HEART DISEASE PREDICTION USING MACHINE LEARNING APPROACH

BY

A.K.M TANVIR ALAM

ID: 201-15-13835

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Aliza Ahmed Khan
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

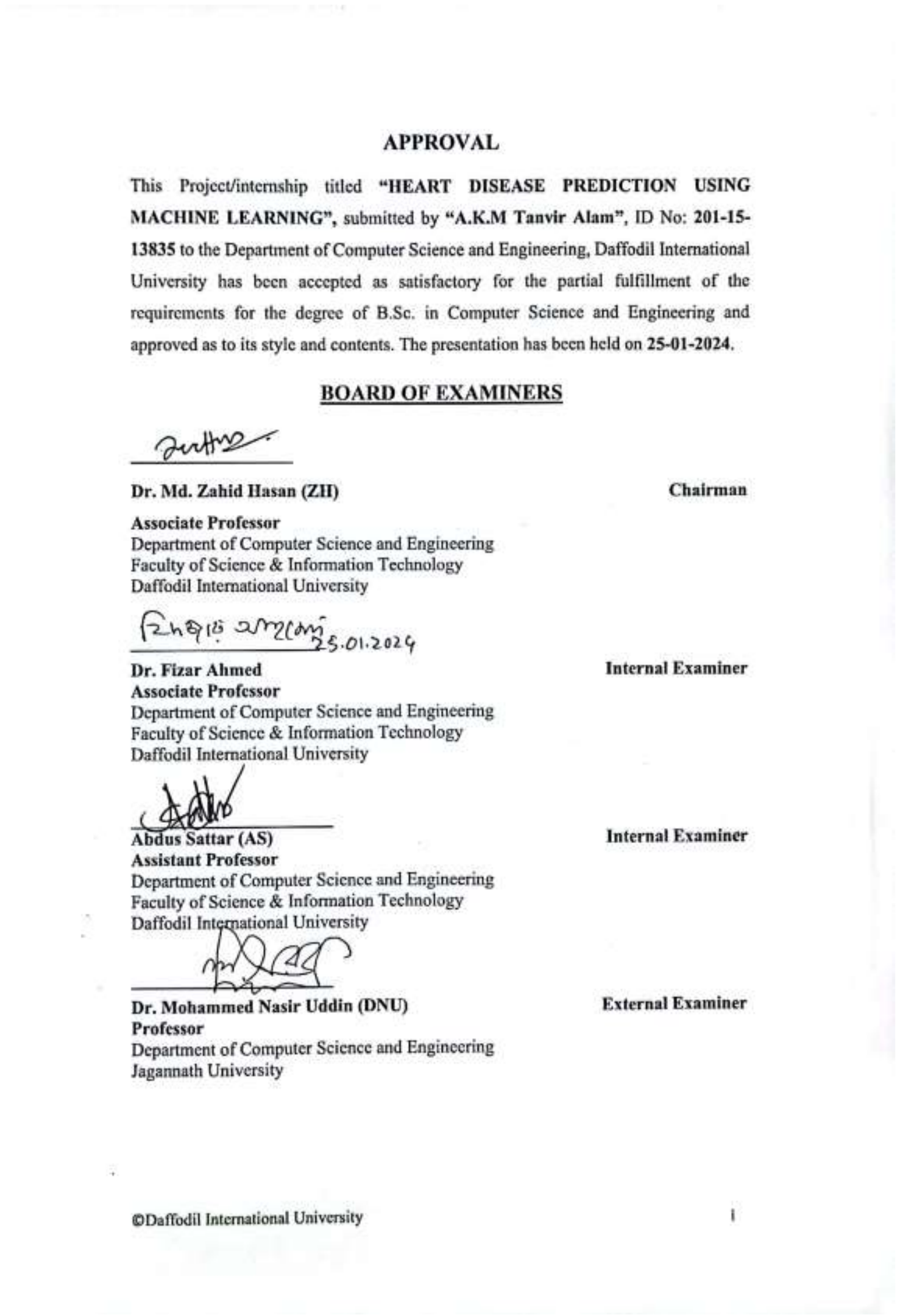
Co-Supervised By

Mr. Nahid Hasan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

25th JANUARY 2024



APPROVAL

This Project/internship titled **"HEART DISEASE PREDICTION USING MACHINE LEARNING"**, submitted by **"A.K.M Tanvir Alam"**, ID No: **201-15-13835** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **25-01-2024.**

BOARD OF EXAMINERS

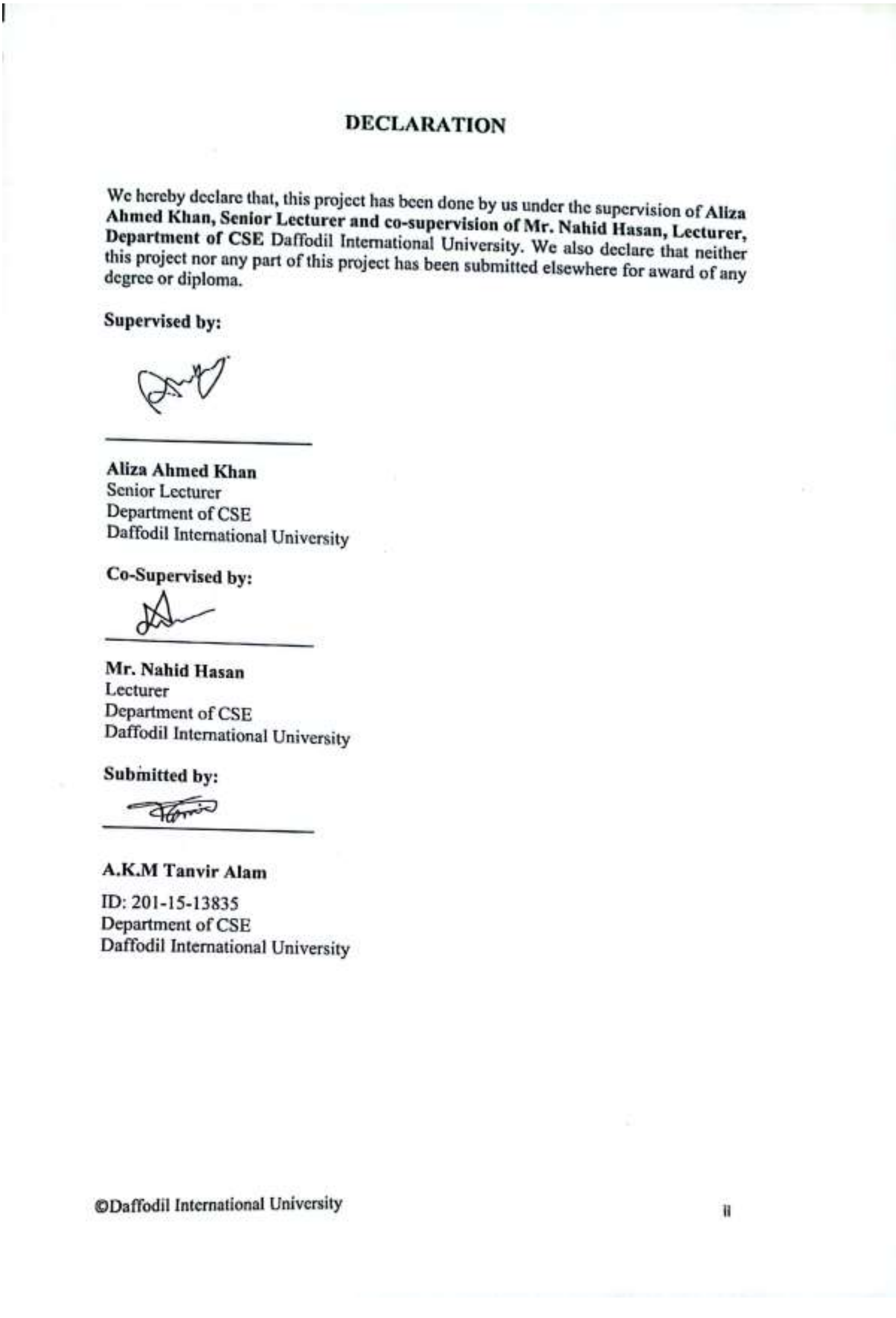
Dr. Md. Zahid Hasan (ZH) — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

