A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR PERSONALITY TRAITS PREDICTION

BY

MD MAHFUZUR RAHMAN
ID: 183-15-2286

AND

MD SHEBLU KHAN
ID: 183-15-2307

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

Supervised By

Amatul Bushra
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Aliza Ahmed Khan
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
FEBRUARY 2023

APPROVAL

This Project titled **"A Comparative Analysis of Machine Learning Algorithms for Personality Traits Prediction"**, submitted by Md Mahfuzur Rahman , ID No: 183-15-2286 and Md Sheblu khan, ID No: 183-15-2307 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 01 February 2023.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Lamia Rukhsara (LR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of Ms. Amatul Bushra, Assistant Professor and 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:

30.01.23

Amatul Bushra
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

30.01.23

Aliza Ahmed Khan
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Mahfuzur Rahman

Md. Mahfuzur Rahman
ID: 183-15-2286
Department of CSE
Daffodil International University

Sheblu khan

MD Sheblu Khan
ID: 183-15-2307
Department of CSE
Daffodil International University

©Daffodil International University iii

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 **Amatul Bushra, Assistant Professor**, Department of CSE Daffodil International University, Dhaka, Bangladesh. Deep Knowledge & keen interest of our supervisor in the field of "Data Mining and Machine Learning (ML)" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor **Dr.Touhid Bhuiyan Head,** Department of Computer Science and Engineering, for his kind help to finish our project and also to other faculty member and the staff of Computer Science and Engineering department of Daffodil International University.

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

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

ABSTRACT

Humans' personalities play a significant psychological role. One characteristic that affects how people interact with the outside environment is personality. A person's personality can be thought of as a crucial component of their behaviour. People's personalities are determined by how they connect with other people. Characteristic thoughts, feelings, and behaviour patterns are reflected in people's personality traits. This project aids in the creation of personality tests and the assessment of an individual's personality. The person can view their personality type and make improvements to it depending on the results of the personality classification. The goal of this study is to evaluate the performance of several classification algorithms for personality trait prediction. In this analysis, a system for predicting personality is built using the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. We provide a comparative analysis of the following data mining algorithms for personality trait prediction: decision tree, random forest, support vector machine, naive bayes algorithm, K-nearest neighbour, and artificial neural networks. We evaluate the performance of these algorithms on the Kaggle dataset. We compare all of the algorithms, the Random Forest algorithm has the highest accuracy (95.60%). Random forest outperformed the other five machine learning algorithms in terms of accuracy.

Keywords: Decision Tree, Random forest, Support vector machine, Naive bayes algorithm, K-nearest neighbour, Artificial neural networks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Another of the subjects that has received the most investigation in recent years is personality classification. A person's personality is the culmination of their behaviour and traits that determines how they react in various situations. Personality is a concept that describes how people behave in various contexts. Important human traits like happiness, emotions, and melancholy are all reflected in it, as well as human behaviour [1]. People's life choices are influenced by their personalities. Utilizing these strategies, user attributes are mined in order to train the model to anticipate future user personalities. Numerous tools, such as online job and book suggestion services, custom search engines, mental health screenings, etc., might be useful for automatically identifying personality traits [2].

This paper's primary goal is to provide an overview of the data mining methods that are used to determine a user's personality. Big five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism [24].

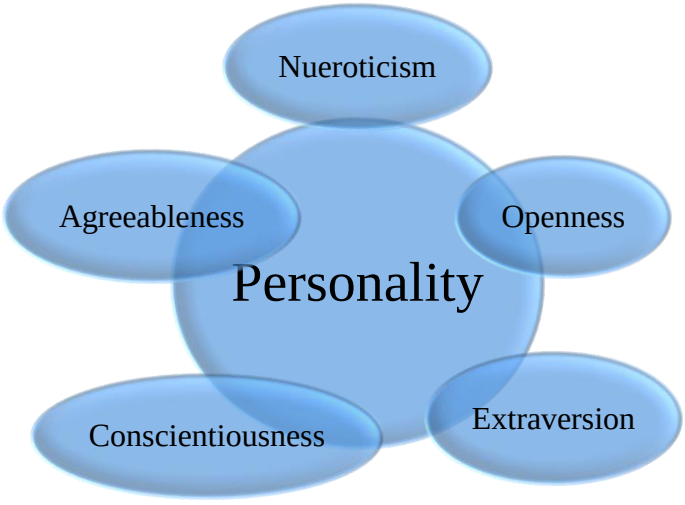

Fig 1.1: Model of Personality Trait

Predicting personality by analysing a person's behaviour is an ancient technique. This manual personality prediction method was time and resource intensive [3]. A lot of human effort must be expended in order to conduct a personality analysis based solely on one's own temperament. This traditional method of predicting personality is time consuming and very limited in scope. Also, this manual analysis yielded inaccurate results when analysing user personality based on user dispositions and behaviours [10]. Analysis was done manually, which affects the accuracy of the results.

The Big Five attributes and personality traits are related in the following ways:

• **Openness**: Openness places the greatest emphasis on creativity and insight among the five personality traits. People with an open mind frequently get a broad spectrum of interests. Their curiosity about the world and other people inspires them to learn new things and enjoy new experiences [1].

• **Conscientiousness:** Deep levels of thought, strong impulse control, and goal-oriented behaviour are characteristics of the human trait known as conscientiousness. Structured and meticulous people tend to be highly conscientious. They are structured, deadline-driven, and sensitive to others' feelings [3]. Those who do poorly on this important personality trait lack organization. They could procrastinate on tasks, utterly missing deadlines as a result.

• **Extraversion:** People who are extroverted express their emotions and make their opinions known to others. People that possess this attribute have good social skills. These folks have the typical traits of being energetic, enthusiastic, and gregarious. Without giving it any consideration, they prefer to act and observe results [23].

• **Agreeableness**: Generosity-based individual behaviour analysis is known as agreeability. This personality trait encompasses prosocial attributes like compassion, trustworthiness, generosity, affection, and other prosocial behaviours [20]. People who exhibit these qualities are immensely trustworthy, caring, and cooperative. People are understanding, kind, and considerate.

• **Neuroticism**: The term "neuroticism" refers to a person's emotional stability, mood swings, and sulkiness. Irritability, impatience, worry, and despair are common traits of high neurotic

people [9]. Less of this personality attribute is associated with more emotional stability and resilience.

1.2 Motivation

There are many reasons why it is important to accurately predict personality traits. Personality prediction has many potential uses, including improving the accuracy of hiring decisions, predicting consumer behaviour, and understanding how people interact with technology. For example, employers may want to know if an applicant is a good fit for a position before making an offer, or educators may want to identify students who may benefit from additional support. There are many different data mining algorithms available for personality prediction, each with their own strengths and weaknesses. This article compares and contrasts his six most commonly used algorithms: Decision trees, Random forests, Support vector machines, Naïve bayes algorithm, K nearest neighbours, artificial neural networks. We evaluate each algorithm based on accuracy, interpretability, and computational efficiency. Ultimately, our goal is to find the best algorithms for predicting personality traits.

1.3 Rationale of the Study

This research compares and contrasts various procedures for data mining predicting personality traits. Data mining techniques have attracted more attention recently as a means of predicting personality. Data mining is to blame for this, as opposed to more conventional approaches like questionnaire-based surveys, can offer a more effective and automated technique to forecast personality traits. We compare and contrast various techniques for personality trait prediction via data mining in this research. In this study, we examine the accuracy of six well-known data mining techniques for predicting personality traits: Decision Tree, Random Forest, Support Vector Machine, Naive Bayes method, K-nearest Neighbour, and Artificial Neural Networks. To train and test our models, we will use a dataset with

information on more than 1024 individuals. Our goal is to find the most accurate algorithm for predicting personality traits.

1.4 Research Questions

As like as research question for this thesis are:

1. What are the data mining techniques used to predict personality traits?
2. What is the most effective Data Mining Algorithm for Personality Traits Prediction?
3. Are there any other factors that could affect the effectiveness of six algorithms?
4. Is there a difference in the effectiveness of different types of data mining algorithms for personality trait prediction between women and men?

