JOB SELECTION USING MACHINE LEARNING ALGORITHMS

ANJAN RHUDRA PAUL

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

JOB SELECTION USING MACHINE LEARNING ALGORITHMS

Anjan Rhudra Paul

This Thesis Report submitted in partial fulfillment of the requirements
for the award of the degree of
Master of Science in Information and Communication Engineering

Department of Information and Communication Engineering
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

August 2022

SUPERVISOR'S DECLARATION

We hereby declare that we have checked this thesis and in our opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science in Department of Information and Communication Engineering of Noakhali Science and Technology University.

_____

Name of Supervisor: DR. MD. ASHIKUR RAHMAN KHAN
Position: PROFESSOR & CHAIRMAN
Date:

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged. This thesis report has not been accepted for any degree and is not concurrently submitted for the award of other degrees.

_____

Name: Anjan Rhudra Paul
ID Number: ASH1811MSc130M
Date:

ACKNOWLEDGEMENTS

I am grateful and would like to express my sincere gratitude to my supervisor Professor Dr. Md. Ashikur Rahman Khan for his germinal ideas, invaluable guidance, continuous encouragement, and constant support in making this research possible. He has always impressed me with his outstanding professional conduct, his strong conviction for science, and his belief that an MSc program is only the start of a life-long learning experience. I appreciate his consistent support from the first day I applied to the graduate program to these concluding moments. I am truly grateful for his progressive vision about my training in science, his tolerance of my naive mistakes, and his commitment to my future carrier. I also sincerely thank you for the time spent proofreading and correcting my many mistakes.

My sincere thanks go to all my lab mates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways and made my stay at NSTU pleasant and unforgettable.

I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream and sacrifice throughout my life. Special thanks should be given to my committee members. I would like to acknowledge their comments and suggestions which were crucial for the successful completion of this study.

ABSTRACT

Suitable job selection is a great challenging task, especially for freshers. As every year millions of students complete their graduation, they come with many options for choosing their jobs. This job selection procedure depends on various factors. In this research, we try to build a model to predict whether a job is suitable for a candidate or not according to their skills, experiences, and their desires for the job. First of all, we collect data from 120 individual people who are currently appointed in a job in various fields. Then we train our model with Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, K-Nearest-Neighbor, Random Forest Tree, decision tree, and multilayer perceptron neural network algorithm to predict whether they are satisfied with their current job or not. Which acquire an 80% accuracy rate for Logistic Regression, Gaussian Naive Bayes gives a 79% accuracy, K-Nearest-Neighbor gives of 87.5% accuracy, Support Vector Machine gives 79% accuracy, Random Forest Tree gives 92% accuracy, Multilayer Perceptron Algorithm gives 91.6% accuracy, Decision 83% accuracy. We also find the other performance matrices and compare them with other algorithms to evaluate the performance of best model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The introductory study of this thesis is carried out in this chapter. Since this research focuses on the prediction model for job selection in Bangladesh; thus, background analysis on these sectors is carried out first. The problem statement is then introduced and the research objectives are described on this basis. At the end of this chapter, the outline of this thesis is presented. In this thesis paper, we study several models based on machine learning and apply a dataset into training and testing phases. Then we compare the algorithm analysis result and found a suitable model for this problem.

1.2 BACKGROUND

Job selection is a crucial part of a person's life. Sometimes people may have proper skill sets but they cannot decide whether they should go for the offered job or not. It's amazing to have alternatives; it's such a tremendous confidence booster! However, deciding which choice is the best puts a lot of pressure on a candidate.

To make the best decision, one must first determine which elements are most essential to one in new employment, and then select the alternative that best addresses these criteria. This, however, works on two levels: a cognitive level and an emotional, gut-level level. Only if these are in sync will one be completely satisfied with their selection.

Choosing a suitable job path is crucial to a successful and secure future. Certain personal elements, as well as the individual's academic history, might impact this selection process, such as innate abilities, technical skills, and academic performance [1]. Some graduates are alert about their career goal. But some individuals find it difficult to decide their career path and their interest according to their skills.

Machine learning is employed in a variety of sectors and businesses in today's digital world, including clinical analysis, image processing, classification, regression, and more. Machine learning is a branch of computer science that has exploded in popularity in recent years and is

likely to continue to do so in the future. Our world is awash with data, and new data is being generated at a breakneck pace all around the globe. Machine learning algorithms can also be used to build a prediction model that can automate the process of job selection for fresh graduates. Machine learning allows humans to acquire insight from huge amounts of data that would otherwise be too difficult or impossible to process.

There are mainly three types of machine learning algorithms. They are:

- Supervised machine learning
- Unsupervised machine learning
- Reinforcement machine learning

The procedure creates a mathematical model from a set of data that includes both the inputs and the expected outputs in supervised learning. Supervised learning methods include classification and regression techniques. Unsupervised learning creates a mathematical model from a collection of data that comprises just inputs and no desired output labels. Unsupervised learning techniques are used to discover structure in data, such as data point grouping or clustering.

Machine learning algorithms comprise different stages. The basic stages of machine learning algorithms are training of dataset, applying classification algorithms and then the final stage is to classify them into specific classes.