ফিজার আহমেদ 25.01.2024

Dr. Fizar Ahmed — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Abdus Sattar (AS) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammed Nasir Uddin (DNU) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University



DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Aliza Ahmed Khan, Senior Lecturer and co-supervision of Mr. Nahid Hasan, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Aliza Ahmed Khan
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Nahid Hasan
Lecturer
Department of CSE
Daffodil International University

Submitted by:

A.K.M Tanvir Alam
ID: 201-15-13835
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Aliza Ahmed Khan, Senior Lecturer and co-supervision of Mr. Nahid Hasan, Lecturer,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning and Data Mining” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori** and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Heart disease is getting increasingly widespread, and it has a high fatality rate throughout the world. Heart disease has become a major health concern for many individuals and the leading cause of mortality worldwide in the previous decade. This is a challenging process that must be completed accurately and effectively. The study report focuses on which people are more prone to acquire heart disease depending on a variety of medical factors. We created a heart disease prediction method based on the patient's medical history that predicts whether the patient is likely to be diagnosed with a heart disease or not. The research title is **"Heart Disease Prediction Using Machine Learning"** and it focuses on the prediction of heart disease as well as showing who is impacted by heart disease and who is not based on the patient's medical data. Machine learning may provide an effective decision-making solution as well as precise forecasts. In the medical field, machine learning techniques are commonly used. Researchers favor models based on supervised learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Decision Trees (DT), and ensemble models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Heart disease continues to be a major worldwide health issue, demanding new approachesfor early diagnosis and prediction. In recent years, the incorporation of machine learning techniques into healthcare has proven to be a game changer, opening up new possibilities for accurate risk assessment and prompt intervention. This research delves into theintricacies of risk factors, predictive characteristics, and the overall potential to improve cardiovascular health outcomes when using machine learning algorithms to forecast heart disease. Heart disease is associated with a variety of symptoms, making it challenging to identify it quickly and accurately. Heart disease, the top cause of death globally, necessitates a proactive and precise strategy to identifying persons at risk. Traditional riskassessment approaches frequently rely on a small number of variables, but machine learning offers a more dynamic and data-driven approach. Machine learning models can identify detailed patterns by exploiting vast datasets and powerful algorithms, contributing to a more comprehensive knowledge of the variables driving cardiac disease. Working ondatabases of heart disease patients is comparable to real-world application.

1.2 Motivation

Considering heart disease has become among the most common causes death among individuals globally, early identification and prevention can improve patient outcomes dramatically. Techniques based on machine learning have the ability to transform the way we predict and treat cardiac disease by evaluating massive amounts of data and detecting variations that human forecasters might not recognize right away. Using machine learning to forecast heart disease, healthcare staff are able to recognize at-risk individuals, allowing them to apply preventative measures or begin medicine earlier in the sickness process. Reducing the requirement for expensive and resource-intensive drugs would not only improve the results for patients, but it will also reduce the overall load on the medical field. Reducing demand for expensive and resource-intensive drugs would not only improve the results for patients, but it will also reduce the overall load on the medical field. Machine learning can assist to improve both the effectiveness and precision of heart disease prediction by automating the evaluation of massive volumes of data. By assuring patients get the best available treatment based on their specific needs and conditions, we may help to minimize prejudice and subjectivity in conventional risk identification approaches. Overall, utilizing machine learning to forecast heart disease presents a major potential to enhance public health and reduce the damage caused by this illness in individuals and populations across the globe.

1.3 Rationale of the Study

The purpose of this research endeavor is to look at the capabilities of algorithms based on machine learning to forecast cardiac illness in order to enhance patient outcomes and reduce the load on the medical sector. This study has resulted in the capacity to generate and choose algorithms with the highest standards of precision and effectiveness. Machine learning methods and algorithms have recently advanced, hastening the development of methodology and approaches for identifying heart illness. This issue has been addressed using a variety of approaches, including categorizing, grouping, and others. The following categorization techniques were employed in this work: The Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and k-nearest neighbors (KNN) are the first four algorithms. This prediction system was created using a five-classifier technique. We looked at prior models in general to determine what was lacking from the current study on this topic as they needed to get updated.

1.4 Research Questions

At the initial stage of our research we got some inquiry. Such as,

1. Is it possible to improve machine learning to provide more effective outcomes?
2. Can we use renowned methods for machine learning such as classifiers, and how can we improve our feature selection approaches?
3. Does every algorithm work perfectly?
4. Does it anticipate an actual outcome based on sample data provided by our system?

1.5 Expected Outcome

Based on data from our dataset, our heart disease prediction method assists by providing a predicted result. The outcome's correctness is entirely reliant on the training dataset. Our machines learning system will get ready after we have met our system's specifications. To acquire accurate findings, we employed a number of methods. On our system, we got 52% accuracy with support vector machine algorithms (SVM) and from the k-nearest- neighbors (KNN) we got 100% accuracy, a precision of 88% with Random Forests (RF), as well as 83% accuracy with Decision Trees (DT).

1.6 Report Layout

Chapter 1 In this chapter, we discussed the cause for our effort in addition to the intended outcome of our project.

Chapter 2 We created the framework for our research in the second chapter, analyzing prior applicable research, similar inquiries, the severity of the problems, and hurdles.

Chapter 3 We are considering our topic of study and tools, as well as our technique of data collecting, statistical analysis, and execution.

Chapter 4 In this chapter, we will provide the results of the experiment and descriptive results from our inquiry.

Chapter 5 We spoke over its effect on society, the environment, and sustainability in this chapter.

Chapter 6 This chapter addresses the conclusion summary as well as further research methods.

CHAPTER 2

BACKGROUND

2.1 Introduction

Significant discoveries, problems faced, a study of relevant literature, and an in-depth explanation of the inquiry are all included in the research. The "Related Works" section delves into research papers written by other researchers, critically assessing the congruence of their methodologies and the veracity of their opinions. The thorough study extends to many scholarly publications in the similar studies area, digging into their substance, techniques, and the validity of their results in relation to my research topic. The research summary section provides a concise review of the interconnected investigations, emphasizing overarching discoveries. The tactics used to overcome problems throughout the inquiry are described, with an emphasis on the advances made in predicted accuracy throughout every phase. It is critical to recognize that heart disease is a serious global health concern, necessitating a thorough investigation of predictive modeling. This context underscores the significance of the research on "Heart Disease Prediction Using Machine Learning," which attempts to address healthcare challenges and enhance clinical decision- making procedures.

2.2 Related Works

Heart disease is a main cause of mortality worldwide. Medical informatics diagnosis advances more swiftly and simply once there are no blanks, replicas, or extraneous data in the data. Feature selection is the process of selecting an appropriate attribute from the initial set of characteristics based on a defined criterion.