1.5 Expected Output

Understanding what is meant by a personality characteristic is required before performing a comparative analysis of various data mining techniques for predicting personality traits. Characteristics that make up a person's personality and can be utilized to forecast future behaviour are known as personality traits. Although there are many alternative techniques to measure personality traits, datasets are the most widely used approach. Once the data has been gathered, a variety of Data mining methods can be utilized to ascertain which model is the best predictive. Decision Tree, Random Forest, Support Vector Machine , Naive Bayes approach, K-nearest Neighbour, and Artificial Neural Networks are a few of the most well-liked algorithms. It is crucial to select the appropriate algorithm for the particular dataset because every algorithm has different advantages and disadvantages. A set of guidelines or a mathematical model that may be used to forecast the likelihood of specific behaviours would be the anticipated result from a data mining algorithm for predicting personality traits. But it's crucial to keep in mind that the objective is to strike the ideal balance between precision and effectiveness for the particular work at hand.

1.6 Project Management and Finance

Numerous data mining algorithms are available for predicting personality characteristics. We will compare the accuracy and efficiency of six alternative algorithms, including decision trees, support vector machines, and logistic regression, in this comparative analysis. Due to their simplicity of use and interpretability, decision trees are a popular choice for predicting personality traits. They are, however, frequently less precise than alternative techniques, especially when there are numerous variables at play. Compared to decision trees, support vector machines (SVMs) are a more sophisticated method that can handle complex data sets. Several studies have demonstrated that SVMs are more accurate than decision trees, despite the fact that their implementation can be more challenging. All six techniques are equally efficient in terms of computer resources needed. The quickest to train are usually decision trees. Another family of supervised learning technique that can be applied to both classification and regression applications are decision trees and random forests. The findings show that support vector machines and decision trees are more precise than neural networks. Overall, when it comes to predicting personality traits, each of these five data mining algorithms has its own advantages and disadvantages. But when the data collection is sizable and the training set contains data that is broadly representative of the population, all six algorithms perform well.

1.7 Report Layout

We organized the complete research into several chapters. First, we go through some background information in chapter one's introduction. We also break it down into subsections for things like goals, inspirations, research questions, and problems. On the other hand, chapter two provides an overview of the available research publications in this area. Research methodology was covered in Chapter three of our discussion. Results and Discussions were the topics of Chapter four while Conclusions and Future Work were the topics of Chapter five.

CHAPTER 2
BACKGROUND

2.1 Preliminaries/Terminologies

In order to investigate the feasibility of using data mining algorithms for predicting personality traits, it is necessary to first understand the preliminaries and terminologies associated with the topic. Data mining algorithms can be used for a variety of purposes, but in the context of personality prediction, they are used to identify Data patterns and correlations that can be employed to forecast a person's personality. These characteristics are thought to remain relatively stable throughout an individual's lifetime and influence his or her thoughts, emotions, and behaviours. There are many different models of personality, but the most popular one is the Big Five model, which includes the following five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism.

2.2 Related Works

Desai, Helly N., and Rakesh Patel, et al. [1] people update their thoughts, opinions and activities on social media which can be analysed to discover the personality of the social media user. To survey the prediction of personality traits more accurately by combining user's demographics, features, likes and activities. Facebook is a widely used social networking site among all the other social networking sites.

William, P., and Abhishek Badholia, et al. [2] there are numerous classification algorithms focused on bayes, trees, functions, or laws that are commonly used. There are several machine learning algorithms for classifying rules and performing various functions. This article compares classification algorithms used to identify and forecast personality using the HEXACO Model dataset and backs it up with implementation performance. Logical

Regression and Naive Bayes algorithms performs better than the other four algorithms with best accuracy of 95%.

Katiyar, Sandhya, Himdweep Walia, et al. [3] provides automated Personality Classification – a system that analyses the personality of a user based on certain features using Data Mining Algorithms. In this paper, a system is proposed which analyses the personality of an applicant. The personality prediction results are based on Big Five Personality traits and the classification is done using Naïve Bayes Algorithm and Support Vector Machine. Naive Bayes algorithm has the best accuray of 60%.

Singh, B., & Singhal, S. et al. [4] proposed the report and discuss in this part how well the categorization algorithms worked in predicting the user's personality. The Naïve Bayes Algorithm, with an average accuracy of about 60%, has the best accuracy of the two evaluated techniques. The study and prediction of personalities has grown significantly in recent years. The current method makes it very easy to extract a user's personality, which is quite useful in many areas including the hiring process and medical counselling, among others. Personality detection in surveys refers to the process of identifying the user's behaviour patterns. This study's main objective is to provide a cutting-edge review of the developing field of personality detection through surveys.

Zhong, Xiao-Feng, et al. [5] their analysis for predicting personality traits, they used supervised learning. First, the Facebook Graph API is used to collect the data. Many pre-processing techniques are used before categorizing personality traits. It entails removing URLs, hashtags, tags, emoticons, and other content. They suggested an approach to remove the data's outliers. They did it by using the Mahalanobis distance measure and the z-score value. Z-score value aids in the detection of outliers. Outliers are then eliminated. Following the removal of outliers, various classification techniques, including KNN and SVM, are used. SVM algorithm has the best accuracy.

Dutt, Ashish, Maizatul Akmar Ismail, et al. [6] had said developing techniques for simulating student behaviour using data from online dialogues, classroom debates, etc. However,

techniques such as Educational Data Mining (EDM) and Intelligent Tutoring System (ITS) utilized a person's personality and behaviour for analysis. A system that can be customized by the user and that also analyses how students interact with one another is created.

Gjurković, Matej, et al. [7] the few datasets that are made publicly available include minimal topic diversity, high labelling costs, and privacy concerns. By introducing a sizable dataset collected from Reddit—a source hitherto ignored for personality prediction—they attempt to solve this issue. The dataset includes a wealth of characteristics for more than 9k users and is tagged with Myers-Briggs Type Indicators (MBTI). The dataset is used to train and assess benchmark personality prediction models, with macro F1-scores on the individual dimensions ranging from 67% to 82% and an accuracy of 82% for exact or one-off correct type prediction.

Golbeck, Jennifer, et al. [8] Social networks are a platform where users tend to expose themselves to the public, providing details about their conduct and letting others inside their life. Personality has a significant role in a variety of interpersonal interactions and can be used to forecast success in both romantic and professional relationships as well as work satisfaction. Until now, they used a survey test to poll a variety of people in an effort to precisely forecast a user's personality. Correct personality analysis was a challenge because this was highly impractical while gathering data from social media networks.

Mehta, Yash, et al. [9] there has been a lot of interest in the automatic prediction of personality traits. In particular, the prediction of personality traits using multimodal data has become a hot topic in the field of emotional computing. With a focus on deep learning-based techniques, they discuss key machine learning models that have been used for personality recognition. gives a general overview of the most widely used methods for automated personality detection, as well as information on different computational datasets, commercial uses, and cutting-edge machine learning models for personality detection with a focus on multimodal methods.

Artissa, YB Nunung Damayanti, et al. [11] People frequently use social media as a platform to represent oneself to the public by disclosing personal information and aspects of their lives. To find out the personality of a person based on the statuses they write on Facebook use text

classification techniques using Multinomial Naive Bayes method. The test result shows that accuracy increases after reducing the number of word variations and variations of prior probability values. The accuracy obtained by using stemming in the pre-processing process is 59.9% and using uniform prior increases 0.3%.

Başaran, Seren, et al. [12] Use of social media, and particularly Facebook, is increasing daily and rapidly. The purpose of this project is to develop a neural network-based predictive model that forecasts the Big 5 personalities using data and behaviour from Facebook users. 7,438 distinct Facebook users were selected for the study using the Personality database. The results of this study demonstrated how much the personality of an individual can be predicted simply by looking at their Facebook activity. An individual's personality could be accurately classified using the suggested artificial neural network model with 85% prediction accuracy.