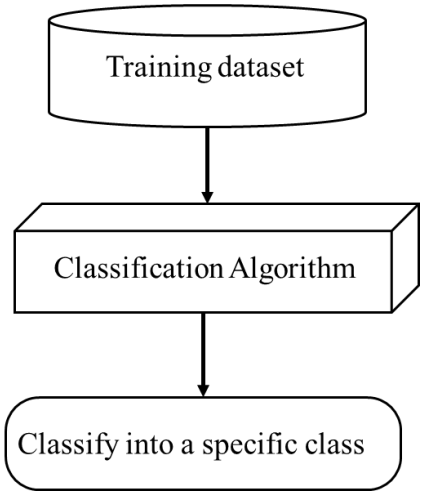


Figure 1.1: Machine Learning Model.

1.3 PROBLEM STATEMENT

In today's world, incredible changes are occurring in every sector of existence, from art to technology, and everything in between. More sectors are emerging, more technologies are being found, and current technologies are being improved. Growth in technology leads to an increase in the need for professionals who can successfully deal with those technologies in their employment. And, in order for these professionals to be successful in their industries, they must make deliberate and intelligent career choices. Such as Information Technology (IT) jobs continue to grow in several disciplines, such as cloud computing, cyber security, mobile applications, and big data analytics [2].

So, it becomes very difficult for a fresh graduate to decide which job offer suit their desire according to their skill sets. Traditionally, people in specialized professions used surveys to determine the most essential criteria influencing their career pathways. However, given of the complexities of each student's goals and desires, predicting a professional path is extremely difficult.

The job selection process depends on several parameters like salary, house facilities, increment, transport facility, medical facility, gratuity etc. One can only select his desired job when he gets positive results on these several parameters.

In this research, we compare several machine learning models to predict whether an offered job will be suitable for the job-seeking candidate or not according to their desire and technical, and professional skills. This models automates the prediction process and make it easy for fresh graduate and also for one who is trying to switch current job.

1.4 RESEARCH OBJECTIVES

This research paper is worked on machine learning algorithms especially decision tree algorithm and provides a better result that will directly help the job-seeking candidate for selecting a suitable job. The main purposes of the research are the following:

i. To analyze the criterion of the job selection process.
ii. To analyze various research works done on career guidelines.
iii. To predict whether one is satisfied with his/her job in the context of Bangladesh.
iv. To make a comparison among distinct methods in the prediction of job selection.
v. To find out the optimal technique for job selection prediction.

1.5 SCOPE OF THE STUDY

The data used in this study to predict is sourced mainly from the individual people and social media such as Facebook, and LinkedIn. The data is collected from people who are currently doing their job in various companies in various locations in Bangladesh. The data are collected from different job sectors for different positions.

1.6 THESIS OUTLINE

This thesis consists of 5 chapters and these chapters are organized as follows, in chapter 1 it provides a detailed discussion on the importance of the work that has been done and why the current topic is selected for the thesis. The next that is chapter 2 is about some related works of this thesis is described in this section. It gives information about the challenges of job selection of different sectors using different classification algorithms and techniques. Next is chapter 3 which introduces the dataset and the methods and explains a deep understanding of the algorithms and evaluating criteria. Then chapter 4 describes experimental plans and prediction test results on evaluation criteria and shows a comparison of different algorithms' performance. Finally, in Chapter 5, paper will end up as allowing us to know why this study is valuable for effective techniques in job selection in the conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

In this chapter, different past works related to job selection criteria, prediction methods, and prediction accuracy are described in short. These three things which are related to job selection prediction are discussed in the literature review. We present a table of the summary of literature at the end of this chapter.

2.2 LITERATURE REVIEW

The literature review is performed on the works which relates on these sectors. The novelty of this thesis compared to the related previous works can be easily understood.

Bolander P. et al try to present an ethnomethodological discourse-analytical real-time study of how selection decisions are made in situ [3]. The main findings suggest that selection decision-making is characterized by ongoing practical deliberation involving four interconnected discursive processes: assembling candidate versions, establishing candidate versions as factual, reaching selection decisions, and using selection tools as sensemaking devices. Furthermore, this research distinguishes between two main types of selection decision-making: one characterized by initial agreement and the other by first disagreement. Selectors reason through the four discursive processes in a systematic, situated, and practical manner in order to develop local versions of the candidates and make reasonable selection judgments in each fundamental form of decision making.

Magesh N et al. proposed a system for evaluating the performance of an employee using decision tree algorithm [4]. In this paper the employee data is analyzed to determine whether or not they should be promoted, given an annual raise, or advance in their careers. They examined utilizing prior historical data of employees in order to generate an annual increase

for an employee. The decision tree method is used to learn the historical data in the table, and the performance is determined by comparing an employee's qualities to the rules created by the decision tree classifier. This study focuses on gathering data on workers, creating a decision tree using historical data, testing the decision tree with employee traits, and determining whether or not to award the promotion. They use a user interface to collect information about an employee. These data are compared to the data that has been learned and stored in the decision tree. The ultimate objective node determines whether or not the employee will receive an annual raise or promotion.

Another model is proposed by Luís Matos Pombo et al. where they test a collection of Machine Learning algorithms and through the usage of cross-validation, we calibrate the most important hyper-parameters of each algorithm [5]. Using previous data from selected/reject applications in the screening phase, the learning algorithms attempt to understand what constitutes a successful match between candidate profile and job criteria. The similarities between the job criteria and the candidate profile in dimensions such as skills, profession, and geography, as well as a set of job attributes that aim to capture the experience level and compensation expectations, are among the features we utilize to develop their models. Their best results in the first round of trials come from using the Multilayer Perceptron technique (also known as Feed-Forward Neural Networks). Following that, they improve the skills-matching feature by employing techniques for semantically embedding required/offered skills to address issues such as synonyms and typos, which artificially reduce the similarity between the job profile and the candidate profile, lowering the overall quality of the results. They achieve their best results by using the word2vec approach for embedding skills and the Multilayer Perceptron to learn the overall matching. They think that by expanding the concept of semantic embedding to other aspects and by identifying candidates with similar job preferences to the target applicant and developing a richer presentation of the candidate profile top of that, their findings might be even better.