M. K. (2018) [1] Heart diseases, referred to as cardiovascular diseases, have historically been responsible for a large number of deaths worldwide in the past few decades and have established as the most dangerous sickness internationally. To diagnose such illnesses in time for appropriate therapy, a reliable, accurate, and practical technique is necessary. To automate the study of vast and complicated medical data sets, machine learning techniques and methods have been used.

Sharma, V. [2] The goal of this study is to develop a machine learning algorithm capable of predicting cardiac disease based on relevant factors. Random Forest outperforms other ML systems in terms of accuracy and prediction time. As a decision-assistance system, this model might be useful to medical professionals in their clinic.

Patel, J., TejalUpadhyay [3] WEKA is used in this study to evaluate several decision-tree classification algorithms in pursuit of improved performance in heart disease detection. Among those being evaluated include the J48 method, a Logistic model tree technique, and the random-forest algorithm. Relevant datasets of cardiac disease patients were used to analyze and clarify the efficacy of decision tree techniques. There are 303 incidences and 76 characteristics in this collection.

Javid, I., Alsaedi [7] A few classification algorithms do well in terms of accuracy prediction, while others perform poorly. Several machine learning models constructed using ANN have been well supported in previous studies for the detection of heart disease.

2.3 Research Summary

Our inquiry activity plans determine if a person is discouraged or not. Because our informative index is based on paired grouping, any of the two outcomes is possible. Because our data provides paired order and we have few missing attributes for some of the queries, we used a handful of the most widely used formulas for both of the cases mentioned. We employed four techniques to determine our prediction model's accuracy, precision, recall, specificity, and f-measure. The higher the precision, the better the system. Another important parameter to examine is precision, which relates to the overall number of actual positives across all yes forecasts. This indicates that a job holder who is truly depressed from all of the projected depressed effects. Remember another essential aspect of the prediction system: the number of genuine positives in real yes outcomes.

2.4 Challenges

Choosing the correct datasets is a big difficulty for this endeavor. Data collection is the most challenging challenge in improving our forecast accuracy. Data collecting is a difficult procedure in the perspective of our country (Bangladesh). We visited many hospitals to gather data manually, but we were unable to obtain information from them due to patient privacy concerns. As a consequence, we approached a well-known doctor for permission to obtain the dataset from the hospital. The other problem was deciding on the algorithms. For training an object detection model, there are several approaches available. To train such a large dataset, we would need a lot of computational power, which we didn't have. This project was accomplished with the help of a GPU-powered Jupyter Notebook and Google Colab.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this part, we will go over our research approach and plan. Data collecting, research tools, research topic, characteristics, pretreatment and processing, data storage, analysis, and execution will also be covered in this course.

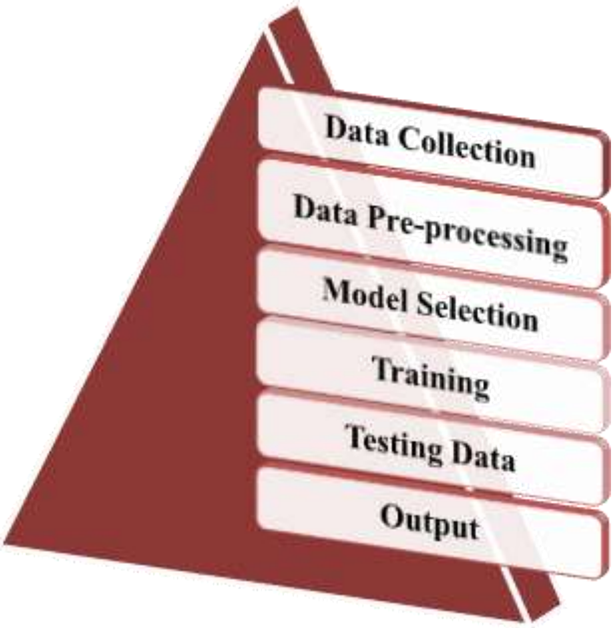


Figure 3.1.1: Methodology at a Glance

3.2 Research Subject and Instrumentation

We ought to stress that information is the most crucial part of the test. It is vital for a specialist to uncover beautiful facts as well as a fantastic strategy or model for our research effort. Comparable test papers must also be examined. At that time, we must select one of the following alternatives:

- What kind of information should be gathered?
- How can we be certain that the information we obtain is correct?
- What structure should each data collection have?

3.3 Data Collection Procedure

We collected data from a non-government hospital with the help of a well-known doctor. We'd like to thank him for assisting us in gathering data from his facility. Some data was obtained from online sources. We collected about 1400+ data. 1330 of these records were included in our study. We also gathered some data from the online sources.

3.3.1 Attributes

The dataset collected has 14 characteristics. age, sex, cp, trestbps, chol, fbs, restecg, thalac, exang, oldpeak, slope, ca, thal, and target are the primary qualities. Moreover, the data was separated into two distinct datasets: one for training and one for testing the model. Random Forest, Decision Tree, Support Vector Machine, and K-Nearest Neighbors.

3.3.2 Data Organizing

We combined all of the obtained data in one CSV file to organize it. While clearing the dataset tends to indicate the cleaning approach for the dataset in question, the null values were carefully deleted from the appropriate dataset while making sure no critical data was lost in the process.

3.3.3 Data Storing

To make our work simpler, we saved a CSV version of the data in Google Drive under this area. We may utilize the Google Drive online storage data in our project. Then we stored them on Google Drive to avoid losing them. Following that, we may use the data in our project work by executing some basic processes or writing code.

3.3.4 Machine Learning Algorithm

We employ the SVM, Random Forest, Decision Trees, and K-Nearest Neighbors algorithms to develop a model for data accuracy rate. NumPy, Pandas, Scikit-learn, and Matplotlib are among the Python algorithms and libraries utilized.

3.4 Workflow

We sought to increase our model's accuracy by utilizing machine learning techniques for instance RF, DT, SVM, and KNN. There are 20% test data & 80% training data in the collected dataset.

