PREDICTING CARDIOVASCULAR DISEASE PATIENTS WITH MACHINE LEARNING: A COMPARATIVE ANALYSIS OF CLASSIFICATION MODELS

Dissertation Submitted
In Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE
in
MATHEMATICS

by

Aarushi Saini
(Roll No. 2K22/MSCMAT/01)

Diya Malhotra
(Roll No. 2K22/MSCMAT/12)

Under the Supervision of
Prof. Anjana Gupta
Delhi Technological University

Department of Applied Mathematics

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2024

ACKNOWLEDGEMENT

The successful completion of this dissertation required significant guidance and support from various individuals and we received assistance throughout the process. It gives immense pleasure to present our research work titled “Prediction of cardiovascular disease patients with Machine Learning: A Comparative Analysis of Classification Models”.

First and foremost, we are deeply grateful to our supervisor Prof. Anjana Gupta, for her helpful suggestions and ideas. Her encouragement and guidance were beneficial to us to explore and understand this topic. We express our gratitude to the college for providing necessary resources. Also, appreciate everyone who helped us throughout the process. We are also grateful to each other for our collaboration and coordination, as it was essential for timely completion of the project.

Sincere Thanks,

Aarushi Saini

Diya Malhotra

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

We Aarushi Saini (2K22/MSCMAT/01) and Diya Malhotra (2K22/MSCMAT/12) hereby certify that the work which is being presented in the thesis entitled "Predicting Cardiovascular Disease Patients with Machine Learning: A Comparative Analysis of Classification Models" in partial fulfilment of the requirements for the award of the degree of Master of Science submitted in the Department of Applied Mathematics, Delhi Technological University is an authentic record of our own work carried out during the period from **August 2023** to **April 2024** under the supervision of Prof. Anjana Gupta.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that Aarushi Saini (2k22/MSCMAT/12) and Diya Malhotra (2k22/MSCMAT/12) has carried out their search work presented in this thesis entitled “Predicting Cardiovascular Disease Patients with Machine Learning: A Comparative Analysis of Classification Models” for the award of the Degree of Master of Science from the Department of Mathematics, Delhi Technological University under my supervision. The thesis embodies results of original work, and studies are carried out by the students themselves and the contents of the thesis do not form the basis for the award of any degree to the candidate or to anybody else from this or any other University.

Signature

Prof. Anjana Gupta

Professor

Department of Applied Mathematics

Delhi Technological University

Date:

Place:

Prediction of cardiovascular disease patients with Machine Learning: A Comparative Analysis of Classification Models

ABSTRACT

Cardiovascular disease is a very serious health issue. So, in order to prevent its spread, we need to understand the reason behind its increase. Different factors like our lifestyle, our genes, the surroundings that we live in and so on all contribute to the risk of getting CVD. So, it is important to make positive changes in our day-to-day life which in the end will make us healthier. This research paper delves into understanding the importance of such factors. We have used classification models like Logistic Regression, Decision Tree Algorithm, Random Forest, KNN, Support Vector Machine and Naïve Bayes to make predictions regarding cardiovascular disease patients. We have used data from the UCI Repository that includes the features (predictor variables) such as age, BMI, gender, cholesterol, alcohol intake, and so on to determine the presence of cardiovascular disease patients (response variable). Different models have been used to find out which model works best and we have done this by estimating various metrics that are essential for the assessment of model performance such as accuracy, precision, recall, etc. The Support Vector Machine model had the highest accuracy, Roc-Auc. So, this shows that the Support Vector Machine (SVM) so far is the best model for making predictions regarding CVD.

Keywords: Cardiovascular Diseases, Machine Learning, Logistic Regression, SVM, KNN, Decision Tree, Random Forest, Naive Bayes.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 COMMENCEMENT

The heart is a vital organ that is present in both human and animal bodies and is crucial for survival. It circulates blood throughout the body via the circulatory system. Cardiovascular disease encompasses a range of conditions affecting the heart and blood vessels. These conditions can influence the heart or blood vessels in various ways. An individual may experience symptoms (showing physical signs of the disease) or remain asymptomatic (not experiencing any noticeable symptoms).

As reported by the World Health Organization (WHO), cardiovascular diseases (CVDs) accounted for the most deaths globally in 2019, representing 32% of all global deaths. The majority of the fatalities around 85% resulted from heart attacks and strokes.[1]. Populations that are growing older and becoming more numerous are making matters more complicated; in 2030, 22.2 million deaths from CVD are predicted [2]. Around 75% of cardiovascular fatalities happen in low- and middle-income nations. Heart disease is caused by multiple risk factors, including lack of physical activity, poor diet, and excessive use of alcohol and tobacco. Most cardiovascular diseases can be averted by adopting a healthy daily lifestyle, which includes limiting salt in the diet, eating more vegetables and fruits, engaging in regular physical activity, and quitting smoking and excessive alcohol consumption [3]. Early detection of cardiovascular disease is essential for the timely initiation of medication and counseling-based treatment.

Currently, computers are being used in various domains including medicine where they enhance decision-support systems. Machine Learning (ML) is becoming more used in a variety of disciplines, including disease diagnosis in health care. Many researchers and healthcare professionals demonstrate the perspective of machine learning-based disease diagnostics (MLBDD), which is both affordable and time-efficient [4]. Healthcare data, including as pictures (X-ray, MRI) and tabular data (patient conditions, age so on), are utilized in developing MLBDD systems. ML is an instance of artificial intelligence (AI) that is becoming popular in the domain of cardiovascular health. It refers to how computers handle data and make decisions or classifications, with or without human intervention. ML framework is based on models that take input data and forecast results leveraging a combination of mathematical optimization and statistical examination. Various ML algorithms have been utilized in daily activities. Numerous causes of heart disease can complicate prediction.

Heart disease can be detected using machine learning techniques. Various techniques are used to reliably and precisely predict heart disease. Feature selection methods such as Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) and many more have been developed to be equally effective in disease prediction [5]. Our goal is to advance the creation of more accurate and efficient models for predicting heart disease by investigating various methodologies.

The remaining section of this paper is structured as follows: chapter 1 elaborates on introduction. chapter 2 elaborates on literature review whereas chapter 3 explains the approach in depth, including data collection, data preparation, and the use of classification models. chapter 4 compares the metrics of different classifiers using experimental results and performance evaluation indicators. Finally, chapter 5 summarizes the important results and their importance in heart disease prediction.

CHAPTER 2 LITERATURE REVIEW

2.1 Related Work

Following section, our focus will be on exploring previously conducted or related research on this area.

The authors of [7] proposed a method for predicting coronary heart disease (CHD) using Logistic Regression Model, demonstrating its suitability for binary as well as multi-class classification. The study is based on training dataset containing 4000 instances along with 15 different attributes. The research aims to improve CHD risk prediction and find out the risk factors to get an estimate of overall risk using Logistic Regression.

P. H. Swain and H. Hauska (1977) [9], following authors researches on concepts of multi-stage classification strategy using decision tree classifiers and with that it compares both manual and optimized design methods hence showing that optimized design yields better results. The research paper shows decision tree classifiers are 30-50% more efficient while maintaining accuracy.

Madhumita Pal and Smita Parija (2021) [10], following authors presents a paper on a random forest data mining algorithm for predicting heart disease, achieving a sensitivity of 90.6%, specificity of 82.7, and accuracy of 86.9%. The model yielded accuracy of 86.9% and diagnosis rate of 93.3%. The system can be applied to other diseases using machine learning algorithms like Naïve Bayes, DT, K-NN, LR, and fuzzy logic for improved predictions.

Authors of [11] aims to find machine learning based cardiovascular disease prediction system. A lot of algorithms such as Random Tree, Naïve Bayes, Linear Regression, etc. have been used to predict cardiovascular disease. The research showed that Random Tree model outperformed all other models having efficiency of 100% with MAE of 0.0011 and RMSE of 0.0231.

Authors of [12] proposed hybrid system, developed using Python and tested on the Cleveland heart disease dataset at UCI, achieved an accuracy of 86.6%, outperforming existing heart disease prediction systems. Various ML algorithms were used to process raw data, providing a novel approach to heart disease prediction. Early detection and early prevention can significantly reduce mortality rates. Further research is needed to apply the hybrid HRFLM approach to real-world datasets, combining the characteristics of Random Forest (RF). Further extension is recommended for real-world datasets.

Aditya Duneja & Thendral Puyalnithi [13], following authors present an intuitive and easy-to-understand modification to KNN, which slightly improves its classification accuracy in some datasets but drastically in others. Out of 16 evaluated datasets, 8 showed a 5% increase in model’s accuracy, while 4 showed an average 3% increase. The worst-case accuracy of the suggested approach is the same as that of the traditional

KNN, suggesting that it is as accurate as or more accurate than the original KNN algorithm.

The authors [15] have developed a Naive Bayes algorithm model for predicting heart disease using clinical data from a Chennai diabetic research institute. The model scored 71% precision, 74% recall, and 71.2% F-measure, accurately classifying 74% of input examples. Regardless of the fact that the variables utilized in this method are not directly predicted by certain factors.
This method outperforms further comparable techniques in the literature.

Authors [16] proposed model on how machine learning algorithms can be used in the diagnosis of heart disease. Here they have used various ML algorithms like SVM, NB, DT, Bagging &boosting, and RF for the prediction of heart disease. The models were trained and tested on UCI repository data including patients with heart disease and those who are not affected by the disease. Thus, random forest algorithm achieved 89.4% accuracy, proving cost-effective diagnosis on large datasets.

Uma N Dulhare [17] author proposed model improves Naive Bayes classifier accuracy using particle swarm optimization for feature subset selection, achieving similar or better classification performance. The predictive model achieved Naïve Bayes accuracy of 79.12%, while the Naive Bayes and PSO model achieved 87.91%. This algorithm maximizes classification performance while minimizing feature number, enabling feature subset selection with fewer features and increasing classification performance compared to using all features in a dataset.

CHAPTER 3 RESEARCH METHODOLOGY

In this section, roadmap of various methods and the techniques that are used to accomplish our research objectives of evaluating and comparing various classification models in the context of cardiovascular diseases are outlined.

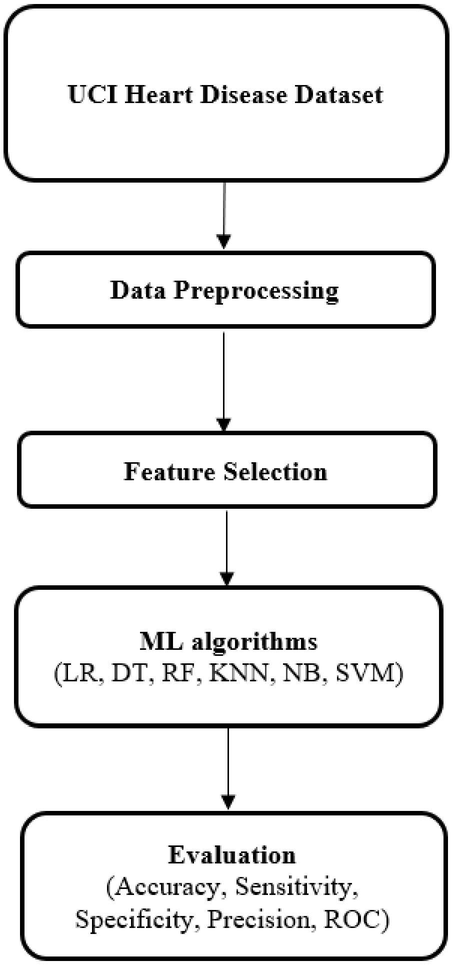


Figure 3.1: Research Methodology

3.1 Data: The aim of the research paper is to predict the people suffering from cardiovascular disease. For that we have taken the dataset regarding cardiovascular disease from UCI Repository and it consists of 70000 records, each representing a patient's information. This dataset includes data on observable traits and risk factors related to heart attacks. Figure I.1 in the Appendix provides an overview of the dataset.