John F. Magee et al. present one recently developed a concept called the “decision tree,” which has tremendous potential as a decision-making tool [6]. He forecast that in the next years, there will be a lot of talk about decision trees. Although they are unfamiliar to most business people today, they will undoubtedly become commonplace in management jargon before long.
Patel H, Sanghavi J, et al. developed a career guidance system for science students studying in the 11th and 12th standards [7]. This project was divided into two parts: the Web Portal, which

serves as a user interface between the student and the system and was built using HTML, CSS, and Bootstrap 4 on the front end, and MySQL and PHP on the back end; and the Recommendation Engine, which uses Machine Learning and Python to recommend suitable career options to the student using the system. Naive Bayes, K-Nearest Neighbour, and Random Forest Classifier were the three Machine Learning Algorithms that were employed.
In another method was proposed by Kiselev P, Kiselev B that explains the social constructivism foundations of machine learning approaches in career advising, as well as the relevance of social networks in psychology research [8]. AUC-ROC measure computation in career advice modeling experimentally confirms the theoretical reasons. There are additional implications for career guidance practice.

Vidya Shreeram N, et al proposed a strategy that addresses whether or not students will continue their studies at a higher level [9]. This was assessed using machine learning ideas, which were a subset of artificial intelligence. Machine learning is based on mathematical and scientific principles. This study uses machine learning methods such as Decision Tree, Random Forest, Support Vector Machine, and Adaboost to predict students' career paths. When compared to other machine learning classifiers, the RF classifier has a 93% accuracy. The Python programming language is used to create machine learning classifiers.
Al-Dossari H, Al-Qahtani Z et al. proposed CareerRec, a recommendation system based on machine learning algorithms, is designed to assist IT graduates in choosing a career path based on their talents [10]. A dataset of 2255 employees in Saudi Arabia's IT industry was used to train and evaluate CareerRec. They compared the accuracy of five machine learning methods in identifying the best-suited job route among three classes. Their tests show that the XGBoost algorithm beats other models and provides the best level of accuracy (70.47%).

A career prediction model was proposed et al. the focus of this article was on computer science domain applicants' career area projection [11]. They used algorithms like Support Vector Machines, OneHot encoding, Decision tree, and XG boost. The first goal was to forecast the output using all of the techniques listed above, then analyze the findings before continuing with the most accurate approach.Finally, this study discusses a variety of sophisticated machine learning methods that involve classification and prediction and are used to increase accuracy, predictability, and analyze the performance of these algorithms. The data was trained and evaluated with all three algorithms, with SVM achieving 90.3% accuracy and XG Boost achieving 88.33% accuracy.

Lou Y, Ren R, Zhao Y et al. proposed a machine learning approach for future career planning [12]. The goal of this research is to simulate people's career pathways. They start by gathering a large number of profiles and extracting characteristics from the descriptive data. To assist avoid the detrimental effects of natural language, hand rules, and a clustering method were used. They use the Markov Chain to represent people's career progression and provide our method for estimating the transition probability matrix. Finally, they address the challenge of determining the greatest career growth advice for a person based on his or her present work path and goals. Finally, they go through the findings and talk about how they might improve their model.

Another research conduct by Nigam A. et al. They introduce a new machine learning model that incorporates the dynamics of a highly unpredictable employment market by using candidates' job preferences over time [13]. In addition, their strategy includes a number of minor suggestions that help with the challenges of a) creating serendipitous recommendations. b) addressing the issue of new jobs and applicant cold-starts. They employed skills as embedded characteristics to extract latent competencies from them, allowing them to increase the skills of jobs and candidates to attain greater skill coverage. Their model was created and tested in a real-world job recommender system, and the greatest click-through rate metric performance was attained by combining machine learning and non machine learning suggestions. Bidirectional Long Short Term Memory Networks (Bi-LSTM) with Attention produced the greatest results for recommending tasks using machine learning, which was a big aspect of their recommendation.

Sah U, Singh M another system et al. This study uses machine learning methods such as Decision Tree, Random Forest, Support Vector Machine, and Adaboost to forecast a student's career [14]. The Python programming language is used to construct machine learning algorithms. They consider the most accurate results provided by the algorithm for their further processing after training and evaluating the data using these. So, the first goal is to forecast the output using all of the strategies listed above, then analyze the findings before moving on to the most accurate approach. As a result, this article discusses a variety of sophisticated machine learning algorithms that involve classification and prediction and are used to increase the accuracy for improved prediction, dependability, and examining the performance of these algorithms. After the data was trained and evaluated with all three algorithms, with Support

Vector Machine achieving 90.3 percent accuracy and XG Boost achieving 88.33 percent accuracy.