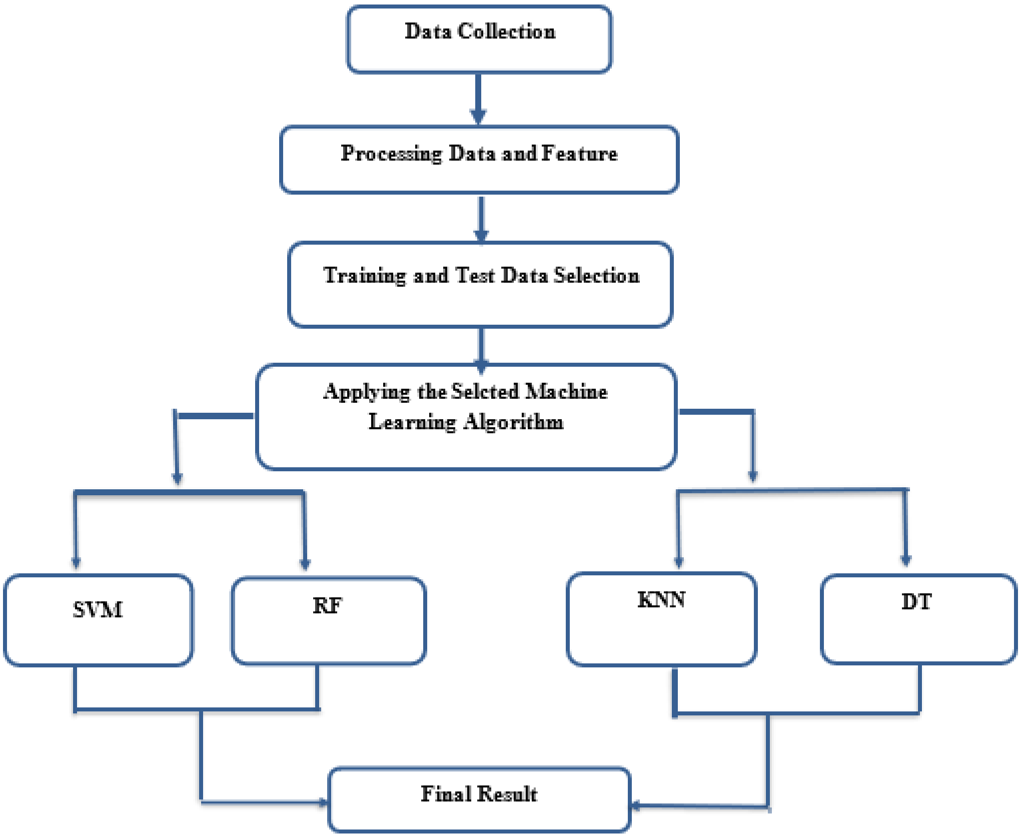

Data Collection
Processing Data and Feature
Training and Test Data Selection
Applying the Selcted Machine Learning Algorithm
SVM
RF
KNN
DT
Final Result

Figure 3.4.1: Proposed Model Structure

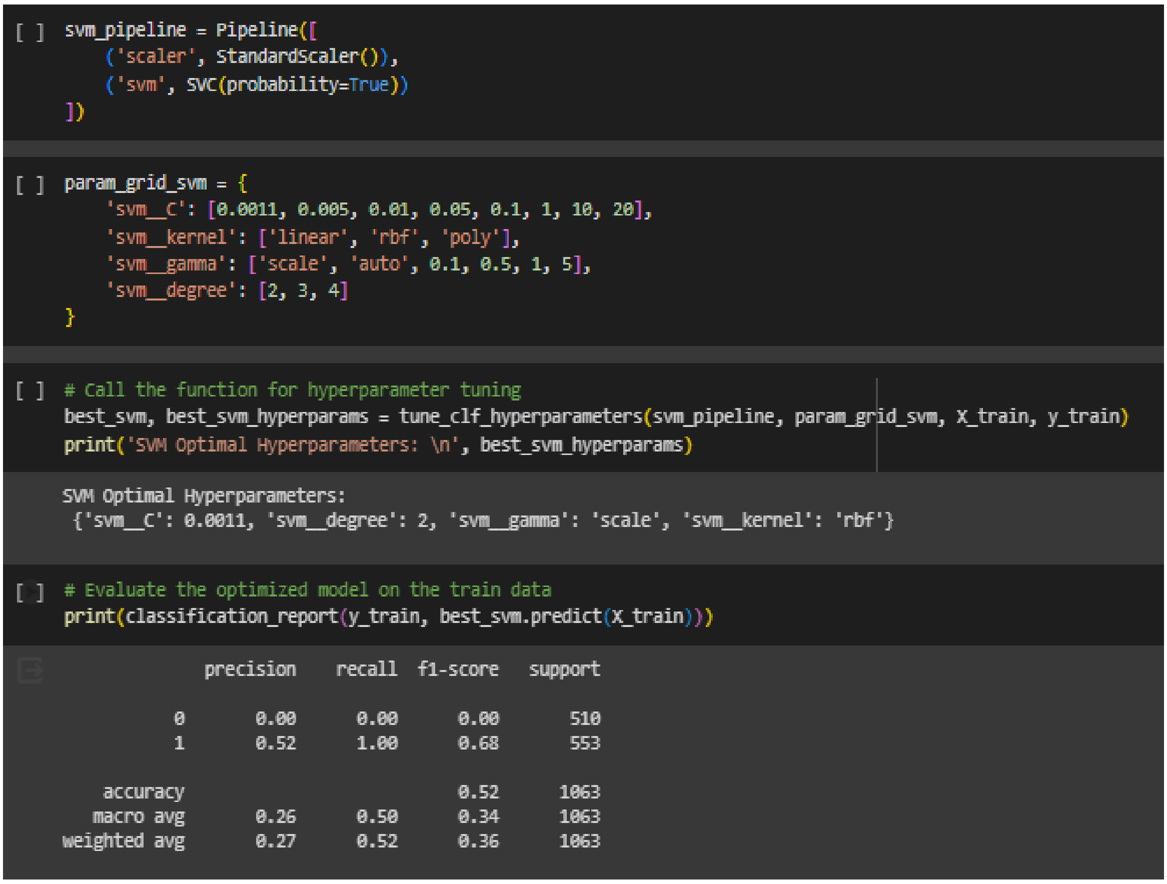
In this diagram, we can observe how to get toward our objective step by step

3.5 Applied Mechanism

I've applied the Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and Decision Tree (DT) methods on the training data set to develop models because of their established accuracy rates. All of the Python algorithms that are used make use of NumPy, Pandas, and Matplotlib packages.

Support Vector Machines (SVM):

The Support Vector Machine (SVM) is a technique for supervised machine learning that can handle both regression and classification tasks. It has mostly been applied to categorization issue solving. Building result boundaries that can split dimensional structure space into categories is the goal of linear SVM. This will allow us to easily add additional data points into the relevant category over time. The perfect choice boundary is a hyperplane. It divides the data collection into two classes, denoted by the hyperplane as 0 and 1. We employed an SVM algorithm for our job, and it provided us with 52% accuracy.



```
[ ] svm_pipeline = Pipeline([
        ('scaler', StandardScaler()),
        ('svm', SVC(probability=True))
    ])
```

```
[ ] param_grid_svm = {
        'svm__C': [0.0011, 0.005, 0.01, 0.05, 0.1, 1, 10, 20],
        'svm__kernel': ['linear', 'rbf', 'poly'],
        'svm__gamma': ['scale', 'auto', 0.1, 0.5, 1, 5],
        'svm__degree': [2, 3, 4]
    }
```