3.2 Data Preprocessing: After the data has been collected, Preprocessing was done that involves dealing with missing values, correcting incorrect records, inconsistencies and so on. We made use of some statistical methods such as mean to fill in any missing information and also removed duplicated values. This was done to ensure that our research is based on complete and accurate dataset.

3.3 Feature Selection: Refers to a process where the most important features are selected from the original dataset and neglecting irrelevant ones. This process helps in improving the performance of model as too many features may lead to overfitting, complexity and so on.

3.4 Classification: Classification is the process of organizing a given dataset into distinct categories, a process which is applicable to both structured and unstructured data. The initial step in classification is predicting the category of the data points given. There are various methods in classification:

3.4.1 Logistic Regression: Logistic Regression (LR) [6] is supervised machine learning algorithm which aims at figuring out the chances of something happening based on different factors. It’s like trying to predict if the person will have the cardiovascular disease or not. The LR technique is implemented in this study to classify heart disease. To boost performance, pre-processing of the data is performed, such as cleaning and detecting missing values. The most important part is feature selection, which improves algorithm accuracy and especially focuses on behaviour. Logistic regression's behaviour is that as training grows, so does prediction accuracy. The dependent variable in logistic regression is binary. It is used for prediction and estimating probability of success.
The sigmoid function is employed as a cost function to limit the logistic regression hypothesis between 0 and 1 (squashing), i.e., 0 h (x) 1 [7]. We classified and predicted the CVD patients in the machine learning LR using the function [5]:

$$f(x) = \begin{cases} 1, & CVD\ present \\ 0, & CVD\ absent \end{cases}$$

In LR, the cost function is referred to as:

$$\log \frac{P}{1-P} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_P x_p \dots .$$

3.4.2 Decision Tree: - The decision tree (DT) stands out as the most important predictive modelling and classification method within learning algorithms, extensively applied in practical approaches in supervised learning techniques [8, 9]. With the help of Decision Tree algorithm, a tree like structure is created which includes internal node (decisions made on the basis of features available), branches (outcomes of the decision) and leaf node (represents final decision).
While making predictions regarding cardiovascular disease through decision tree algorithm we first divided the dataset into training and testing datasets and then by calculating entropy which determines which feature to be used at various nodes. This process continues until a specific criterion is met i.e. all the decision nodes have been made.

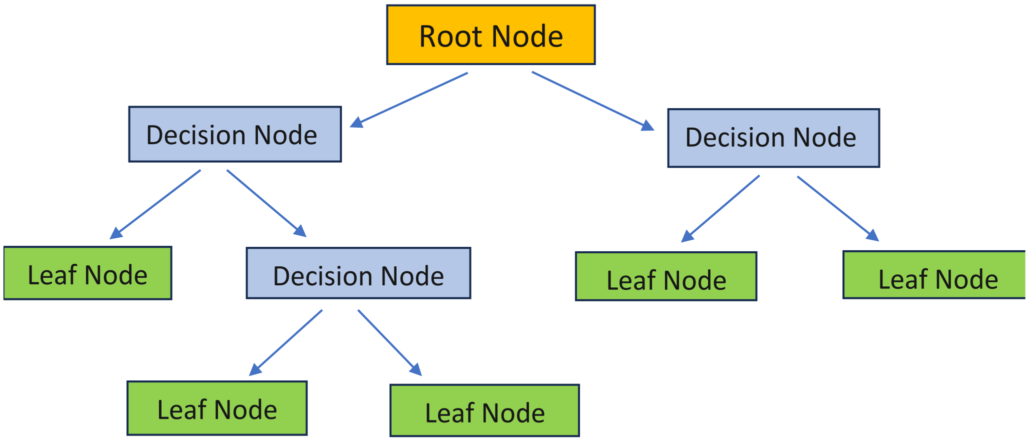


Figure 3.2: Decision Tree

3.4.3 Random Forest: The Random Forest (RF) is a supervised machine learning technique that uses decision trees to perform classification and prediction by averaging several independent base models [10]. Random Forest creates a collection of decision trees by randomly selecting a subset. It is essentially a set of decision trees (DT) drawn from a randomly chosen subset of the training set, and it then aggregates the votes from the many decision trees to determine the final prediction [11]. As the number of trees in the model increases, accuracy rises while risk of overfitting diminishes. The working of Random Forest algorithm is demonstrated in the below diagram [12]:

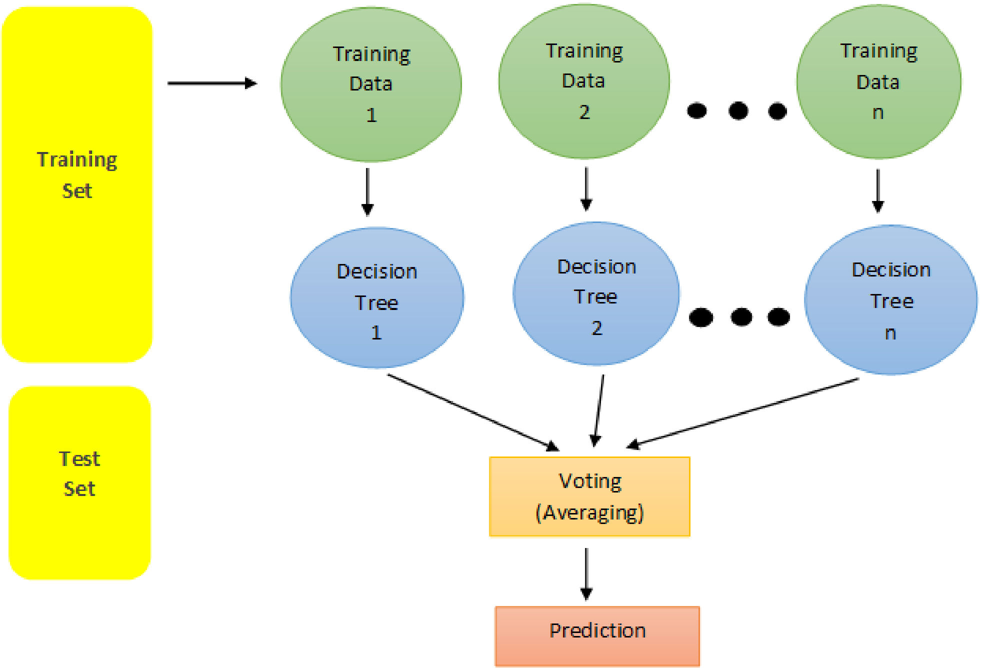


Figure 3.3: Random Forest algorithm

3.4.4 K-Nearest Neighbor (KNN): The K-Nearest Neighbors algorithm is a type of machine learning approach used for classification tasks. It works by determining the class of given data point based on the distances between the point and the points in the training dataset. In KNN we need to find out the k neighbors in the space and then with the help of k neighbors we perform the classification i.e. KNN makes prediction on the basis of what majority of k neighbors indicate. For example, in case of predicting cardiovascular disease, if a lot of neighbors have the disease, then the new person also has the disease.

Sure, here's a revised version:

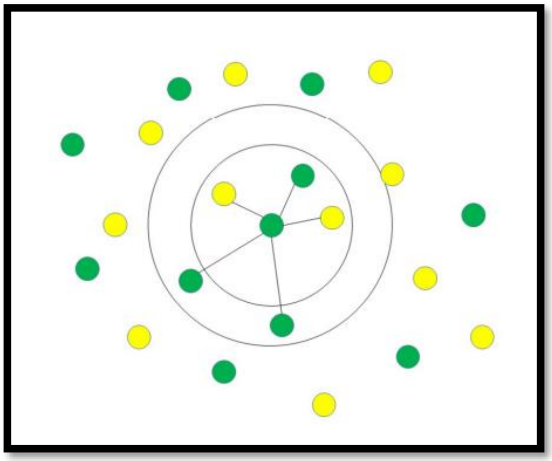


Figure 3.4: Demonstration of K-NN identifying unknown pattern by assigning a value to the K, where the nearest neighbor category of the K training sample is considered the same as the classification [13].

3.4.5 Support Vector Machine (SVM): SVM is used for the purpose of classification. The aim of the SVM is to find a hyperplane that well separates the data points that belongs to different classes. SVM aims to maximize the margin i.e. the distance between the hyperplane and the nearest support vectors from each set. Larger the margin means there is gap between the classes and in turn we are able to draw the best possible line i.e. hyperplane that best separates the data as shown in the figure.

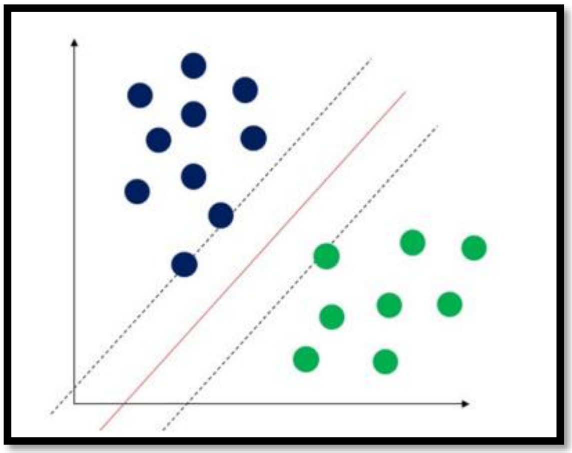
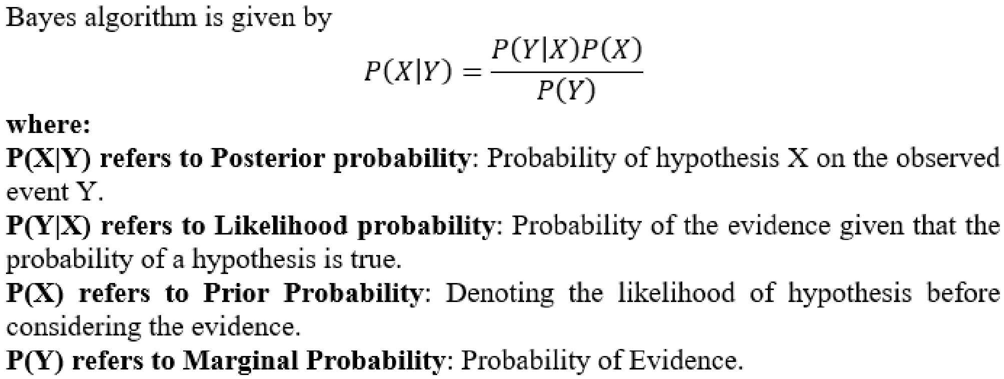


Figure 3.5: Demonstration of support vector machine. In the illustration, the solid red line represents the separating hyperplane, while the distance between two dotted lines depicts the maximum margin for separating distinct classes [14].

3.4.6 Naïve Bayes: The Naive Bayes algorithm is a supervised learning technique that uses the Bayes theorem with strong independence assumptions between features to solve problems regarding classification. It is a probabilistic classifier i.e. it predicts on the basis of an object's probability [15]. The Naïve Bayes classifier stands out as simple and efficient classification algorithm, hence making it a popular choice for quickly creating machine learning models making predictions swiftly.

Internal calculations will be performed to determine the probability of each value in relation to the feature. We take into consideration the value with the highest probability once we have computed the probability of each value of the output attributes [16,17].



Bayes algorithm is given by

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

where:

P(X|Y) refers to Posterior probability: Probability of hypothesis X on the observed event Y.