Awujoola J Olalekan et al. provide a current, accurate, and deserving machine learning classification model that can be deployed, implemented, and used to make predictions and evaluations on job candidate qualities based on their academic performance records in order to match industry selection standards [15]. This research looked at both supervised and unsupervised machine learning classifiers. Support Vector Machine, Nave Bayes, Logistic Regression. Although Random Forest and Decision Tree fared well, Logistic Regression surpassed the others with a 93% accuracy rate. The summary of past works is described in short in the table 2.1 which is shown in below:

Table 2.1: Summary of literature

Author (Year)	Implemented Methods	Considered area	Remark
P. Bolander and J. Sandberg (2013)	Assembling candidate versions, establishing candidate versions as factual, reaching selection decisions, and using selection tools as sense making devices	IT Consultant & IT Bank. IT Consultant was the Swedish subsidiary of a global technology consulting group, while IT bank supplies all it solution to this group.	
N. Magesh et al. (2013)	Decision Tree Algorithm	Private data of a company's employee	ROC-AUC value 0.838
Luís Matos Pombo et al. (2019)	Feed-Forward Neural Networks	An international IT-specialized online job marketplace	ROC-AUC Sample Mean 0.814
John F. Magee et al. (1964)	Decision Tree Algorithm	Stygian Chemical Industries, Ltd	Provide Methodology
Patel H, Sanghavi J, et al. (2021)	Naive Bayes, K-Nearest Neighbour, and Random Forest Classifier	Web portal	Accuracy for Naive Bayes 57%, K-NN 64%, Random Forest 81%.
Kiselev P, Kiselev B (2020)	Machine Learning approaches , AUC-ROC	www.digitalhuman.ru	The AUC-ROC range from 0.68

			(media sphere) to 0.85.
Vidya Shreeram N, et al (2021)	Decision Tree, Random Forest, Support Vector Machine, and Adaboost	Undergraduate level students	Accuracy of Adaboost 87%, SVM 88.5%, RF 91% and DT 93%
Al-Dossari H, Al-Qahtani Z et al. (2020)	XGBoost algorithm	Saudi Arabia	XGBoost has highest accuracy (70.47%).
K. S. Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, and J. Priyanka (2018)	Support Vector Machine, OneHot encoding, Decision tree, and XG boost	Different organizations, LinkedIn api, college alumni database	Support Vector Machine gave more accuracy with 90.3%
Y. Lou, R. Ren, and Y. Zhao (2010)	K-means clustering algorithm	LinkedIn	Has generated plausible and logical career path recommendations
A. Nigam, A. Roy, H. Singh, and H. Waila (2019)	Bidirectional Long Short Term Memory Networks (Bi-LSTM)	Author Company's database	80% from BLSTM-A,
U. K. Sah, M. A. Singh, and M. Tech Scholar (2022)	Decision Tree, Random Forest, Support Vector Machine, and Adaboost	Collected through questionaries	XG Boost with 88.33 percent accuracy
O. Awujoola, Philip O Odion, Martins E Irhebhude, and Halima Aminu (2021)	Support Vector Machine, Nave Bayes, Logistic Regression	Survey	Logistic Regression outperformed others with 93% accuracy

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter presents an outline of the research methods that were followed in the study. It provides information on the workflow data collection and preprocessing, Questionnaire, summary of the dataset, machine learning modeling (Logistic Regression, Gaussian naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest Tree, Multilayer Perception Algorithm, Decision Tree). We train the model and do performance analysis matrix. The requirements of this work and the procedures that were followed to carry out this study are included.

3.2 DATA COLLECTION AND DATA PREPROCESSING

To reduce the computational cost of the several models, we collect and preprocess the data. Here we describe parameter selection process, represent data value and attribute short names to speed up the computational process.

3.2.1 Parameter Selection

For the proper analysis, it is necessary to evaluate some of the main attributes of the job selection dataset. Accurate prediction of job selection is a challenging task for fresh graduates. Different classification techniques are used to classify the data and predict the appropriate job by using the datasets of job selection. Attributes of a proper job are the type of current job, the gross salary range of the job, the company rating, Job environment rating, behavior rating of the management of that company, How many bonuses are provided by the company per year, yearly bonus range, Interval of salary review, Leave Facilities Provided by the company, Transport/car facilities, Lunch Facilities, Medical Facilities, Educational Facilities for children, Provident fund facilities, Gratuity facilities, Higher study leave facilities, House/House rent facilities.

Table 3.1: Questionnaire & Collected Data from the Dataset

Sl. No.	Queries	Data									
i	Gender	Male	Male	Male	Male	Male	Male	Male	Male	…	Male
ii	Age	27	32	32	26	28	25	24	24		31
iii	Living Location	Comilla, Bangladesh	Dhaka	Dhaka	Dhaka	Dhaka	Dhaka	Dhaka	Uttara, Dhaka	…	Dhaka
iv	Graduation degree	IT/ICT/ICE	ME (Mechanical Engineering)/IPE/CHE/	Physical Science (Physics/Mathematics etc.)	IT/ICT/ICE	CSE/EEE/ETE/ECE	IT/ICT/ICE	IT/ICT/ICE	IT/ICT/ICE	…	Commerce related (BBA etc.)
v	Name of the educational institute	NSTU	SU	DU	NSTU	CoU	NSTU	NSTU	NSTU	…	NU
vi	Result of the graduation degree (C.G.P.A. out of 4)	3.51-3.75	2.26-2.5	3.26-3.5	3.01-3.25	3.51-3.75	3.01-3.25	3.01-3.25	3.01-3.25	…	3.26-3.5
vii	Highest educational degree	-	Bsc	MS	Bsc	BSc	BSc	Bsc	B. Sc.	…	BSc
viii	Name of the last educational institute	-	SU	DU	NSTU	CoU	NSTU	NSTU	NSTU	…	-
ix	Result of the last education degree	-	2.26-2.5	3.01-3.25	3.01-3.25	3.51-3.75	3.01-3.25	3.01-3.25	3.01-3.25	…	-
x	Problem-solving skill	Moderate	Professional	Proficient	Do not have programming skill	Moderate	Moderate	Proficient	Beginner	…	Do not have programming skill
xi	No. of known programming languages	5	-	2	0	3	2	3	1	…	-