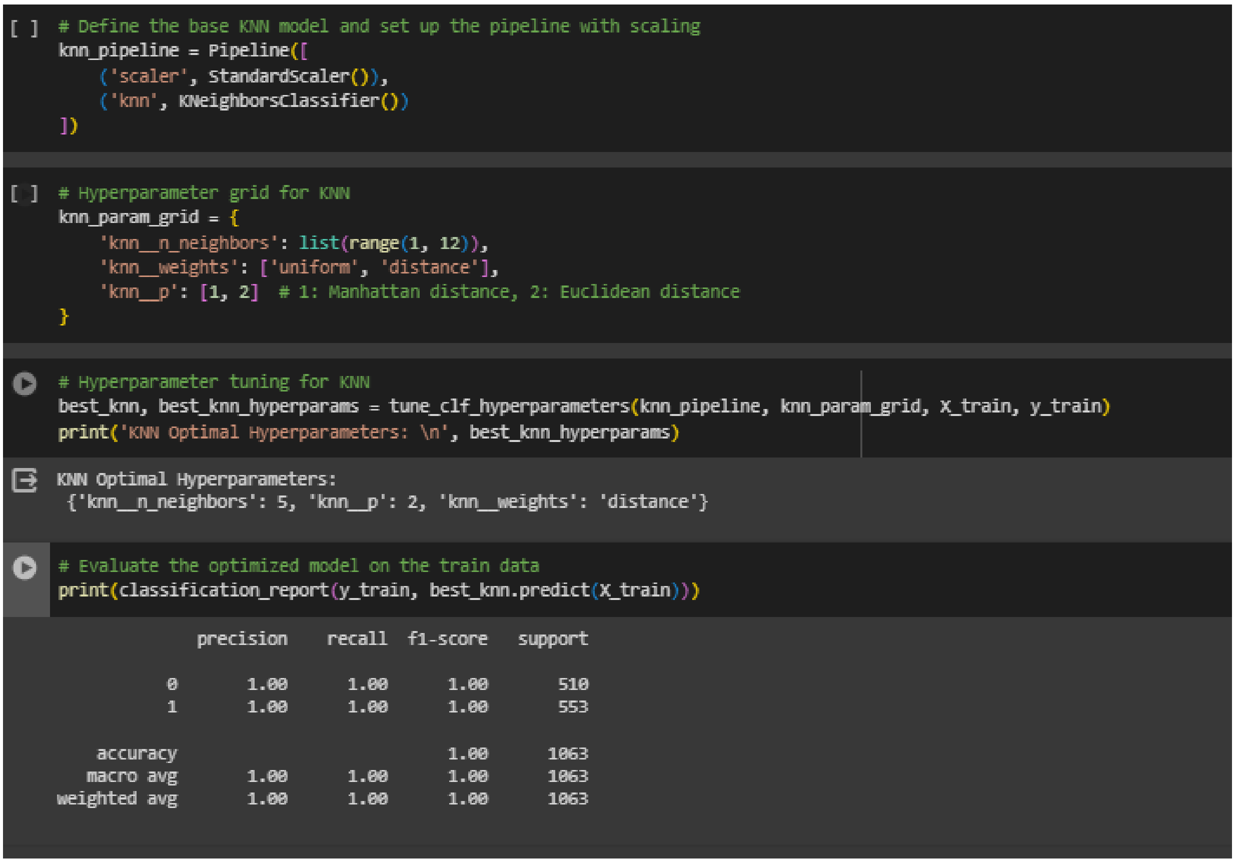
```
[ ] # Call the function for hyperparameter tuning
    best_svm, best_svm_hyperparams = tune_clf_hyperparameters(svm_pipeline, param_grid_svm, X_train, y_train)
    print('SVM Optimal Hyperparameters: \n', best_svm_hyperparams)

    SVM Optimal Hyperparameters:
     {'svm__C': 0.0011, 'svm__degree': 2, 'svm__gamma': 'scale', 'svm__kernel': 'rbf'}
```

```
[ ] # Evaluate the optimized model on the train data
    print(classification_report(y_train, best_svm.predict(X_train)))

                  precision    recall  f1-score   support

               0       0.00      0.00      0.00       510
               1       0.52      1.00      0.68       553

        accuracy                           0.52      1063
       macro avg       0.26      0.50      0.34      1063
    weighted avg       0.27      0.52      0.36      1063
```

Figure 3.5.1: SVM Model

K-Nearest Neighbor (KNN):

Finding series and recurrence may be done non-parametrically using K-nearest neighbor analysis. The k-nearest prepared models from both situations in the composition space make up the data set. KNN is a type of gradual deployment or occurrence learning where capacity is only calculated locally and all analyses are maintained current until the task assessment is finished. The accuracy of this study may be substantially enhanced by standardizing preparation data because it relies upon group segregation. Either classify or repeat; distributing the burden among neighbors' responsibilities allows the nearest neighbors to offer more consistent services than the neighbors who live further away. We employed the machine learning model KNN in our job, and it provided us with 100% accuracy.



```
[ ] # Define the base KNN model and set up the pipeline with scaling
    knn_pipeline = Pipeline([
        ('scaler', StandardScaler()),
        ('knn', KNeighborsClassifier())
    ])
```

```
[ ] # Hyperparameter grid for KNN
    knn_param_grid = {
        'knn__n_neighbors': list(range(1, 12)),
        'knn__weights': ['uniform', 'distance'],
        'knn__p': [1, 2]  # 1: Manhattan distance, 2: Euclidean distance
    }
```

```
# Hyperparameter tuning for KNN
best_knn, best_knn_hyperparams = tune_clf_hyperparameters(knn_pipeline, knn_param_grid, X_train, y_train)
print('KNN Optimal Hyperparameters: \n', best_knn_hyperparams)
```

```
KNN Optimal Hyperparameters:
 {'knn__n_neighbors': 5, 'knn__p': 2, 'knn__weights': 'distance'}
```

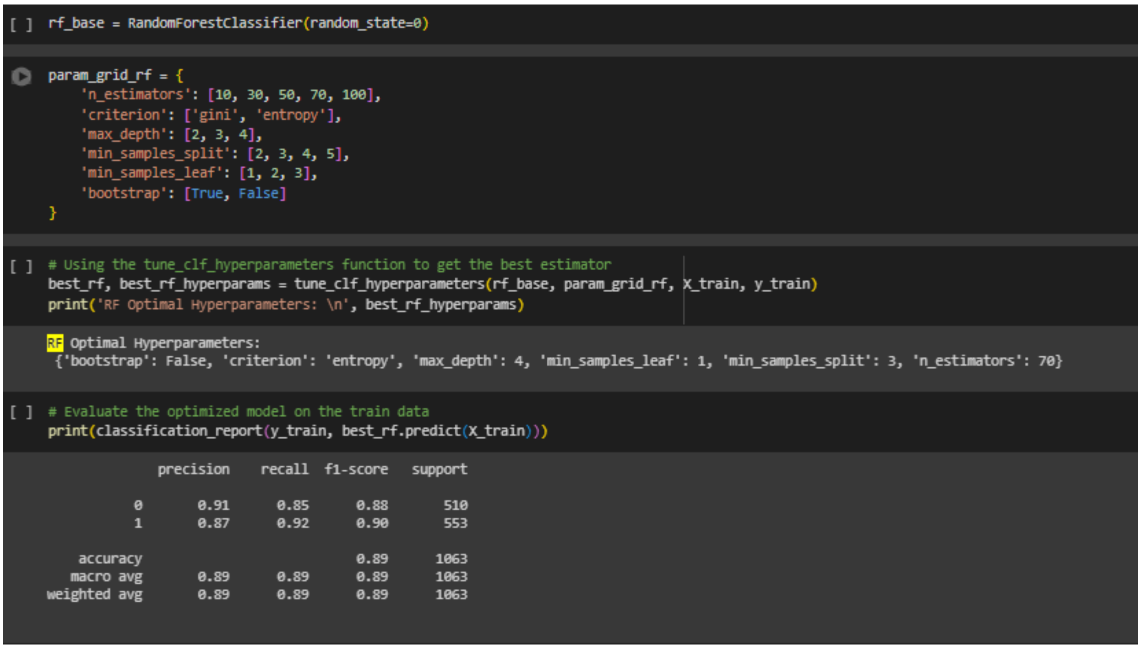
```
# Evaluate the optimized model on the train data
print(classification_report(y_train, best_knn.predict(X_train)))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       510
           1       1.00      1.00      1.00       553

    accuracy                           1.00      1063
   macro avg       1.00      1.00      1.00      1063
weighted avg       1.00      1.00      1.00      1063
```

Figure 3.5.2: KNN Model

Random Forest (RF):