P(Y|X) refers to Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(X) refers to Prior Probability: Denoting the likelihood of hypothesis before considering the evidence.

P(Y) refers to Marginal Probability: Probability of Evidence.

Figure 3.6: Formula for Bayes Algorithm

3.5 Assessment of model performance

To find out whether the model is performing successfully and making right predictions on the basis of new unseen data we use metrics which are used for almost all machine learning algorithms.
These are: -

1) Confusion Matrix:

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predictive Negative	False Negative	True Negative

True positive are the outcomes that are correctly predicted as positive. For example, in case of Cardiovascular disease, True positive finds out that the person tested with cardiovascular disease actually have the disease whereas True Negative finds out that the person tested do not have the disease and indeed they do not have disease.

False Positive: Outcomes that are actually negative but incorrectly predicted as positive. E.g. the test says person have the disease but actually they do not have and False Negative is opposite of False Positive.

2) Accuracy:
Accuracy is the ratio of sum of total number of true positives and true negatives to total number of instances.

i.e.

$$Accuracy = \frac{\text{Total number of true positives } + \text{ Total number of false positives}}{\text{Total number of instances}}$$

for example, in case of predicting cardiovascular disease, how well the machine learning model correctly identify patients with or without disease. The more the accuracy means the model is able to make correct predictions most of the time.

3) Precision and Recall:
Precision is the ratio of total number of true positives to the sum of total number of true positives and total number of false positives.

i.e.

$$Precision = \frac{\text{Total number of true positives}}{\text{Total number of true positives } + \text{ Total number of false positives}}$$

Precision measures positive outcome, for example, in the dataset taken above regarding cardiovascular diseases, it measures whether the person has disease. The higher the precision means the model is accurate meaning if the model is predicting that a person has cardiovascular disease, it is likely to be true.

Recall also known as Sensitivity represents proportion of actual positives that are correctly identified. It is total number of true positives to the sum of total number of true positives and total number of false negatives.

i.e.

$$Recall = \frac{\text{Total number of true positives}}{\text{Total number of true positives } + \text{ Total number of false negatives}}$$

For example, recall identifies number of patients who actually have the cardiovascular disease. Higher the recall means our ML model is able to identify all the persons who actually have cardiovascular disease.

4) Specificity

Specificity also known as true negative rate is ratio of total number of true negatives to the sum of total number of true negatives and total number of false positives.

i.e.

$$Specificity = \frac{\text{Total number of true negatives}}{\text{Total number of true negatives } + \text{ Total number of false positives}}$$

For example, in case of predicting whether the patients have cardiovascular disease or not, specificity helps in finding out how well the model is predicting the patients that do not have disease.

5) Receiver Operating Characteristic (ROC)

The ROC curve is a probability graph that illustrates how well a classification model performs at various threshold values. It plots the true positive rate (TPR) against the false positive rate (FPR) across various classification thresholds.

TPR or true positive rate also known as Recall, can be calculated as:

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

FPR or False Positive Rate can be calculated as:

$$FPR = \frac{FP}{FP + TN}$$

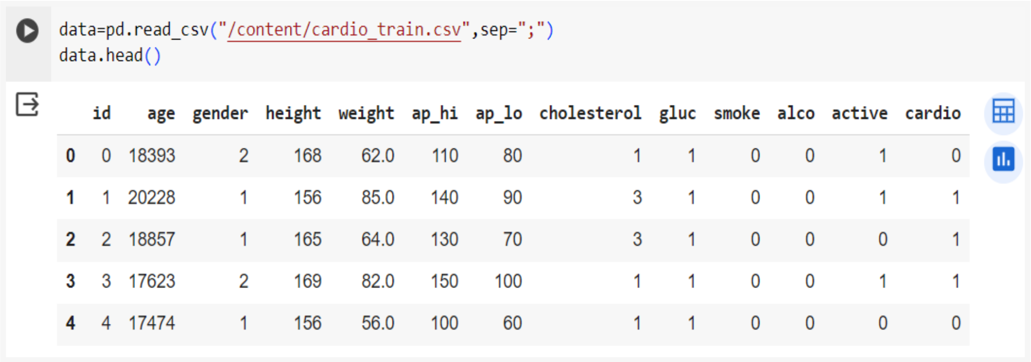
Where:
TP: True Positive; **FP**: False Positive; **TN**: True Negative; **FN**: False Negative

CHAPTER 4 RESULTS

We have used various classification models such as Logistic Regression, Decision Tree Algorithm, SVM, Naive Bayes, etc., to make predictions related to cardiovascular patients. First, we commence by initiating the data cleaning process that involves removing null values, removing duplicates and filtering out irrelevant information. The following is done to ensure that our research is based on complete and accurate dataset.

4.1 ML Algorithms in Python

1) Data Understanding



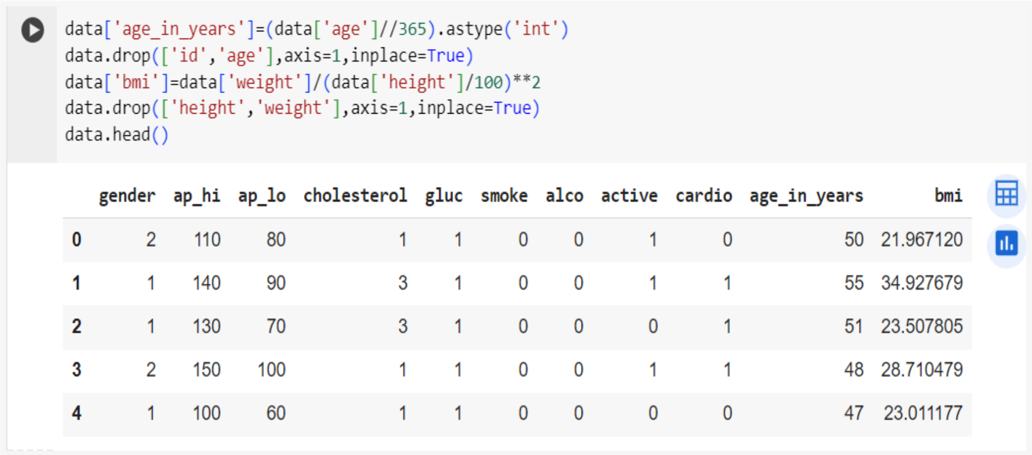
```
data=pd.read_csv("/content/cardio_train.csv",sep=";")
data.head()
```

```
   id    age  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc  smoke  alco  active  cardio
0   0  18393       2     168    62.0    110     80            1     1      0     0       1       0
1   1  20228       1     156    85.0    140     90            3     1      0     0       1       1
2   2  18857       1     165    64.0    130     70            3     1      0     0       0       1
3   3  17623       2     169    82.0    150    100            1     1      0     0       1       1
4   4  17474       1     156    56.0    100     60            1     1      0     0       0       0
```

Figure 4.1: Part of code for importing data

2) Data Preprocessing

i) Age Transformation, BMI Calculation, Column Removal

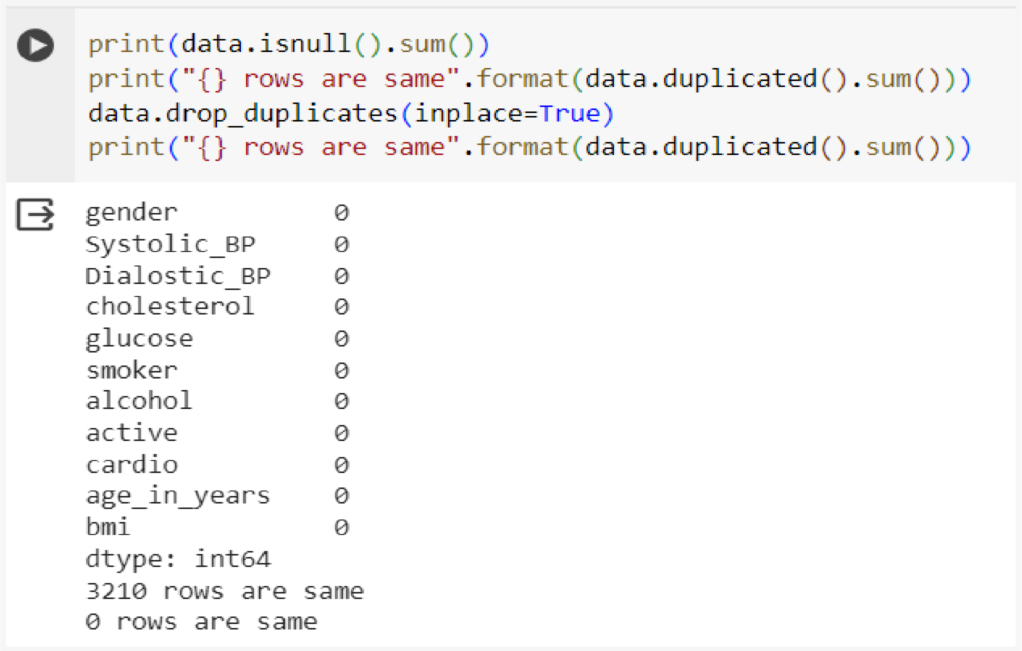


```
data['age_in_years']=(data['age']//365).astype('int')
data.drop(['id','age'],axis=1,inplace=True)
data['bmi']=data['weight']/(data['height']/100)**2
data.drop(['height','weight'],axis=1,inplace=True)
data.head()
```

```
   gender  ap_hi  ap_lo  cholesterol  gluc  smoke  alco  active  cardio  age_in_years        bmi
0       2    110     80            1     1      0     0       1       0            50  21.967120
1       1    140     90            3     1      0     0       1       1            55  34.927679
2       1    130     70            3     1      0     0       0       1            51  23.507805
3       2    150    100            1     1      0     0       1       1            48  28.710479
4       1    100     60            1     1      0     0       0       0            47  23.011177
```

Figure 4.2: Part of code for transforming age(years), calculation of BMI & removing column

ii) Dealing with null values and duplicated values

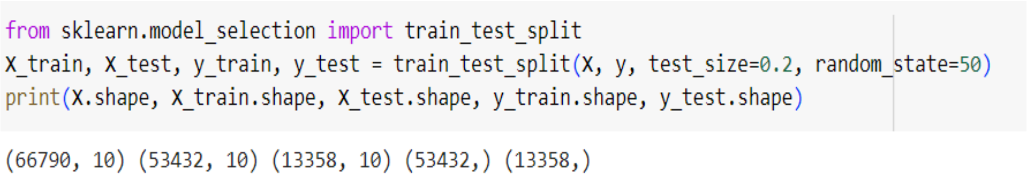


```
print(data.isnull().sum())
print("{} rows are same".format(data.duplicated().sum()))
data.drop_duplicates(inplace=True)
print("{} rows are same".format(data.duplicated().sum()))
```

```
gender          0
Systolic_BP     0
Dialostic_BP    0
cholesterol     0
glucose         0
smoker          0
alcohol         0
active          0
cardio          0
age_in_years    0
bmi             0
dtype: int64
3210 rows are same
0 rows are same
```

Figure 4.3: Part of code for removing Null values & duplicate values

3) Training and Testing Data

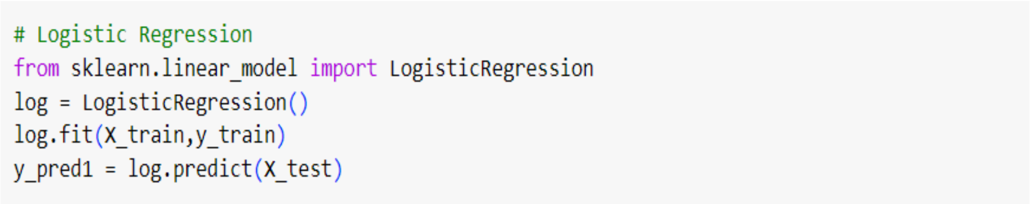


```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=50)
print(X.shape, X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(66790, 10) (53432, 10) (13358, 10) (53432,) (13358,)
```

