Family History	Qualitative
BP	Blood Pressure	Quantitative
PR	Pulse Rate	Quantitative
Edema	Fluid trapped in patient's body	Qualitative
Systolic Murmur	Heart murmurs heard during systole	Qualitative
Chest Pain	The presence of substernal chest pain	Qualitative
Dyspnea	Breathing problem	Qualitative
LVH	Left Ventricular Hypertrophy	Qualitative
FBS	Fasting Blood Sugar	Quantitative

Table 1. (Continued)

Variables	Explanation	Variable Type
CR	Creatinine	Quantitative
TG	Triglyceride	Quantitative
LDL	Low-Density Lipoprotein	Quantitative
HDL	High-Density Lipoprotein	Quantitative
BUN	Blood Urea Nitrogen	Quantitative
Hb	Hemoglobin	Quantitative
K	Potassium	Quantitative
Na	Sodium	Quantitative
WBC	White Blood Cell	Quantitative
Lymph	Lymphocyte	Quantitative
Neut	Neutrophil	Quantitative
PLT	Platelet	Quantitative
EF-TTE	Ejection-Fraction	Quantitative
Region-RWMA	Regional Wall Motion Abnormality	Qualitative
VHD	Breathing problem	Qualitative

Software

IBM SPSS software (Demo Version 21.0 for Windows) was used for the statistical analysis. For hypothesis testing of datasets, descriptive, Simple LR, Multiple and ROC were used in IBM SPSS program. ML algorithms were used in Orange 3-3.29.3 program for data mining and cross validation (Bioinformatics Laboratory, 2022).

Analysis Workflow

The research involves univariate, bivariate, multivariate statistical methods and machine learning algorithms.

Univariate, bivariate, and multivariate statistical analyzes were applied to each data set separately with SPSS. Then, multivariate LR analysis was performed on the combined data set, and the ROC graph was given.

In the second part, ML algorithms are applied. The algorithms were applied to each data set separately. Afterwards, there were applied to the combined data. The classification of algorithms were compared in the final stage. Both groups were assigned as test and train data for two separate runs. All variables were included in the Orange program to determine how accurate the ML approach produced the classification and to learn the accuracy of this classification. It has been utilized as a classification approach, particularly in LR, a machine learning methodology. This study aims to see which algorithm is more successful in classifying CAD patients. Figure 3 depicts a schematic representation of the work completed.

The Orange software provides the equated classification performance or the classification result of the target class. Algorithm performances are obtained from and evaluated concerning the design shown in Figure 3.

The metrics obtained for the evaluation of classification performances of each ML algorithm are; AUC, Accuracy Classification Score (CA), Weighting depending on the average parameter (F1), Precision, and Recall. AUC results are shown with ROC curves.

$$Accuracy (CA) = (Tp + TN) / (TP + TN + FP + FN)$$

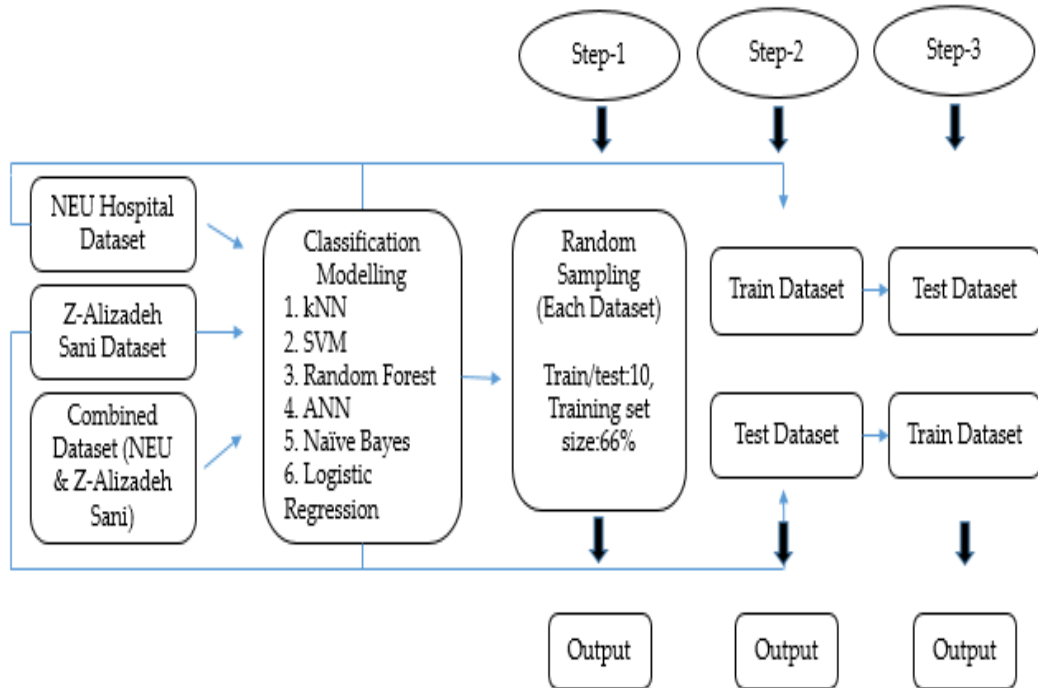
$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Figure 3.

Classification workflow.



In step 1, ML classification techniques were applied with sampling settings (train/test 10, training set size 66%) made in each dataset. After separately applying the ML algorithms to each dataset, two datasets were merged, and the same algorithms were applied to the combined dataset (Fig 4.5&6).

In step 2, NEU Hospital dataset was determined as the training dataset and Z-Alizadeh Sani as the test dataset (Fig.7).

In step 3, NEU Hospital data was assigned as tests as Z-Alizadeh Sani data was assigned as a training dataset, and ML algorithms were applied. The purpose is to observe the performance of ML algorithms through trained independent datasets (Fig.8).

Figure 4.

Step-1 Classification workflow for NEU Hospital Dataset

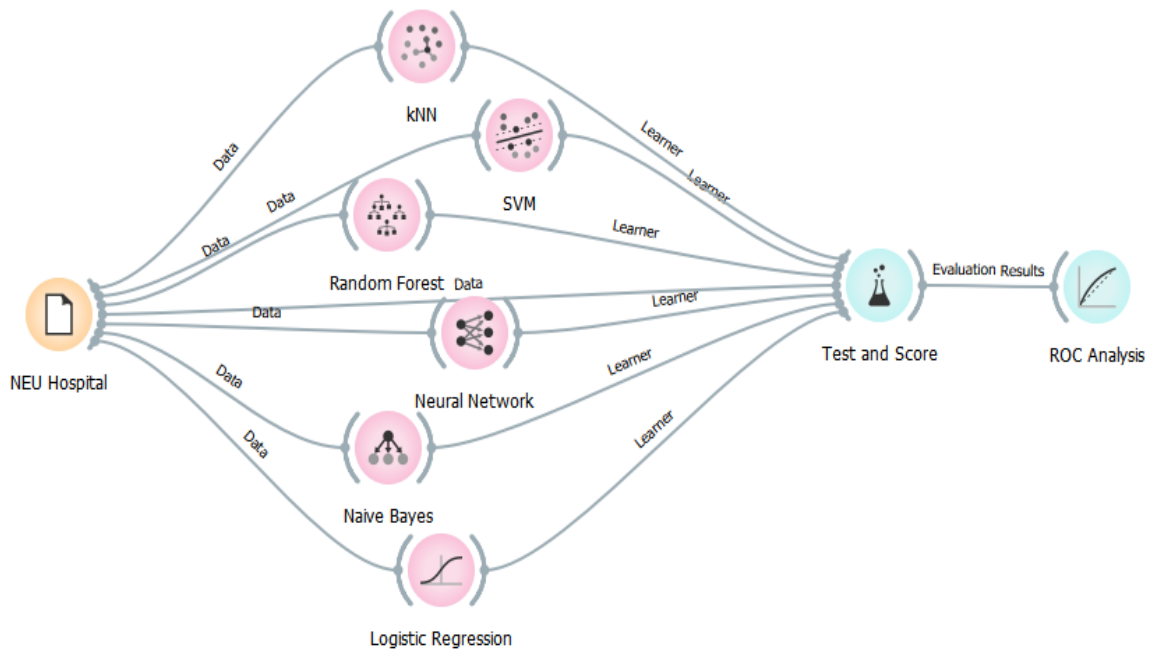


Figure 5.

Step-1 Classification workflow for Z-Alizadeh Sani Dataset

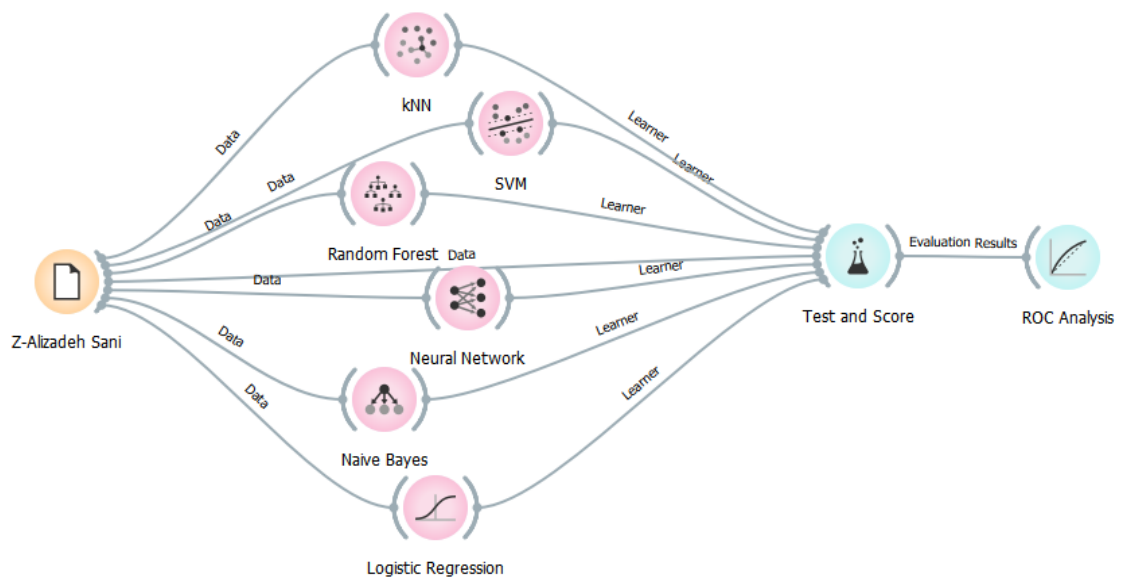


Figure 6.

Step-1 Classification workflow for Combined (Neu Hospital & Z-Alizadeh Sani Dataset)

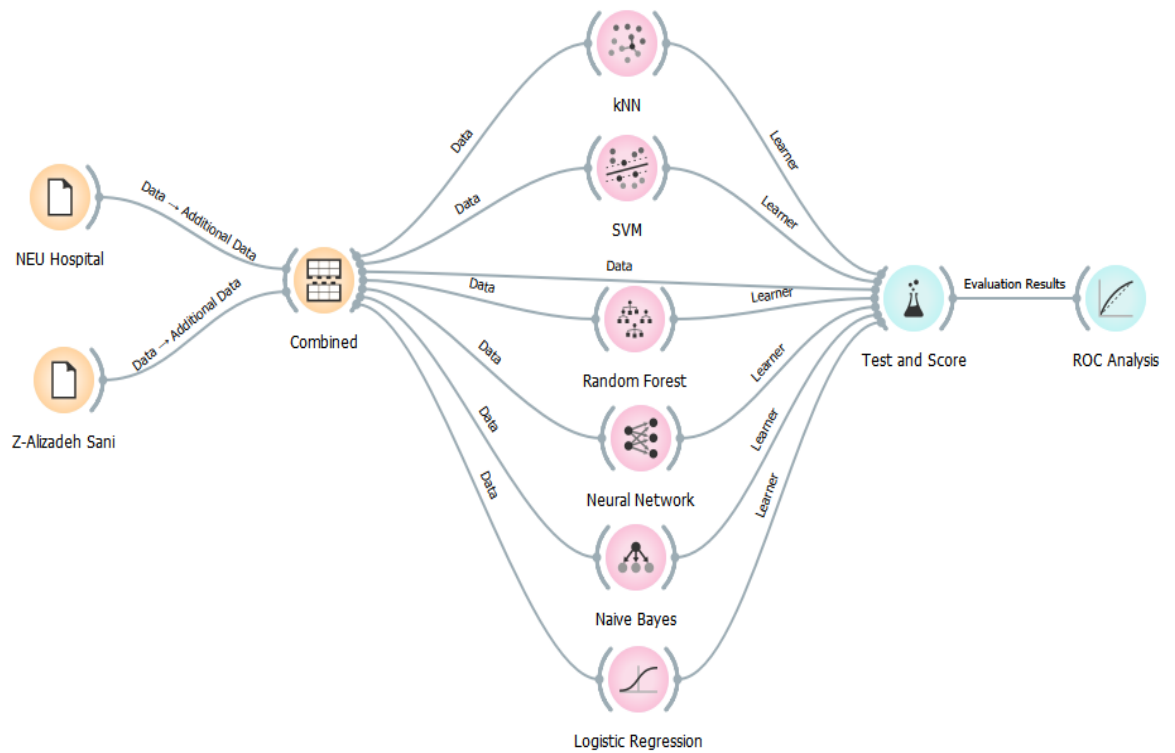


Figure 7.

Step-2 Classification workflow (NEU Hospital as the Training Dataset)

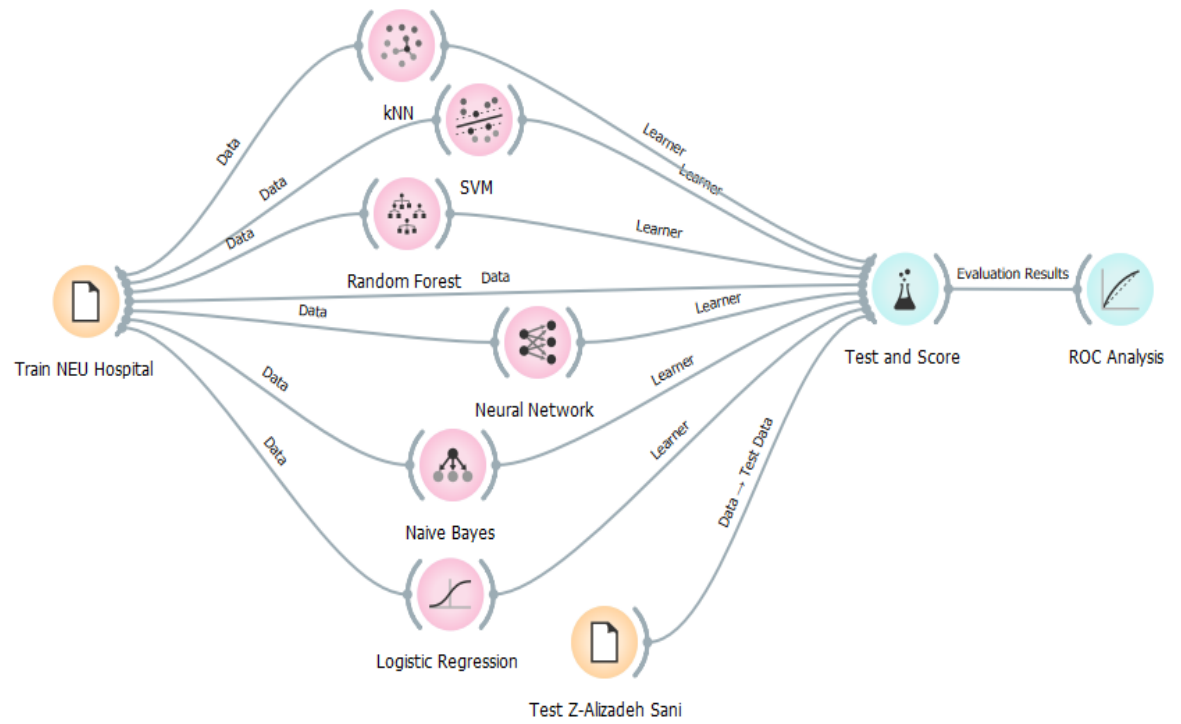
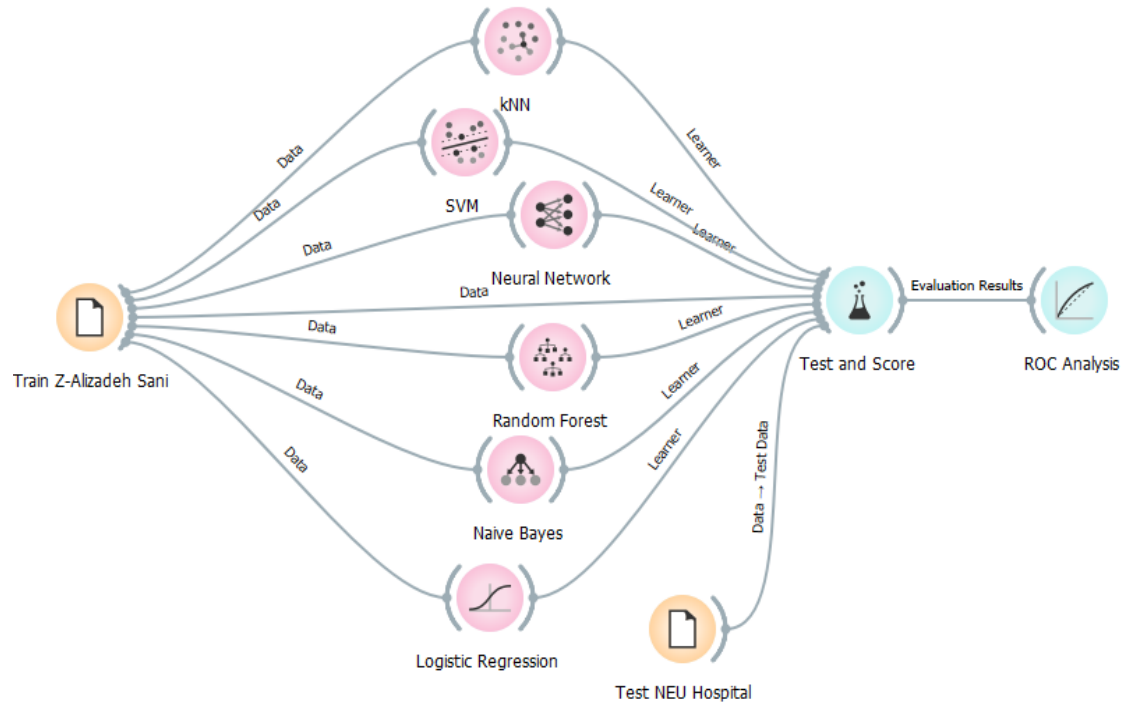


Figure 8.