The foundation of RF is the multi-model learning approach, which involves combining many classifiers to improve mode performance and resolve challenging issues. The most accurate predictive model can surpass the predictions of any single model since it is built by merging many learning algorithms. Using the RF methodology, the standard deviation of all Decision Tree methods is calculated. It is therefore just an organized set of decision trees. In this endeavor, we obtain an accuracy rating of 88% by applying the RF model.



```
rf_base = RandomForestClassifier(random_state=0)
```

```
param_grid_rf = {
    'n_estimators': [10, 30, 50, 70, 100],
    'criterion': ['gini', 'entropy'],
    'max_depth': [2, 3, 4],
    'min_samples_split': [2, 3, 4, 5],
    'min_samples_leaf': [1, 2, 3],
    'bootstrap': [True, False]
}
```

```
# Using the tune_clf_hyperparameters function to get the best estimator
best_rf, best_rf_hyperparams = tune_clf_hyperparameters(rf_base, param_grid_rf, X_train, y_train)
print('RF Optimal Hyperparameters: \n', best_rf_hyperparams)
```

```
RF Optimal Hyperparameters:
 {'bootstrap': False, 'criterion': 'entropy', 'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 70}
```

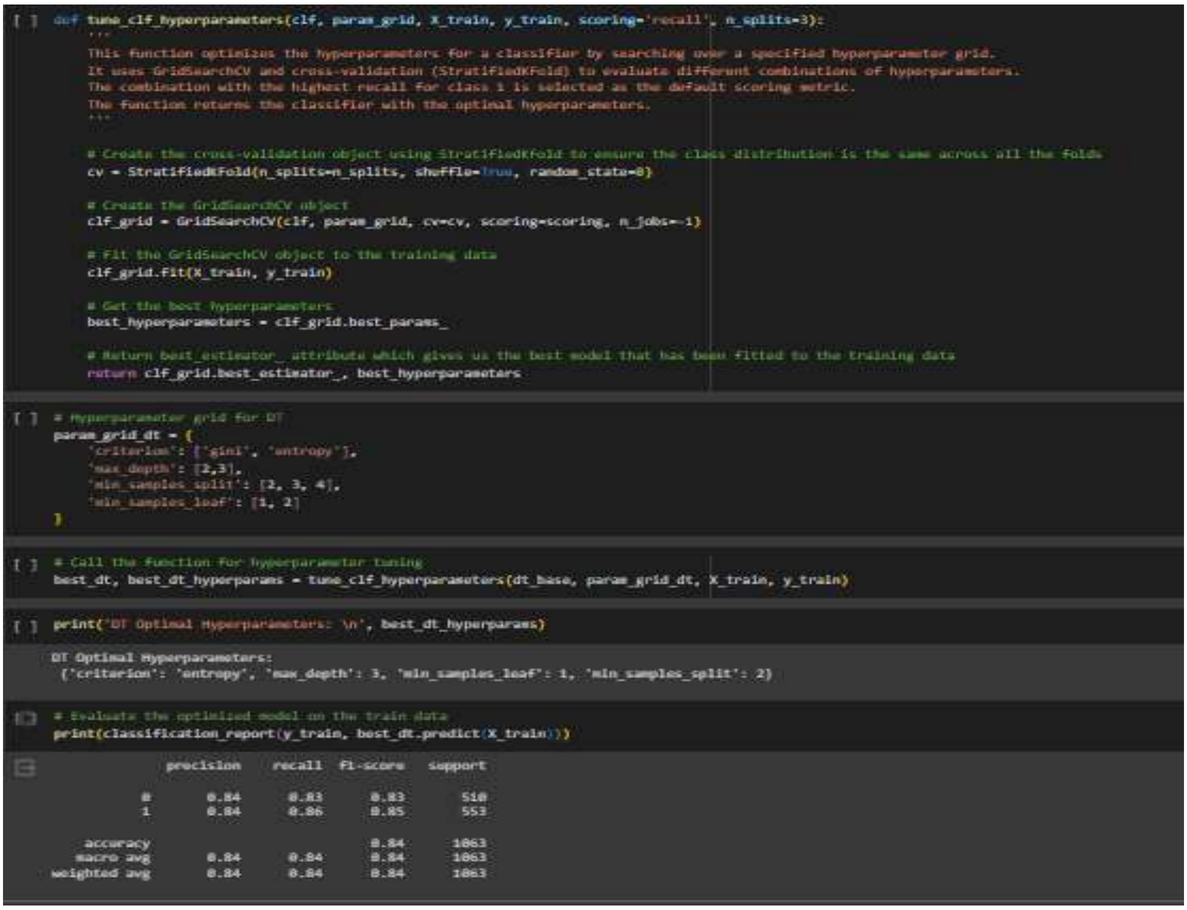
```
# Evaluate the optimized model on the train data
print(classification_report(y_train, best_rf.predict(X_train)))
```

```
              precision    recall  f1-score   support

           0       0.91      0.85      0.88       510
           1       0.87      0.92      0.90       553

    accuracy                           0.89      1063
   macro avg       0.89      0.89      0.89      1063
weighted avg       0.89      0.89      0.89      1063
```

Figure 3.5.3: Random Forest Model

Decision Tree (DT):

A decision tree is an illustration of supervised machine learning and a graphical data representation. The design of the DT is a flowchart, where each leaf center indicates the cycle classifier, each fork chooses test findings, and each core location analyzes quality. In decision-making research, the relationship between representation rules and the root-to- leaf path is used to estimate the usual benefits of competing options. Strongly connected impact outlines and decision trees are used as statistical and visual aids for choosing. The journey from the root to the leaf is connected to the rules of representation. In decision- making research, decision trees and substantially connected impact sketches are used as graphical and statistical selection aids to evaluate the average benefits of competing options. The accuracy obtained from the DT model was 83%.



```
def tune_clf_hyperparameters(clf, param_grid, X_train, y_train, scoring='recall', n_splits=3):
    '''
    This function optimizes the hyperparameters for a classifier by searching over a specified hyperparameter grid.
    It uses GridSearchCV and cross-validation (StratifiedKFold) to evaluate different combinations of hyperparameters.
    The combination with the highest recall for class 1 is selected as the default scoring metric.
    The function returns the classifier with the optimal hyperparameters.
    '''

    # Create the cross-validation object using StratifiedKFold to ensure the class distribution is the same across all the folds
    cv = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=0)

    # Create the GridSearchCV object
    clf_grid = GridSearchCV(clf, param_grid, cv=cv, scoring=scoring, n_jobs=-1)

    # Fit the GridSearchCV object to the training data
    clf_grid.fit(X_train, y_train)

    # Get the best hyperparameters
    best_hyperparameters = clf_grid.best_params_

    # Return best_estimator_ attribute which gives us the best model that has been fitted to the training data
    return clf_grid.best_estimator_, best_hyperparameters
```

```
# Hyperparameter grid for DT
param_grid_dt = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [2,3],
    'min_samples_split': [2, 3, 4],
    'min_samples_leaf': [1, 2]
}
```

```
# Call the function for hyperparameter tuning
best_dt, best_dt_hyperparams = tune_clf_hyperparameters(dt_base, param_grid_dt, X_train, y_train)
```

```
print('DT Optimal Hyperparameters: \n', best_dt_hyperparams)
```

```
DT Optimal Hyperparameters:
 {'criterion': 'entropy', 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 2}
```

```
# Evaluate the optimized model on the train data
print(classification_report(y_train, best_dt.predict(X_train)))
```

```
              precision    recall  f1-score   support

           0       0.84      0.83      0.83       510
           1       0.84      0.86      0.85       553

    accuracy                           0.84      1063
   macro avg       0.84      0.84      0.84      1063
weighted avg       0.84      0.84      0.84      1063
```

Figure 3.5.4: Decision Tree Model

3.6 Implementation Requirements

Python 3.7

It's a version made in the advanced programming language Python. Most researchers use it to carry out their study. It is a strongly suggested language for programming for machine learning-based algorithms work due to its ease of understanding and acquisition, and it is particularly well-liked by younger programmers.