Figure 4.4: Part of code for splitting data into train and test datasets

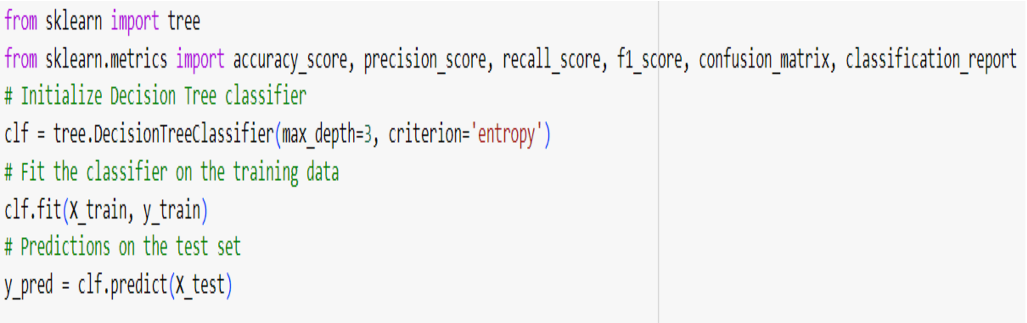
i) Logistic Regression



```
# Logistic Regression
from sklearn.linear_model import LogisticRegression
log = LogisticRegression()
log.fit(X_train,y_train)
y_pred1 = log.predict(X_test)
```

Figure 4.5: Part of code for Logistic Regression

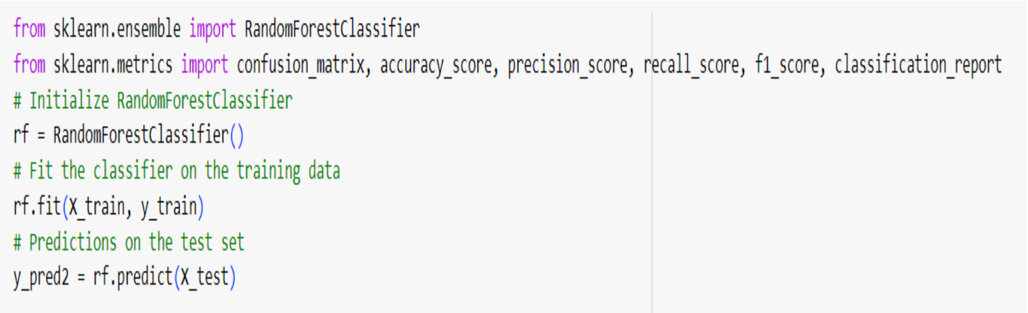
ii) Decision Tree



```python
from sklearn import tree
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report
# Initialize Decision Tree classifier
clf = tree.DecisionTreeClassifier(max_depth=3, criterion='entropy')
# Fit the classifier on the training data
clf.fit(X_train, y_train)
# Predictions on the test set
y_pred = clf.predict(X_test)
```

Figure 4.6: Part of code for Decision Tree

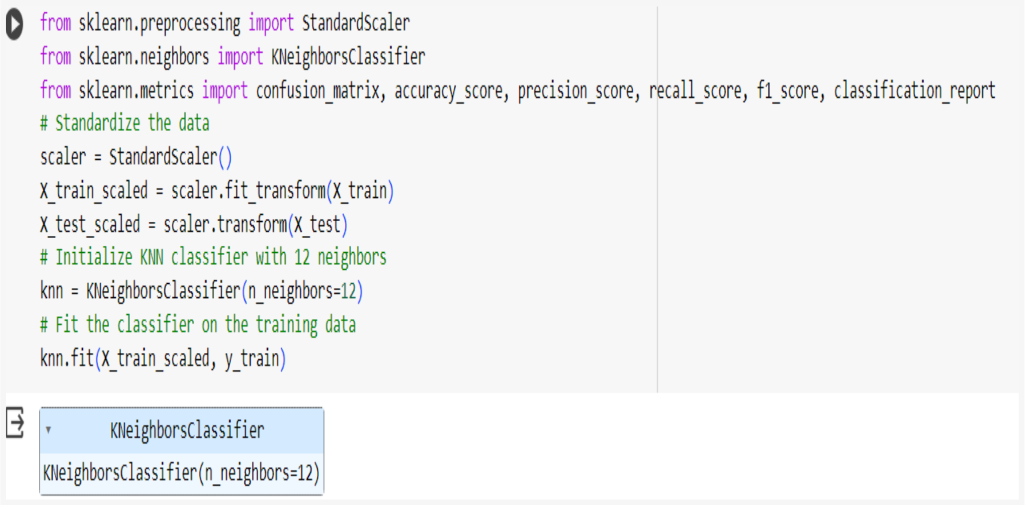
iii) Random Forest



```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score, classification_report
# Initialize RandomForestClassifier
rf = RandomForestClassifier()
# Fit the classifier on the training data
rf.fit(X_train, y_train)
# Predictions on the test set
y_pred2 = rf.predict(X_test)
```

Figure 4.7: Part of code for Random Forest

iv) KNN

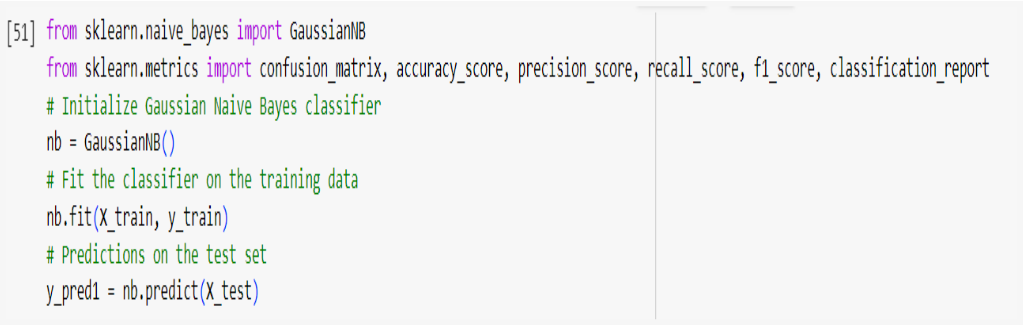


```python
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score, classification_report
# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
# Initialize KNN classifier with 12 neighbors
knn = KNeighborsClassifier(n_neighbors=12)
# Fit the classifier on the training data
knn.fit(X_train_scaled, y_train)
```

```
▾ KNeighborsClassifier
KNeighborsClassifier(n_neighbors=12)
```

Figure 4.8: Part of code for K-Nearest Neighbors

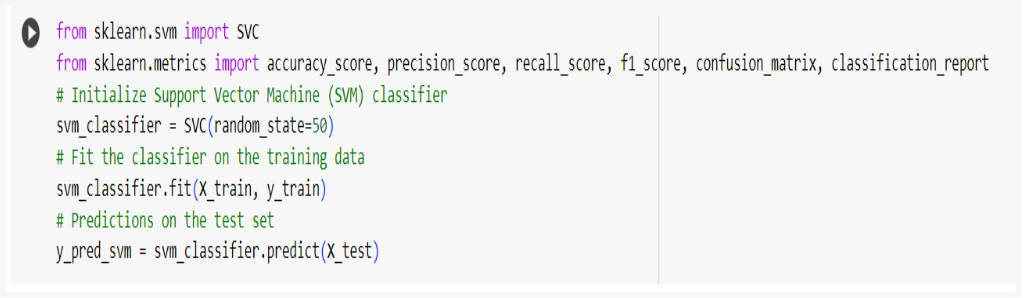
v) Naïve Bayes



```
[51] from sklearn.naive_bayes import GaussianNB
     from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score, classification_report
     # Initialize Gaussian Naive Bayes classifier
     nb = GaussianNB()
     # Fit the classifier on the training data
     nb.fit(X_train, y_train)
     # Predictions on the test set
     y_pred1 = nb.predict(X_test)
```

Figure 4.9: Part of code for Naïve Bayes

vi) SVM

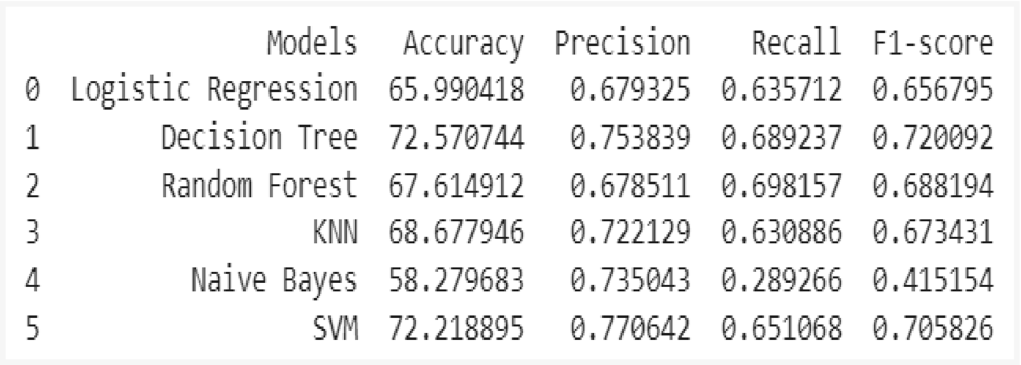


```
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report
# Initialize Support Vector Machine (SVM) classifier
svm_classifier = SVC(random_state=50)
# Fit the classifier on the training data
svm_classifier.fit(X_train, y_train)
# Predictions on the test set
y_pred_svm = svm_classifier.predict(X_test)
```

Figure 4.10: Part of code for SVM

4.2 Comparison of Model Metrics

We have also calculated various metrics that are essential for the assessment of model performance such as accuracy, precision, recall, etc. that can be seen from the figure below.



```
                Models   Accuracy  Precision    Recall  F1-score
0  Logistic Regression  65.990418   0.679325  0.635712  0.656795
1        Decision Tree  72.570744   0.753839  0.689237  0.720092
2        Random Forest  67.614912   0.678511  0.698157  0.688194
3                  KNN  68.677946   0.722129  0.630886  0.673431
4          Naive Bayes  58.279683   0.735043  0.289266  0.415154
5                  SVM  72.218895   0.770642  0.651068  0.705826
```

Figure 4.11: Assessment of model performance

From all the models we have found that the Decision Tree and Support Vector Machine have approximately same accuracy i.e. 72%. Higher Accuracy is important as it indicates the ability of model to make correct predictions. But accuracy alone might not be able to provide complete picture. Our main objective is to correctly identify the number of patients who actually have the cardiovascular disease.

So, sensitivity also known as recall is the metric that holds significant importance in the above context. Higher the sensitivity means our Machine Learning model is better at finding out larger proportion of persons who actually have the disease. We can also see the visualization of comparison of Model Metrics in Figure 4.12.