xii	Name of known programming languages	C,C++,Java ,PHP,Html	-	R, SAS	Null and void	C/C++/C#	cpp, java	C++,jav a,python	Javascr ipt	…	-
xiii	Known Frameworks	-	-	-	0	Dot.Net	none	None	React	…	-
xiv	Gained software skills	10	10	4	0	0	2	0	10	…	2
xv	Professional certificates	2	-	3	0	0	-	0	12	…	-
xvi	Projects done before your first job in the related field	-	-	10	0	3	0	0	1	…	-
xvii	Internship Timeframe (days/months)	3	-	-	0	0	0	15	3	…	3 months
.xvi ii	Preferred job type	Related with graduation degree	Clerica l/Offici al	Any type	Teachi ng	Technical	Related with graduati on degree	Related with graduati on degree	Techni cal	…	Related with graduation degree
xix	Switched Jobs	0	1	0	0	0	0	0	1	…	<1
xx	Jobs Designation/posi tion	-	AE	-	-	-	-	-	Junior Design er	…	Sales Executive
xxi	Name of the company		MICFL	-	-	-	-	-	RPSI Limited : Tech-Teem	…	Meghna Group
xxii	Obligation to serve the company	-	Obligat ion of more than 5 years	-	-	-	-	-	No obligati on	…	No obligation
xxii i	Location of the company	-	Divisio nal city (Chatto gram, Khulna etc.)	-	-	-	-	-	Capital city (Dhaka)	…	Divisional city (Chattogram, Khulna etc.)

xxiv	Job category	-	No governmental	-	-	-	-	-	No governmental	…	No governmental
xxv	Type of first job	-	Remote	-	-	-	-	-	Office / onsite	…	Hybrid
xxvi	First job gross salary range	-	25001-35000	-	-	-	-	-	25001-35000	…	35001-50000
xxvii	Company rating of your first job (out of five)	-	4	-	-	-	-	-	4	…	4
xxviii	Environment rating of your first company (out of five)	-	4	-	-	-	-	-	4	…	4
xxix	Behavior rating of the management (out of five)	-	4	-	-	-	-	-	3	…	4
xxx	Bonuses are provided by the company per year in your first job	-	2 festival bonus+ 1 profit bonus	-	-	-	-	-	Two festival bonus	…	2 festival bonus+1 profit bonus
xxxi	Yearly bonus range of first job	-	25001-35000	-	-	-	-	-	50001-75000	…	15000-25000
xxxii	Interval of salary review (mention year/months)	-	6 months	-	-	-	-	-	1 year	…	6 months
xxxiii	Leave Facilities Provided by the company per year (mark the suitable options)	-	30-40 days	-	-	-	-	-	20-30 days	…	10-15 days
xxxiv	Transport/car facilities Provided by the company (mark the suitable options)	-	Provided personal car, fuel and driver	-	-	-	-	-	No 14ransport facility	…	No transport facility

xxxv	Lunch Facilities Provided by the company (mark the suitable options)	-	Yes	-	-	-	-		No	…	No
xxxvi	Medical Facilities Provided by the company (mark the suitable options)	-	Medical support for employee	-	-	-	-		No Medical facility	…	No Medical facility
xxxvii	Educational Facilities for children Provided by the company (mark the suitable options)	-	Yes	-	-	-	-	-	No	…	No
xxxviii	Provident fund facilities Provided by the company (mark the suitable options)	-	Yes	-	-	-	-	-	No	…	Yes
xxxix	Gratuity facilities Provided by the company (mark the suitable options)	-	Yes	-	-	-	-	-	No	…	Yes
xl	Higher study facilities Provided by the company (mark the suitable options)	-	Higher study leave facility with salary	-	-	-	-	-	No Higher study leave facility	…	No Higher study leave facility
xli	House/House rent facilities Provided by the company (mark the suitable options)	-	Yes	-	-	-	-		Yes	…	Yes

xlii	First job satisfaction	-	-	-	-	-	-	-	-	…	No
xliii	Reason to leave the first job (mark the suitable options)	-	Get offer from higher rank company	-	-	-	-	-	New job provides more facilities than the present one.	…	Get offer from higher rank company
xliv	Designation/position in current job	Graphics Designer	AE	manager	Officer	Senior software engineer	Software Engineer	Software Engineer	Digital Marketing Strategist	…	Sales Officer
xlv	Name of the current company	Canva	MICFL	Walton Hi-tech Industries PLC	Sonali Bank Ltd	Enosis solutions	Samsung R&D institute Bangladesh	Samsung R&D Institute Bangladesh	Asiatic Mindshar Ltd.	…	RFL
xlvi	Job obligation to serve the company for a certain duration and otherwise you cannot switch to another job (in current job)	No obligation	No obligation	No obligation	No obligation	No obligation	No obligation	No obligation	No obligation	…	Obligation of 1-2 years
xlvii	Location of current company	Overseas	Capital city (Dhaka)	Capital city (Dhaka)	Capital city (Dhaka)	Capital city (Dhaka)	Capital city (Dhaka)	Capital city (Dhaka)	Capital city (Dhaka)	…	Divisional city (Chattogram, Khulna etc.)
xlviii	Current job category	No governmental	No governmental	No governmental	Governmental	No governmental	No governmental	No governmental	No governmental	…	No governmental
xlix	Type of current job	Remote	Remote	Office / onsite	Office / onsite	Hybrid	Office / onsite	Office / onsite	Hybrid	…	Office / onsite