Step-3 Classification workflow (Z-Alizadeh Sani as the Training Dataset)



CHAPTER IV

Findings and Discussion

This section firstly presents the application findings of Univariate, Bivariate, Multivariate and ROC curve analysis. Secondly, the results obtained from ML algorithms in 3 stages.

Table 2.

Descriptive statistics for quantitative variables from NEU Hospital dataset.(n=475)

Variables	Mean±SD	Median (Min-Max)
Age	60.98 ± 10.95	62 (32 - 89)
BP (mm/Hg)	124.33 ± 17.03	120 (80 -220)
PR (ppm)	74.92 ± 12.50	74 (45 - 171)
FBS (mg/dl)	120.02 ± 42.68	107 (69 - 362)
CR (mg/dl)	0.90 ± 0.45	0.82 (0.5 - 7.6)
TG (mg/dl)	148.96 ± 82.30	131 (7 - 686)
LDL (mg/dl)	117.75± 39.32	114 (40 - 337)
HDL (mg/dl)	45.77 ± 13.76	44 (13 - 169)
Bun (mg/dl)	36.76 ± 15.22	34 (13 - 182)
Hb (g/dl)	13.70 ± 1.68	13.90 (8.2 - 17.5)
K (mEq/lit)	4.39 ± 0.44	4.4 (3 - 5.7)
Na (mEq/lit)	139.87 ± 2.70	140 (129 - 147)
WBC (cells/ml)	7.85 ± 5.36	7280 (2300 - 11240)
Lymph (%)	30.73 ± 9.30	30.3 (1.58 - 86.4)
Neut (%)	59.74 ± 9.67	60.16 (3.27 - 90.1)
PLT (1000/ml)	241.95 ± 77.12	232 (66 - 778)
EF-TTE (%)	57.75 ± 7.35	60 (30 - 72)

Table 2. provides an overlook on descriptive statistics of the quantitative variables of NEU Hospital dataset. The patients in this dataset have an average age of 60.98 ± 10.95 yrs,with a minimum age of 32 and maximum age of 89.

While the average PR value is 74.92 ± 12.50 ppm, FBS is 120.02 ± 42.68 mg/dL and the average CR is 0.90 ± 0.45 mg/dL and the maximum CR value is 7.6 mg/dL. The mean value of TG is 148.96 ± 82.30 mg/dL and the maximum TG value is 686 mg/dL, the average LDL value is 117.75 ± 39.32 mg/dl, and the mean HDL is 45.77 ± 13.76 mg/dl. The average Bun is 36.76 ± 15.22 mg/dL. Hb mean value is 13.70 ± 1.68 gm/dL, K mean value is 4.39 ± 0.44 mEq/lit, and Na average is 139.87 ± 2.70 mEq/lit, WBC is 7.85 ± 5.36 cells/mL while Lymph mean value is 30.73 ± 9.30 %, Neut is 59.74 ± 9.67 %. The average PLT value is 241.95 ± 77.12 (1000/mL)and EF-TTE mean value is 57.75 ± 7.35 %.

Table 3.

Descriptive statistics for qualitative variables from NEU Hospital dataset (n=475)

Variables		n (%)
Gender	Female	148 (31.2%)
	Male	327 (68.8%)
DM	Absent	361 (76.0%)
	Present	114 (24.0%)
HT	Absent	259 (54.5%)
	Present	216 (45.5%)
Smoking Status	Absent	336 (70.7%)
	Present	139 (29.3%)
FH	Absent	410 (86.3%)
	Present	65 (13.7%)
Edema	Absent	460 (96.8%)
	Present	15 (3.2%)
Systolic Murmur	Absent	456 (96.0%)
	Present	19 (4.0%)
Chest Pain	Absent	259 (54.5%)
	Present	216 (45.5%)
Dyspnea	Absent	400 (84.2%)
	Present	75 (15.8%)
LVH	Absent	402 (84.6%)
	Present	73 (15.4%)
Region RWMA	Absent	405 (85.3%)
	Present	70 (14.7%)
VHD	Absent	295 (62.1%)
	Present	180 (37.9%)

In Table 3, frequencies and percentages of categorical variables are given. 114 (24.0%) patients have DM problem and 216 (45.5%) patients have HT problem. The number of people who smoke actively is 139 (29.3%) and the number with a Family History (FH) of the disease is 65 (13.7%). Furthermore, the number of patients with

Edema is 15 (3.2%) people and 19 (4.0%) people have a Systolic Murmur. On the other hand, 216 (45.5%) people have Chest Pain, 73 (15.4%) people have LVH, and 75 (15.8%) people have Dyspnea problem. The number of patients with Region RWMA problem is 70 (14.7%) and 180 (37.9%) patients have VHD health problem.

Table 4.

Comparison of quantitative variables between patients and controls in NEU Hospital dataset.(Mann Whitney U test)

Variable	CAD	Mean±SD	Median(Min-Max)	Z	P
Age	Absent	60.41±10.57	61.00 (34.00 - 89.00)	-1.086	0.277
	Present	61.30±11.15	62.00 (32.00 - 89.00)		
Systolic BP	Absent	123.76±16.07	120.00 (95.00 - 200.00)	-0.694	0.488
	Present	124.65±17.56	120.00 (80.00 - 220.00)		
PR	Absent	73.15±13.23	72.00 (52.00 - 168.00)	-3.073	0.002
	Present	75.90±11.98	75.00 (45.00 - 171.00)		
FBS	Absent	117.16±40.75	105 (78.00 - 362.00)	-1.367	0.172
	Present	121.61±43.70	107.00 (69.00 - 338.00)		
CR	Absent	0.92±0.57	0.830 (0.560 - 7.590)	-0.552	0.581
	Present	0.89±0.37	0.810 (0.53 - 5.60)		
TG	Absent	140.31±70.06	130.00 (7.00 - 388.00)	-1.265	0.206
	Present	153.78±88.13	131.00 (24.00 - 686.00)		
LDL	Absent	113.19±32.34	109.50 (42.00 - 200.00)	-1.425	0.154
	Present	120.29±42.55	118.00 (40.00 - 337.00)		
HDL	Absent	45.65±11.76	44.50 (13.00 - 82.00)	-1.031	0.303
	Present	45.84±14.77	43.00 (23.00 - 169.00)		
BUN	Absent	37.02±15.63	34.00 (17.00 - 182.00)	-0.551	0.582
	Present	36.61±15.02	34.00 (13.00 - 141.00)		
Hb	Absent	13.84±1.55	13.90 (10.30 - 17.30)	-0.920	0.358
	Present	13.63±1.74	13.80 (8.20 - 17.50)		
K	Absent	4.381±0.42	4.35 (3.30 - 5.50)	-0.572	0.567
	Present	4.39±0.44	4.40 (3.00 - 5.70)		

Table 4. (Continued)

Variable	CAD	Mean±SD	Median(Min-Max)	Z	P
Na	Absent	139.86±2.72	140.00 (129.00 - 146.00)	-0.086	0.931
	Present	139.88±2.70	140.00 (130.00 - 147.00)		
WBC	Absent	8.01±8.35	7.05 (3.87 - 112.40)	-1.689	0.091
	Present	7.77±2.47	7.43 (2.30 - 27.71)		
Lymph	Absent	31.03±10.10	30.71 (3.37 - 86.38)	-0.474	0.636
	Present	30.56±8.84	30.29 (1.58 - 73.87)		
Neut	Absent	59.58±10.15	59.28 (9.06 - 90.08)	-0.623	0.533
	Present	59.83±9.41	60.28 (3.27 - 87.78)		
PLT	Absent	234.33±65.65	227.5 (60.00 -492.00)	-1.401	0.161
	Present	246.19±82.62	235.00 (79.00 - 778.00)		
EF-TTE	Absent	59.42±5.51	60.00 (30.00 - 68.00)	-3.450	0.001
	Present	56.81±8.05	60.00 (30.00 - 72.00)		

Table 4 shows the comparison of the quantitative variables in NEU dataset. Mann-Whitney U test was used in this study because variables are not normally distributed. The mean age of people without the disease is 60.41 ± 10.57 yrs and 61.30 ± 11.15 yrs in people with the disease. The mean Systolic BP of people who are patient is 124.65 ± 17.56 mm/Hg, and CAD absent group are 123.76 ± 16.07 mm/Hg. There is a statistically significant difference of the PR ($p = 0.002$) and EF-TTE ($p = 0.001$) between CAD patients and CAD absent group. The median PR in patients with CAD is 75.00 ppm (45.00 - 171.00), but in CAD absent group, it is 72.00 ppm (52.00 - 168.00).

Table 5.

Comparison of qualitative variable between patients and controls from NEU Hospital dataset.(Chi-Squared test)

Variable		Category				χ^2	P
		Normal		CAD			
		n	%	n	%		
Gender	Female	47	31.8	101	68.2	1.521	0.217
	Male	123	37.6	204	62.4		
DM	No	137	38.0	224	62.0	3.056	0.080
	Yes	33	28.9	81	71.1		
HT	No	97	37.5	162	62.5	0.685	0.408
	Yes	73	33.8	143	66.2		
Smoking	No	134	39.9	202	60.1	8.364	0.004
Status	Yes	36	25.9	103	74.1		
FH	No	151	36.8	259	63.2	1.410	0.235
	Yes	19	29.2	46	70.8		
Edema	No	168	36.5	292	63.5	3.399	0.065
	Yes	2	13.3	13	86.7		
Systolic	No	168	36.8	288	63.2	5.497	0.019
Murmur	Yes	2	10.5	17	89.5		
Chest Pain	No	141	54.4	118	45.6	86.212	<0.001
	Yes	29	13.4	187	86.6		
Dyspnea	No	158	39.5	242	60.5	15.178	<0.001
	Yes	12	16.0	63	84.0		
LVH	No	150	37.3	252	62.7	2.644	0.104
	Yes	20	27.4	53	72.6		
Region	No	161	39.8	244	60.2	18.788	<0.001
RWMA	Yes	9	12.9	61	87.1		
VHD	No	107	36.3	188	63.7	0.079	0.779
	Yes	63	35.0	117	65.0		

Table 5 shows the qualitative variables outline of the Bivariate analysis. The Chi-Squared statistics states Gender is not showing a statistically significant difference between the patients' CAD or absent group. But Smoking status ($\chi^2 = 8.364$, $p < 0.05$), Systolic Murmur ($\chi^2 = 5.497$, $p < 0.05$), Chest Pain ($\chi^2 = 86.212$, $p < 0.001$), Dyspnea ($\chi^2 = 15.178$, $p < 0.001$), Region RWMA ($\chi^2 = 18.788$, $p < 0.001$) categories are significantly different relative to the patients of CAD or absent group. In the table, it has been shown that, out of 305 patients with CAD problems 202 of them are non-smokers, 288 have no Systolic Murmur problems, only 63 have Dyspnea, 187 have Chest Pain and 61 have Region RWMA.

As per percentages 103 (74.1%) out of 139 active smokers, 17 (89.5%) out of 19 Systolic Murmur patients, 187 (86.6%) out of 216 patients with Chest Pain, 63 (84,0%) out 75 patients with Dyspnea, and 61 (87.1%) out of 70 patients with Region RWMA have CAD.

Table 6.

Bivariate Logistic Regression results of each variables in NEU Hospital dataset

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β)		R ²	p
					Lower	Upper		
Age	0.007	0.009	0.722	1.007	0.990	1.025	0.002	0.396
Systolic BP	0.003	0.006	0.295	1.003	0.992	1.014	0.001	0.587
PR	0.020	0.009	5.186	1.020	1.003	1.037	0.016	0.023
FBS	0.003	0.002	1.173	1.003	0.998	1.007	0.004	0.279
CR	-0.116	0.206	0.318	0.890	0.594	1.334	0.001	0.573
TG	0.002	0.001	2.881	1.002	1.000	1.005	0.009	0.090
LDL	0.005	0.003	3.543	1.005	1.00	1.010	0.010	0.060
HDL	0.001	0.007	0.021	1.001	0.987	1.015	0.000	0.885
BUN	-0.002	0.006	0.082	0.998	0.986	1.010	0.000	0.775
Hb	-0.075	0.058	1.686	0.928	0.828	1.039	0.005	0.194
K	0.072	0.220	0.107	1.075	0.698	1.654	0.000	0.743
Na	0.002	0.035	0.003	1.002	0.935	1.074	0.000	0.957
WBC	-0.008	0.017	0.218	0.992	0.959	1.026	0.001	0.641
Lymph	-0.005	0.010	0.273	0.995	0.975	1.015	0.001	0.601
Neut	0.003	0.010	0.074	1.003	0.983	1.022	0.000	0.786
PLT	0.002	0.001	2.551	1.002	1.000	1.005	0.008	0.110
EF-TTE	-0.060	0.017	12.604	0.942	0.911	0.973	0.044	<0.001
Gender	-0.259	0.210	1.518	1.296	0.858	1.956	0.004	0.218
DM	0.406	0.233	3.034	1.501	0.950	2.371	0.009	0.082
HT	0.159	0.193	0.684	1.173	0.804	1.712	0.002	0.408
Smoking Status	0.641	0.223	8.229	1.898	1.225	2.941	0.025	0.004
FH	0.345	0.291	1.400	1.412	0.797	2.498	0.004	0.237
Edema	1.319	0.766	2.967	3.740	0.834	16.773	0.011	0.085
Systolic Murmur	1.601	0.754	4.511	4.958	1.132	21.727	0.019	0.034

Table 6. (Continued)

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower - Upper	R²	p
Chest Pain	2.042	0.235	75.262	7.705	4.858 - 12.221	0.242	<0.001
Dyspnea	1.232	0.331	13.838	3.428	1.791 - 6.560	0.048	<0.001
LVH	0.456	0.282	2.613	1.577	0.908 - 2.741	0.008	0.106
Region RWMA	1.498	0.371	16.280	4.472	2.160 - 9.258	0.061	<0.001
VHD	0.055	0.198	0.079	1.057	0.717 - 1.557	0.000	0.779

In Table 6, the results of separate simple LR regression results for each variable are given. It has been shown that 2 quantitative variables and 5 qualitative variables were statistically significant.

These are; PR, EF-TTE, Smoking Status, Systolic Murmur, Chest Pain, Dyspnea and Region RWMA.

As seen in the table, the PR variable was estimated from the model as 0.020. The odds value was found to be 1.020. The probability of each unit increase being CAD increases 1.020 times.

The parameter value of EF-TTE variable was calculated as -0.060, Odds value was found as 0.942. For each unit increase of the variable, the probability of CAD decreases by 0.942 times.

The parameter value of the Smoking Status variable was calculated as 0.641, Odds value was found as 1.898. The probability of CAD risk is 1.898 times higher than that of those who have a Smoking Status problem compared to the absent group.

The parameter value of the Systolic Murmur variable was calculated as 1.601, Odds value was found to be 4.958. The probability of CAD risk is 4.958 times higher than that of those who have a Systolic Murmur problem compared to the absent group.

The parameter value of the Chest Pain variable was calculated as 2.042, Odds value was found as 7.705. The probability of CAD risk is 7.705 times higher than that of those who have a Chest Pain problem compared to the absent group.

The parameter value of the Dyspnea variable was calculated as 1.232, Odds value was found as 3.428. The probability of CAD risk is 3.428 times higher than that of those who have a Dyspnea problem compared to the absent group.

The parameter value of Region RWMA variable was calculated as 1.498, Odds value was found to be 4.472. The probability of CAD risk is 4.472 times higher than that of those who have a Region RWMA problem compared to the absent group.

Table 7.

Multivariate Logistic Regression Equations Summary (NEU Hospital dataset)

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower - Upper	p
Age	0.045	0.015	9.676	1.046	1.017 - 1.077	0.002
Systolic BP	0.000	0.008	0.003	1.000	0.983 - 1.016	0.960
PR	0.016	0.010	2.779	1.016	0.997 - 1.036	0.096
FBS	0.005	0.003	2.105	1.005	0.998 - 1.012	0.147
CR	-0.078	0.440	0.032	0.925	0.391 - 2.189	0.859
TG	0.002	0.002	1.335	1.002	0.999 - 1.006	0.248
LDL	0.011	0.004	8.216	1.011	1.003 - 1.018	0.004
HDL	0.005	0.012	0.151	1.005	0.981 - 1.029	0.698
BUN	-0.010	0.012	0.679	0.990	0.966 - 1.014	0.410
Hb	-0.124	0.098	1.579	0.884	0.728 - 1.072	0.209
K	0.361	0.315	1.315	1.435	0.774 - 2.660	0.252
Na	0.081	0.051	2.525	1.084	0.981 - 1.198	0.112
WBC	-0.038	0.030	1.639	0.963	0.909 - 1.020	0.200
Lymph	-0.038	0.032	1.440	0.963	0.905 - 1.024	0.230
Neut	-0.055	0.031	3.197	0.946	0.890 - 1.005	0.074
PLT	0.000	0.002	0.056	1.000	0.997 - 1.004	0.813
EF-TTE	0.034	0.039	0.765	1.035	0.959 - 1.116	0.382
Gender	0.012	0.346	0.001	1.012	0.513 - 1.996	0.971
DM	0.004	0.365	0.000	1.004	0.490 - 2.054	0.992
HT	0.083	0.285	0.084	1.086	0.621 - 1.901	0.772

Table 7. (Continued)

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower - Upper	p
Smoking Status	0.787	0.293	7.237	2.197	1.238 - 3.897	0.007
FH	0.464	0.397	1.367	1.590	0.731 - 3.461	0.242
Edema	1.482	0.866	2.933	4.404	0.807 - 24.020	0.087
Systolic Murmur	2.424	0.870	7.769	11.292	2.053 - 62.099	0.005
Chest Pain	2.959	0.304	94.531	19.270	10.614 - 34.987	<0.001
Dyspnea	1.797	0.430	17.484	6.034	2.598 - 14.012	<0.001
LVH	0.607	0.364	2.775	1.834	0.898 - 3.745	0.096
Region RWMA	2.450	0.866	8.012	11.591	2.125 - 63.231	0.005
VHD	-0.564	0.278	4.116	0.569	0.330 - 0.981	0.042

Table 7, shows the Multivariate Logistic Regression results. It has been shown that 2 quantitative variables and 6 qualitative variables were statistically significant. These are; Age, LDL, Smoking Status, Systolic Murmur, Chest Pain, Dyspnea, Region RWMA and VHD.