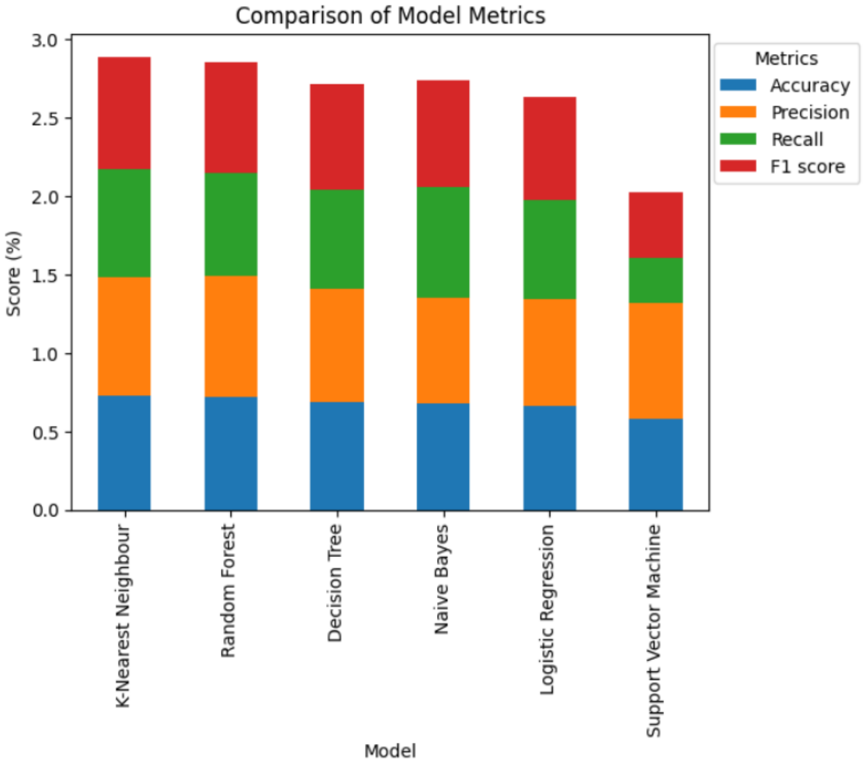


Figure 4.12: Visualization of metrics

4.3 ROC-AUC Curve

Receiver operating characteristic (ROC) is also an important metrics that can be used to evaluate the performance of the machine learning model i.e. ROC curve looks like a graph that can also show us what model is performing better. The Area under the curve (AUC) quantifies overall effectiveness of the model. A higher score means that ROC-AUC is better at identifying the people who actually suffer from cardiovascular disease instead of making predictions that healthy people have the disease.

In the analysis we have done, AUC (area under the curve) for Support Vector Machine (SVM) is highest i.e. at 79% value as illustrated in figure 4.13.

So, on the basis of our analysis, we have found that even though the Decision Tree model has slightly higher recall of 68% of class 1(with disease) whereas SVM achieves a slightly lower recall of 65% for the same class. But since our main objective is to correctly identify the number of patients who actually have the cardiovascular disease, it’s important to consider the various performance metrics including ROC-AUC. Clearly from figure 4.13 we can see that SVM model outperforms all the machine learning models indicating that SVM is better at finding out larger proportion of persons who actually have the cardiovascular disease, making it a favorable choice for the data we have taken. Therefore, considering both accuracy and ROC-AUC, Support Vector Machine model appears to be more suitable for the dataset that we have taken.

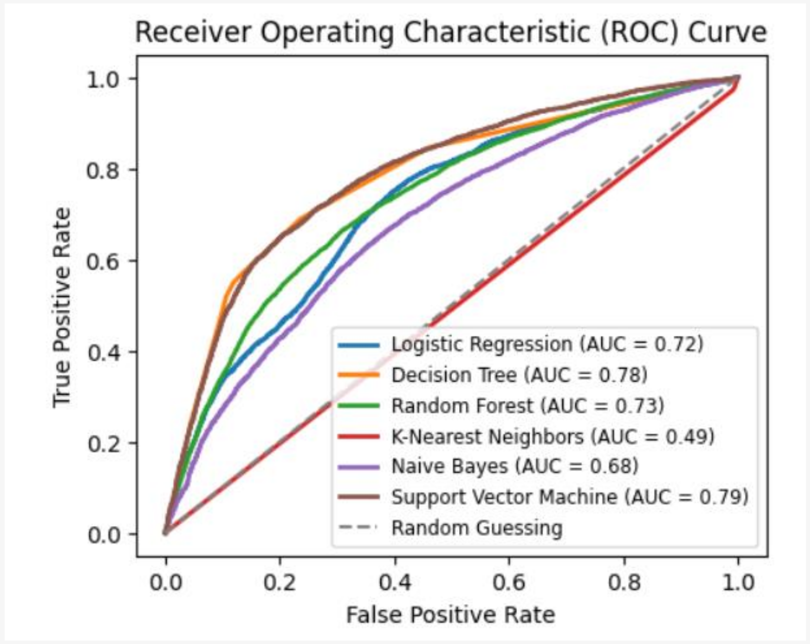


Figure 4.13: ROC-AUC

CHAPTER 5 CONCLUSION

In medical data analysis, heart diseases are seen as a significant source of concern. Heart disease is a potentially fatal disease affecting lot of people worldwide each year. As therefore, early detection of heart disease benefit both patients and healthcare professionals by providing the statistical information required to avert deaths and expenditures. This research paper delves into the effectiveness of ML algorithms for predicting CVD patients. This research addressed the prediction problem of the UCI heart disease dataset including features such as age, BMI, gender, cholesterol, alcohol intake, and so on. The dataset consisted of labelled 70000 patients, including both diagnosed heart disease patients and normal patients. To classify and predict CVD, six machine learning classification algorithms—DT, RF, LR, NB, KNN and SVM—were put into practice. It adhered to a suitable data science workflow, from data analysis and preparation to model creation, training, and evaluation. In order to identify the most appropriate ML algorithm within the model class, the performance of the selected ML method was estimated under a variety of conditions. SVM algorithm outperformed other methods in terms of the model’s accuracy, precision, Roc-Auc according to the results.

The SVM algorithm had the highest accuracy of 72% for CVD. According to the above results, the SVM algorithm is the most suitable method for predicting cardiovascular disease patients for the dataset that we have considered. The ML algorithms used are restricted to research on heart disease prediction but we can use them for predicting a lot of other things. So, in general on the basis of our research we can see from ‘Figure 5.1’ that 56% of the people are suffering from cardiovascular disease while 44% of the people are healthy according to the dataset that we have used.

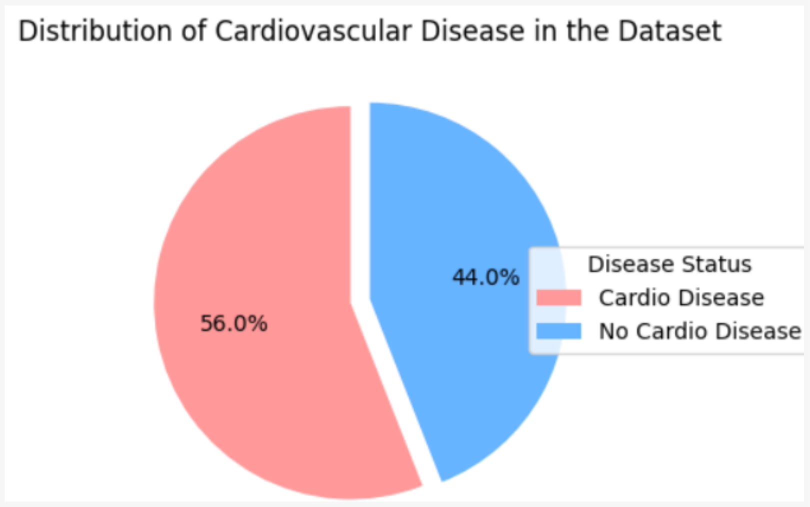


Figure 5.1: People having the disease and being healthy

Hence, we can conclude that machine learning has a significant role to play in our healthcare system. Historically, medical professionals used traditional processes and their intuition to diagnose diseases, which had drawbacks and resulted in higher expenses. However, with the use of machine learning models, diagnosis can be economically performed on large datasets.

REFERENCES

[1] Y. Mamani-Ortiz, M. San Sebastián, A. X. Armaza et al., 2019. "Prevalence and determinants of cardiovascular disease risk factors using the WHO STEPS approach in Cochabamba, Bolivia," *Bmc Public Health*, vol. 19, no. 1, p. 786, 2019.

[2] G. A. Roth, M. H. Forouzanfar, A. E. Moran et al., 2015. "Demographic and epidemiologic drivers of global cardiovascular mortality," *New England Journal of Medicine*, vol. 372, no. 14, pp. 1333–1341, 2015.

[3] R. Ndejjo, G. Musinguzi, F. Nuwaha, H. Bastiaens, and R. K. Wanyenze, 2022. "Understanding factors influencing uptake of healthy lifestyle practices among adults following a community cardiovascular disease prevention programme in Mukono and Buikwe districts in Uganda: A qualitative study," PLoS One, vol. 17, no. 2, p. e0263867, Feb. 2022.

[4] Ahsan M.M., Siddique Z., 2021. "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review".

[5] Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, and Kassim Tawiah, 2023. "A Novel Study on Machine Learning Algorithm-Based Cardiovascular Disease Prediction". Volume 2023 | Article ID 1406060 | https://doi.org/10.1155/2023/1406060.

[6] M. S. Homaid, I. Moulitsas, and K. W. Jenkins, 2023. "The Effect of Stopwords Removal and Feature Engineering on Analysing the Sentiment of Air-traveller". *5th International Conference on Advances in Signal Processing and Artificial Intelligence* (ASPAI' 2023).

[7] Kavya S M, PrathanyaSree C, Deepasindhu M, Nowshika B, Shijitha R, 2022. "Heart Disease Prediction Using Logistic Regression", 30 December 2022, *Journal of Coastal Life Medicine.*

[8] T. G. Dietterich, 1990. "Machine learning, Annual Review of Computer Science", vol. 4, no. 1, pp. 255–306, 1990.

[9] P. H. Swain and H. Hauska, 1977. "The decision tree classifier: design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[10] Madhumita Pal and Smita Parija, 2021, "Prediction of Heart Diseases using Random Forest".

[11] Rajkumar Gangappa Nadakinamani, A. Reyana, Sandeep Kautish, A. S. Vibith, Yogita Gupta, Sayed F. Abdelwahab, and Ali Wagdy Mohamed, 2022 Jan 11, "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques".

[12] Sharath Pokala, Bandari Nithya, 2021. "Random Forest Classifier based on Heart Disease Prediction", Journal of Cardiovascular Disease Research, VOL12, ISSUE01,2021, pp.799-801.

[13] Duneja, A.; Puyalnithi, T. Enhancing (1385–1388) "classification accuracy of k-nearest neighbours algorithm using gain ratio". *Int. Res. J. Eng. Technol* 2017, 4.

[14] Yuan, R.; Li, Z.; Guan, X.; Xu, L, 2010. “An SVM-based machine learning method for accurate internet traffic classification. Inf. Syst”. Front. 2010, 12, 149–156.

[15] K. Vembandasamy, R. Sasipriya and E. Deepa, 2015. “Heart Diseases Detection Using Naive Bayes Algorithm”, Vol. 2 Issue 9, September 2015.

[16] Intisar Ahmed, (12-2022). “A study of heart disease diagnosis using machine learning and data mining”, pp. 17-19.