Google Collab

Google Colab is a Python programming language distributor that is free and open source. We may work online here using both our browser and our Jupiter notebook. Google Collab's main benefit is that it provides us with free online virtual GPU access.

- **Hardware/Software Requirements**

❖ Windows OS (Windows 10 or above)

❖ Web Browser (Firefox, Chrome, Edge).

- Google Colab.
- HDD (8GB).
- RAM (4GB or above).

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

We began by gathering data for our model's utilization and code execution. The procedure is as follows:

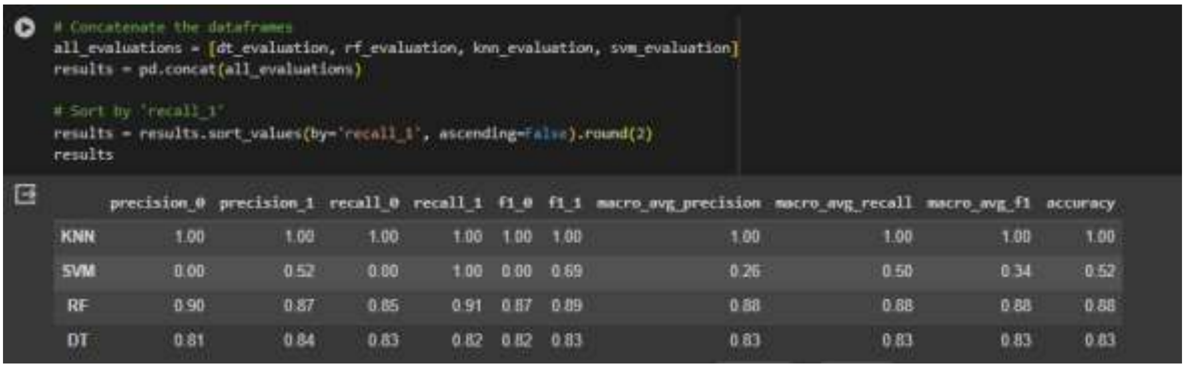
- While working on the prediction of heart disease, we gathered data from a variety of government and non-government facilities.
- The vast majority of our exploratory research time was spent acquiring information regarding cardiovascular disease from various types of hospitals.
- We have also gathered some data from online source.
- After marking the data, we found that they are ready for additional use.
- We had finalized and standardized the material at that time, so we could begin the preparation.
- We had already preprocessed our data at that moment.

4.2 Experimental result and Analysis

With a diversified and big quantity of numerical data, we had to design four distinct ML algorithms as like as SVM, KNN, RF, and DT. These models have been built and evaluated on our proprietary dataset to reach the maximum accuracy possible. Because the model's results and outputs are in statistical format. We may compare the values by arranging them in a table. The table below compares the four models and their respective assessment scores:

Table 4.1: Accuracy Table

Algorithm Name	**Accuracy of Models**
Random Forest(RF)	88%
Decision Tree(DT)	83%
Support Vector Machine(SVM)	52%
K-Nearest Neighbors(KNN)	100%



```
# Concatenate the dataframes
all_evaluations = [dt_evaluation, rf_evaluation, knn_evaluation, svm_evaluation]
results = pd.concat(all_evaluations)

# Sort by 'recall_1'
results = results.sort_values(by='recall_1', ascending=False).round(2)
results
```

```
     precision_0  precision_1  recall_0  recall_1  f1_0  f1_1  macro_avg_precision  macro_avg_recall  macro_avg_f1  accuracy
KNN         1.00         1.00      1.00      1.00  1.00  1.00                 1.00              1.00          1.00      1.00
SVM         0.00         0.52      0.00      1.00  0.00  0.69                 0.26              0.50          0.34      0.52
 RF         0.90         0.87      0.85      0.91  0.87  0.89                 0.88              0.88          0.88      0.88
 DT         0.81         0.84      0.83      0.82  0.82  0.83                 0.83              0.83          0.83      0.83
```

Figure 4.2.1: Model Accuracy Output

4.3 Discussion

We changed the dataset and mode. Based on the update, we can assume that our algorithm for classification is suitable to a diverse group of datasets and can predict their precision. We've made strides toward defining the precision of 100% impact expectations.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

With heart disease being the leading cause of disability and mortality worldwide, early detection and prevention can significantly improve patient outcomes. The use of machine learning methods might significantly improve both the precision and efficacy of heart disease prediction, with significant social ramifications. One significant advantage could be a decrease in the economic impact of cardiovascular disease on the medical field. By identifying people who have a high risk of developing heart disease, preventative measures can be taken to reduce the risk and postpone the onset of the illness. Because treating heart illness is frequently lengthy and expensive, this might result in cost reductions for both individuals and healthcare institutions. Furthermore, greater cardiac disease prognosis might lead to better patient outcomes. Through recognizing people who have a high risk of developing heart disease as soon as possible, healthcare workers can take early steps to avoid or reduce the burden of the illness. As a consequence, patients' quality of life may improve, and the number of early heart-related deaths may decrease. Another possible benefit of machine learning-based heart disease prediction is the discovery of previously identified risk variables. In general, applying machine learning to forecast cardiac illness has an opportunity to improve disease detection, diagnosis, and medication, with far- reaching social consequences.

5.2 Impact on Environment

It's unknown how the environment might be impacted by machine learning used to forecast cardiac illness. On the other side, there might be a number of unintended environmental consequences from the possible health benefits of improved cardiac illness prediction. For example, preventing and treating heart disease may lead to a decrease in the number of persons who have become incapable to go to work or participate in physical exercise. People's capacity to be not so reliant on cars and more active might lead to a decrease in the amount of fossil fuels used for transportation. Furthermore, because people with heart disease can live and work for longer periods of time, their overall resource consumption may decrease as a result of their improved health and longevity. On the other side, using machine learning to forecast heart illness may have detrimental environmental repercussions. The development and deployment of machine learning systems, for example, which may need the use of resources such as power and water, may have a negative influence on the environment. Moreover, the use of highly energy-intensive server platforms may be required for the collection and processing of enormous amounts of data in order to apply machine learning. Overall, the potential health benefits of learning algorithms for forecasting heart disease may have positive or negative indirect impacts on the environment, even if it is unclear how this technology could impact the environment.

5.3 Ethical Aspects

There are several ethical concerns to consider when using machine learning to predict heart illness. A critical ethical concern in the development and deployment of machine learning models is the risk of bias. Machine learning models are trained using data, and if the input is biased, the model may be skewed as well. A heart disease forecasting algorithm may be inaccurate for members of other racial or ethnic groups if it is predominantly built on data pertaining to one group. The model predicts that individuals from communities that are underrepresented are more unlikely to receive appropriate therapeutic and preventive care, which might result in unequal access to healthcare.

5.4 Plan for Sustainability

Sustainability challenges for heart disease prediction using machine learning include taking environmental, also social factors into account to ensure responsible and ethical techniques. To keep the model accurate and up to date, it is critical to collect fresh data and update the training dataset on a regular basis. This entails adding recent heart disease data to the training dataset on a regular basis, categorizing it, and making any required modifications. It is also critical to analyze the model's performance on a regular basis and discover any potential flaws or improvements. To stay up with the latest data, the algorithm will need to be retrained on a regular basis using the most recent datasets. The rate of data change and the level of accuracy desired will decide the frequency of retraining intervals.