l	Gross salary range of your current job	50001-75000	25001-35000	75001-100000	25001-35000	75001-100000	35001-50000	50001-75000	25001-35000	…	50001-75000
li	Company rating of your current job (out of five)	4	4	4.5	3	4.5	4	4	5	…	5
lii	Environment rating of your current company (out of five)	4	4	4.5	1	4	4	4	5	…	5
liv	Behavior rating of the management of current company (out of five)	4	3	5	3	5	4	5	4	…	5
lv	Bonuses are provided by the company per year in your current job	others	2 festival bonus+ 1 profit bonus	others	others	Two festival bonus	others	2 festival bonus+1 profit bonus	2 festival bonus+ 1 profit bonus	…	2 festival bonus+1 profit bonus
lvi	Yearly bonus range of your current job	-	25001-35000	>100000	75001-100000	35001-50000	-	25001-35000	50001-75000	…	15000-25000
lvii	Interval of salary review of current job (mention year/months)	6	Jun	Year	1	1 year	-.	Year	1 year	…	6 Months
lviii	Leave Facilities Provided by the current company per year (mark the suitable options):	10-15 days	30-40 days	30-40 days	30-40 days	20-30 days	10-15 days	30-40 days	20-30 days	…	20-30 days
lviii	Transport/car facilities Provided by the current company (mark the suitable options):	No transport facility	Provided personal car, fuel and driver	Provided personal car, fuel and driver	No 17ransport facility	No transport facility	No transport facility	Provided a certain amount	No 17ransport facility	…	Provided a certain amount

lix	Lunch Facilities Provided by the current company (mark the suitable options):	No	Yes	Yes	No	Yes	Yes	Yes	Yes	…	Yes
lx	Medical Facilities Provided by the current company (mark the suitable options):	No Medical facility	Medical support for employee	Medical support for employee	No Medical facility	No Medical facility	Medical support for family members of employee	Medical support for employee	Medical support for employee	…	Medical support for employee
lxi	Educational Facilities for children Provided by the current company (mark the suitable options):	No	Yes	No	No	No	No	No	No	…	No
lxii	Provident fund facilities Provided by the current company (mark the suitable options):	No	Yes	Yes	Yes	No	Yes	Yes	Yes	…	Yes
lxiii	Gratuity facilities Provided by the current company (mark the suitable options):	Yes	Yes	No	Yes	Yes	Yes	Yes	No	…	Yes
lxiv	Higher study facilities Provided by the current company (mark the suitable options):	No Higher study leave facility	Higher study leave facility with no salary	No Higher study leave facility	Higher study leave facility with no salary	No Higher study leave facility	Higher study leave facility with no salary	Higher study leave facility with salary	No Higher study leave facility	…	No Higher study leave facility

lxv	House/House rent facilities Provided by the current company (mark the suitable options):	No	Yes	Yes	Yes	No	No	No	No	…	Yes
lxvi	Current job satisfaction	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	…	Yes
lxvii	Still searching for a new job	No	No	Yes	Yes	No	No	No	No	…	No

Data Elaboration:

NSTU = Noakhali Science & Technology University

DU = Dhaka University

SU = Sonargaon University

CoU = Comilla University

NU = National University

Table 3.1 describes Questionnaire & Collected Data from the Dataset. Here we show 9 persons answer for 67 Questions.

Table 3.2: Represented Data Value from the Dataset

Sl. No.	Queries	Data									
i	Gender	1	1	1	1	1	1	1	1	…	1
ii	Age	4	9	9	3	5	2	1	1		8
iii	Living Location	7	9	9	9	9	9	9	35	…	9
iv	Graduation degree	7	9	10	7	3	7	7	7	…	4
v	Name of the educational institute	31	42	21	40	15	32	36	31	…	29
vi	Result of the graduation degree (C.G.P.A. out of 4)	4	0	3	2	4	2	2	2	…	3
vii	Highest educational degree	20	10	12	10	9	9	10	1	…	9
viii	Name of the last educational institute	42	35	18	32	13	27	30	24	…	42
ix	Result of the last education degree	6	0	2	2	4	2	2	2	…	6
x	Problem-solving skill	2	3	4	1	2	2	4	0	…	1
xi	Known programming languages	5	8	2	0	3	2	3	1	…	8
xii	Name of the programming languages	10	57	22	20	15	49	1	19	…	57
xiii	Known Frameworks	30	30	1	15	31	22	24	14	…	30
xiv	Gained software skills	2	2	8	0	0	5	0	2	…	5
xv	Professional certificates	3	6	4	0	0	6	0	2	…	6

xvi	Projects done before your first job in the related field	10	10	2	0	5	0	0	1	…	10
xvii	Internship Timeframe (days/months)	5	11	11	0	0	0	3	5	…	8
.xviii	Preferred job type	2	1	0	3	4	2	2	4	…	2
xix	Switched Jobs	0	1	0	0	0	0	0	1	…	2
xx	Jobs Designation/position	19	0	19	19	19	19	19	4	…	12
xxi	Name of the company	23	13	23	23	23	23	23	18	…	14
xxii	Obligation to serve the company	5	3	5	5	5	5	5	0	…	0
xxiii	Location of the company	6	1	6	6	6	6	6	0	…	1
xxiv	Job category	1	0	1	1	1	1	1	0	…	0
xxv	Type of first job	4	3	4	4	4	4	4	2	…	1
xxvi	First job gross salary range	3	1	3	3	3	3	3	1	…	2
xxvii	Company rating of your first job (out of five)	5	3	5	5	5	5	5	3	…	3
xxviii	Environment rating of your first company (out of five)	5	3	5	5	5	5	5	3	…	3
xxix	Behavior rating of the management (out of five)	5	3	5	5	5	5	5	1	…	3
xxx	Bonuses are provided by the company per year in your first job	3	0	3	3	3	3	3	2	…	0
xxxi	Yearly bonus range of first job	4	1	4	4	4	4	4	3	…	0
xxxii	Interval of salary review (mention year/months)	7	6	7	7	7	7	7	1	…	2
xxxiii	Leave Facilities Provided by the company per year (mark the suitable options)	4	2	4	4	4	4	4	1	…	0
xxxiv	Transport/car facilities Provided by the company (mark the suitable options)	4	3	4	4	4	4	4	0	…	0