Pletzer, Jan Luca, et al. [20] they compared the Big Five domains' and HEXACO domains' predictive accuracy of workplace misbehaviour. The strongest correlation between workplace deviation and HEXACO Honesty-Humility, Big Five and HEXACO Consciousness, and Agreeableness. These findings largely confirm the conceptual distinctions between the HEXACO personality model and the Big Five personality model. Compared to the Big Five domains (19.05%), the HEXACO domains (31.97%) explain more variation in workplace deviance.

Subramanian, Ramanathan, et al. [21] Using market physiological sensors, they propose ASCERTAIN, a multimodal database for implicit personality and affect recognition. Big-five personality measures, emotional self-ratings, electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), and facial activity data from 58 users are all included in ASCERTAIN. These measurements were taken while the participants watched affective movie clips. Before studying linear and non-linear physiological correlates of emotion and personality, they first investigate correlations between users' emotional assessments and personality scales in the context of earlier observations. According to their analysis, non-linear statistics are preferable to linear ones for describing the link between emotions and personality.

Utami, Ninda Anggoro, et al. [23] Focuses on categorizing Facebook users' personalities into one of the Big Five Personality Traits. 170 Facebook users who volunteered to participate in the study were invited to complete the Big Five Inventory form. The classifier is created using data mining techniques employing Support Vector Machine (SVM) based on the collected data with the goal of determining a person's personality based solely on their Facebook account without the need for a questionnaire. SVM's best accuracy rating is 87.5%.

Maharani, Warih, et al. [24] Researchers have conducted interdisciplinary studies looking at the connection between personality traits and social media activity in response to the growing popularity of social media. Examines a number of machine learning methods based on semantic characteristics including mood, sentiment, and publicly accessible Twitter profiles, such as naïve Bayes (NB), K-nearest neighbours (KNN), and support vector machine (SVM). The Big Five personality model, which is the best model for predicting user personality in social media, can be used to predict personality. The accuracy test outcomes are SVM (59.45%), NB (45.92%), and KNN (48.02%).

2.3 Comparative Analysis and Summary

Personality prediction has many potential uses, such as better understanding how people interact and behave in social situations, predicting consumer behaviour, or providing personalized recommendations application. However, before using personality predictions effectively, it is important to understand what personality is and how it can be measured. Personality can be defined as the consistent set of behaviours, thoughts, and emotions that define a person. It is often seen as a set of stable traits that influence how someone interacts with the world. There are many ways to measure personality, but the most common is the Big Five model of his personality. This model includes her five broad areas of personality. Neuroticism, extroversion, agreeableness, conscientiousness, openness to experience. Analysing patterns in data allows data mining systems to predict personality traits. For example, if we want to predict a person’s neurotic score (which measures how upset or stressed they are), whether they frequently post negative emotions on social media, or whether they

have a history of anxiety disorders. We can examine factors such as whether there is a by finding patterns in your data, we can create models that accurately predict individual outcomes in the Big Five personality model. For this objective, a variety of data mining algorithms are available, but the most common options include Decision trees, Random forests, Support vector machines, Naïve bayes algorithm, K nearest neighbours, Artificial neural networks. Which of the following algorithms is the best for personality prediction? It depends on the specific situation. Each algorithm has its own strengths and weaknesses, so choosing the right algorithm for the task at hand is important

2.4 Scope of the Problem

One must first comprehend the extent of the issue in order to conduct a thorough and objective analysis of data mining algorithms for personality traits prediction. This thesis compares and contrasts several data mining techniques in order to assess how well they can forecast personality traits. This will be accomplished by using a dataset made up of ratings for various people's personality traits. The training and test sets of this dataset will be separated, and various techniques will be used to each set in order to create a prediction model. In order to identify which method is most efficient for this work, the models' accuracy will be assessed using a variety of criteria, and the results will be compared.

2.5 Challenges

The study of personality has benefited greatly from the development of data mining techniques. Using their personality traits, behaviour, and surroundings, they can be used to predict people’s personality attributes. Data mining is the process of removing information from enormous amounts of data collected from numerous sources and using it to analyse, synthesize, and foresee different patterns. Each algorithm uses a unique set of guidelines to assess data and make predictions. A sample of 1024 people’s data traits were gathered. The outcomes demonstrated that Random Forest had a higher accuracy in predicting personality traits.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

Advanced technology requires more complex and refined instruments due to the growing amount of data. Although advances in data mining technologies have made huge data collecting much more prudent, there is always a constant need for innovative methods and tools that may assist in turning this data into insightful knowledge. The user is provided with an understanding and practical application of the theory and practice of finding patterns concealed in the vast data set by concept and techniques, continuing a long heritage. We are utilizing SPSS, Python 3.7 and Microsoft Excel in this work, and our study topic is "A Comparative Analysis of Data Mining Algorithms for Personality Traits Prediction."

3.2 Data Collection Procedure

Information is gathered from Kaagle for Personality Traits Prediction. The place where data scientists spend their evenings and weekends is called Kaggle. To address issues using predictive analytics, machine learning, and data science it is a crowdsourced platform that draws, trains, educates, and tests several data scientists over the world. Around 536,000 of its users are active, and it receives about 150,000 submissions each month. On Kaggle, enthusiasts for data science from all around the world fight for rewards and to advance in the rankings. Data is always the prior issue to have some result in any field of work. To complete this process we needed the datasets and we collected it from the kaggle website. The data collection procedure is a survey questionnaire and interview conducted with the peoples based on their knowledge of Data Mining Algorithms, personality traits prediction and practical experience. The dataset with eight attributes that we used in our study. These include gender, age, openness, neuroticism, conscientiousness, agreeableness, extraversion, and personality.

3.3 Statistical Analysis

The process of reviewing, purifying, manipulating, and modelling data for statistical analysis is done in order to get meaningful information, insights, or conclusions. A comparison of data mining algorithms for predicting personality traits is essential to understand the potential of each technique and to know how they can be improved. There are a variety of data mining algorithms that can be applied to predict personality qualities. In this article, we emphasize six state-of-the-art supervised learning algorithms for personality traits prediction: Decision Trees, Random Forests, Support Vector Machines, Naive Bayes Algorithm, K-Nearest Neighbours, and Artificial Neural Networks. Each algorithm has its own strengths and weaknesses, so it is important to select the right algorithm for the specific data set and prediction task at hand. Research comparing various data mining techniques for predicting personality qualities have shown that some methods are more accurate than others. We compare these methods using a dataset of over 1024 people's personality traits. The study found that there are a number of different ways that Personality traits can be predicted using data mining methods and that the most accurate predictions come from using a combination of algorithms. The study also found that the accuracy of predictions varied depending on the type of personality trait being predicted.

3.4 Proposed Methodology

The dataset for this study is split into two parts: a training set and a testing set. 80% of the data are saved in the training set, while 20% are in the testing set. In this process first, we make a clear diagram to show us the path that how we are going to do our work. The path is pre-processing that is seen in the diagram first we start the process then collected the data from Kaggle. We have tried to find the real dataset from various institution but we have failed then we plan if we can work on the demo data’s we will have the experience to work on the real that’s why we didn’t stop/limit ourselves we keep working as per the flowchart diagram.

Figure 1 below shows the basic methodology.