CHAPTER 6
SUMMARY, CONCLUSION, FUTURE STUDY

6.1 Summary of the Study

We developed this study project to forecast both male and female cardiac disease. The project has proved that the SVM, KNN, DT, and RF classifier calculations are the best for its study. We obtained medical information from a non-government hospital with the assistance of a well-known doctor. After that, we carefully implemented data pre-handling decisions to make them appropriate for the system climate. Data gathering should be organized ahead of time in order to accomplish the data handling objective. We prepare a portion of the data sets while experimenting with others in our study.

6.2 Conclusion

Our research title is Prediction of Heart Disease, and the purpose of our work is to predict whether or not a person has cardiac disease. Each algorithm in our work is meticulously designed. We analyze the performance of four different algorithms to see which one is the most successful at identifying heart disease. We found various challenges while doing the work on the purpose of our research. The biggest challenging part was gathering datasets. After collecting all the data, we preprocessed and then processed the data. In this work we used four models from which we get preferred accuracy.

6.3 Recommendations

There are a few prominent possibilities.

- Compare and contrast several machine learning methods for heart diseaseprediction.
- Conduct a feature importance analysis to determine factors that are most significantin predicting heart disease.
- Discuss the clinical significance of the highlighted characteristics.
- Examine how data preprocessed affects model performance.
- Discuss the potential for generalization to other cardiovascular conditions.

6.4 Implication for Further Study

We will collect additional data to improve the quality of our examination.

- Investigate the incorporation of other data types, such as genetic information, lifestyle characteristics, or patient-reported outcomes, to improve the model's prediction capabilities.
- Collaborate with numerous healthcare institutions to obtain varied information, encouraging a collaborative model development and validation strategy.
- Analyze the financial effects of using machine learning to predict heart disease by assessing how it affects long-term cost-effectiveness, resource allocation, and healthcare costs.
- Investigate the ethical implications of predictive model deployment, including data privacy, informed consent, and possible biases.

REFERENCES

[1] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

[2] Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In 2020 2nd international conference on advances in computing, communication control and networking (ICACCCN) (pp. 177-181). IEEE.

[3] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.

[4] Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. International Journal of Research in Engineering, Science and Management, 2(2), 352-355.

[5] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing.

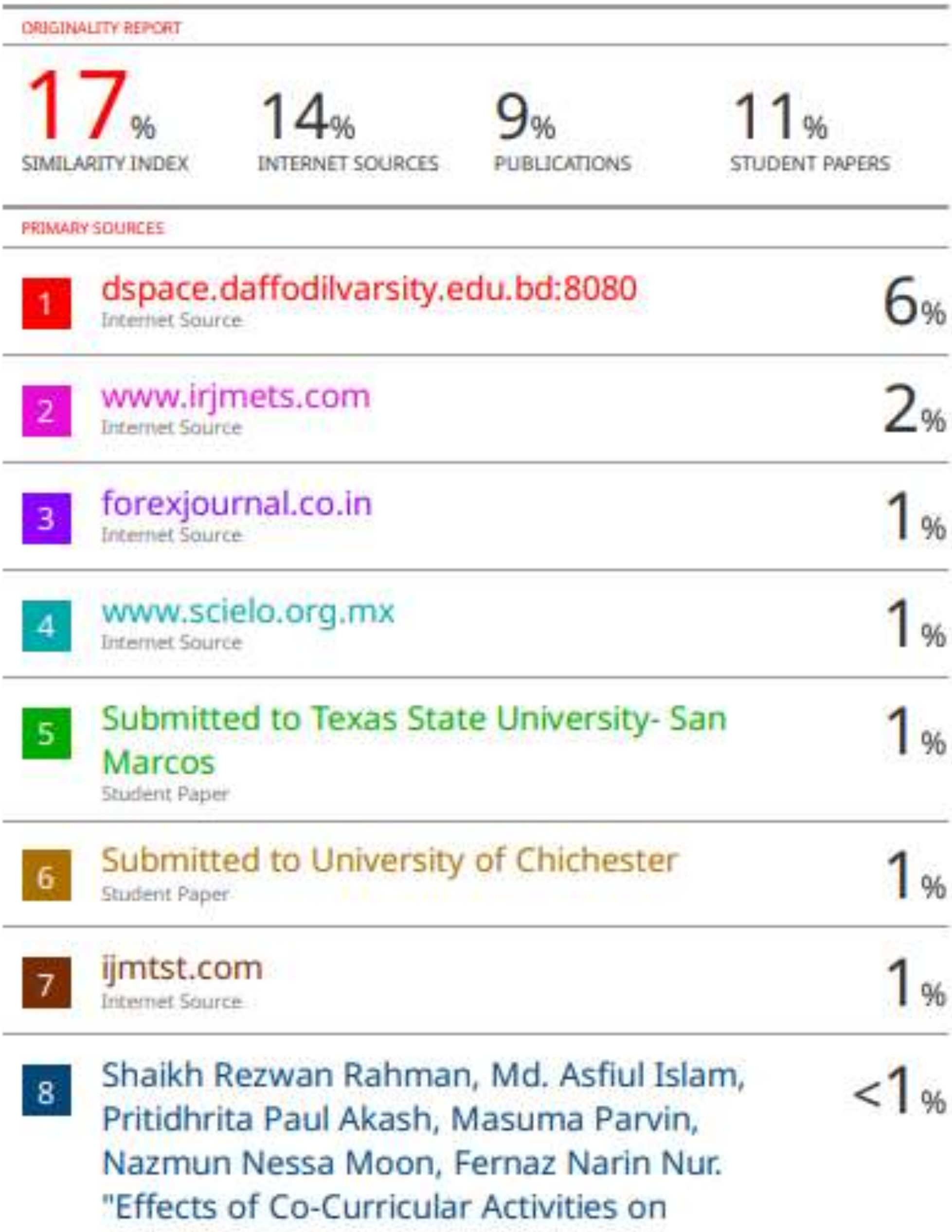
[6] Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. Health and Technology, 11, 87-97.

[7] Javid, I., Alsaedi, A. K. Z., & Ghazali, R. (2020). Enhanced accuracy of heart disease prediction using machine learning and recurrent neural networks ensemble majority voting method. International Journal of Advanced Computer Science and Applications, 11(3).

APPENDIX RESEARCH REFLECTION

We discovered numerous sorts of issues while working on this investigation. Among these were a few big issues. The first step is to gather data, next to create survey questions, and last to choose the best models. We were unable to achieve the best results utilizing other models prior to working with SVM and KNN models. We discovered several issues with creating and selecting survey questions while attempting to make them. We had a lot of trouble gathering data. Because collecting data locally takes too much time and is too difficult for students. We must rely on web sources for information as a result. And, after a lengthy period of struggle, we were successful.

Plagiarism Report



ORIGINALITY REPORT

17%	14%	9%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	www.irjmets.com Internet Source	2%
3	forexjournal.co.in Internet Source	1%
4	www.scielo.org.mx Internet Source	1%
5	Submitted to Texas State University- San Marcos Student Paper	1%
6	Submitted to University of Chichester Student Paper	1%
7	ijmtst.com Internet Source	1%
8	Shaikh Rezwan Rahman, Md. Asfiul Islam, Pritidhrita Paul Akash, Masuma Parvin, Nazmun Nessa Moon, Fernaz Narin Nur. "Effects of Co-Curricular Activities on	<1%