xxxv	Lunch Facilities Provided by the company (mark the suitable options)	2	1	2	2	2	2	2	0	…	0
xxxvi	Medical Facilities Provided by the company (mark the suitable options)	2	0	2	2	2	2	2	1	…	1
xxxvii	Educational Facilities for children Provided by the company (mark the suitable options)	2	1	2	2	2	2	2	0	…	0
xxxviii	Provident fund facilities Provided by the company (mark the suitable options)	0	1	2	2	2	2	2	0	…	1
xxxix	Gratuity facilities Provided by the company (mark the suitable options)	2	1	2	2	2	2	2	0	…	1
xl	Higher study facilities Provided by the company (mark the suitable options)	3	1	3	3	3	3	3	2	…	2
xli	House/House rent facilities Provided by the company (mark the suitable options)	2	1	2	2	2	2	2	1	…	1
xlii	First job satisfaction	2	2	2	2	2	2	2	2	…	30
xliii	Reason to leave the first job (mark the suitable options)	6	3	6	6	6	6	6	5	…	3
xliv	Designation/position in current job	22	3	54	35	44	46	46	15	…	39
xlv	Name of the current company	23	42	83	70	32	65	65	9.	…	57
xlvi	Job obligation to serve the company for a certain duration and otherwise you cannot switch to another job (in current job)	0	0	0	0	0	0	0	0	…	1
xlvii	Location of current company	3	0	0	0	0	0	0	0	…	1
xlviii	Current job category	1	1	1	0	1	1	1	1	…	1

xlix	Type of current job	3	3	2	2	1	2	2	1	…	2
l	Gross salary range of your current job	3	1	4	1	4	2	3	1	…	3
li	Company rating of your current job (out of five)	2	2	3	0	3	2	2	4	…	4
lii	Environment rating of your current company (out of five)	3	3	4	0	3	3	3	5	…	5
liv	Behavior rating of the management of current company (out of five)	2	2	3	0	3	2	2	4	…	3
lv	Bonuses are provided by the company per year in your current job	3	0	3	3	2	3	0	0	…	0
lvi	Yearly bonus range of your current job	6	1	5	4	2	6	1	3	…	0
lvii	Interval of salary review of current job (mention year/months)	7	11	14	2	4	.0	14	4	…	8
lviii	Leave Facilities Provided by the current company per year (mark the suitable options):	0	2	2	2	1	0	2	1	…	1
lviii	Transport/car facilities Provided by the current company (mark the suitable options):	1	4	4	1	1	1	2	1	…	2
lix	Lunch Facilities Provided by the current company (mark the suitable options):	0	1	1	0	1	1	1	1	…	1
lx	Medical Facilities Provided by the current company (mark the suitable options):	2	0	0	2	2	1	0	0	…	0
lxi	Educational Facilities for children Provided by the current company (mark the suitable options):	0	1	0	0	0	0	0	0	…	0
lxii	Provident fund facilities Provided by the current company (mark the suitable options):	0	1	1	1	0	1	1	1	…	1

lxiii	Gratuity facilities Provided by the current company (mark the suitable options):	1	1	0	1	1	1	1	0	…	1
lxiv	Higher study facilities Provided by the current company (mark the suitable options):	2	0	2	0	2	0	1	2	…	2
lxv	House/House rent facilities Provided by the current company (mark the suitable options):	0	1	1	1	0	0	0	0	…	1
lxvi	Current job satisfaction	1	1	1	1	1	1	1	1	…	1
lxvii	Still searching for a new job	0	0	1	1	0	0	0	0	…	0

Table 3.2 describes Represented Data Value from the Dataset. Each answer got a digit for identification. It will reduce the computational time.

Table 3.3: Attributes Short Name list for job assessment.

No.	Feature Name	Feature Name
1	your gender	Gender
2	Your Age	Age
3	Your Living Location	LLocation
4	What is the name of Graduation degree that you...	Gdegree
5	Name of the educational institute?	NeInstitute?
6	What was the result of the graduation degree (...	CGPA
7	What was the highest educational degree that y...	Hdegree
8	Name of the last educational institute:	LeInstitute:
9	What was the result of the last education degree?	LCGPA
10	How is your Problem-solving skill?	PsSkill?
11	How many programming languages do you know ?	#PL
12	what are they?	Planguage
13	Which frameworks you have skilled on?	Frameworks
14	How many software skills that you have gained?	Swskill
15	How many professional certificates that you ha...	#Pcertificates
16	How many projects you have done before your fi...	#Projects
17	How many days/months was in your internship (I..	Internship
18	What type of job do you prefer?	Jtype
19	How many jobs have been switched by you till now?	#Jswitched
20	What was your designation/position?	FJdes
21	What was the name of the company?	Fcname
22	Is there any obligation to serve the company f...	FJobligation
23	Where was the location of the company?	FClocation
24	Which following category the first job was?	FJcategory
25	What was the type of first job?	FJtype