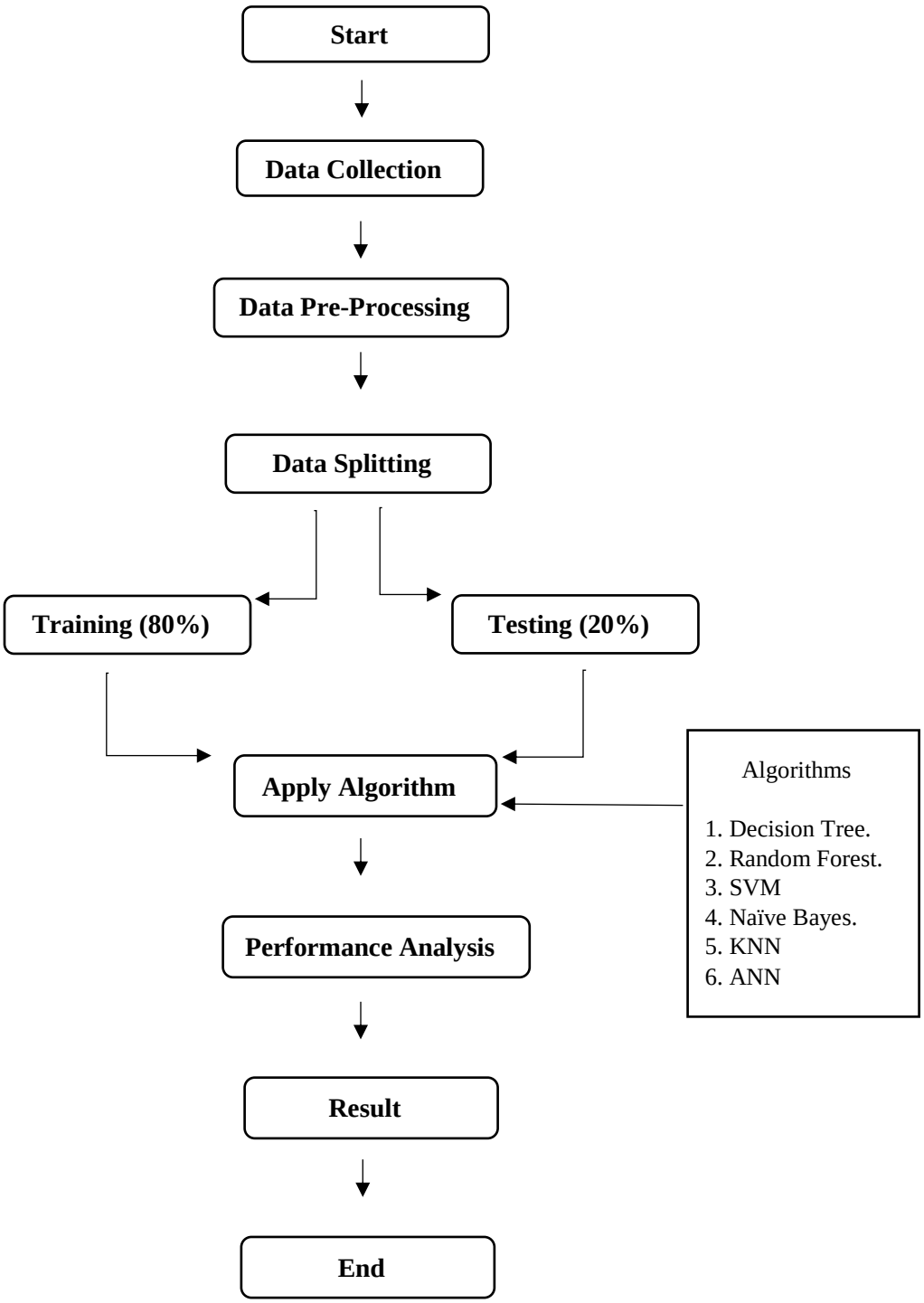

Fig 3.4.1: Research Methodology

After collected the datasets, the data was not ready to apply the algorithms so we pre-processing of data. Then we splitting the data sets into two parts: training and testing. The training (80%) percentage of the data and the rest (20%) of the data as testing dataset. After completing the process, we have applied six algorithms to find the result.
All machine learning classifiers have been studied in detail before evaluating the ease of use and to obtain better optimal results. Detailed study summary given below:

• **Decision Tree:** Data is categorized using attributes and categories using decision trees. We must treat little variance as significant since categorization is difficult and decision trees may produce contradictory results as a result of the different types of input. Information in numbers and categories [4].
A decision tree is the best solution for classification issues since it uses trained rules. It divides the data into "tree trunks, followed by branches and leaves".

• **Random Forest:** The forest-based estimator fits a variety of decision trees to a range of sub - samples while using the average to prevent model overfitting. Except by replacement, the subsample nearly always closely resembles the originating sample. It is a challenging and drawn-out project [1]. Sometimes using a random forest and less overfitting will result in more accurate results.
The random forest algorithm is a version on the decision tree technique in that it builds some kind of real-world decision trees from your training dataset and then contrasts the fresh results. The distance to connect to the nearest tree is set on the data scale. Since random forests don't force data points into arbitrary categories, they are effective.

• **Support Vector Machine (SVM):** Auxiliary vectors are used to describe machine learning machine data as points in space. The subsequent sorting of new examples takes into account which ones are located in the same area [2]. The high-dimensional space uses a subset of the training points in the evaluation, as the time-consuming 5-fold cross validation is not included in the algorithm. Algorithms are used to teach and classify data by degree of polarity. Red and blue tags, as well as X and Y data features, are used in the presentation. The farthest hyperplane

is then located via SVM. This is only in two dimensions. A double blue emphasis is applied to anyone on either side of the red line. Feelings that are both "pleasant" and "awful."

• **K-Nearest Neighbours (KNN):** Since it doesn't aim to create an all-encompassing model, this subset-based learning approach to classification is a subset-based method. A resounding majority of the k nearest points decides on the formation of the regiment. There is a large computational expense associated with calculating the distance between each instance and the entire training set [4]. The technique works well with large amounts of data, is simple to use, and can tolerate noisy data.

KNN uses training data to determine a trend in family members' places. A classification algorithm, it can find information nearby its companions. It will go to the class having the best chance of being number one. By using majority rule, cautious decisions are made.

• **Naïve Bayes:** A Nave Bayes model is based on the Bayes theorem and infers independence for each function. Nave Bayes models excel in practical spam filtering and document recognition are examples of practical uses. A minimal amount of information is required for parameter estimation [20]. Naïve classifiers operate very quickly. To determine if a data point belongs to a specific group, Nave Bayes is utilized. In textual processing, a name can be recognized by the words and phrases that make up it.

• **Artificial Neural networks (ANN):** Machine learning algorithms of a specific kind are called artificial neural networks that get their inspiration from how the real neural circuits. In other words, ANNs are able to learn from the data in a manner similar to how the neurons in our nervous system are able to do so and create solutions in the form of predictions or classifications.

By revealing a complex link between inputs and outputs, ANNs aim to detect unique patterns. Many different activities, including speech recognition, machine translation, image recognition, and medical diagnosis, are carried out using these artificial neural networks.

A significant advantage is that ANN can learn from sample data sets [5]. The use of ANN to approximate a random function is its most popular use. These technologies make it possible to define the distribution's solutions in a practical way.

3.5 Implementation Requirements

The current study compares and contrasts various data mining techniques for predicting personality traits. It is critical to consider the many implementations that are needed for each algorithm in order to create a reliable prediction. There are several factors to consider here, including as the type of data used as input, the size and complexity of the data collection, the quantity of personality traits that need to be predicted, and the level of precision that is required. The available processing resources and the time restrictions must also be taken into account. Six distinct algorithms—support vector machines, decision trees, artificial neural networks, k-nearest neighbours, Naive Bayes Algorithm, and Random Forests—were chosen to be used for this study after taking into account all of these factors.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

The task of predicting personality traits from data can be approached in a variety of ways. We outline the experimental setting that we employed in our comparison of data mining algorithms for predicting personality traits. we discuss the evaluation measures that were employed to contrast the effectiveness of the various algorithms. six widely used data mining methods are compared and contrasted for this purpose: Decision Tree, Random forest, Support vector machine, Naive bayes algorithm, K-nearest neighbour and artificial neural networks. We evaluate the performance of each algorithm on a publicly available dataset of personality profiles. The dataset for this study is split into two parts: a training set and a testing set. 80% of the data are saved in the training set, while 20% are in the testing set. This study employed the K Neighbour Classifiers approach, which divides the training dataset into k equal-sized chunks. For each iteration, one split is reserved for while the remaining k-1 splits are kept as training data, the evaluation dataset.