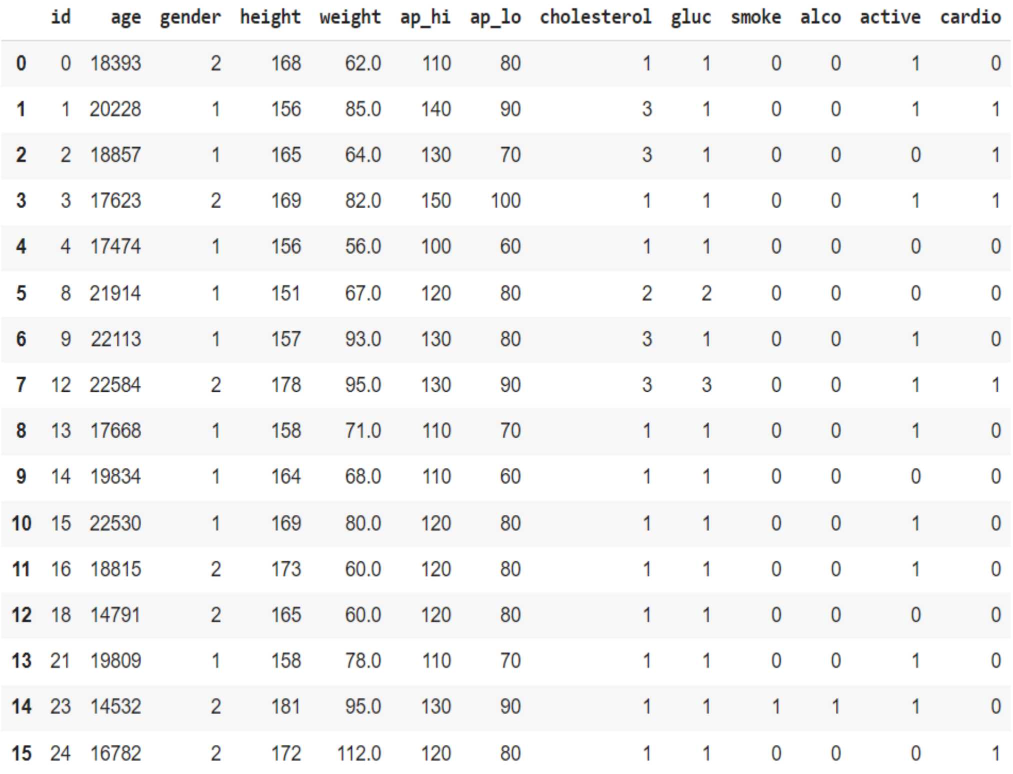
[17] Uma N Dulhare, 2018, “Prediction system for heart disease using Naive Bayes and particle swarm optimization”, May 21,2018.

[18] Ebrahim Ardeshir-Larijani and Mehdi Nasiri, 2024. “Quantum Machine Intelligence”.

Appendix-I

Data features:

- Age | Objective Feature | age | int (years)
- Height | Objective Feature | height | int (cm) |
- Weight | Objective Feature | weight | float (kg) |
- BMI | Objective Feature | BMI | int |
- Gender | Objective Feature | gender | categorical code |
- Systolic blood pressure | Examination Feature | ap_hi | int |
- Diastolic blood pressure | Examination Feature | ap_lo | int |
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoking | Subjective Feature | smoke | binary |
- Alcohol | Subjective Feature | alco | binary |
- Physical Activity | Subjective Feature | active | binary |
- Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

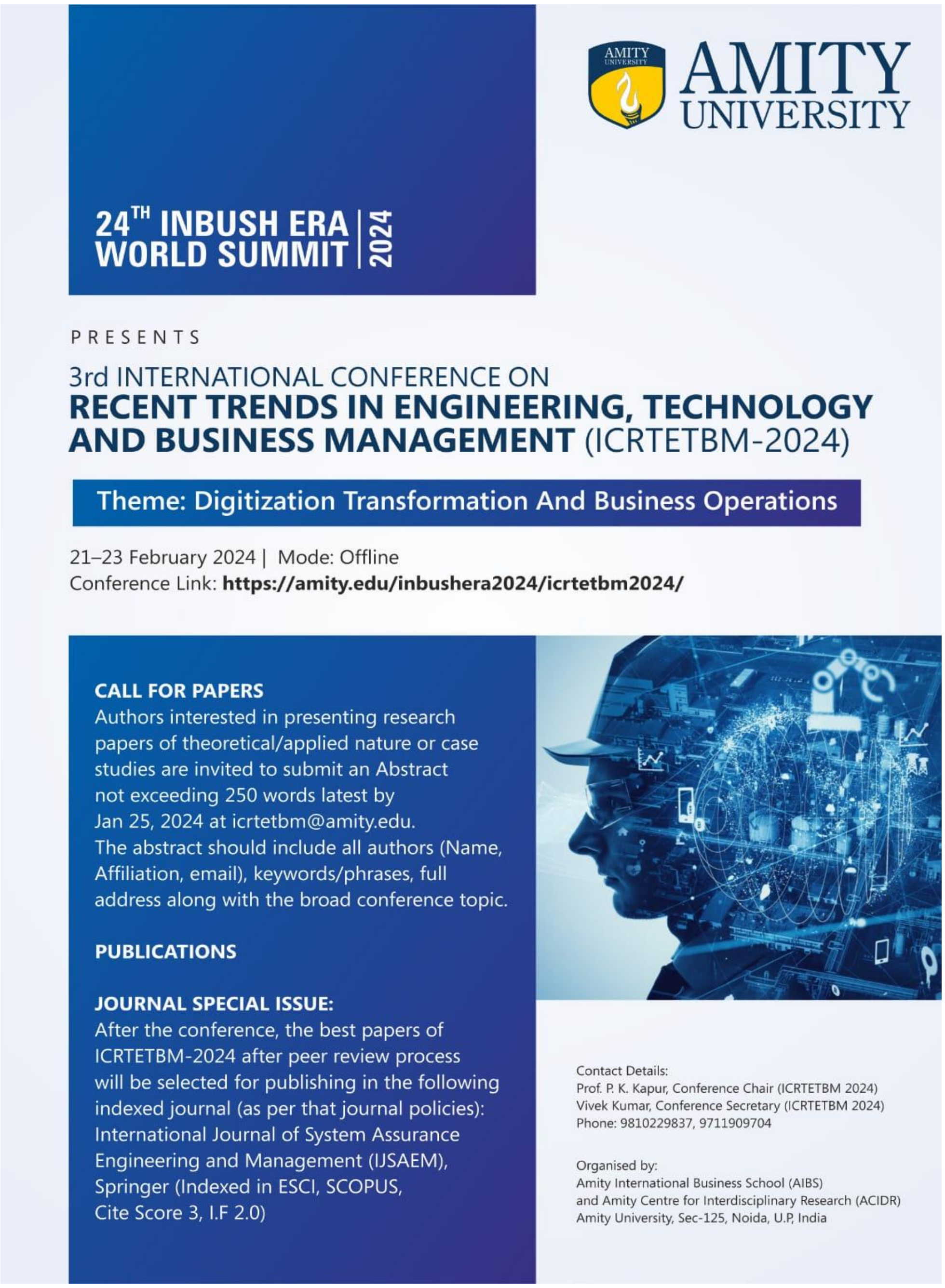


	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
5	8	21914	1	151	67.0	120	80	2	2	0	0	0	0
6	9	22113	1	157	93.0	130	80	3	1	0	0	1	0
7	12	22584	2	178	95.0	130	90	3	3	0	0	1	1
8	13	17668	1	158	71.0	110	70	1	1	0	0	1	0
9	14	19834	1	164	68.0	110	60	1	1	0	0	0	0
10	15	22530	1	169	80.0	120	80	1	1	0	0	1	0
11	16	18815	2	173	60.0	120	80	1	1	0	0	1	0
12	18	14791	2	165	60.0	120	80	1	1	0	0	0	0
13	21	19809	1	158	78.0	110	70	1	1	0	0	1	0
14	23	14532	2	181	95.0	130	90	1	1	1	1	1	0
15	24	16782	2	172	112.0	120	80	1	1	0	0	0	1

Figure I.1: Overview of the Dataset

LIST OF PUBLICATIONS

ICRTEBM, Amity University Conference Brochure



AMITY UNIVERSITY

24TH INBUSH ERA WORLD SUMMIT | 2024

PRESENTS

3rd INTERNATIONAL CONFERENCE ON
RECENT TRENDS IN ENGINEERING, TECHNOLOGY AND BUSINESS MANAGEMENT (ICRTETBM-2024)

Theme: Digitization Transformation And Business Operations

21–23 February 2024 | Mode: Offline
Conference Link: **https://amity.edu/inbushera2024/icrtetbm2024/**

CALL FOR PAPERS
Authors interested in presenting research papers of theoretical/applied nature or case studies are invited to submit an Abstract not exceeding 250 words latest by Jan 25, 2024 at icrtetbm@amity.edu.
The abstract should include all authors (Name, Affiliation, email), keywords/phrases, full address along with the broad conference topic.

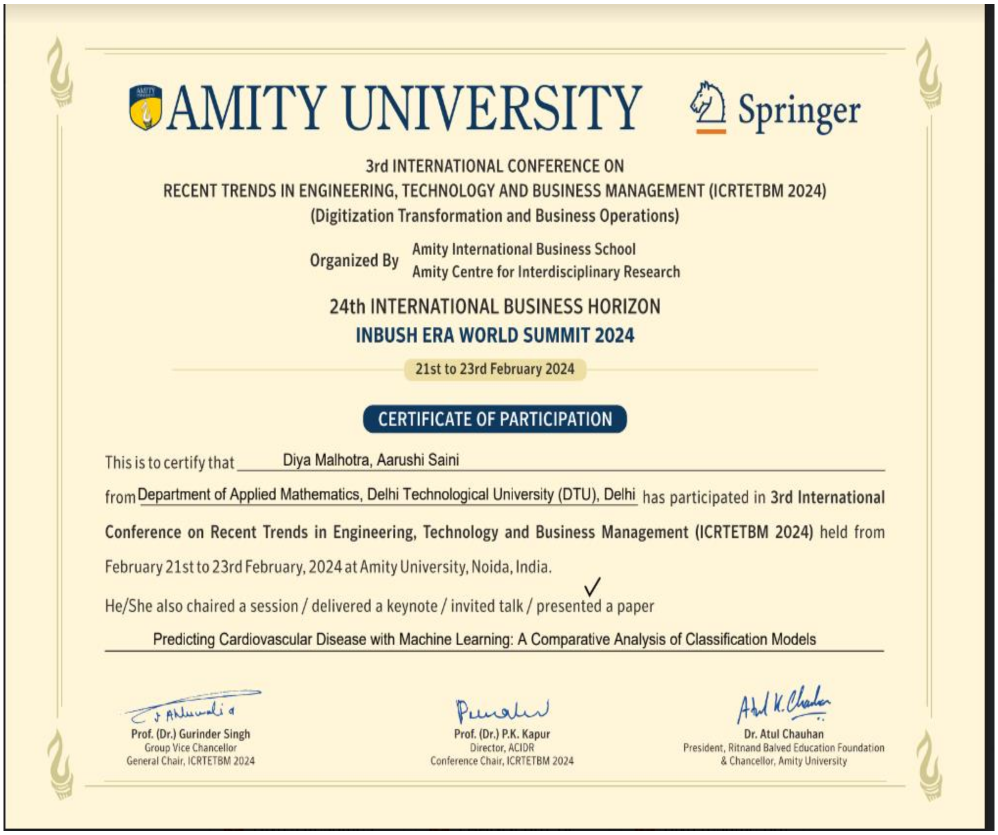
PUBLICATIONS

JOURNAL SPECIAL ISSUE:
After the conference, the best papers of ICRTETBM-2024 after peer review process will be selected for publishing in the following indexed journal (as per that journal policies): International Journal of System Assurance Engineering and Management (IJSAEM), Springer (Indexed in ESCI, SCOPUS, Cite Score 3, I.F 2.0)

Contact Details:
Prof. P. K. Kapur, Conference Chair (ICRTETBM 2024)
Vivek Kumar, Conference Secretary (ICRTETBM 2024)
Phone: 9810229837, 9711909704

Organised by:
Amity International Business School (AIBS)
and Amity Centre for Interdisciplinary Research (ACIDR)
Amity University, Sec-125, Noida, U.P, India

CERTIFICATE



AMITY UNIVERSITY Springer

3rd INTERNATIONAL CONFERENCE ON
RECENT TRENDS IN ENGINEERING, TECHNOLOGY AND BUSINESS MANAGEMENT (ICRTETBM 2024)
(Digitization Transformation and Business Operations)

Organized By Amity International Business School
Amity Centre for Interdisciplinary Research

24th INTERNATIONAL BUSINESS HORIZON
INBUSH ERA WORLD SUMMIT 2024

21st to 23rd February 2024

CERTIFICATE OF PARTICIPATION

This is to certify that Diya Malhotra, Aarushi Saini

from Department of Applied Mathematics, Delhi Technological University (DTU), Delhi has participated in **3rd International Conference on Recent Trends in Engineering, Technology and Business Management (ICRTETBM 2024)** held from February 21st to 23rd February, 2024 at Amity University, Noida, India.