As seen in the table, the Age variable was estimated from the model as 0.045. The odds value was found to be 1.046. The probability of each unit increase being CAD increases 1.046 times.

The parameter value of the LDL variable was calculated as 0.011, Odds value was found to be 1.011. For each unit increase of the variable, the probability of CAD increases by 1.011 times.

The parameter value of the Smoking Status variable was calculated as 0.787, Odds value was found as 2.197. The probability of CAD risk is 2.197 times higher than that of those who have a Smoking Status problem compared to the absent group.

The parameter value of the Systolic Murmur variable was calculated as 2.424, Odds value was found to be 11.292. The probability of CAD risk is 11.292 times higher than that of those who have a Systolic Murmur problem compared to the absent group.

The parameter value of the Chest Pain variable was calculated as 2.959, Odds value was found as 19.270. The probability of CAD risk is 19.270 times higher than that of those who have a Chest Pain problem compared to the absent group.

The parameter value of the Dyspnea variable was calculated as 1.797, Odds value was found as 6.034. The probability of CAD risk is 6.034 times higher than that of those who have a Dyspnea problem compared to the absent group.

The parameter value of the Region RWMA variable was found to be 2.450, whilst the Odds value was 11.591. The probability of CAD risk is 11.591 times higher than that of those with Region RWMA problems compared to the absent group.

The parameter value of VHD variable was calculated as -0.564, Odds value was found as 0.569. The probability of CAD risk is 0.569 times lower than those with VHD problems compared to the absent group.

Table 8.

Omnibus tests of Model Coefficients for the Multivariate Logistic Regression (NEU Hospital dataset)

	Chi-square	Df	Sig.
Step	208.695	29	< 0.001
Block	208.695	29	< 0.001
Model	208.695	29	< 0.001

In Table 8, the Omnibus test result is based on Chi-square and is obtained according to the probability of real data being observed, assuming the model is correct. The result shows that the Multivariate Logistic Model is statistically significant because the p-value of 0.001 is less than the significance level of 0.05.

Table 9.

Model Summary (Multivariate for the Multivariate Logistic Regression comprising of all Logistic Regression Models, NEU Hospital dataset)

-2 Log-likelihood	Cox & Snell R Square	Nagelkerke R Square
410.893 ^a	0.356	0.488

The Multivariate Logistic Regression Model illustrates between 0.356 and 0.488 variations in the influences on the risk of CAD formation in patients.

Table 10.

Hosmer and Lemeshow Test to Assesses the Model Fit (NEU Hospital dataset)

Chi-square	df	Sig.
9.275	8	.320

The Hosmer Lemeshow test is one of the methods of evaluating LR Model fit. The p-value was found to be 0.320. According to this result, it is seen that the model and data fit well, and the predictive power of the model is high since there is no significant difference between the expected value and the observed value.

Table 11.

Classification Table for the Multivariate Logistic Regression Model (NEU Hospital dataset)

Observed		Predicted		
		CAD		Percentage
		Absent	Present	Correct
CAD	Absent	119	51	70.0
	Present	36	269	88.2
Overall				81.7
Percentage				

a. The cut value is 0.5

The classification Table 11 describes well the model categorizes the dependent results. 51 of absent patients were incorrectly separated as CAD by the model and 36 CAD patients designated as absent group. The cases of the study have been 81.7% classified correctly by the model.

Table 12.

Area Under the Curve for the ROC for the quantitative variables (NEU Hospital dataset)

Variable	Area Under the Curve	S.E	p-Value	CI (95%)
Age	0.530	0.027	0.278	0.476 - 0.584
Systolic BP	0.519	0.028	0.502	0.464 - 0.573
PR	0.585	0.027	0.002	0.531 - 0.639
FBS	0.538	0.027	0.172	0.484 - 0.591
CR	0.515	0.028	0.581	0.461 - 0.569
TG	0.535	0.027	0.206	0.481 - 0.589
LDL	0.539	0.027	0.154	0.487 - 0.592
HDL	0.528	0.028	0.303	0.475 - 0.582
Bun	0.515	0.027	0.582	0.463 - 0.568
Hb	0.525	0.027	0.358	0.472 - 0.579
K	0.516	0.028	0.568	0.461 - 0.570
Na	0.502	0.028	0.932	0.448 - 0.556
WBC	0.547	0.028	0.091	0.493 - 0.601
Lymph	0.513	0.028	0.636	0.458 - 0.568
Neut	0.517	0.028	0.533	0.462 - 0.572
PLT	0.539	0.028	0.161	0.485 - 0.593
EF-TTE	0.587	0.027	0.002	0.535 - 0.639

Table 12 presents the AUC result of the quantitative variables analyzed separately for ROC. The ROC shows that the best performance variable is the EF-TTE variable with an AUC of 0.587 (58.7%), closely followed by the PR variable with an AUC of 0.585 (58.5%) and they are statistically significant. The lowest performing variable is the Na variable with an AUC of 50.2%.

Figure 9.

ROC Curve for the Quantitative Variables (NEU Hospital dataset)

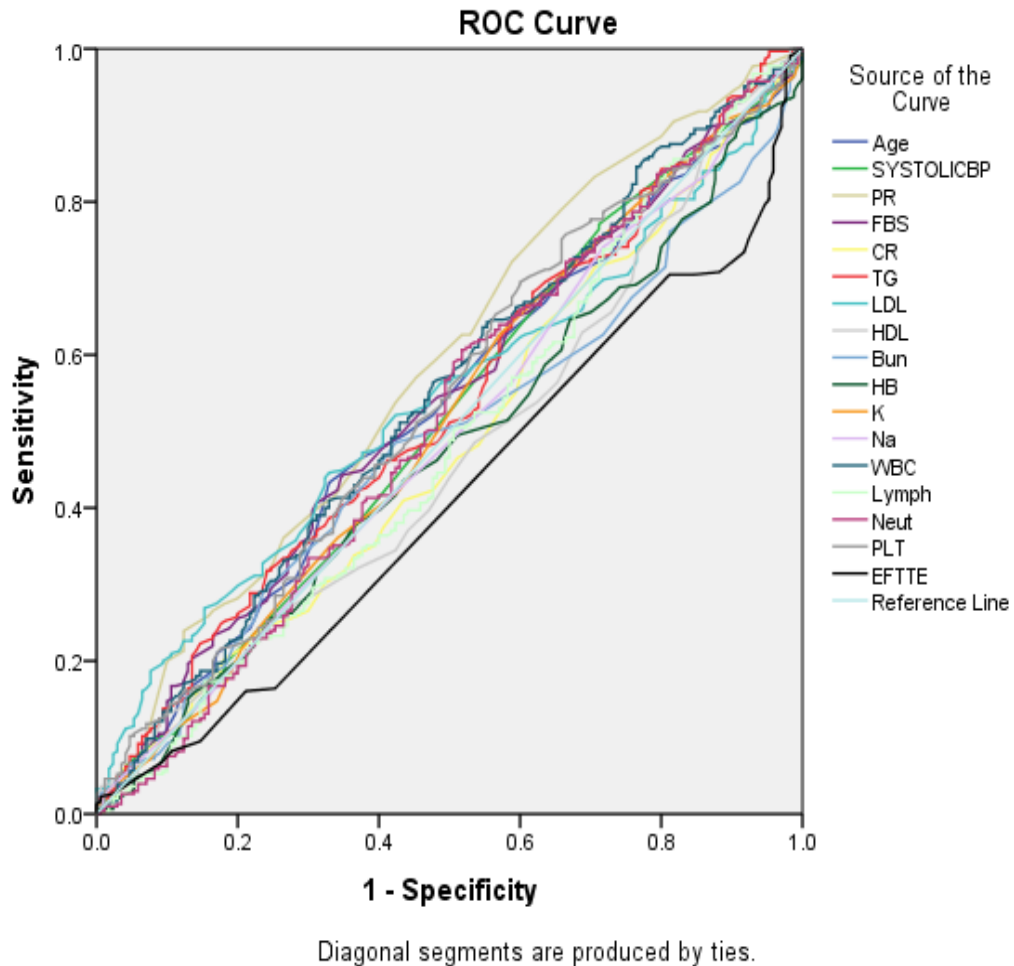


Table 13.

Area Under the Curve for the ROC for the Multivariate Logistic Regression (NEU Hospital dataset)

Area Under The Curve	S.E	p-value	CI (95%)
0.868	0.017	<0.001	0.834 - 0.902

In Table 13, Multivariate Logistic Regression analysis was performed with all variables and model performance was evaluated with ROC using probability values. The AUC result is 0.868 and the confidence interval (CI) is between 0.834 - 0.902.

In Fig 10., the area under the curve is 0.868 (86.8%). This area represents the area where LR correctly classified patients. The p-value of <0.001 demonstrates that the curve is statistically essential.

Figure 10.

ROC curve for the Final Multivariate Logistic Regression Model (NEU Hospital dataset)

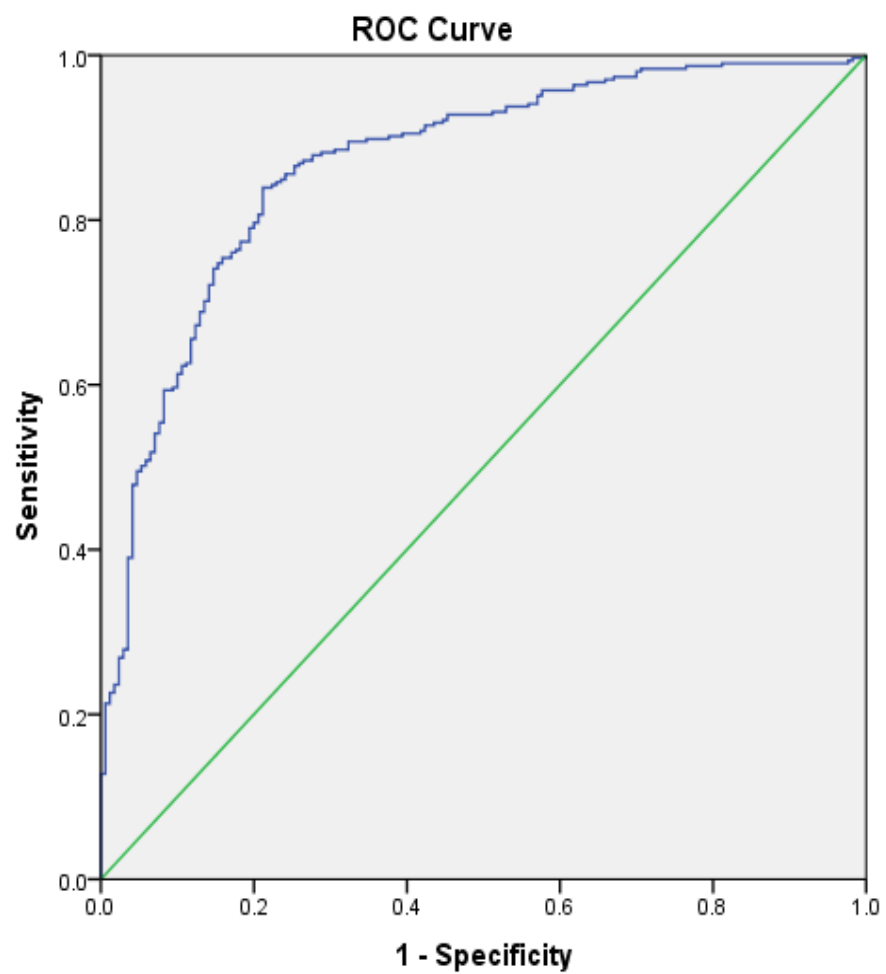


Table 14.

Descriptive statistics for quantitative variables from Z-Alizadeh Sani dataset. (n=303)

Variables	Mean±SD	Median (Min-Max)
Age	58.89 ± 10.39	58 (30 - 86)
BP (mm/Hg)	129.55 ± 18.94	130 (90 - 190)
PR (ppm)	75.14 ± 8.91	70 (50 - 110)
FBS (mg/dl)	119.18 ± 52.08	98 (62 - 400)
CR (mg/dl)	1.05 ± 0.26	1 (0.5 - 2.2)
TG (mg/dl)	150.34 ± 97.96	122 (37 - 1050)
LDL (mg/dl)	104.64 ± 35.40	100 (18 - 232)
HDL (mg/dl)	40.23 ± 10.56	39 (15.9 - 111)
Bun (mg/dl)	17.50 ± 6.96	16 (6 - 52)
Hb (g/dl)	13.53 ± 1.61	13.2 (8.9 - 17.60)
K (mEq/lit)	4.23 ± 0.46	4.2 (3 - 6.60)
Na (mEq/lit)	140.1 ± 3.81	141 (128 - 156)
WBC (cells/ml)	7562.06 ± 2413.74	7100 (3700 - 18000)
Lymph (%)	32.4 ± 9.97	32 (7 - 60)
Neut (%)	60.15 ± 10.18	60 (32 - 89)
PLT (1000/ml)	221.49 ± 60.79	210 (25- 742)
EF-TTE (%)	47.23 ± 8.93	50 (15 - 60)

Table 14. provides an overlook on descriptive statistics of the quantitative variables of NEU Hospital dataset. The patients in this dataset have an average age of 58.89 ± 10.39 yrs, with a minimum age of 30 and maximum age of 89. While the average PR value is 75.14 ± 8.91 ppm, FBS is 119.18 ± 52.08 mg/dl and the average CR is 1.05 ± 0.26 mg/dl and the maximum CR value is 2.2 mg/dl. The mean value of TG is 150.34 ± 97.96 mg/dL and the maximum TG value is 1050mg/dL, the average LDL value is 104.64 ± 35.40 mg/dl, and the mean HDL is 40.23 ± 10.56 mg/dl. The average Bun is 17.50 ± 6.96 mg/dl. Hb mean value is 13.53 ± 1.61 gm/dl, K mean value is 4.23 ± 0.46 mEq/lit, and Na average is 140.1 ± 3.81 mEq/lit, Lymph mean value is 32.4 ± 9.97 %, and Neut mean value is 60.15 ± 10.18 %.

Neut is 60.15 ± 10.18 %. The average PLT value is 221.49 ± 60.79 (1000/ml) and EF-TTE mean value is 47.23 ± 8.93 %.

Table 15.

Descriptive statistics for qualitative variables from Z-Alizadeh Sani dataset (n=303)

Variables		n (%)
Gender	Female	127 (41.9%)
	Male	176 (58.1%)
DM	Absent	213 (70.30%)
	Present	90 (29.7%)
HT	Absent	124 (40.90%)
	Present	179 (59.10%)
Smoking Status	Absent	240 (79.2%)
	Present	63 (20.8%)
FH	Absent	255 (84.2%)
	Present	48 (15.8%)
Edema	Absent	291 (96.0%)
	Present	12 (4.0%)
Systolic Murmur	Absent	262 (86.5%)
	Present	41 (13.5%)
Chest Pain	Absent	139 (45.9%)
	Present	164 (54.1%)
Dyspnea	Absent	169 (55.8%)
	Present	134 (44.2%)
LVH	Absent	283 (93.4%)
	Present	20 (6.6%)
Region RWMA	Absent	217 (71.6%)
	Present	86 (28.4%)
VHD	Absent	116 (38.3%)
	Present	187 (61.7%)

In Table 15, frequencies and percentages of categorical variables are given. 90 (29.7%) patients have DM problem and 179 (59.10%) patients have HT problem. The number of people who smoke actively is 63 (20.8%) and the number with a Family History of the disease is 48 (15.8%). Furthermore, the number of patients with Edema is 12 (4.0%) people and 41 (13.5%) people have a Systolic Murmur. On the other hand, 164 (54.1%) people have Chest Pain, 20 (6.6%) people have LVH, and 134 (44.2%) people have Dyspnea problem. The number of patients with Region RWMA problem is 86 (28.4%) and 187 (61.7%) patients have VHD health problem.

Table 16.

Comparison of quantitative variables between patients and controls in Z-Alizadeh Sani dataset.(Mann Whitney U test)

Variable	CAD	Mean±SD	Median(Min-Max)	Z	p
Age	Absent	53.06 ± 9.32	52.00 (30.00 – 79.00)	-6.102	<0.001
	Present	61.25 ± 9.88	61.50 (36.00 – 86.00)		
Systolic BP	Absent	122.47 ± 18.30	120 (90.00 – 180.00)	-4.455	<0.001
	Present	132.41 ± 18.48	130 (90.00 – 190.00)		
PR	Absent	72.78 ± 8.08	70 (50.00 – 100.00)	-2.944	0.003
	Present	76.09 ± 9.07	74 (50.00 – 110.00)		
FBS	Absent	102.34 ± 34.79	92 (65.00 – 300.00)	-4.121	<0.001
	Present	125.97 ± 56.26	103 (62.00 – 400.00)		
CR	Absent	1.02 ± 0.19	1.0 (0.60 - 1.60)	-0.985	0.325
	Present	1.07± 0.29	1.0 (0.50 - 2.20)		
TG	Absent	128.68 ± 75.54	110 (37.00 – 450.00)	-3.214	0.001
	Present	159.07 ± 104.55	130 (50.00 – 1050.00)		
LDL	Absent	105.95 ± 35.41	101 (18.00 – 232.00)	-0.512	0.608
	Present	104.12 ± 35.46	100 (30.00 – 213.00)		
HDL	Absent	40.94 ± 11.59	42 (15.90 - 83.00)	-0.669	0.503
	Present	39.95 ± 10.13	39 (18.00 – 111.00)		
BUN	Absent	16.53 ± 6.15	15 (6.00 – 41.00)	-1.518	0.129
	Present	17.89 ± 7.23	16 (8.00 – 52.00)		
Hb	Absent	13.26 ± 1.51	13.40 (9.00 - 17.50)	-0.802	0.423
	Present	13.11 ± 1.65	13.10 (8.90 - 17.60)		

Table 16. (Continued)

Variable	CAD	Mean±SD	Median(Min-Max)	Z	p
K	Absent	4.10 ± 0.38	4.10 (3.00 - 5.20)	-3.133	0.002
	Present	4.28 ± 0.48	4.30 (3.10 - 6.60)		
Na	Absent	141.51 ± 3.35	141 (131.00 – 153.00)	-1.686	0.092
	Present	140.79 ± 3.97	141 (128.00 – 156.00)		
WBC	Absent	7293.10 ± 2115.33	7100 (3800 – 17800)	-0.902	0.367
	Present	7670.37 ± 2520.48	7150 (3700 – 18000)		
Lymph	Absent	34.39 ± 9.533	34 (9.00 – 60.00)	-2.171	0.030
	Present	31.60 ± 10.06	31.50 (7.00 – 60.00)		
Neut	Absent	58.16 ± 9.817	58 (32.00 – 89.00)	-2.156	0.031
	Present	60.95 ± 10.24	60 (33.00 – 86.00)		
PLT	Absent	230.56 ± 76.02	217 (129.00 – 742.00)	-1.203	0.229
	Present	217.83 ± 53.23	208 (25.00 – 410.00)		
EF-TTE	Absent	50.52 ± 8.04	55 (15.00 – 60.00)	-5.238	<0.001
	Present	45.91 ± 8.94	45.50 (15.00 – 60.00)		

Table 16 shows the comparison of the quantitative variables in Z-Alizadeh Sani dataset. Mann-Whitney U test was used in this study because variables are not normally distributed.