The figure of Training set and testing set is given below:

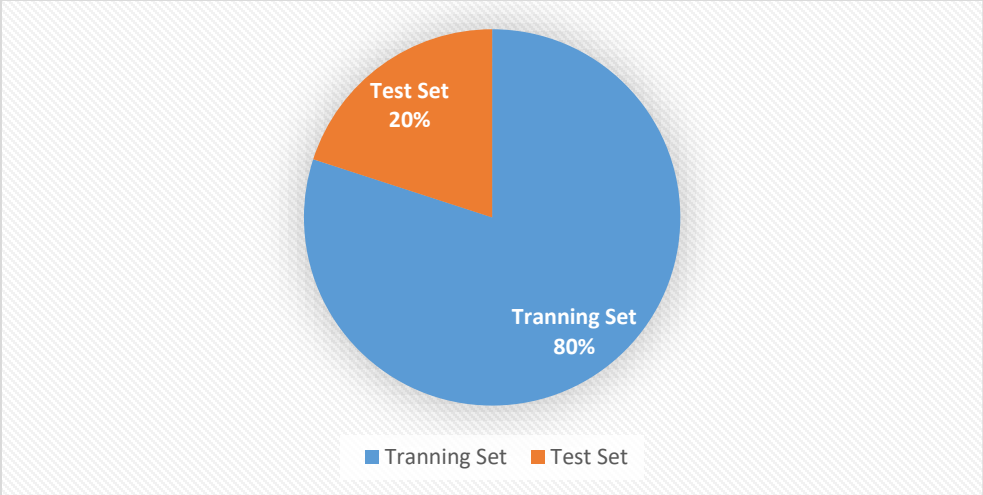

Fig 4.1.1: Training set and testing set

4.2 Experimental Results & Analysis

Error Rate:

A confusion matrix provides an overview of categorization issue prediction results. The number of correct and incorrect predictions is counted and divided down by class using count values. The confusion matrix's secret can be found here. The confusion matrix illustrates the various ways that your categorization model could be interpreted incorrectly.

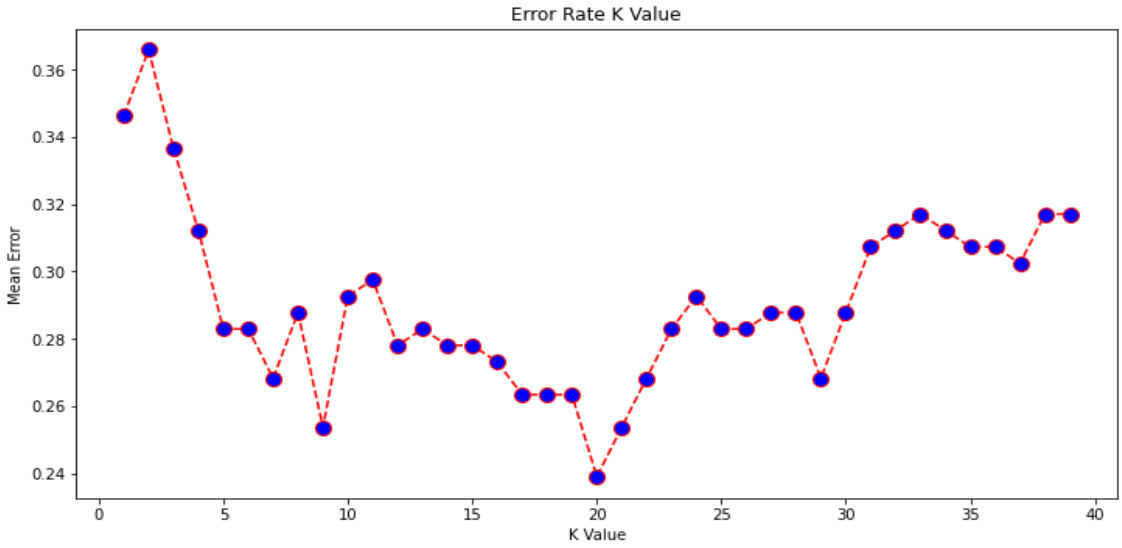

Fig 4.2.1: Error rate K value.

Table 4.2.1: Accuracy Table of Algorithms

	Used Algorithm	Accuracy (%)
1	Random Forest	95.60
2	Support Vector Machine	93.65
3	Artificial Neural Network	90.24
4	Naive Bayes	86.82

5	K- Nearest Neighbour	74.63
6	Decision Tree	16.58

The Confusion matrix shows the intended and actual categorization. As a result, a number of standards, including accuracy, precision, recall, and F1 Score, are developed using a confusion matrix.

Figure shows formulas using a 2 x 2 confusion matrix.

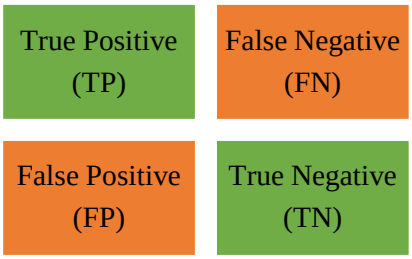

Figure 4.2.2: Confusion Matrix

• **TP: True Positive:** The model expected a positive value, and the actual result was positive.

• **FP: False Positive:** Your statement is both accurate and wrong.

• **FN: False Negative:** Your prediction is negative, and result it is also false.

• **TN: True Negative**: As predicted by the model, the actual value was negative.

Precision, Recall, and F1 score are all measures of performance for classification tasks.

Precision is calculated by dividing the total number of positive outcomes by the number of correct positive results. Recall is calculated by dividing the total number of true positive results by the total number of positive results that were expected. The harmonic mean of recall and precision, or the F1 score, provides a balance between the two.

Precision, recall, and F1 score are all measures of performance for classification problems.

Precision is a measure of how many of the items that were predicted to be positive are actually positive. In other words, it is the proportion of true positives to all items that is calculated were predicted to be positive.

Recall is a measurement of the proportion of positive items that were actually anticipated to be positive. In other words, it is calculated by dividing the number of genuine positives by the total number of things that are actually positive.

F1 score is a metric that combines recall and precision. It represents the harmonic mean of recall and precision. The harmonic mean is a type of average that is used when there is a need to balance two conflicting quantities.

The precision, recall, and F1 score can be calculated using the following formulas:

Precision = True Positives / (True Positives + False Positives)

Recall = True Positives / (True Positives + False Negatives)

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

The performance of K-nearest neighbour:

Table 4.2.2: For K-nearest neighbour

	0	1	2	3	4
Precision	0.69	0.72	0.75	0.74	0.71
Recall	0.73	0.89	0.38	0.74	0.51
F1 Score	0.71	0.80	0.50	0.74	0.60

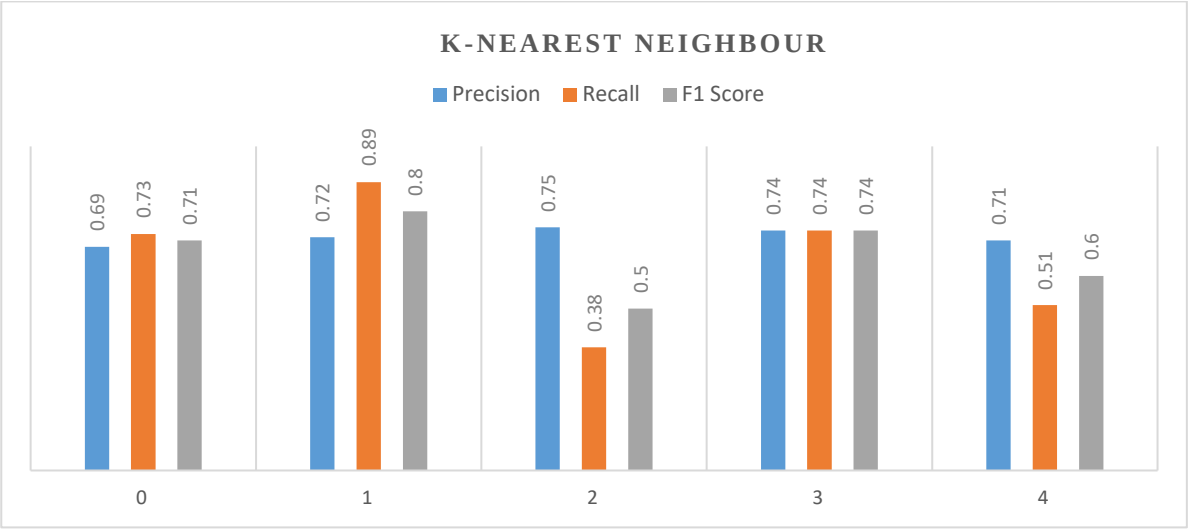

Figure 4.2.3: K-Nearest Neighbour

Performance of Decision Tree:

Table 4.2.3: For Decision tree:

	0	1	2	3	4
Precision	0.69	0.72	0.75	0.74	0.71
Recall	0.73	0.89	0.38	0.74	0.51
F1 Score	0.71	0.80	0.50	0.74	0.60

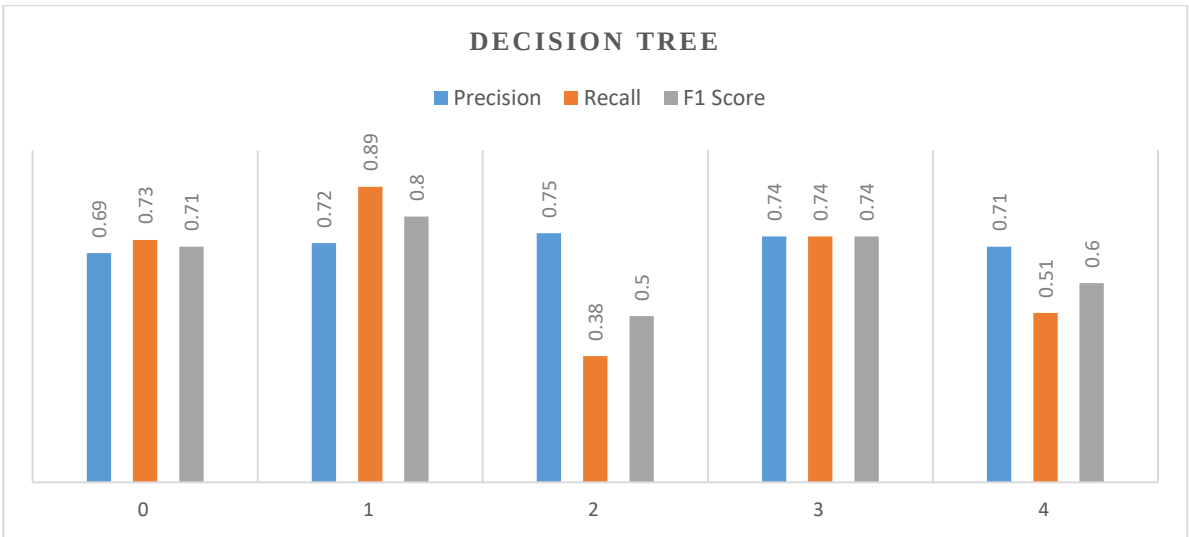

Figure 4.2.4: Decision Tree

Performance of Naïve Bayes:

Table 4.2.4: For Naïve Bayes:

	0	1	2	3	4
Precision	0.69	0.72	0.75	0.74	0.71
Recall	0.73	0.89	0.38	0.74	0.51
F1 Score	0.71	0.80	0.50	0.74	0.60

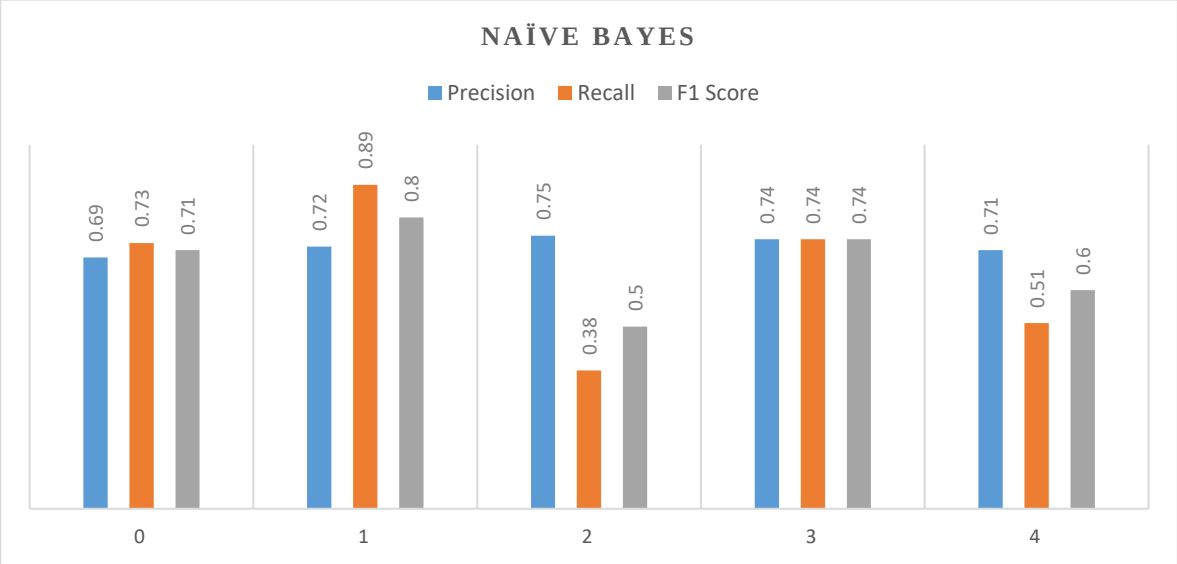

Figure 4.2.5: Naïve Bayes

Performance of Support Vector Machine:

Table 4.2.5: For Support Vector Machine:

	0	1	2	3	4
Precision	0.69	0.72	0.75	0.74	0.71
Recall	0.73	0.89	0.38	0.74	0.51
F1 Score	0.71	0.80	0.50	0.74	0.60

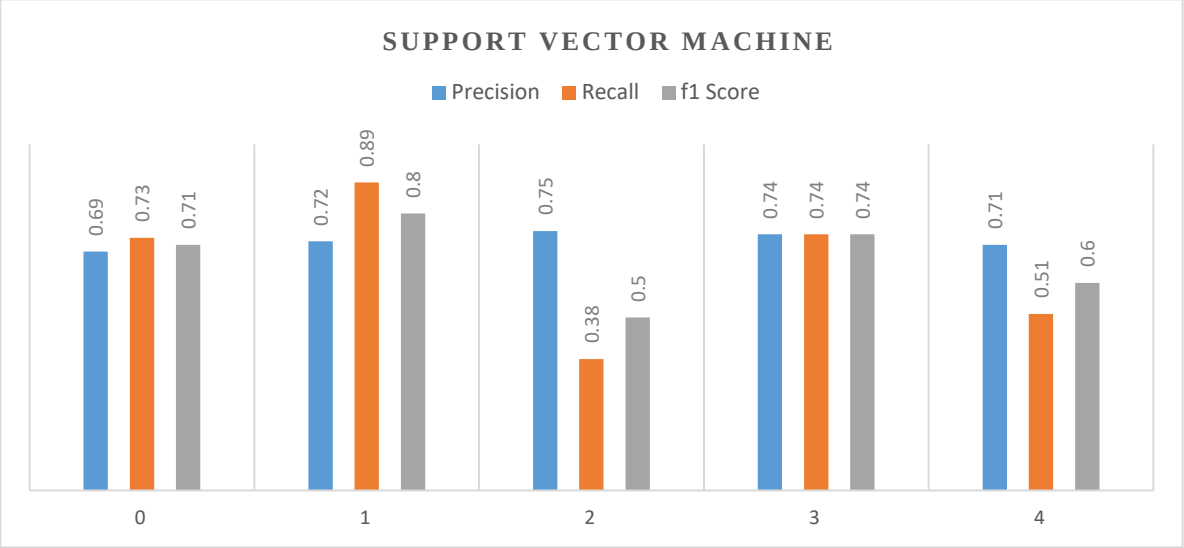

Figure 4.2.6: Support Vector Machine

Performance of Random Forest:

Table 4.2.6: For Random Forest:

	0	1	2	3	4
Precision	0.69	0.72	0.75	0.74	0.71
Recall	0.73	0.89	0.38	0.74	0.51
F1 Score	0.71	0.80	0.50	0.74	0.60