He/She also chaired a session / delivered a keynote / invited talk / [✓ presented a paper]

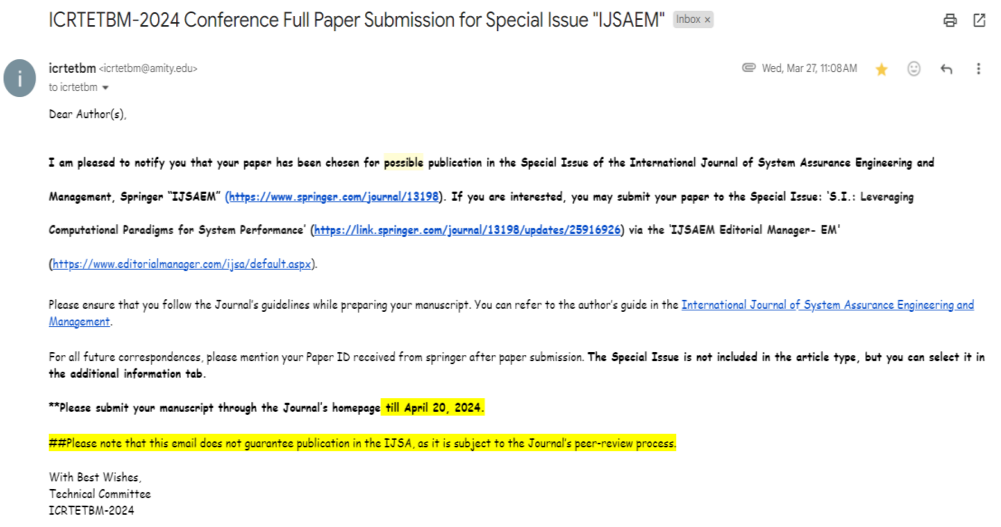
Predicting Cardiovascular Disease with Machine Learning: A Comparative Analysis of Classification Models

Prof. (Dr.) Gurinder Singh
Group Vice Chancellor
General Chair, ICRTETBM 2024

Prof. (Dr.) P.K. Kapur
Director, ACIDR
Conference Chair, ICRTETBM 2024

Dr. Atul Chauhan
President, Ritnand Balved Education Foundation
& Chancellor, Amity University

Publication Under Special Issue



ICRTETBM-2024 Conference Full Paper Submission for Special Issue "IJSAEM"

icrtetbm <icrtetbm@amity.edu>
to icrtetbm

Wed, Mar 27, 11:08 AM

Dear Author(s),

I am pleased to notify you that your paper has been chosen for possible publication in the Special Issue of the International Journal of System Assurance Engineering and Management, Springer "IJSAEM" (https://www.springer.com/journal/13198). If you are interested, you may submit your paper to the Special Issue: 'S.I.: Leveraging Computational Paradigms for System Performance' (https://link.springer.com/journal/13198/updates/25916926) via the 'IJSAEM Editorial Manager- EM'

(https://www.editorialmanager.com/ijsa/default.aspx).

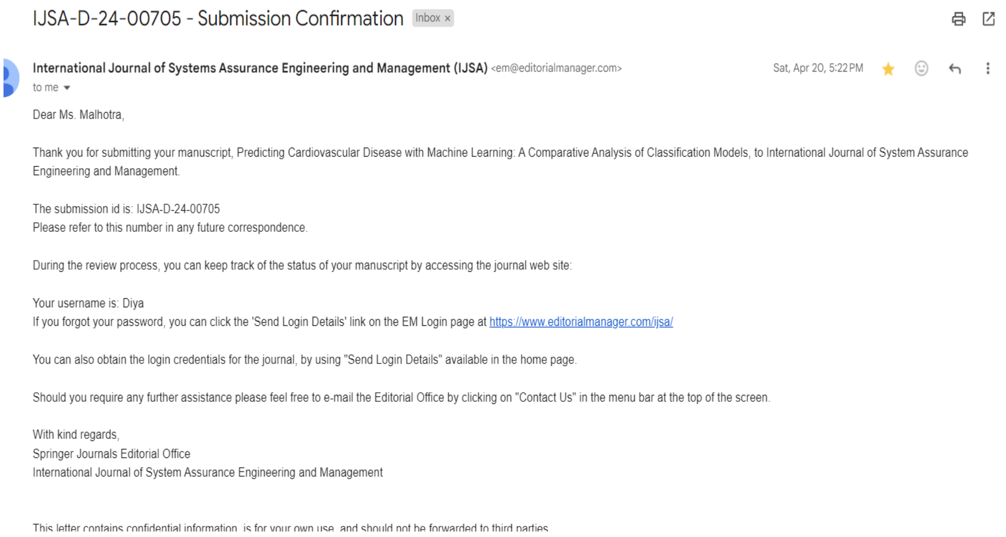
Please ensure that you follow the Journal's guidelines while preparing your manuscript. You can refer to the author's guide in the International Journal of System Assurance Engineering and Management.

For all future correspondences, please mention your Paper ID received from springer after paper submission. **The Special Issue is not included in the article type, but you can select it in the additional information tab.**

****Please submit your manuscript through the Journal's homepage till April 20, 2024.**

##Please note that this email does not guarantee publication in the IJSA, as it is subject to the Journal's peer-review process.

With Best Wishes,
Technical Committee
ICRTETBM-2024



IJSA-D-24-00705 - Submission Confirmation

International Journal of Systems Assurance Engineering and Management (IJSA) <em@editorialmanager.com>
to me

Sat, Apr 20, 5:22 PM

Dear Ms. Malhotra,

Thank you for submitting your manuscript, Predicting Cardiovascular Disease with Machine Learning: A Comparative Analysis of Classification Models, to International Journal of System Assurance Engineering and Management.

The submission id is: IJSA-D-24-00705
Please refer to this number in any future correspondence.

During the review process, you can keep track of the status of your manuscript by accessing the journal web site:

Your username is: Diya
If you forgot your password, you can click the 'Send Login Details' link on the EM Login page at https://www.editorialmanager.com/ijsa/

You can also obtain the login credentials for the journal, by using "Send Login Details" available in the home page.

Should you require any further assistance please feel free to e-mail the Editorial Office by clicking on "Contact Us" in the menu bar at the top of the screen.

With kind regards,
Springer Journals Editorial Office
International Journal of System Assurance Engineering and Management

This letter contains confidential information, is for your own use, and should not be forwarded to third parties

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE OF THESIS SUBMISSION FOR EVALUATION

1. Name: Aarushi Saini and Diya Malhotra

2. Roll No.: 2k22/MSCMAT/01 and 2k22/MSCMAT/12

3. Thesis title: "Prediction of cardiovascular disease patients with Machine Learning: A Comparative Analysis of Classification Models".

4. Degree for which the thesis is submitted: M.Sc. Mathematics

5. Faculty of the University to which the thesis is submitted: Anjana Gupta

6. Thesis Preparation Guide was referred to for preparing the thesis.
YES □ NO □

7. Specifications regarding thesis format have been closely followed.
YES □ NO □

8. The contents of the thesis have been organized based on the guidelines.
YES □ NO □

9. The thesis has been prepared without resorting to plagiarism. YES □ NO □

10. All sources used have been cited appropriately. YES □ NO □

11. The thesis has not been submitted elsewhere for a degree. YES □ NO □

12. Submitted 2 spiral bound copies plus one CD. YES □ NO □

(Signature of Candidate)
Name(s): Aarushi Saini and Diya Malhotra
Roll No.: 2k22/MSCMAT/01 and 2k22/MSCMAT/12

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE OF FINAL THESIS SUBMISSION

1. Name: Aarushi Saini and Diya Malhotra
2. Roll No.: 2k22/MSCMAT/01 and 2k22/MSCMAT/12
3. Thesis title: "Prediction of cardiovascular disease patients with Machine Learning: A Comparative Analysis of Classification Models".
4. Degree for which the thesis is submitted: M.Sc. Mathematics
5. Faculty of the University to which the thesis is submitted: Anjana Gupta
6. Thesis Preparation Guide was referred to for preparing the thesis.
   YES □ NO □
7. Specifications regarding thesis format have been closely followed.
   YES □ NO □
8. The contents of the thesis have been organized based on the guidelines.
   YES □ NO □
9. The thesis has been prepared without resorting to plagiarism. YES □ NO □
10. All sources used have been cited appropriately. YES □ NO □
11. The thesis has not been submitted elsewhere for a degree. YES □ NO □
12. All the correction has been incorporated. YES □ NO □
13. Submitted 2 hard bound copies plus one CD. YES □ NO □

(Signature of Candidate)
Name(s): Aarushi Saini and Diya Malhotra
Roll No.: 2k22/MSCMAT/01 and 2k22/MSCMAT/12

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: “**Predicting Cardiovascular Disease Patients with Machine Learning: A Comparative Analysis of Classification Models.”**

Total Pages: __

Name of the Scholar:

1. **Aarushi Saini**
2. **Diya Malhotra**

Supervisor(s):