The mean age of people without the disease is 53.06 ± 9.32 yrs and 61.25 ± 9.88 yrs in people with the disease. The mean Systolic BP of people who are patient is 132.41 ± 18.48 mm/Hg, and CAD absent group are 122.47 ± 18.30 mm/Hg.

There is a statistically significant difference of the Age ($p < 0.001$), Systolic BP ($p < 0.001$), PR ($p = 0.003$), FBS ($p < 0.001$), TG ($p = 0.001$), K ($p = 0.002$), Lymph ($p = 0.030$), Neut ($p = 0.031$), and EF-TTE ($p < 0.001$) between CAD patients and CAD absent group. The median PR in patients with CAD is 74.00 ppm (50.00 - 110.00), but in CAD absent group, it is 70.00 ppm (50.00 - 100.00).

Table 17.

Qualitative variable distributions between patients and controls from Z-Alizadeh Sani dataset (Chi-Squared test)

Variable		Category				χ^2	P
		Normal		CAD			
		N	%	N	%		
Gender	Female	41	32.30%	86	67.70%	1.362	0.243
	Male	46	26.10%	130	73.90%		
DM	No	77	36.20%	136	63.80%	19.379	<0.001
	Yes	10	11.10%	80	71.30%		
HT	No	55	44.40%	69	55.60%	25.090	<0.001
	Yes	32	17.90%	147	82.10%		
Smoking Status	No	73	30.40%	167	69.60%	1.637	0.201
	Yes	14	22.20%	49	77.80%		
FH	No	75	29.40%	180	70.60%	0.384	0.535
	Yes	12	25.00%	36	75.00%		
Edema	No	85	29.20%	206	70.80%	0.886	0.519
	Yes	2	16.70%	10	83.30%		
Systolic	No	75	28.60%	187	71.40%	0.007	0.933
Murmur	Yes	12	29.30%	29	70.70%		
Chest Pain	No	77	55.40%	62	44.60%	89.328	<0.001
	Yes	10	6.10%	154	93.90%		
Dyspnea	No	40	23.70%	129	76.30%	4.750	0.029
	Yes	47	35.10%	87	64.90%		
LVH	No	83	29.30%	200	70.70%	0.794	0.373
	Yes	4	20.00%	16	80.00%		
Region	No	83	38.20%	134	61.80%	33.966	<0.001
RWMA	Yes	4	4.70%	82	95.30%		
VHD	No	40	34.5%	76	65.5%	3.057	0.080
	Yes	47	25.1%	140	74.9%		

Table 17 shows the qualitative variables outline of the Bivariate analysis. The Chi-Squared statistics states Gender is not showing a statistically significant difference between the patients' CAD or absent group.

But DM ($\chi^2 = 19.379$, $p < 0.001$), HT ($\chi^2 = 25.090$, $p < 0.0001$), Chest Pain ($\chi^2 = 89.328$, $p < 0.001$), Dyspnea ($\chi^2 = 4.750$, $p < 0.05$), Region RWMA ($\chi^2 = 33.966$, $p < 0.001$) categories are significantly different relative to the patients of CAD or absent group. In the table, it has been shown that, out of 216 patients with CAD problems 80 of them are DM problems, 69 have HT problems, 87 have Dyspnea, 154 have Chest Pain and 134 have no Region RWMA.

As per percentages 80 (71.3%) out of 90 DM patients, 147 (82.1%) out of 179 HT patients, 154 (93.9%) out of 164 patients with Chest Pain, 87 (64.9%) out 134 patients with Dyspnea, and 82 (95.3%) out of 86 patients with Region RWMA have CAD.

Table 18.

Bivariate Logistic Regression results of each variables in Z-Alizadeh Sani dataset.

Variable	β	S.E	Wald	Exp (β)	%95 C.I for Exp(β)		R ²	p
					Lower	Upper		
Age	0.090	0.016	33.598	1.094	1.061	-1.128	0.186	< 0.001
Systolic BP	0.032	0.008	16.187	1.032	1.016	- 1.048	0.085	< 0.001
PR	0.048	0.017	8.249	1.049	1.015	- 1.084	0.043	0.004
FBS	0.013	0.004	11.632	1.013	1.006	- 1.021	0.075	0.001
CR	0.764	0.509	2.258	2.148	0.793	- 5.821	0.011	0.133
TG	0.005	0.002	6.108	1.005	1.001	- 1.008	0.036	0.013
LDL	-0.001	0.004	0.168	0.999	0.992	- 1.006	0.001	0.682
HDL	-0.009	0.012	0.546	0.991	0.969	- 1.015	0.003	0.460
BUN	0.031	0.020	2.353	1.032	0.991	- 1.074	0.012	0.125
Hb	-0.059	0.079	0.544	0.943	0.807	- 1.102	0.003	0.461
K	0.953	0.306	9.682	2.595	1.423	- 4.730	0.049	0.002
Na	-0.049	0.034	2.163	0.952	0.891	- 1.017	0.010	0.141
WBC	0.000	0.000	1.507	1.000	1.000	- 1.000	0.007	0.220
Lymph	-0.029	0.013	4.798	0.972	0.947	- 0.997	0.023	0.028
Neut	0.028	0.013	4.593	1.028	1.002	- 1.054	0.022	0.032
PLT	-0.003	0.002	2.553	0.997	0.993	- 1.001	0.012	0.110
EF-TTE	-0.076	0.019	15.298	0.927	0.892	- 0.963	0.089	< 0.001

Table 18. (Continued)

Variable	β	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower – Upper	R ²	p
Gender	0.298	0.256	1.358	1.347	0.816 – 2.225	0.006	0.244
DM	1.511	0.364	17.178	4.529	2.217 – 9.253	0.099	<0.001
HT	1.298	0.266	23.818	3.662	2.174 – 6.167	0.113	<0.001
Smoking Status	0.425	0.334	1.621	1.530	0.975 – 2.944	0.008	0.203
FH	0.223	0.361	0.383	1.250	0.617 – 2.534	0.002	0.536
Edema	0.724	0.785	0.851	2.063	0.443 – 9.615	0.005	0.356
Systolic Murmur	-0.031	0.369	0.007	0.969	0.470 – 1.999	0.000	0.933
Chest Pain	2.951	0.368	64.218	19.126	9.293 – 39.362	0.392	<0.001
Dyspnea	-0.555	0.256	4.704	0.574	0.348 – 0.948	0.022	0.030
LVH	0.507	0.574	0.779	1.660	0.539 – 5.114	0.004	0.377
Region RWMA	2.541	0.531	22.928	12.698	4.487 – 35.934	0.186	<0.001
VHD	0.450	0.258	3.037	1.568	0.945 - 2.60	0.014	0.081

In Table 18, the results of separate simple LR regression results for each variable are given. It has been shown that 9 quantitative variables and 5 qualitative variables were statistically significant.

These are; Age, Systolic BP, PR, FBS, TG, K, Lymph, Neut, EF-TTE, DM, HT, Chest Pain, Dyspnea and Region RWMA.

The table demonstrates the Age variable that was estimated as 0.090. The odds value was found to be 1.094. The probability of each unit increase being CAD increases 1.094 times.

The parameter value of the Systolic BP variable was calculated as 0.032, the Odds value was found as 1.032. For each unit increase of the variable, the probability of CAD increases by 1.032 times.

The parameter value of PR variable was calculated as 0.048, Odds value was found to be 1.049. For each unit increase of the variable, the probability of CAD increases by 1.049 times.

The parameter value of FBS variable was calculated as 0.013, Odds value was found as 1.013. For each unit increase of the variable, the probability of CAD increases by 1.013 times.

The parameter value of TG variable was calculated as 0.005, Odds value was found to be 1.005. For each unit increase of the variable, the probability of CAD increases by 1.005 times.

The parameter value of K variable was calculated as 0.953, Odds value was found to be 2.595. For each unit increase of the variable, the probability of CAD increases by 2.595 times.

The parameter value of the Lymph variable was calculated as -0.029, Odds value was found as 0.972. For each unit increase of the variable, the probability of CAD decreases by 0.972 times.

The parameter value of Neut variable was calculated as 0.028, Odds value was found to be 1.028. For each unit increase of the variable, the probability of CAD increases by 1.028 times.

The parameter value of EF-TTE variable was calculated as -0.076, Odds value was found as 0.927. For each unit increase of the variable, the probability of CAD decreases by 0.927 times.

The parameter value of DM variable was calculated as 1.511, Odds value was found to be 4.529. The probability of CAD risk is 4.529 times higher than that of those who have DM problems compared to the absent group.

The parameter value of HT variable was calculated as 1.298, Odds value was found to be 3.662. The probability of CAD risk is 3.662 times higher than that of those who have an HT problem compared to the absent group.

The parameter value of Chest Pain variable was calculated as 2.951, Odds value was found to be 19.126. The probability of CAD risk is 19.126 times higher than that of those who have a Chest Pain problem compared to the absent group.

The parameter value of the Dyspnea variable was calculated as -0.555, Odds value was found as 0.574. The probability of CAD risk is 0.574 times lower than that of those who have a Dyspnea problem compared to the absent group.

The parameter value of Region RWMA variable was calculated as 12.541, Odds value was found as 12.698. The probability of CAD risk is 12.698 times higher than that of those who have a Region RWMA problem compared to the absent group.

Table 19.

Multivariate Logistic Regression Equations Summary (Z-Alizadeh Sani dataset)

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower – Upper	p
Age	0.142	0.030	21.950	1.153	1.086 - 1.224	<0.001
Systolic BP	0.003	0.018	0.022	1.003	0.969 - 1.038	0.883
PR	0.069	0.035	3.863	1.071	1.000 - 1.147	0.049
FBS	0.005	0.007	0.475	1.005	0.992 - 1.018	0.491
CR	0.608	1.318	0.213	1.836	0.139 - 24.298	0.645
TG	0.010	0.004	7.010	1.010	1.003 - 1.018	0.008
LDL	-0.005	0.008	0.393	0.995	0.979 - 1.011	0.531
HDL	0.006	0.022	0.073	1.006	0.964 - 1.050	0.786
BUN	-0.035	0.048	0.539	0.965	0.879 - 1.061	0.463
Hb	-0.567	0.238	5.650	0.567	0.356 - 0.905	0.017
K	-0.220	0.627	0.123	0.802	0.235 - 2.743	0.725
Na	0.028	0.085	0.108	1.028	0.871 - 1.214	0.742
WBC	0.000	0.000	0.026	1.000	1.000 - 1.000	0.871
Lymph	0.003	0.070	0.002	1.003	0.875 - 1.149	0.969
Neut	0.010	0.069	0.020	1.010	0.883 - 1.155	0.886
PLT	-0.002	0.005	0.176	0.998	0.989 - 1.008	0.675
EF-TTE	-0.091	0.041	4.993	0.913	0.843 - 0.989	0.025
Gender	1.258	0.727	2.997	3.519	0.847 - 14.625	0.083
DM	2.194	0.849	6.684	8.972	1.700 - 47.347	0.010
HT	2.108	0.698	9.113	8.233	2.095 - 32.357	0.003
Smoking Status	0.968	0.668	2.099	2.634	0.711 - 9.760	0.147
FH	2.234	0.774	8.343	9.341	2.051 - 42.549	0.004
Edema	-1.655	1.522	1.182	0.191	0.010 - 3.775	0.277

Table 19. (Continued)

Variable	B	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower – Upper	p
Systolic Murmur	0.819	0.857	0.913	2.268	0.423 - 12.166	0.339
Chest Pain	4.058	0.690	34.619	57.843	14.970 - 223.499	<0.001
Dyspnea	-1.548	0.608	6.486	0.213	0.065- 0.700	0.011
LVH	1.028	1.066	0.931	2.796	0.346 - 22.578	0.335
Region RWMA	2.468	0.805	9.399	11.800	2.436 - 57.171	0.002
VHD	-1.049	0.638	2.702	0.350	0.100 - 1.224	0.100

Table 19, shows the Multivariate Logistic Regression results. It has been shown that 5 quantitative variables and 6 qualitative variables were statistically significant.

These are; Age, PR, TG, Hb, EF-TTE, DM, HT, FH, Chest Pain, Dyspnea and Region RWMA.

The Age variable was estimated from the model as 0.142. The odds value was found to be 1.153.

The probability of each unit increase being CAD increases 1.153 times.

The parameter value of PR variable was calculated as 0.069, Odds value was found as 1.071. For each unit increase of the variable, the probability of CAD increases by 1.071 times.

The parameter value of TG variable was calculated as 0.010, Odds value was found to be 1.010. For each unit increase of the variable, the probability of CAD increases by 1.010 times.

The parameter value of Hb variable was calculated as -0.567, Odds value was found as 0.567. For each unit increase of the variable, the probability of CAD decreases by 0.567 times.

The parameter value of EF-TTE variable was calculated as -0.091, Odds value was found as 0.913. For each unit increase of the variable, the probability of CAD decreases by 0.913 times.

The parameter value of DM variable was calculated as 2.194, Odds value was found as 8.972. The probability of CAD risk is 8.972 times higher than that of those who have a DM problem compared to the absent group.

The parameter value of HT variable was calculated as 2.108, Odds value was found as 8.233. The probability of CAD risk is 8.233 times higher than that of those who have an HT problem compared to the absent group.

The parameter value of FH variable was calculated as 2.234, Odds value was found as 9.341. The probability of CAD risk is 9.341 times higher than that of those who have an FH problem compared to the absent group.

The parameter value of the Chest Pain variable was calculated as 4.058, Odds value was found as 57.843. The probability of CAD risk is 57.843 times higher than that of those who have a Chest Pain problem compared to the absent group.

The parameter value of the Dyspnea variable was calculated as -1.548, Odds value was found as 0.213. The probability of CAD risk is 0.213 times lower than that of those who have a Dyspnea problem compared to the absent group.

The parameter value of Region RWMA variable was calculated as 2.468, Odds value was found to be 11.800. The probability of CAD risk is 11.800 times higher than that of those who have a Region RWMA problem compared to the absent group.

Table 20.

Omnibus tests of Model Coefficients for the Multivariate Logistic Regression (Z-Alizadeh Sani dataset)

	Chi-square	Df	Sig.
Step	231.948	29	<0.001
Block	231.948	29	<0.001
Model	231.948	29	<0.001

In Table 20, the Omnibus test result is based on Chi-Square and is obtained according to the probability of real data being observed, assuming the model is correct. The result shows that the Multivariate Logistic Model is statistically significant because the p-value of 0.001 is less than the significance level of 0.05.

Table 21.

Model Summary (Multivariate for the Multivariate Logistic Regression comprising of all Logistic Regression Models, Z-Alizadeh Sani dataset)

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
131.385 ^a	.535	.766

The Multivariate Logistic Regression Model illustrates between 0.535 and 0.766 variations in the influences on the risk of CAD formation in patients.

Table 22.

Hosmer and Lemeshow Test to Assesses the Model Fit (Z-Alizadeh Sani dataset)

Chi-square	df	Sig.
2.445	8	.964

The Hosmer Lemeshow test is one of the methods of evaluating LR Model fit. The p-value was found to be 0.964.

According to this result, it is seen that the model and data fit well, and the predictive power of the model is high since there is no significant difference between the expected value and the observed value.

Table 23.

Classification Table for the Multivariate Logistic Regression Model(Z-Alizadeh Sani dataset)

Observed		Predicted		
		CAD		Percentage
		Absent	Present	Correct
CAD	Absent	73	14	83.9
	Present	11	205	94.9
Overall				91.7
Percentage				

a. The cut value is 0.5

The classification Table 23 describes well the model categorizes the dependent results. 14 of absent patients were incorrectly separated as CAD by the model and 11 CAD patients designated as absent group. The cases of the study have been 91.7% classified correctly by the model.

Table 24.