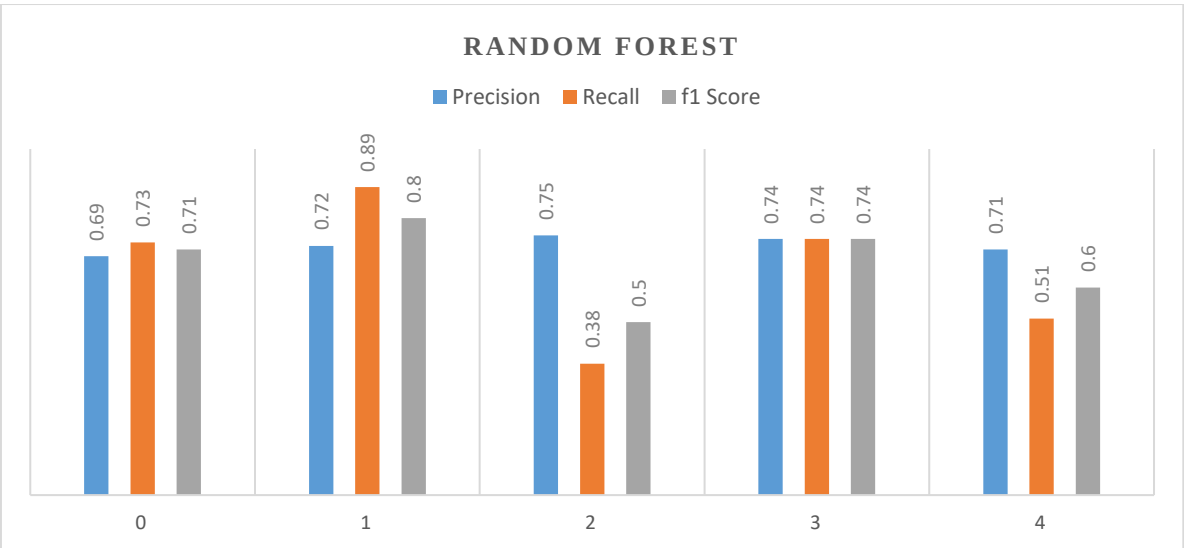

Figure 4.2.7: Random Forest

Performance of Artificial Neural Network:

Table 4.2.7: For Artificial Neural Network:

	0	1	2	3	4
Precision	0.89	0.94	0.80	0.82	0.97
Recall	0.91	0.92	0.75	0.97	0.87
F1 Score	0.90	0.93	0.77	0.89	0.92

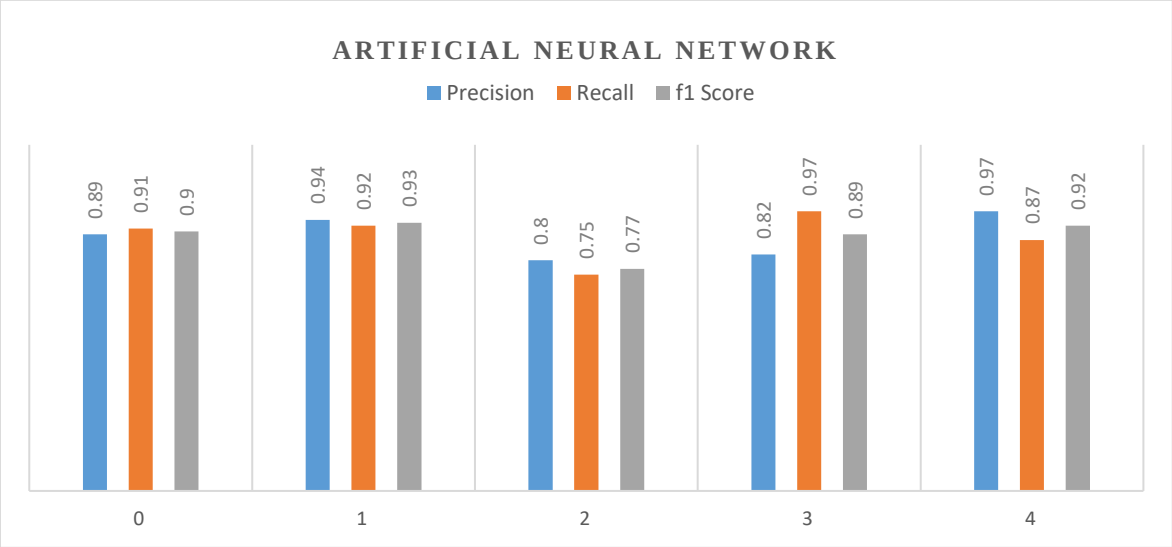

Figure 4.2.7: Artificial Neural Network

4.3 Discussion

The aim of this study was to comparatively analyse different data mining algorithms for their ability to predict personality traits. The dataset used in this study contained information on a variety of individuals, including their personality scores on the Big Five personality dimensions. Six different data mining algorithms were applied to the data, and the results were compared. Overall, all six of the algorithms performed well in terms of accuracy. However, some algorithms fared better than others. The research conducted a comparative analysis of six different data mining algorithms for predicting personality traits, including Decision Trees, Random Forests, Support Vector Machines, Naive Bayes Algorithm, K-Nearest Neighbours,

and Artificial Neural Networks. The result shows that Random forest outperforms the other five algorithms in terms of accuracy, of 95.60%. Support Vector Machines comes in second place with an average accuracy of 93.65%. ANN comes in third place with an average accuracy of 90.24%. Naive bayes comes in fourth place with an average accuracy of 86.82%. Followed by K-Nearest Neighbours (74.63%) and Decision Tree (16.58%). In terms of further study, it would be interesting to investigate why Random forest performs better than the other five algorithms in this particular case. Overall, this research provides valuable insight into which data mining algorithms are most effective for predicting personality traits.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

A comparison of personality trait prediction data mining algorithms can have a big impact on society. For instance, the five-factor model, which predicts personality based on five main categories: neuroticism, extraversion, openness, agreeableness, and conscientiousness, is the most widely used algorithm. Although this technique is frequently regarded as being highly accurate, it does have some restrictions. For instance, it disregards particular actions or environmental conditions that may affect a person's personality. The Big Five Inventory is a well-known program that is similar to the five-factor model but contains a greater range of distinct features. Although the five-factor model is thought to be more accurate and sophisticated, this method requires more time and money to develop. In the end, the particular requirements and objectives of the customer will determine which data mining method is best for personality prediction. The necessity for vast volumes of data and the possibility of overfitting are two issues that still exist when utilizing data mining algorithms to predict personality.

5.2 Impact on Environment

To determine how the environment affects the prediction of personality traits, a comparative examination of data mining algorithms for personality traits was done. False positives or false negatives may result from an algorithm that is not accurate, which may have a detrimental impact on people's life. The study discovered that there are six primary data mining techniques, including support vector machines, decision trees, artificial neural networks, Random Forests, Naive Bayes approaches, and K-nearest Neighbours, that are utilized for this purpose. Each algorithm has features and drawbacks of its own. The Random Forest algorithm was determined to be the most accurate, while all six were capable of correctly predicting

personality traits. The Random Forest method, however, had the biggest effect on the environment, according to the study. This is because, compared to the other five methods, the Random Forest approach requires more training data.

5.3 Ethical Aspects

Data mining algorithms can be used to predict personality traits, however there are difficult ethical issues to take into account. The advantages of this technology could be enormous, but there are also a lot of possible risks and problems that need to be considered. Personality prediction using data mining algorithms has the potential to be a significant tool for both individuals and businesses. Numerous uses for being able to precisely forecast personality features include helping people discover their ideal job or romantic partner as well as spotting potential security hazards. A lot of possible hazards are connected to this technology, though. Personal data has the potential to be exploited to manipulate or oppress people. Employers might choose only those workers who suit their intended profile if they could reliably forecast personality attributes using data mining algorithms, potentially rejecting qualified individuals who don't fit the requisite personality type. Similar to this, insurance providers may utilize predictive analytics to identify clients who are more likely to file claims, and then either raise their premiums or deny them coverage completely. Additionally, there is a chance that incorrect forecasts will breed prejudice and discrimination. Inaccuracies in data mining algorithms could result in false positives, such as erroneously predicting that a person will likely commit a crime when they are actually innocent. As a result, people might be unfairly singled out for attention by the police or other authorities. Another important matter to take into account while using data is privacy issues.