(1) Prof. Anjana Gupta

Department of Applied Mathematics

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

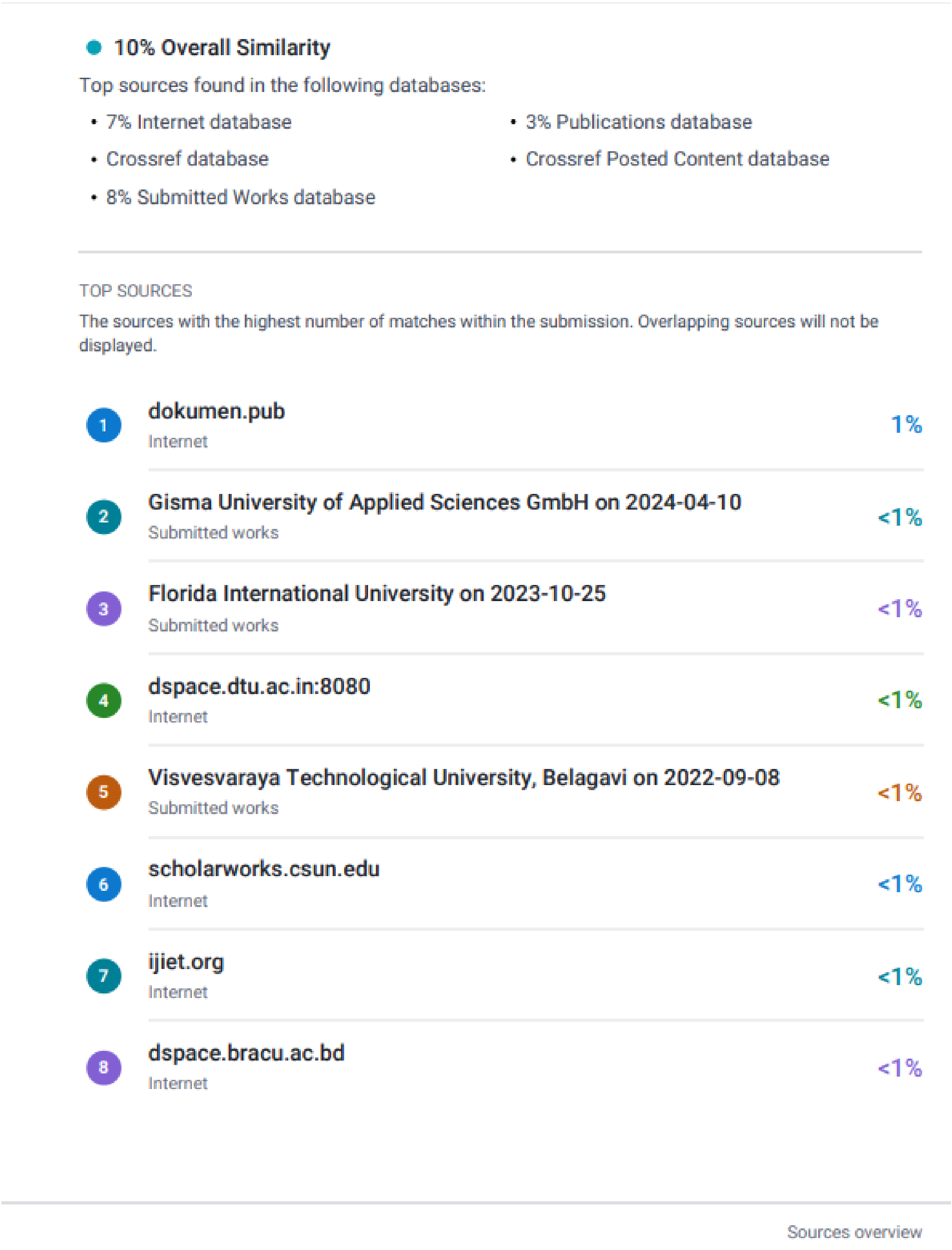
Software used: **Turnitin**, Similarity Index: **10%**,

Total Word Count: ______ **Words**

Date:

Candidate’s Signature(s)s **Signature of Supervisor(s)**

PLAGIARISM REPORT



● 10% Overall Similarity

Top sources found in the following databases:

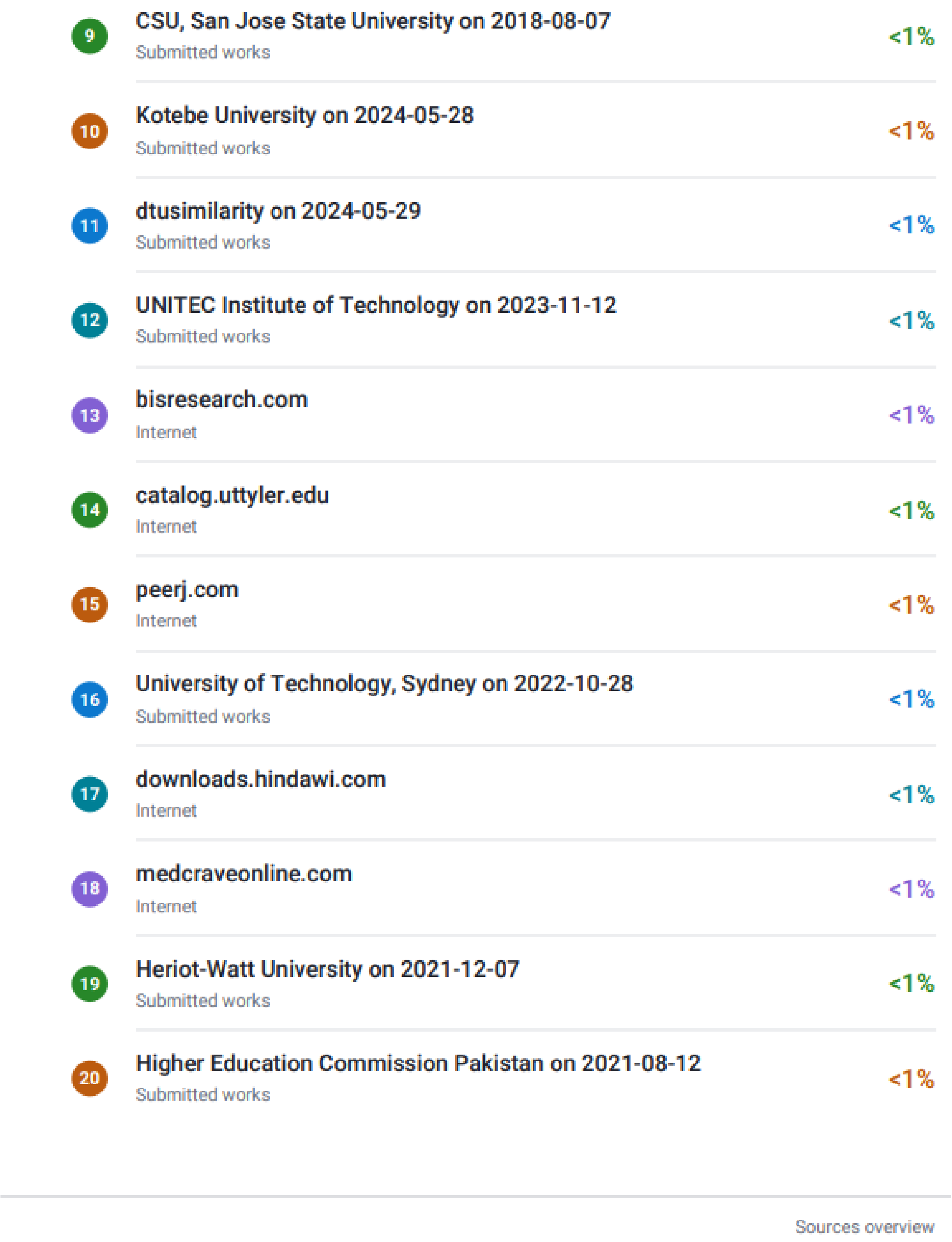
- 7% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

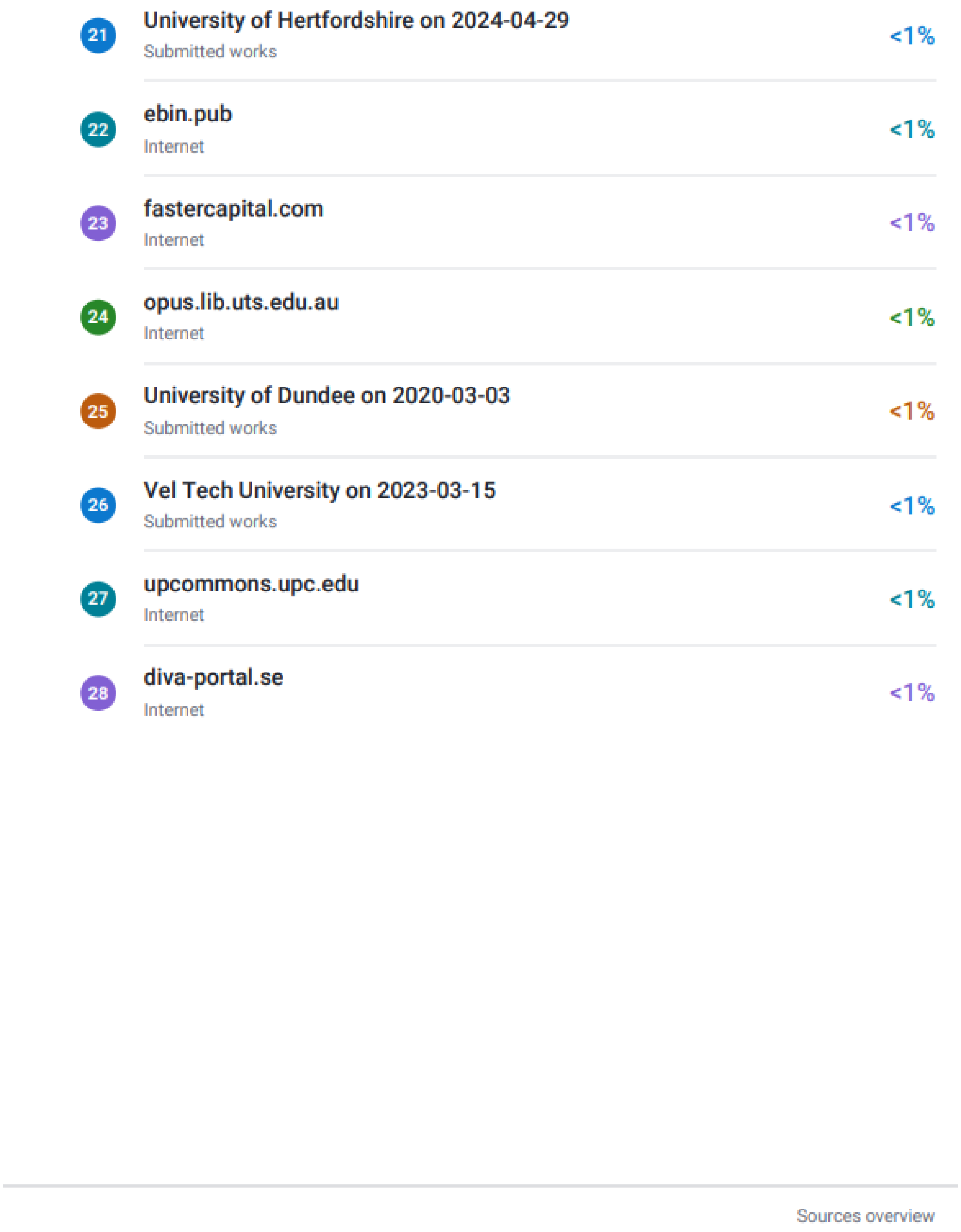
1	dokumen.pub Internet	1%
2	Gisma University of Applied Sciences GmbH on 2024-04-10 Submitted works	<1%
3	Florida International University on 2023-10-25 Submitted works	<1%
4	dspace.dtu.ac.in:8080 Internet	<1%
5	Visvesvaraya Technological University, Belagavi on 2022-09-08 Submitted works	<1%
6	scholarworks.csun.edu Internet	<1%
7	ijiet.org Internet	<1%
8	dspace.bracu.ac.bd Internet	<1%

Sources overview



9	CSU, San Jose State University on 2018-08-07 Submitted works	<1%
10	Kotebe University on 2024-05-28 Submitted works	<1%
11	dtusimilarity on 2024-05-29 Submitted works	<1%
12	UNITEC Institute of Technology on 2023-11-12 Submitted works	<1%
13	bisresearch.com Internet	<1%
14	catalog.uttyler.edu Internet	<1%
15	peerj.com Internet	<1%
16	University of Technology, Sydney on 2022-10-28 Submitted works	<1%
17	downloads.hindawi.com Internet	<1%
18	medcraveonline.com Internet	<1%
19	Heriot-Watt University on 2021-12-07 Submitted works	<1%
20	Higher Education Commission Pakistan on 2021-08-12 Submitted works	<1%

Sources overview



21	University of Hertfordshire on 2024-04-29 Submitted works	<1%
22	ebin.pub Internet	<1%
23	fastercapital.com Internet	<1%
24	opus.lib.uts.edu.au Internet	<1%
25	University of Dundee on 2020-03-03 Submitted works	<1%
26	Vel Tech University on 2023-03-15 Submitted works	<1%
27	upcommons.upc.edu Internet	<1%
28	diva-portal.se Internet	<1%

Sources overview