Area Under the Curve for the ROC for the quantitative variables (Z-Alizadeh Sani dataset)

Variable	Area Under the Curve	S.E	p-Value	CI (95%)
Age	0.724	0.032	< 0.001	0.662 - 0.786
Systolic BP	0.661	0.035	< 0.001	0.592 - 0.730
PR	0.602	0.036	0.005	0.532 - 0.673
FBS	0.651	0.033	< 0.001	0.586 - 0.717
CR	0.536	0.034	0.329	0.469 - 0.602
TG	0.618	0.036	0.001	0.548 - 0.688
LDL	0.519	0.036	0.608	0.448 - 0.589
HDL	0.525	0.038	0.504	0.451 - 0.599
Bun	0.556	0.037	0.130	0.484 - 0.628
Hb	0.529	0.036	0.423	0.459 - 0.600
K	0.615	0.034	0.002	0.548 - 0.682
Na	0.562	0.035	0.093	0.494 - 0.629
WBC	0.533	0.036	0.367	0.463 - 0.603
Lymph	0.580	0.036	0.030	0.509 - 0.650
Neut	0.579	0.036	0.031	0.509 - 0.649
PLT	0.544	0.036	0.229	0.473 - 0.615
EF-TTE	0.687	0.034	< 0.001	0.620 - 0.754

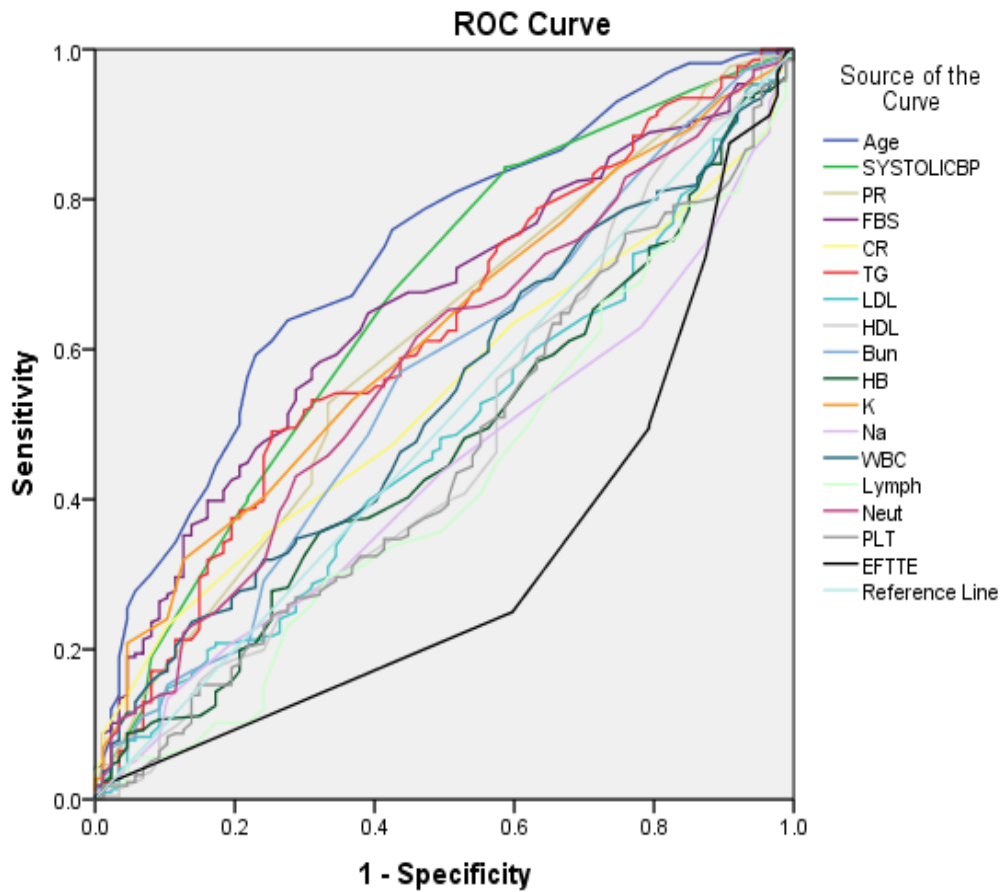
Table 24 presents the AUC result of the quantitative variables analyzed separately for ROC. It has been shown that 9 quantitative variables were statistically significant.

Those are; Age, Systolic BP, PR, FBS, TG, K, Lymph, Neut and EF-TTE.

The ROC shows that the best performance variable is the Age variable with an AUC of 0.724 (72.4%), closely followed by the EF-TTE variable with an AUC of 0.687 (68.7%). The lowest performing variable is the LDL variable with an AUC of 51.9% (Fig. 11).

Figure 11.

ROC Curve for the Quantitative Variables (Z-Alizadeh Sani dataset)



Diagonal segments are produced by ties.

Table 25.

Area Under the Curve for the ROC for the Multivariate Logistic Regression (Z-Alizadeh Sani dataset)

Area Under The Curve	S.E	p-value	CI (95%)
0.964	0.010	<0.001	0.945 - 0.983

In Table 25, Multivariate Logistic Regression analysis was performed with all variables and model performance was evaluated with ROC using probability values. The AUC result is 0.964, and the confidence interval (CI) is between 0.945 - 0.983. In Fig 12., the area under the curve is 0.964 (96.4%).

This area represents where LR correctly classified patients. The p-value of <0.001 evaluates the statistical importance of the curve.

Figure 12.

ROC curve for the Final Multivariate Logistic Regression Model (Z-Alizadeh Sani dataset)

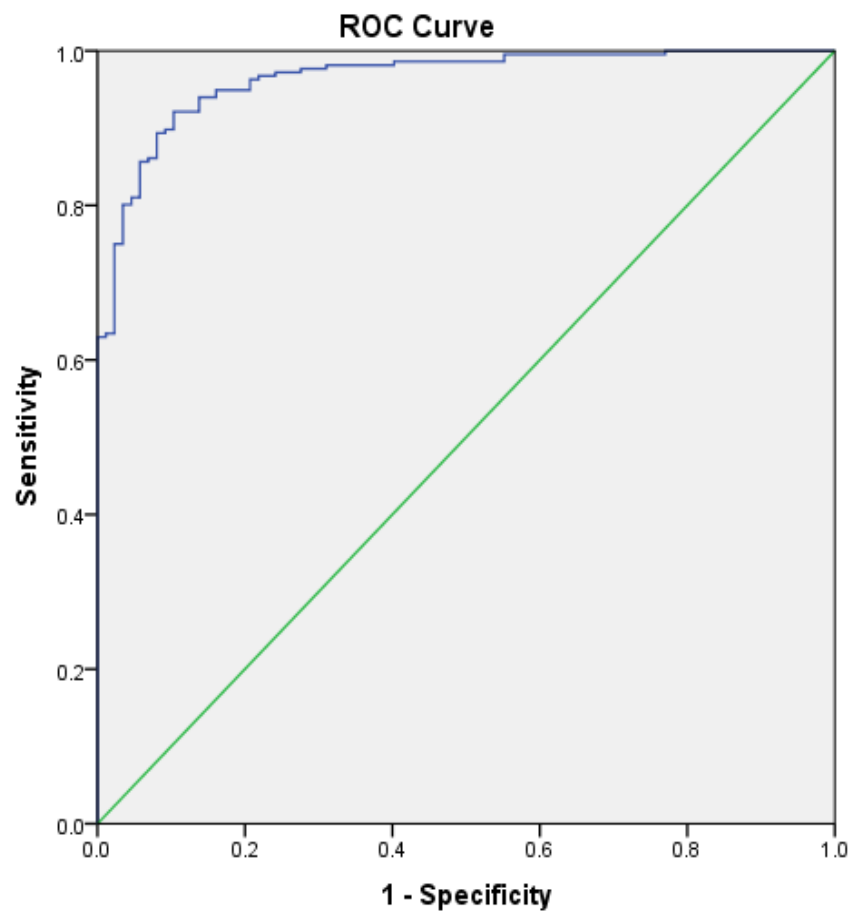


Table 26.

Multivariate Logistic Regression Equations Summary (Combined dataset)

Variable	β	S.E	Wald	Exp (β)	%95 C.I for		P
					Exp(β) Lower - Upper		
Age	0.064	0.012	30.537	1.066	1.042 - 1.091		< 0.001
Systolic BP	0.004	0.007	0.397	1.004	0.991 - 1.017		0.528
PR	0.021	0.009	5.176	1.021	1.003 - 1.039		0.023
FBS	0.003	0.003	1.388	1.003	0.998 - 1.009		0.239
CR	0.032	0.356	0.008	1.033	0.514 - 2.075		0.928
TG	0.003	0.001	4.173	1.003	1.000 - 1.006		0.041
LDL	0.007	0.003	5.071	1.007	1.001 - 1.013		0.024
HDL	0.004	0.009	0.204	1.004	0.986 - 1.023		0.652
BUN	-0.018	0.011	2.807	0.982	0.962 - 1.003		0.094
Hb	-0.087	0.078	1.239	0.917	0.787 - 1.068		0.266
K	0.359	0.250	2.056	1.432	0.877 - 2.338		0.152
Na	0.026	0.036	0.495	1.026	0.955 - 1.101		0.482
WBC	0.000	0.000	0.000	1.000	1.000 - 1.000		0.995
Lymph	-0.035	0.024	2.040	0.966	0.921 - 1.013		0.153
Neut	-0.037	0.024	2.355	0.963	0.918 - 1.010		0.125
PLT	-0.001	0.002	0.115	0.999	0.996 - 1.002		0.735
EF-TTE	-0.011	0.018	0.352	0.989	0.954 - 1.025		0.553
Gender	0.197	0.268	0.539	1.218	0.720 - 2.061		0.463
DM	0.420	0.301	1.940	1.522	0.843 - 2.747		0.164
HT	0.339	0.236	2.070	1.403	0.884 - 2.227		0.150
Smoking Status	0.665	0.247	7.227	1.945	1.198 - 3.159		0.007
FH	0.673	0.308	4.774	1.960	1.072 - 3.584		0.029
Edema	0.956	0.675	2.010	2.602	0.694 - 9.760		0.156

Table 26. (Continued)

Variable	β	S.E	Wald	Exp (β)	%95 C.I for Exp(β) Lower - Upper	P
Systolic Murmur	0.908	0.432	4.418	2.478	1.063 - 5.777	0.036
Chest Pain	2.941	0.242	147.514	18.926	11.775 - 30.419	<0.001
Dyspnea	0.535	0.267	3.999	1.707	1.011 - 2.882	0.046
LVH	0.536	0.330	2.630	1.709	0.894 - 3.266	0.105
Region RWMA	1.967	0.418	22.148	7.147	3.151 - 16.212	<0.001
VHD	-0.552	0.226	5.942	0.576	0.369 - 0.897	0.015

Table 26 shows the Multivariate Logistic Regression results. It has been shown that 4 quantitative variables and 7 qualitative variables were statistically significant.

These are; Age, PR, TG, LDL, Smoking Status, FH, Systolic Murmur, Chest Pain, Dyspnea, Region RWMA and VHD.

The Age variable was estimated from the model as 0.064.

The odds value was found to be 1.066. The probability of each unit increase being CAD increases 1.066 times.

The parameter value of PR variable was calculated as 0.021, Odds value was found as 1.021. For each unit increase of the variable, the probability of CAD increases by 1.021 times.

The parameter value of TG variable was calculated as 0.003, Odds value was found as 1.003. For each unit increase of the variable, the probability of CAD increases by 1.003 times.

The parameter value of LDL variable was calculated as 0.007, Odds value was found to be 1.007. For each unit increase of the variable, the probability of CAD increases by 1.007 times.

The parameter value of the Smoking Status variable was calculated as 0.665, Odds value was found as 1.945. The probability of CAD risk is 1.945 times higher than that of those who have a Smoking Status problem compared to the absent group.

The parameter value of FH variable was calculated as 0.673, Odds value was found as 1.960. The probability of CAD risk is 1.960 times higher than that of those who have a Smoking Status problem compared to the absent group.

The parameter value of the Systolic Murmur variable was calculated as 0.908, Odds value was found as 2.478. The probability of CAD risk is 2.478 times higher than that of those who have a Systolic Murmur problem compared to the absent group.

The parameter value of the Chest Pain variable was calculated as 2.941, Odds value was found as 18.926. The probability of CAD risk is 18.926 times higher than that of those who have a Chest Pain problem compared to the absent group.

The parameter value of the Dyspnea variable was calculated as 0.535, Odds value was found as 1.707. The probability of CAD risk is 1.707 times higher than that of those who have a Dyspnea problem compared to the absent group.

The parameter value of Region RWMA variable was calculated as 1.967, Odds value was found as 7.147. The probability of CAD risk is 7.147 times higher than that of those who have a Region RWMA problem compared to the absent group.

The parameter value of VHD variable was calculated as -0.552, Odds value was found as 0.576. The probability of CAD risk is 0.576 times lower than that of those who have a VHD problem compared to the absent group.

Table 27.

Omnibus tests of Model Coefficients for the Multivariate Logistic Regression
(Combined dataset)

	Chi-square	Df	Sig.
Step	362.669	29	<0.001
Block	362.669	29	<0.001
Model	362.669	29	<0.001

In Table 27, the Omnibus test result is based on Chi-square and is obtained according to the probability of real data being observed, assuming the model is correct. The result shows that the Multivariate Logistic Model is statistically significant because the p-value of 0.001 is less than the significance level of 0.05.

Table 28.

Model Brief (The Multivariate Logistic Regression comprising of all Logistic Regression Models, Combined dataset)

-2 Log-likelihood	Cox & Snell R Square	Nagelkerke R Square
624.481 ^a	.373	.518

The Multivariate Logistic Regression Model illustrates between 0.373 and 0.518 variations in the influences on the risk of CAD formation in patients.

Table 29.

Hosmer and Lemeshow Test to Assess the Model Fit (Combined dataset)

Chi-square	df	Sig.
4.314	8	.828

The Hosmer Lemeshow test is one of the methods of evaluating LR Model fit. The p-value was found to be 0.828. According to this result, it is seen that the model and data fit well, and the predictive power of the model is high since there is no significant difference between the expected value and the observed value.

Table 30.

Classification Table for the Multivariate Logistic Regression Model (Combined dataset)

Observed		Predicted		
		CAD		Percentage
		Absent	Present	Correct
CAD	Absent	187	70	72.8
	Present	66	455	87.3
Overall				82.5
Percentage				

a. The cut value is 0.5

The classification Table 30 describes well the model categorizes the dependent results. 70 of absent patients were incorrectly separated as CAD by the model and 66 CAD patients designated as absent group. The cases of the study have been 82.5% classified correctly by the model.

Table 31.

Area Under the Curve for the ROC for the Multivariate Logistic Regression
(Combined dataset)

Area Under The Curve	S.E	p-value	CI (95%)
0.882	0.013	<0.001	0.857 - 0.907

In Table 31, Multivariate Logistic Regression analysis was performed with all variables and model performance was evaluated with ROC using probability values. The AUC result is 0.882 and the confidence interval (CI) is between 0.857 - 0.907.

In Fig 13., the area under the curve is 0.882 (88.2%). This area represents where LR correctly classified patients. The p-value of <0.001 evaluate the statistical importance of the curve.

Figure 13.

ROC curve for the Final Multivariate Logistic Regression Model (combined dataset)

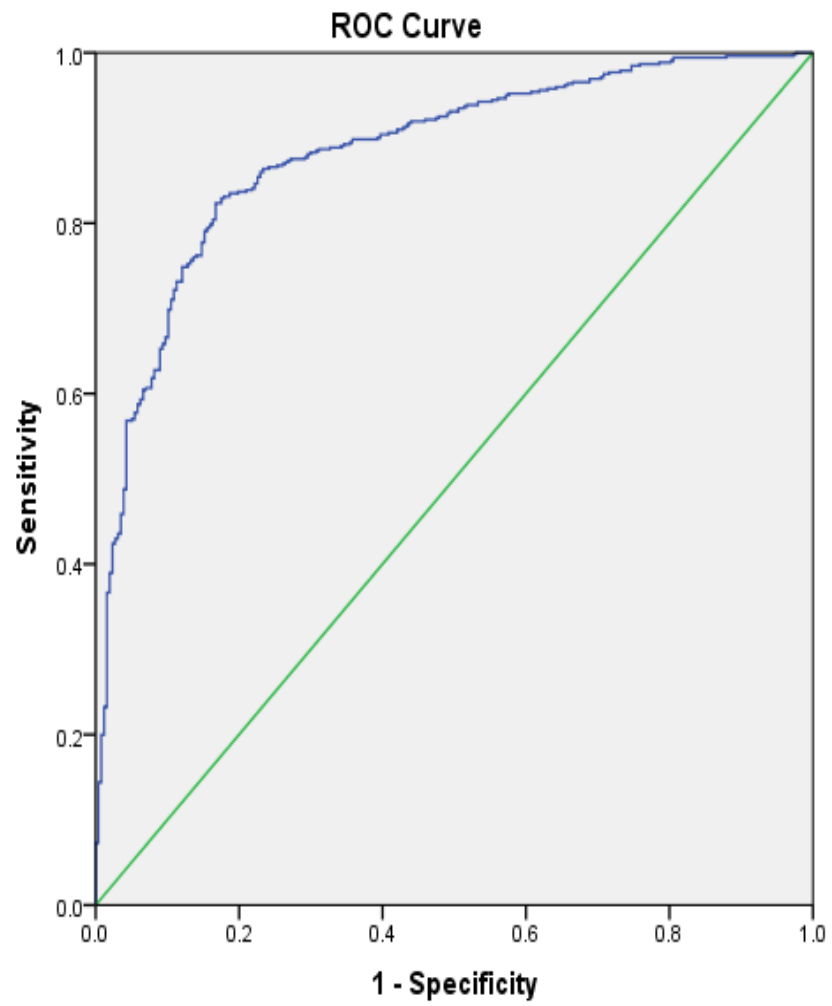


Table 32.

Machine Learning Random Sampling Results for Step-1

	Classifier	AUC	CA	F1	Precision	Recall
NEU Hospital Dataset (475) (a)	kNN	0.527	0.567	0.678	0.649	0.709
	SVM	0.811	0.811	0.857	0.834	0.882
	RF	0.780	0.738	0.805	0.773	0.839
	ANN	0.798	0.754	0.813	0.794	0.834
	Naïve Bayes	0.758	0.710	0.772	0.782	0.762
	LR	0.813	0.765	0.820	0.807	0.834
Z-Alizadeh Sani Dataset (303) (b)	kNN	0.468	0.647	0.770	0.718	0.830
	SVM	0.908	0.856	0.903	0.869	0.939
	RF	0.890	0.832	0.886	0.854	0.922
	ANN	0.896	0.844	0.892	0.880	0.904
	Naïve Bayes	0.914	0.845	0.889	0.906	0.873
	LR	0.924	0.865	0.907	0.895	0.919
Combined Dataset (778) (c)	kNN	0.522	0.605	0.722	0.682	0.767
	SVM	0.826	0.786	0.847	0.813	0.833
	RF	0.815	0.776	0.839	0.807	0.875
	ANN	0.834	0.782	0.842	0.814	0.872
	Naïve Bayes	0.821	0.758	0.816	0.827	0.806
	LR	0.851	0.795	0.851	0.828	0.876

According to AUC results in NEU Hospital dataset LR (81.3%) gave the best results of classification whilst kNN (52.7%) gave the worst results. SVM gave the most accurate outcome out of the five classification results for NEU Hospital, while kNN algorithms gave the least accurate. The classification algorithms that were successful in Z-Alizadeh Sani dataset; LR, SVM and NB. According to the AUC results, LR classification is 92.4% successful, whilst it has a success rate of 86.5% from the CA results. LR (90.7%) and SVM (90.3%) performed in the F1 results. According to the precision results, NB is 90.6%, and LR is 89.5%. The SVM result was successful in the Recall results with 93.9%.