5.4 Sustainability Plan

The goal of this study is to build a sustainability strategy for the best accurate algorithm and to compare and contrast the data mining algorithms for personality trait prediction. Big data has become more crucial in recent years across several industries, including business, industry,

and academia. The need for efficient ways to extract relevant information from data is increasing exponentially along with the volume of data. One such technique that has shown to be especially helpful in revealing hidden patterns and relationships in huge datasets is data mining. In this article, we first review the body of knowledge about data mining algorithms for predicting personality traits. We then evaluate the accuracy, effectiveness, and interpretability of the various algorithms. We then create a sustainability strategy for the most precise algorithm. We anticipate that this study will offer insightful information on the benefits and drawbacks of each algorithm as well as recommendations for how to use them most effectively for predicting personality traits.

CHAPTER 6
SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this study, we evaluated the predictive power of four data mining algorithms for personality traits. Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, K-nearest Neighbour, and Artificial Neural Networks were the four algorithms. We made use of a dataset of 1,023 users from the kaggle website, which also contained data on the Big Five personality traits of the individuals. All six of the algorithms we tested were successful in correctly predicting personality traits, with Random Forest slightly outperforming the others. But when it came to predicting extraversion, agreeableness, and conscientiousness, all six algorithms fared similarly. Naive Bayes outperformed the other five algorithms when it came to predicting neuroticism and openness to experience. In general, we discovered that personality characteristic prediction can be done well using all six techniques. Our findings demonstrated that Random Forest fared better on datasets in terms of accuracy and interpretability than the other five methods.

6.2 Conclusions

The purpose of this work was to analyse various data mining algorithms in terms of their ability to predict personality traits. Using data from the Kaggle Personality data collection, the effectiveness of the selected categorization algorithms is assessed. Using a number of personality models, candidate replies may help in predicting a person's traits. In the past, surveys were employed, but this process was expensive and time-consuming. The purpose of this research is to make personality predictions based on interviewee responses. The research provides examples of the various techniques and templates that were employed. In this work, we compared six different data mining algorithms for their ability to predict personality traits

Decision Trees, Random Forests, Support Vector Machines, Naive Bayes, K nearest Neighbours, Artificial Neural Networks. Although each algorithm has a varied accuracy rate for the different personality traits in the data collection, experimental study shows that the random forest algorithm was found to be the most accurate in predicting personality traits. The Support Vector Machines, Artificial Neural Networks, Naive Bayes algorithms have poor predictive accuracy for personality traits. Decision tree algorithms were the worst at predicting personality traits. Therefore, the Random forest algorithm can be used as an effective tool for predicting personality traits. When assessing the performance of classification algorithms, time can be introduced as another parameter.

6.3 Implication for Further Study

There are many different algorithms that can be used for data mining, and each has its own strengths and weaknesses. The data was subjected to six different data mining methods, and the outcomes were compared. In terms of accuracy, all six algorithms performed admirably. Some algorithms, nevertheless, performed better than others. Six alternative data mining techniques, including Decision Trees, Random Forests, Support Vector Machines, Naive Bayes Algorithm, K-Nearest Neighbours and Artificial Neural Networks, were compared in the study to identify personality traits. Further research into techniques for extrapolating personality traits from internet data may be affected by this conclusion. Selecting the method that is most appropriate for the specific dataset and task at hand is crucial because different algorithms will behave differently on various datasets.

REFERENCES

[1] Desai, Helly N., and Rakesh Patel. "A study of data mining methods for prediction of personality traits." 2020 International Conference on Smart Electronics and Communication (ICOSEC). IEEE, 2020.

[2] William, P., and Abhishek Badholia. "Evaluating Efficacy of Classification Algorithms on Personality Prediction Dataset." Ilkogretim Online 19.4 (2020): 3362-3375.

[3] Katiyar, Sandhya, Himdweep Walia, and Sanjay Kumar. "Personality Classification System using Data Mining." 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2020.

[4] Singh, Bhawna, and Swasti Singhal. "Automated Personality Classification Using Data Mining Techniques." Proceedings of the International Conference on Innovative Computing & Communications (ICICC). 2020.

[5] Zhong, Xiao-Feng, et al. "A general personality prediction framework based on facebook profiles." Proceedings of the 2018 10th International Conference on Machine Learning and Computing. 2018.

[6] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." Ieee Access 5 (2017): 15991-16005.

[7] Gjurković, Matej, and Jan Šnajder. "Reddit: A gold mine for personality prediction." Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media. 2018.

[8] Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." CHI'11 extended abstracts on human factors in computing systems. 2011. 253-262.

[9] Mehta, Yash, et al. "Recent trends in deep learning based personality detection." Artificial Intelligence Review 53 (2020): 2313-2339.

[10] Sun, Xiangguo, et al. "Who am I? Personality detection based on deep learning for texts." 2018 IEEE international conference on communications (ICC). IEEE, 2018.

[11] Y. Artissa, I. Asror, S.A. Faraby.,“ Personality Classification based on Facebook status text using Multinomial Naïve Bayes method,”In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012003.IOP Publishing, 2019.

[12] Başaran, Seren, and Obinna H. Ejimogu. "A neural network approach for predicting personality from Facebook data." Sage Open 11.3 (2021): 21582440211032156.

[13] Huang, Jin, Aiqin Zhu, and Qi Luo. "Personality mining method in web based education system using data mining." 2007 IEEE International Conference on Grey Systems and Intelligent Services. IEEE, 2007.

[14] Bharadwaj, Srilakshmi, et al. "Persona traits identification based on Myers-Briggs Type Indicator (MBTI)-a text classification approach." 2018 international conference on advances in computing, communications and informatics (ICACCI). IEEE, 2018.

[15] Anglim, Jeromy, et al. "Predicting employee attitudes to workplace diversity from personality, values, and cognitive ability." Journal of Research in Personality 83 (2019): 103865.

[16] Huang, Han-Chung, et al. "Impact of online gamers' personality traits on interdependence, network convergence, and continuance intention: Perspective of social exchange theory." International Journal of Information Management 38.1 (2018): 232-242.

[17] Mehta, Yash, et al. "Recent trends in deep learning based personality detection." Artificial Intelligence Review 53.4 (2020): 2313-2339.

[18] More, Ajinkya. "Survey of resampling techniques for improving classification performance in unbalanced datasets." arXiv preprint arXiv:1608.06048 (2016).

[19] Christian, Hans, et al. "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging." Journal of Big Data 8.1 (2021): 1-20.

[20] Pletzer, Jan Luca, et al. "A meta-analysis of the relations between personality and workplace deviance: Big Five versus HEXACO." Journal of Vocational Behavior 112 (2019): 369-383.

[21] Subramanian, Ramanathan, et al. "ASCERTAIN: Emotion and personality recognition using commercial sensors." IEEE Transactions on Affective Computing 9.2 (2016): 147-160.

[22] Ashton, Michael C., and Kibeom Lee. "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure." Personality and social psychology review 11.2 (2007): 150-166.

[23] Utami, Ninda Anggoro, Warih Maharani, and Imelda Atastina. "Personality classification of facebook users according to big five personality using SVM (support vector machine) method." Procedia Computer Science 179 (2021): 177-184.

[24] Maharani, Warih, and Veronikha Effendy. "Big five personality prediction based in Indonesian tweets using machine learning methods." International Journal of Electrical & Computer Engineering (2088-8708) 12.2 (2022).

[25] Tadesse, Michael M., et al. "Personality predictions based on user behavior on the facebook social media platform." IEEE Access 6 (2018): 61959-61969.

PLAGIARISM REPORT

ORIGINALITY REPORT

21%	17%	11%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	www.ilkogretim-online.org Internet Source	1%
4	Sandhya Katiyar, Himdweep Walia, Sanjay Kumar. "Personality Classification System using Data Mining", 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020 Publication	1%
5	fenix.tecnico.ulisboa.pt Internet Source	1%
6	www.researchgate.net Internet Source	1%
7	link.springer.com Internet Source	1%