In NEU hospital dataset, the algorithm with the lowest classification success in Z-Alizadeh Sani dataset is kNN. The classification algorithms applied to the data set created by combining both data sets. In five measurements LR algorithm gave the best results while the kNN algorithm gave the worst result. As a result of the precision classification, LR showed 82.8% success, while NB showed success with 82.7%. Figure 14-19 shows the results of the ROC graphics below.

Figure 14 (a).

ROC for Table 32. (NEU Hospital dataset)

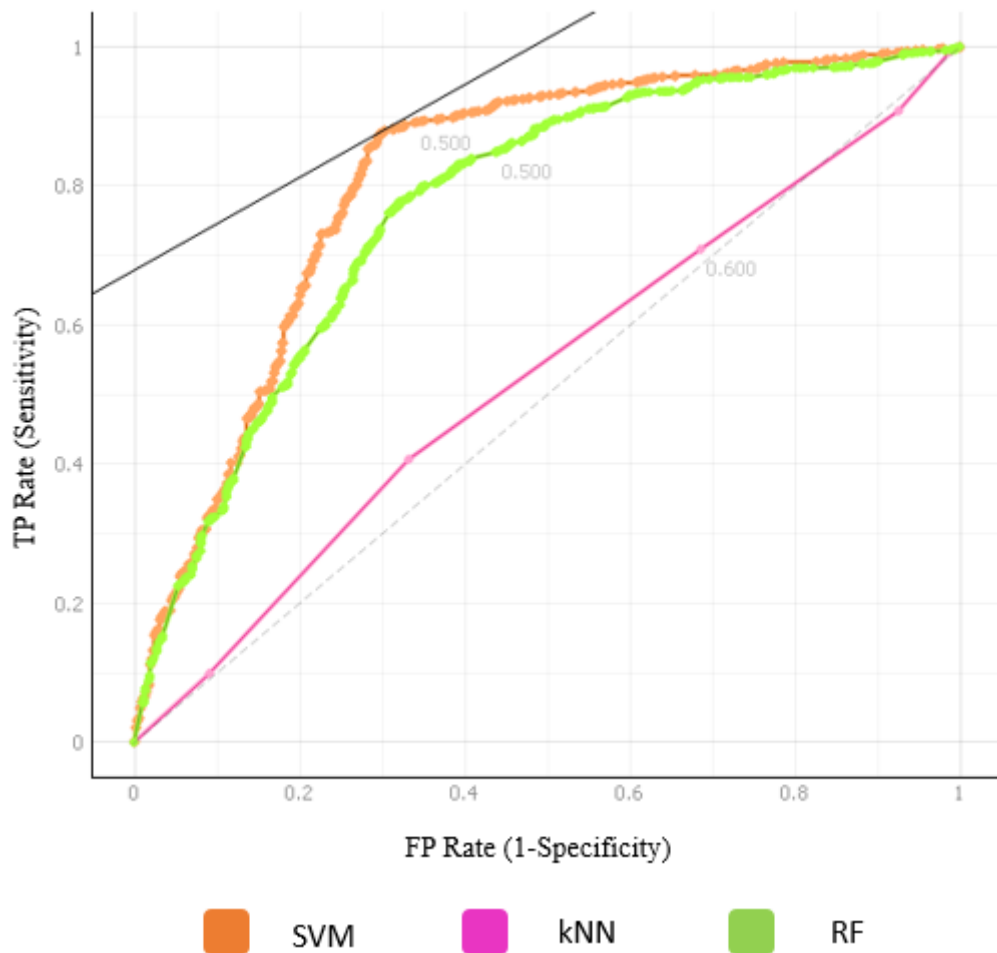


Figure 15 (a).

ROC for Table 32. (NEU Hospital dataset)

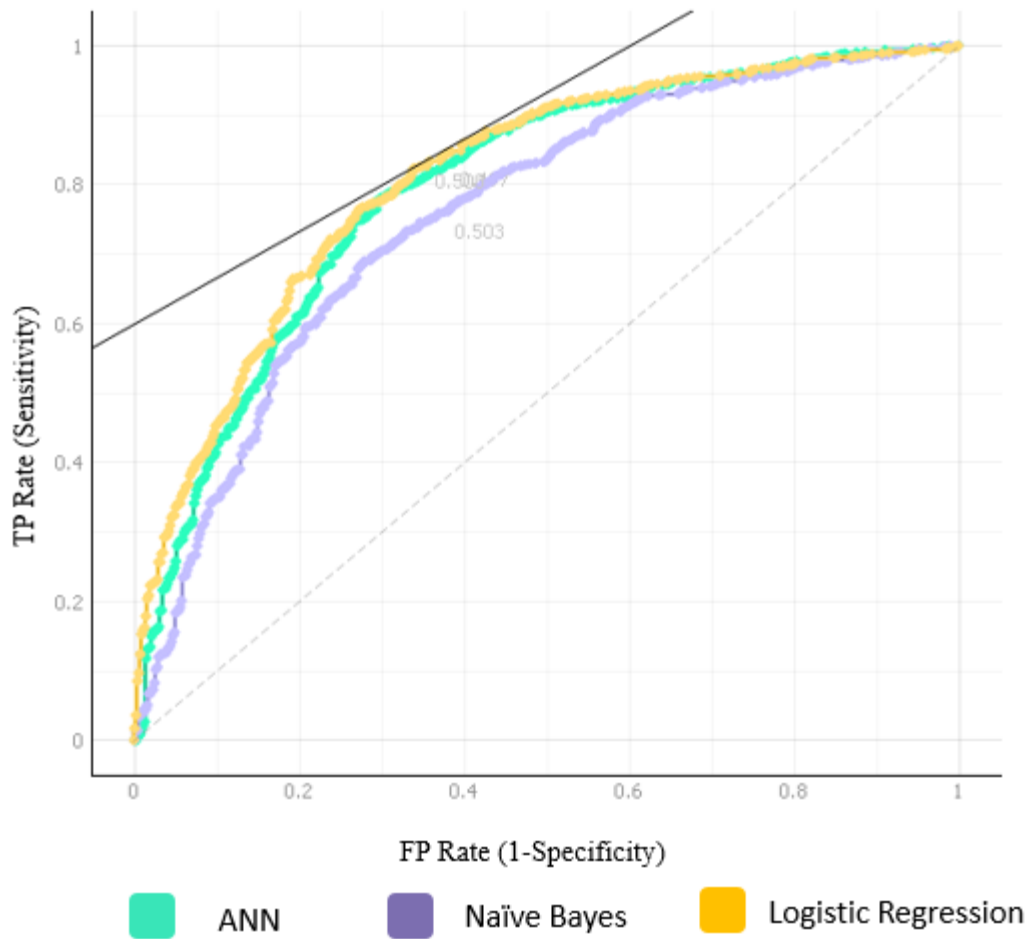


Figure 16 (b).

ROC for Table 32. (Z-Alizadeh Sani dataset)

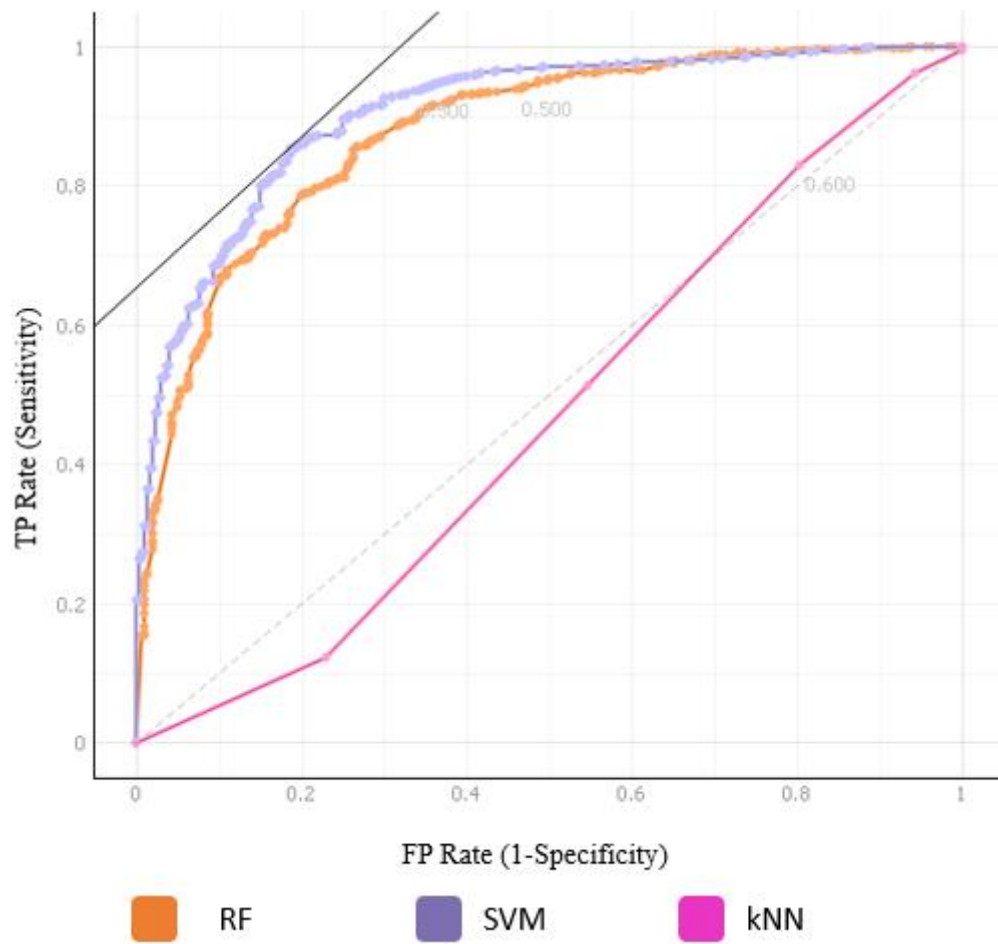


Figure 17 (b).

ROC for Table 32. (Z-Alizadeh Sani dataset)

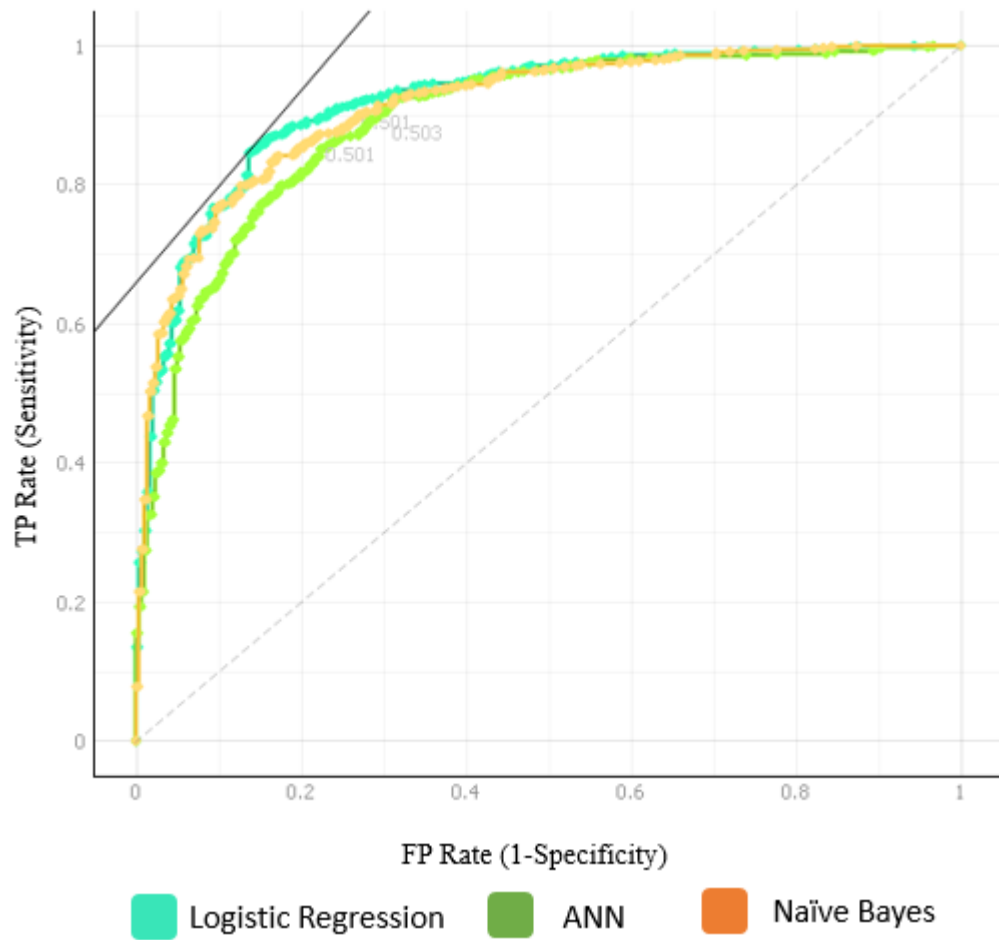


Figure 18 (c).

ROC for Table 32. (Combined Dataset)

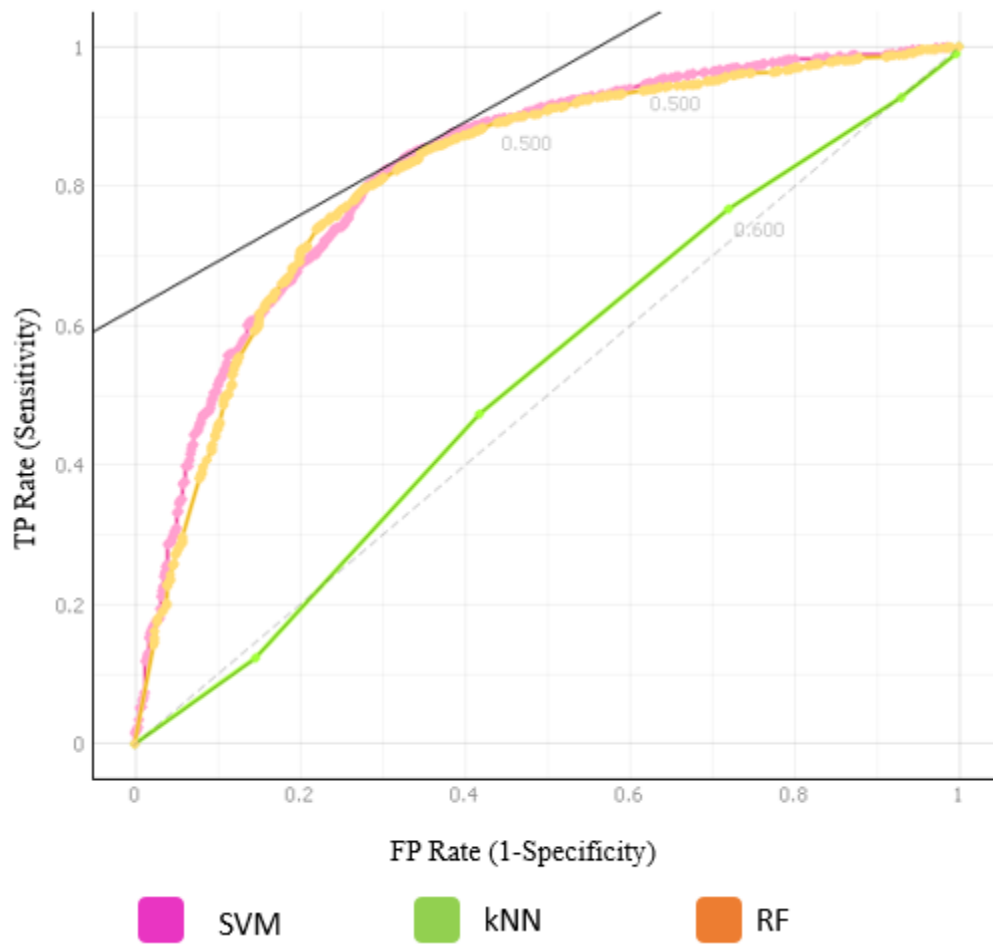


Figure 19 (c).

ROC for Table 32. (Combined Dataset)

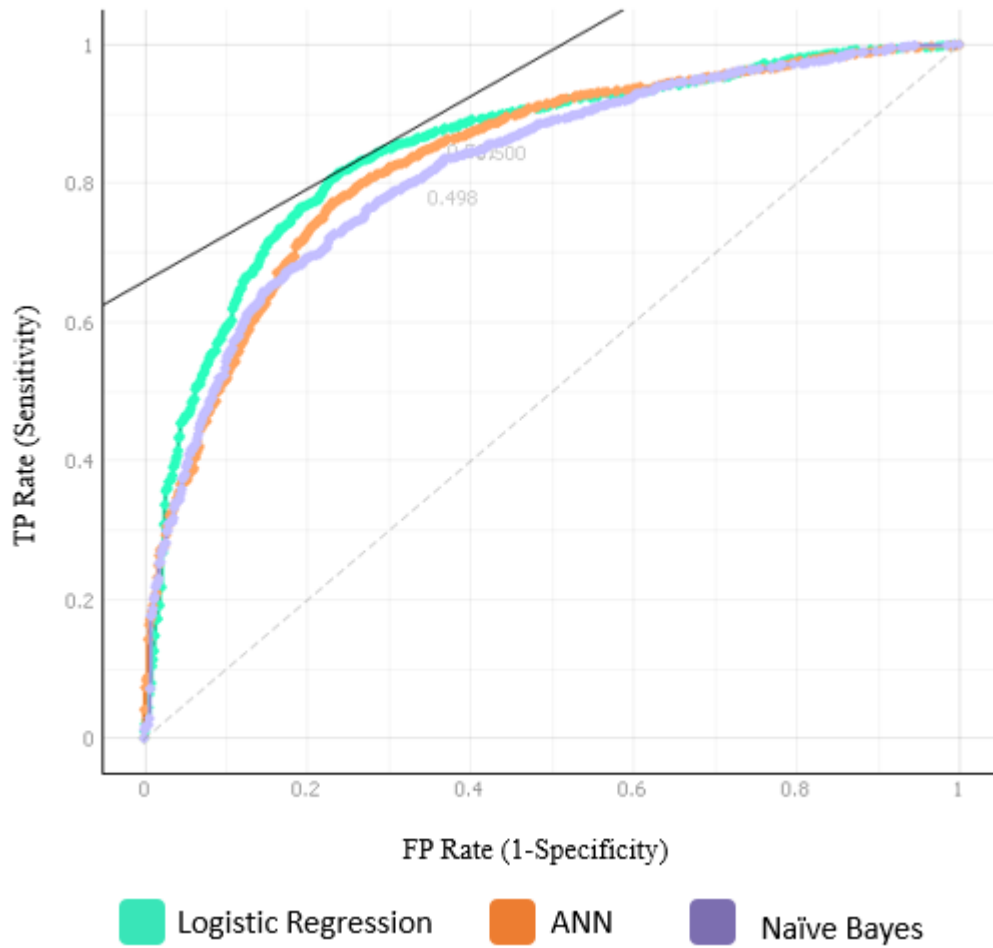


Table 33.

Machine Learning Classification Results for Step-2

Classifier	AUC	CA	F1	Precision	Recall
kNN	0.584	0.657	0.758	0.762	0.755
SVM	0.500	0.713	0.832	0.713	1.000
RF	0.795	0.776	0.858	0.780	0.954
ANN	0.498	0.287	-	-	-
Naïve Bayes	0.861	0.756	0.850	0.755	0.972
LR	0.479	0.287	-	-	-

The second step (Fig.3), NEU Hospital data set was defined as training data and Z-Alizadeh Sani data was defined as test data. According to the AUC classification results, NB (86.1%) gave the best result and LR (47.9%) gave the worst result. The ROC graph of the AUC results is given in Figure 20 (a) and Figure 21 (b).

Figure 20 (a).

ROC graph of ML algorithm (SVM, kNN, RF) results of Step-2.

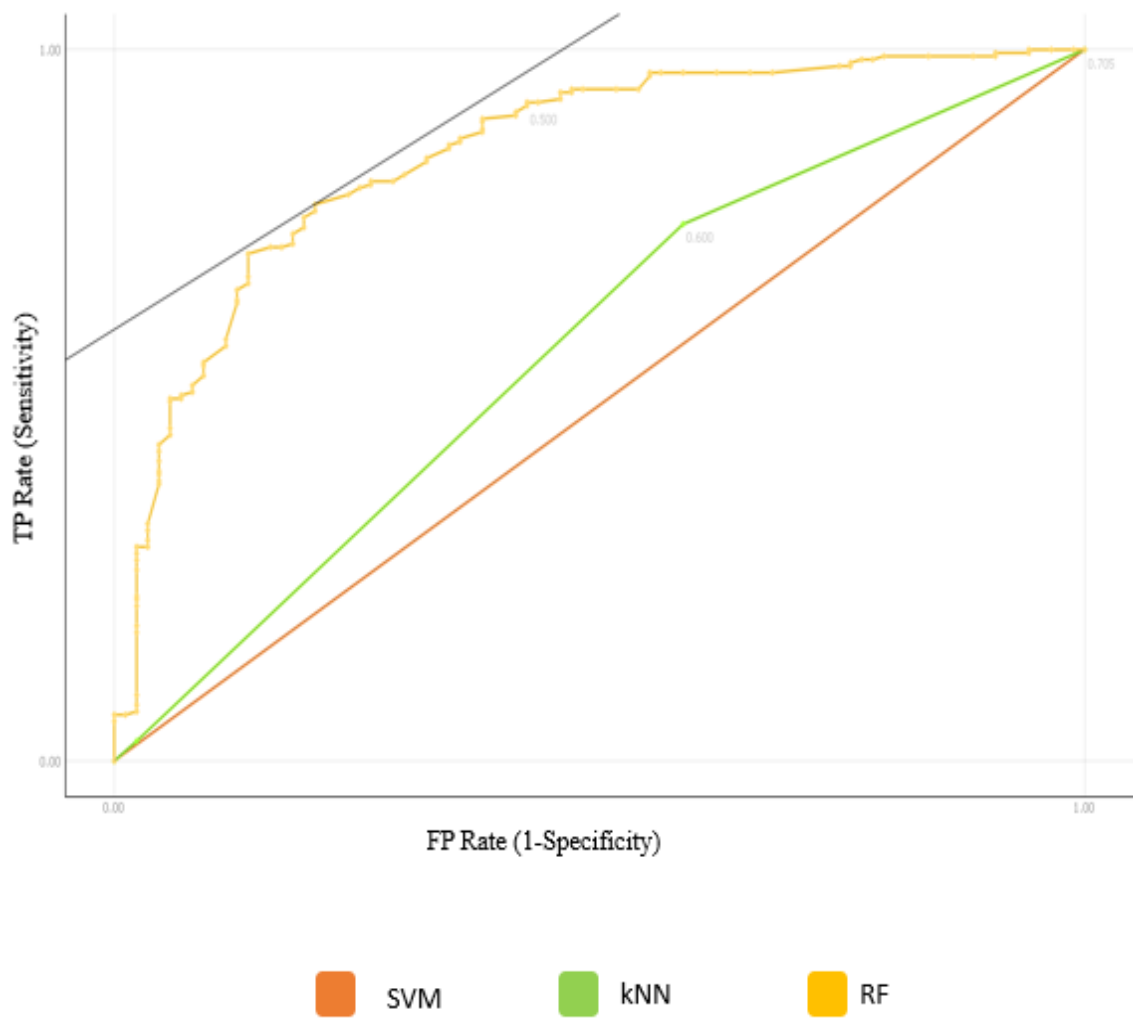


Figure 21 (b).

ROC graph of ML algorithm (Logistic Regression, Naïve Bayes, ANN) results of Step-2.

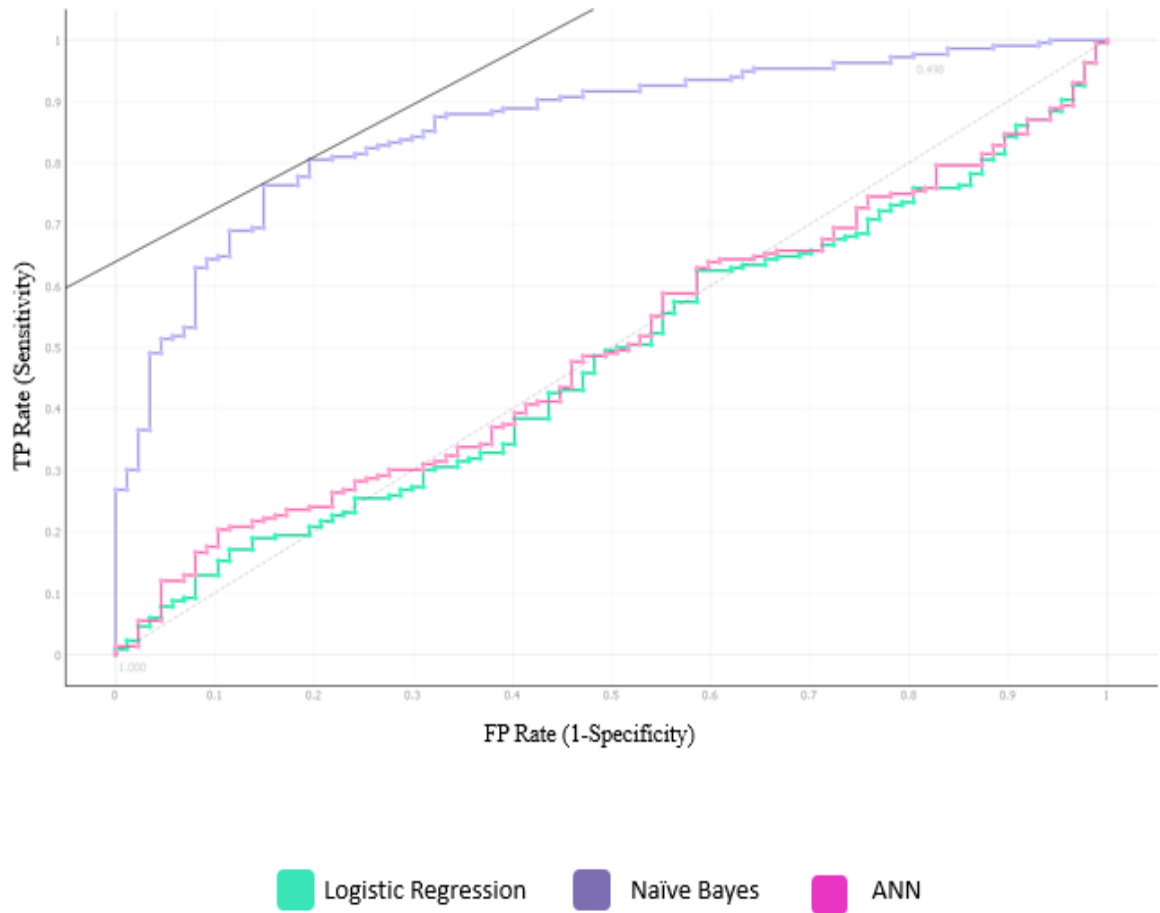


Table 34.

Machine Learning Classification Results for Step-3

Classifier	AUC	CA	F1	Precision	Recall
kNN	0.512	0.642	0.782	0.642	1.000
SVM	0.763	0.716	0.810	0.709	0.944
RF	0.777	0.737	0.786	0.824	0.751
ANN	0.761	0.718	0.772	0.802	0.744
Naïve Bayes	0.729	0.686	0.749	0.771	0.728
LR	0.752	0.716	0.763	0.822	0.711

The third stage (Fig.3), Z-Alizadeh Sani data set was defined as training data and NEU Hospital data was defined as test data. According to the AUC classification results, RF (77.7%) gave the best results, and kNN (51.2%) gave the worst results. The ROC graph of the AUC results is given in Figure 22 (a) and Figure 23 (b).

Figure 22 (a).

ROC graph of ML algorithm (kNN, RF, SVM) results of Step-3.

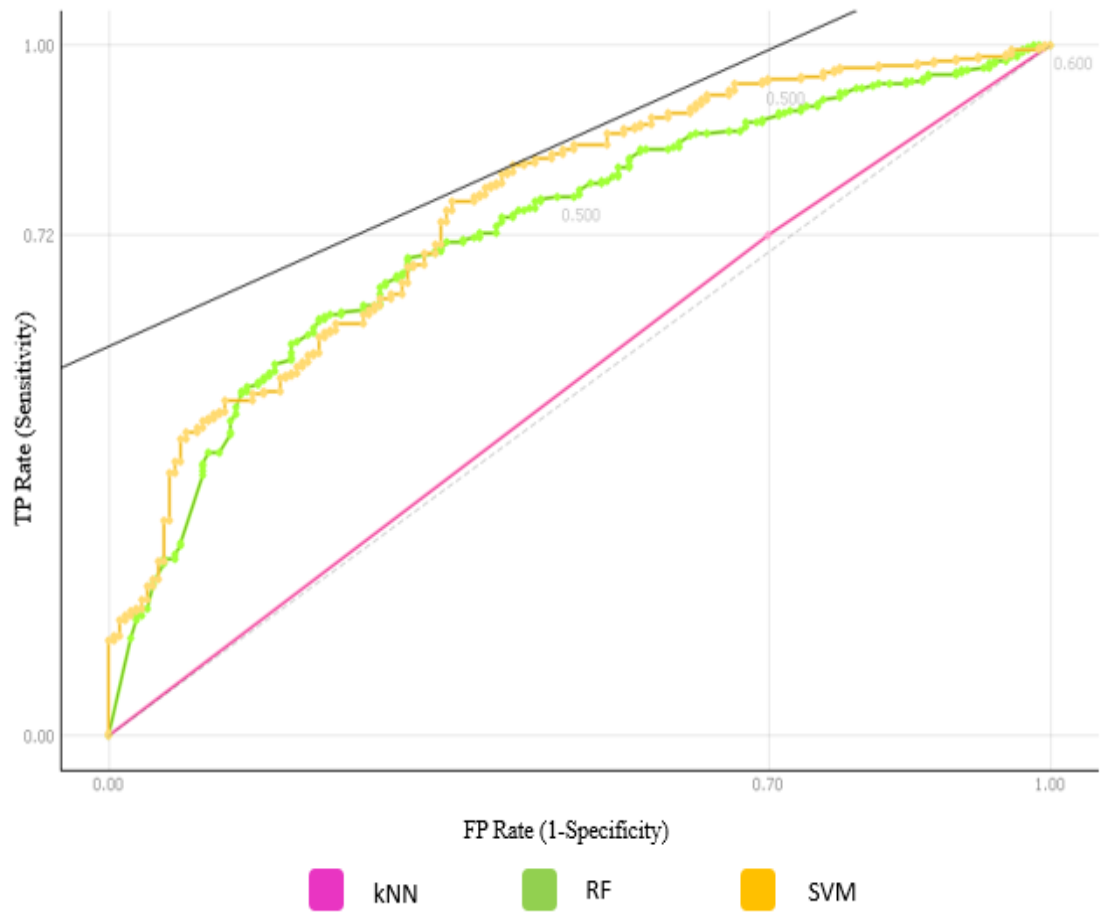
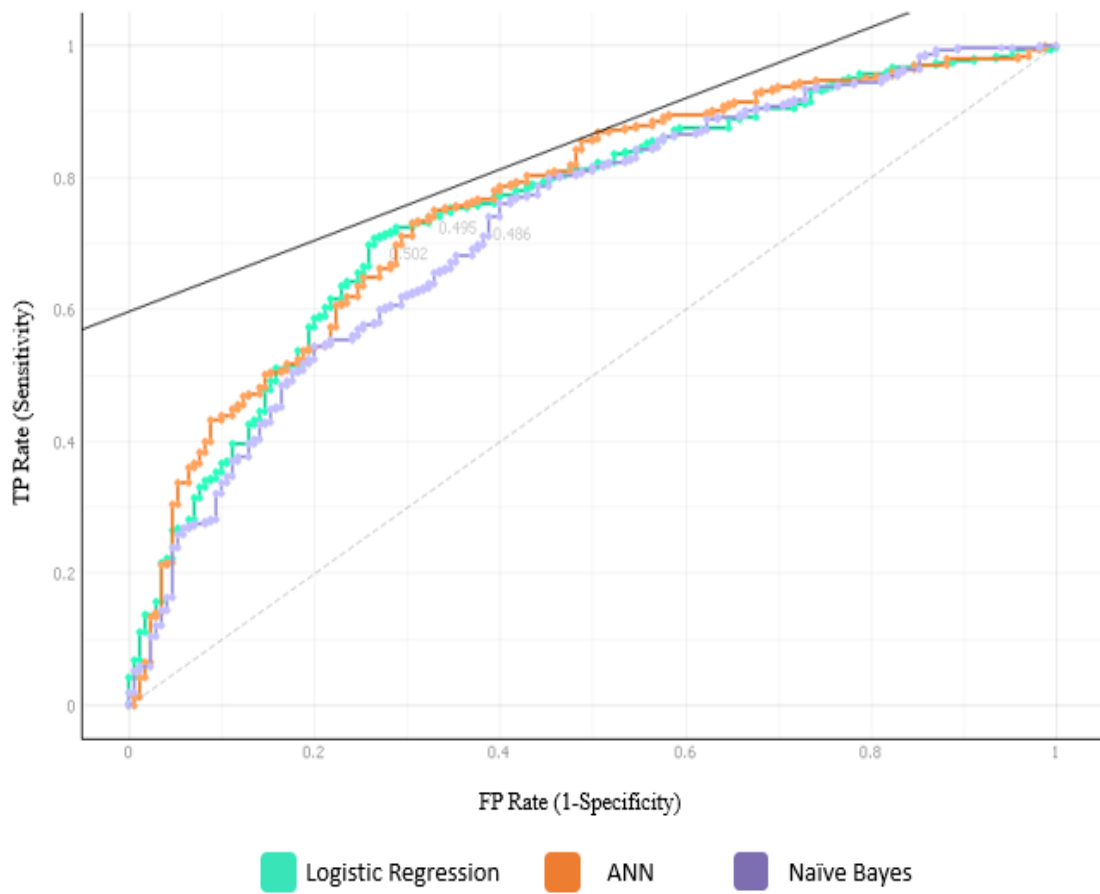


Figure 23 (b).

ROC graph of ML algorithm (Logistic Regression, ANN, Naïve Bayes) results of Step-3.



CHAPTER V

Discussion

Numbers of people are affected by heart disease, which is a common and serious health condition. It's a major health problem with a high death rate, especially among middle-aged and senior citizens. It deposits itself on the inner surfaces of the vessels that supply blood to the heart. A blood clot in the coronary artery is the most common cause of cardiac arrest (Shaima et al., 2016).

The purpose of this research was to test principles that justified the presence of CAD using both traditional statistical approaches and machine learning algorithms. The idea was to employ approaches to use data as training and testing datasets, to identify rules, and compare the results of various statistical and data ML algorithms. To assess the validity of identified rules across diverse data sets, validation approaches were used. It was to investigate to what extent the variables in the data set are successful in determining the dependent variable, and how precise the rule learned in a dataset with ML algorithms makes when applied to an independent and different dataset. Both datasets have suitable variables for this study. Multiple statistical analyse tests were used in the research, such as; Descriptive statistics, Mann Withney U test, Chi-Squared test, Bivariate Logistic Regression, Multivariate Logistic Regression, and ROC followed by ML algorithms; kNN, SVM, RF, ANN, Naïve Bayes and Logistic Regression as a part of the second step.

Dahal and Gautam (2020), obtained the results of LR 90.32% and SVM 88.68% following the AUC result in the classification. According to Kolukisa et. Al. (2020), the AUC results in both data sets, the ML algorithm that makes the best classification is the LR algorithm with 90.6% and 92.0%. Moreover, the research conducted by M. d. Idris (2020), showed the success of three different ML algorithms; LR 95.5%, ANN 96.6%, and kNN 96.6%. In addition, according to the study held by Nazli, et. al. (2020), the results of Accuracy, Recall, Specificity and Precision, Multilayer Perceptron methods have been the most successful. Another algorithm that achieved the closest result is SVM.

Whereas Jinjri et al. (2021) pointed out that Recall, LR was the most successful with 67.99%, thus, SVM was successful again in Precision with 77.35% and accuracy with 72.66%. Intercalarily the ROC results of Muhammad et al. (2021)'s study suggested that the most successful algorithm was RF with 92.20%. A result of the research approach by Dwivedi (2016), shows that LR achieved a high success rate of 89% in sensitivity. The results of the research executed by Tasnim & Habiba, (2021) illustrate that per classification, the RF algorithm estimated heart disease with 92.85% in classification accuracy (CA).

In this research, ML algorithms were applied to each data set one by one with the Random Sampling method, LR was the algorithm that made the most successful classification according to AUC results in both datasets and combined datasets. According to the AUC result, the LR algorithm achieved classification success of 81.3% in the NEU data, 92.4% in the Z-Alizadeh Sani dataset and 85.1% in the combined dataset. Considering the result, it is seen that the LR algorithm has achieved high success, as in most of the other studies. This result is the biggest factor in increasing the reliability of the LR algorithm in general.

Indeed, the results of the study supported the aim and the expected hypothesis. Considering the results and findings of many important studies, we see that the Logistic Regression algorithm has significant success in classification. This shows that ML algorithms perform well in making predictions and noticing biases.

Table 35.

Classification success of the research

AUC	Step-1	Step-2	Step-3
Lower than 60%	kNN	kNN, SVM, ANN, LR	kNN
Higher than 75%	SVM, RF, ANN, Naïve Bayes,LR	RF, Naïve Bayes	SVM, RF, ANN, LR

At each stage of the study, the kNN algorithm failed to successfully classify individuals. The LR algorithm, which performed well in the first step (AUC ranging from 0.813 to 0.924), performed poorly in the second step calculations (AUC = 0.479).

Although the LR algorithm showed poor performance in step 2, it showed high success in two of the classification applied in 3 steps. Other reference studies used in this paper have also shown that the LR algorithm achieves high performance for individual classification. Researchers used Z-Alizadeh Sani data, which was made in 2020 and used in this study, was the most successful LR (90.32%) algorithm among the three algorithms they used in their classification study on a single data set (Dahal et al., 2020). Sametime in this study, in the first stage, it was seen that the LR (92.4%) algorithm obtained successful results on the same data set in the classification made one by one on each data set.

One of the distinguishing features of intelligence is the ability to learn from experience. When machines can identify patterns in data, they can use those patterns to generate insights or predictions about new data. This working principle is the basic idea behind machine learning technology. As a result, The findings, precision and usefulness of machine learning and deep learning algorithms are directly dependent on the relevance of the data they are trained on. ML algorithms, which give very good and reliable results, are very promising for the future and the development of artificial intelligence.

Limitation

There are some limitations to this research. The model's effect predictions are based on research. Data from two different geographies were used. kNN algorithm, which is known as lazy learner, gave bad results in general.

Z-Alizadeh Sani data is divided into two categories as; completely healthy and unhealthy people. On the other hand, NEU hospital data, the people who do not have CAD still have different health problems. As a result, they are prone to biases and confounding, which may have impacted our model's results.

The fact that the data applied for analysis in the thesis is a certain number undermines its reliability, but one of the biggest reasons for this is that the study was taken from the database of a private university hospital in a small island country.

CHAPTER VI

Conclusion and Recommendation

Nowadays, deaths due to heart problems and heart diseases are rapidly increasing. Scientists are constantly researching treatment methods and the factors that could cause this ailment.

Data mining is now a requirement, particularly in the health field, and data transforms into information using ML algorithms to predict the best results in terms of accuracy. The classification success of CAD patients is the target variable. It is illustrated by all ML algorithm results and applied in three steps. This is a significant challenge in the medical field which motivates researchers to work harder to develop ML methods and use information intelligently and extract the best knowledge. The standard models' outputs were assumed to be simple to understand and explain to non-machine learning readers. It has been observed that the algorithms applied for classification in ML and data mining programs gave very close results when the same data set is used, even in different programs. This study targets other researchers to direct them to make the right choices in the future. If we consider that artificial intelligence learns by an experience like the human brain, we can say that the number of data and variables affects the classification results. Increasing the number of data and variables, it can be ensured that ML algorithms can increase their experience on the subject and make a highly successful classification of the newly entered data. Looking at the analyzes applied in the thesis, it has been determined that data mining and artificial intelligence can play a great role in the diagnosis and treatment of diseases within the scope of strengthening the database. However, the limited number of data obtained does not give a definite result, and it may question the reliability of the test. In this case, it is necessary to train the artificial intelligence for a long time to strengthen the system.

The core concepts of machine learning are embodied in the ideas of classification, regression, and clustering. Machine learning algorithms are created to perform these tasks across diverse and large datasets. This research can be extended by concentrating on different ML algorithms and Artificial intelligence programs.

REFERENCES

- Chen, X., Fu, Y., Lin, J., Ji, Y., Fang, Y. and Wu, J., 2020. CVD Detection by Machine Learning with Coronary Bifurcation Features. *Applied Sciences*, 10(21), p.7656.
- The Society of Thoracic Surgeons. (2018). *CVD* [Image]. Retrieved 1 March 2022, from <https://ctsurgerypatients.org/adult-heart-disease/coronary-artery-disease#symptoms-of-coronary-artery-disease->.
- Akella, A. and Akella, S., 2021. Machine learning algorithms for predicting CVD: efforts toward an open-source solution. *Future Science OA*, 7(6).
- Alizadehsani, R., Hosseini, M., Khosravi, A., Khozeimeh, F., Roshanzamir, M., Sarrafzadegan, N. and Nahavandi, S., 2018. Non-invasive detection of CVD in high-risk patients based on the stenosis prediction of separate coronary arteries. *Computer Methods and Programs in Biomedicine*, 162, pp.119-127.
- Ayatollahi, H., Gholamhosseini, L. and Salehi, M., 2019. Predicting CVD: a comparison between two data mining algorithms. *BMC Public Health*, 19(1).
- Blaus, B. (2014). *How the coronary arteries get clogged up with plaque* [Image]. Retrieved 1 March 2022, from https://simple.wikipedia.org/wiki/Coronary_artery_disease#/media/File:Blausen_0259_CoronaryArteryDisease_02.png.
- Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. *Health And Technology*, 10(5), 1137-1144. DOI: 10.1007/s12553-020-00438-1
- Kutrani, H.; Eltalhi, S. Cardiac Catheterization Procedure Prediction Using Machine Learning, and Data Mining Techniques. 2019. Available online: [https://www.semanticscholar.org/paper/Cardiac-Catheterization-Procedure-Prediction-Using-Kutrani Eltalhi/763ac488da8a97c19170ecff36a2e8dbdffe64c6](https://www.semanticscholar.org/paper/Cardiac-Catheterization-Procedure-Prediction-Using-Kutrani%20Eltalhi/763ac488da8a97c19170ecff36a2e8dbdffe64c6)
- Shaima, C.; Moorthi, P.; Shaheen, N. CVDs: Traditional and non-traditional risk factors. *J. Med. Allied Sci.* 2016, 6, 46.

- Naushad, S.; Hussain, T.; Indumathi, B.; Samreen, K.; Alrokayan, S.; Kutala, V. Machine learning algorithm-based risk prediction model of CVD. *Mol. Biol. Rep.* 2018, 45, 901–910.
- Amin, M., Chiam, Y., & Varathan, K. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics And Informatics* 36, 82-93. <https://doi.org/10.1016/j.tele.2018.11.007>
- Cuvitoglu, A., & Isik, Z. (2018). Classification of CAD dataset by using principal component analysis and machine learning approaches. *2018 5Th International Conference On Electrical And Electronic Engineering (ICEEE)*. <https://doi.org/10.1109/iceee2.2018.8391358>
- Kim, H. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2), 152. <https://doi.org/10.5395/rde.2017.42.2.152>
- Zar, J.H. (1999) *Biostatistical Analysis*. (4th ed.) Pearson Prentice Hall, Upper Saddle River.
- Zar, J.H. (2010). *Biostatistical Analysis* (5th ed.). Pearson Prentice Hall, Upper Saddle River.
- Peng, C.-Y., J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1), 3-14. <http://dx.doi.org/10.1080/00220670209598786>.
- Usman, A. (2014). *Statistical methods for biometric and medical research*.
- G.Lebanon (2011). *Maximum Likelihood Estimation*.
- Ramachandran, K.M. and Tsokos, C.P. (2009) *Mathematical Statistics with Applications*. Academic Press
- Kodati, S., & Vivekanandam, R. (2018). Analysis of Heart Disease using in Data Mining Tools Orange and Weka. *Global Journal Of Computer Science And Technology: C Software & Data Mining*, 18(1). Retrieved 12 April 2022, from <https://computerresearch.org/index.php/computer/article/view/1663/1647>.

Nilsson, N., 2005. [online] Available at:

<<https://ojs.aaai.org/index.php/aimagazine/article/view/1850>> [Accessed 11 May 2022].

Kubat, M. (2021). *An Introduction to Machine Learning* (3rd ed., pp. 41-44, 117-121). Springer.,

C. Sammut and G. I. Webb, eds., *Encyclopedia of Machine Learning*. Springer, 2010.

Dwivedi, A., 2016. Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, 29(10), pp.685-693.

Ashraf, S.; Saleem, S.; Ahmed, T.; Aslam, Z.; Muhammad, D. Conversion of adverse data corpus to shrewd output using sampling metrics. *Vis. Comput. Ind. Biomed. Art* 2020, 3, 19.

L. Breiman. Random Forests. *Machine learning*, 45(1):5–32, 2001

Abdulqader, D.M.; Abdulazeez, A.M.; Zeebaree, D.Q. Machine Learning Supervised Algorithms of Gene Selection: A Review. *Technol. Rep. Kansai Univ.* 2020, 62.

Dahal, K. and Gautam, Y., 2020. Argumentative Comparative Analysis of Machine Learning on CVD. *Open Journal of Statistics*, 10(04), pp.694-705.

Kolukisa, B., Yavuz, L., Soran, A., Burcu, B., Tuncer, D., Onen, A., & Gungor, V. (2020). CVD Diagnosis Using Optimized Adaptive Ensemble Machine Learning Algorithm. *International Journal Of Bioscience, Biochemistry, And Bioinformatics*, 10(1), 58-65. <https://doi.org/10.17706/ijbbb.2020.10.1.58-65>

Martins, B., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2021). Data Mining for CVD Prediction. *Journal Of Medical Systems*, 45(1). <https://doi.org/10.1007/s10916-020-01682-8>

Md Idris, N., Chiam, Y., Varathan, K., Wan Ahmad, W., Chee, K., & Liew, Y. (2020). Feature selection and risk prediction for patients with CVD using data mining. *Medical & Biological Engineering & Computing*, 58(12), 3123-3140. <https://doi.org/10.1007/s11517-020-02268-9>.

- Nazlı, B., Yasemin, G. and Altural, H., 2020. Classification of CVD Using Different Machine Learning Algorithms. *International Journal of Education and Management Engineering*, 10(4), pp.1-7.
- Muhammad, L., Al-Shourbaji, I., Haruna, A., Mohammed, I., Ahmad, A., & Jibrin, M. (2021). Machine Learning Predictive Models for CVD. *SN Computer Science*, 2(5). doi: 10.1007/s42979-021-00731-4
- Muhammad, L., Abba Haruna, A., Mohammed, I., Abubakar, M., Badamasi, B., & Musa Amshi, J. (2019). Performance Evaluation of Classification Data Mining Algorithms on CVD Dataset. *2019 9Th International Conference On Computer And Knowledge Engineering (ICCCKE)*. <https://doi.org/10.1109/iccke48569.2019.8964703>
- Imamovic, D., Babovic, E., & Bijedic, N. (2020). Prediction of mortality in patients with CVD using data mining methods. doi: 10.1109/INFOTEH48170.2020.9066297
- Jinjri, W., Keikhosrokiani, P., & Abdullah, N. (2021). Machine Learning Algorithms for The Classification of CVD- A Comparative Study.
- Ricciardi, C., Valente, A., Edmund, K., Cantoni, V., Green, R., & Fiorillo, A. et al. (2020). Linear discriminant analysis and principal component analysis to predict CVD. *Health Informatics Journal*, 26(3), 2181-2192. <https://doi.org/10.1177/1460458219899210>
- Tasnim, F., & Habiba, S. (2021). A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection. *2021 2Nd International Conference On Robotics, Electrical And Signal Processing Techniques (ICREST)*. doi: 10.1109/icrest51555.2021.9331158
- UCI Machine Learning Repository: Z-Alizadeh Sani Data Set. Archive.ics.uci.edu. (2020). Retrieved, 2020, from <https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani>.
- Release Notes -- IBM SPSS Statistics 21.0. Ibm.com. (2022). Retrieved 12 April 2022, from <https://www.ibm.com/support/pages/release-notes-ibm-spss-statistics-210>.
- Bioinformatics Laboratory, U. (2022). Visual Programming. Retrieved 20 January 2021, from https://orangedatamining.com/home/visual-_programming/

APPENDICES

Appendix A
Ethical Approval Document



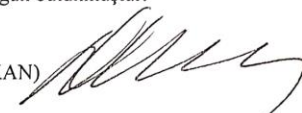

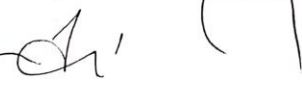




YAKIN DOĞU ÜNİVERSİTESİ
BİLİMSEL ARAŞTIRMALAR ETİK KURULU

EK:1001-2019

ARAŞTIRMA PROJESİ DEĞERLENDİRME RAPORU

Toplantı Tarihi : 21.11.2019
Toplantı No : 2019/74
Proje No :931

Yakin Doğu Üniversitesi Tıp Fakültesi öğretim üyelerinden Yrd. Doç. Dr. Özgür Tosun'un sorumlu araştırmacısı olduğu, YDU/2019/74-931 proje numaralı ve "Application of Data Mining Algorithms on Coronary Artery Disease (CAD) Data for Rule Discovery and Evaluation" başlıklı proje önerisi kurulumuzca değerlendirilmiş olup, etik olarak uygun bulunmuştur.

- | | | |
|-------------------------------------|----------|--|
| 1. Prof. Dr. Rüştü Onur | (BAŞKAN) |  |
| 2. Prof. Dr. Nerin Bahçeciler Önder | (ÜYE) | KATILMADI |
| 3. Prof. Dr. Tamer Yılmaz | (ÜYE) |  |
| 4. Prof. Dr. Şahan Saygı | (ÜYE) |  |
| 5. Prof. Dr. Şanda Çalı | (ÜYE) |  |
| 6. Prof. Dr. Nedim Çakır | (ÜYE) | KATILMADI |
| 7. Prof. Dr. Ümran Dal Yılmaz | (ÜYE) |  |
| 8. Doç. Dr. Nilüfer Galip Çelik | (ÜYE) | KATILMADI |
| 9. Doç. Dr. Emil Mammadov | (ÜYE) |  |
| 10. Doç. Dr. Mehtap Tınazlı | (ÜYE) |  |



NAER EAST UNIVERSITY
SCIENTIFIC RESEARCH ETHICS COMMITTEE

19.12.2019

Dear Assoc. Prof. Dr. Özgür TOSUN

Your application titled “**Application of Data Mining Algorithms on Coronary Artery Disease (CAD) Data for Rule Discovery and Evaluation**” with the application number NEU/2019/74/931 has been evaluated by the Scientific Research Ethics Committee and granted approval.

Prof. Dr. Rüştü ONLAR

Near East University

Scientific Research Ethics Committee Director

Appendix B

Signed Similarity Report

13 Temmuz

ORIGINALITY REPORT

11 %	9 %	7 %	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	docs.neu.edu.tr Internet Source	2 %
2	doaj.org Internet Source	1 %
3	www.tandfonline.com Internet Source	<1 %
4	Jikuo Wang, Changchun Liu, Liping Li, Wang Li, Lianke Yao, Han Li, Huan Zhang. "A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease", IEEE Access, 2020 Publication	<1 %
5	Liaqat Ali, Shafqat Ullah Khan, Muhammad Anwar, Muhammad Asif. "Early Detection of Heart Failure by Reducing the Time Complexity of the Machine Learning based Predictive Model", 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2019 Publication	<1 %
6	searchbusinessanalytics.techtarget.com Internet Source	<1 %

Assist. Prof. Dr. Özgür Tosun

Appendix C

Curriculum Vitae

PERSONAL INFORMATIONS

Surname, Name: Yuvalı, Meliz
 Date of Birth: 21 May 1992
 Place of Birth: Nicosia, Cyprus

EDUCATION

Degree	Department/Program	University	Year of Graduation
M.Sc.	Biostatistics	Near East University	2018
B.Sc.	Biomedical Engineering	Near East University	2016

Master Thesis Title: Application of Multivariate Statistical Methods and Model Evaluation with ROC Curve Analysis on Coronary Artery Disease (CAD) Data.

WORK EXPERIENCE

Title	Place	Year
Lecturer	NEU, Faculty of Engineering, Department of Biostatistics	2017-present
Research Assistant	NEU, Faculty of Engineering, Department of Materials Science and Nanotechnology	2016-2017

FOREIGN LANGUAGES

Fluent spoken and written English

PUBLICATIONS IN INTERNATIONAL REFEREED JOURNALS

American Journal of Biostatistics: General Bearing of Students with Sustainable Satisfaction in Higher Institution of Learning (03.10.2017)

American Journal of Biostatistics: The Extent of Community Awareness Using Natural Element's Instead of Using Chemical Drugs (24.10.2017)

Biometrics & Biostatistics International Journal: Influence of Residential Setting on Student Outcome (22.11.2017)

Biometrics & Biostatistics International Journal: Review of Prostatic Tumor Using Kaplan Meier and Cox Regression (06.12.2017)

Transylvanian Review Journal: Statistical Study of Epidemiological Profile of Pulmonary Tuberculosis at Bandudu Province from 2008 to 2016, in the Democratic Republic of the Congo. (2018)

Global Journal of Reproductive Medicine: Life Expectancy; Factors, Malaria the Most Common Disease Affecting Pregnant Women in Africa [Nigeria and Cameroon]. (21.06.2018)

Biometrics&Biostatistics International Journal: Choosing statistical tests for survival analysis (17.10.2018)

Annals of Biostatistics & Biometric Applications - ABBA: Determine Significant Factors Related to Malaria among Pregnant Women in Nigeria by Logistic Regression Analysis. (15.02.2019)

Annals of Biostatistics & Biometric Applications - ABBA: Application of Multivariate Statistical Methods of Patient Surviving ART Follow-up. (08.04.2019)

Real-Time Intelligence for Heterogeneous Networks: Classification of Solid Wastes Using CNN Method in IoT Het-Net Era. (03.09.2021) - DOI: 10.1007/978-3-030-75614-7_8

Mathematics-MDPI: Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. (20.01.2022)

BULLETING PRESENTED IN INTERNATIONAL ACADEMIC MEETINGS AND PUBLISHED IN PROCEEDINGS BOOK

ICSCCW 2019: 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019: Fuzzy

Ordination of Breast Tissue with Electrical Impedance Spectroscopy Measurements. (27.08.2019)

11th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions and Artificial Intelligence - ICSCCW-2021: Evaluation of HCV Infection Laboratory Test Results Using Machine Learning Methods (01.01.2022)