26	What was your gross salary range of your first...	FJsalary
27	Give a company rating of your first job (out o...	FJrating
28	Give the Job environment rating of your first ...	FJErating
29	Give the behavior rating of the management (ou...	FJMrating
30	How many bonuses are provided by the company p...	FJbonuses
31	What was your yearly bonus range of your first...	FJbonusRange
32	Interval of salary review (mention year/months)?	FJsalaryInt.
33	Leave Facilities Provided by the company per y...	FJleave
34	Transport/car facilities Provided by the compa...	FJtransport
35	Lunch Facilities Provided by the company (mark...	FJlunch
36	Medical Facilities Provided by the company (ma...	FJmedical
37	Educational Facilities for children Provided b...	Fjeducation
38	Provident fund facilities Provided by the comp...	FJpFund
39	Gratuity facilities Provided by the company (m...	FJgratuity
40	Higher study facilities Provided by the compan...	FJhStudy
41	House/House rent facilities Provided by the co...	FJHouseRent
42	Were you satisfied with your first job?	Fjsatisfaction
43	Reason to leave the first job (mark the suitable options)	
44	What is your designation/position in current job?	designation
45	What is the name of the current company?	Cname
46	Is there any obligation to serve the company f...	Cobligation
47	Where is the location of current company?	Clocation
48	Which following category current job is?	CJ category
49	What is the type of current job?	CJ type
50	What is your gross salary range of your curren...	CJsalary
51	Give a company rating of your current job (out...	CCrating
52	Give the Job environment rating of your curren...	CJeRating
53	Give the behavior rating of the management of ...	CJmRating

54	How many bonuses are provided by the company p...	#CJbonuses
55	What is your yearly bonus range of your curren...	CJbonusRange
56	Interval of salary review of current job (ment...	CJsrInter.
57	Leave Facilities Provided by the current compa...	CJleaveF
58	Transport/car facilities Provided by the curre...	CJtransport
59	Lunch Facilities Provided by the current compa...	CJlunchF
60	Medical Facilities Provided by the current com...	CJmedicalF
61	Educational Facilities for children Provided b...	CJeducationF
62	Provident fund facilities Provided by the curr...	CJpFund
63	Gratuity facilities Provided by the current co...	CJgratuity
64	Higher study facilities Provided by the curren...	CJhStudy
65	House/House rent facilities Provided by the cu...	CJhouseRent
66	Are you satisfied with your current job?	CJJobSatis
67	Are you still searching for a new job?	CJSearchJob

Table 3.3 describes attributes short name list for job assessment. We have made this for simplification of the process.

3.2.2 Data Preprocessing

The data used in this study to predict sourced from different locations in Bangladesh. The data is collected from people who are currently doing their job in different companies from different sectors. Data is collected by individual persons using hard copies from, social media messages like Facebook, WhatsApp, and LinkedIn through a google form. We were able to manage data from 120 individual people through these mediums from different occupations.

It should be noticed that the range and measure units of the variables differ before entering the gathered data into the classification algorithm for training. Using this data to train a classification system might result in a significant delay in convergence. Preprocessing the training and validation data before utilizing them is one approach to this problem. First of all, after taking all collected raw data samples are included in the excel file.

(a) Label Encoding

There are several parameters whose values are a string. Because machine learning models demand features to be floats or integers, categorical characteristics such as color, gender, and so on must be translated to numerical values. The label encoder is a program that translates category features into integers.

(b) Data Normalization

Many algorithms compare attributes of data points in order to discover trends in the data. When the features are on significantly different scales, however, there is a problem. For the optimal performance of the classification algorithm, data normalization is required. For normalizing, we apply the Min-Max normalization formula. One of the most prevalent methods of data normalizing is min-max normalization. The minimum value of each feature is converted to a 0, the highest value is converted to a 1, and all other values are converted to a decimal between 0 and 1. The data normalization formula is as follows:

$$Xnormalized = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.5}$$

Where $Xnormalized$ is the new normalized value.

All supervised learning methods start with an input data matrix, called X here. Each tuple represents one observation. Each column represents one variable. For classification, Y can be any data type, here used numeric type- 0 or 1 for the response respectively two-class ‘Yes’ or ‘No’.

Index	your gender	Your Age	Your Living Location	What is the name of Graduation degree that you have achieved?	Name of the educational institute?
0	Male	27	Comilla, Bangladesh	IT/ICT/ICE	Noakhali Science & Technology University
1	Male	32	Dhaka	ME (Mechanical Engineering)/IPE/CHE/	Sonargaon University
2	Male	32	Dhaka	Physical Science (Physics/Mathematics etc.)	Dhaka University
3	Male	26	Dhaka	IT/ICT/ICE	Nstu
4	Male	28	Dhaka	CSE/EEE/ETE/ECE	Comilla University
5	Male	25	Dhaka	IT/ICT/ICE	Noakhali Science & Technology University

Figure 3.1: Before Data Preprocessing.

	0	1	2	3	4	5	6	7	8
0	0.52915	-0.623483	-0.478435	-0.789956	-0.295475	-1.8655	-0.617048	-0.547415	1.42383
1	0.52915	0.836379	2.06095	2.29638	0.159588	0.253884	0.463912	-0.085759	-0.596187
2	0.52915	0.349758	1.2881	1.41457	0.311276	1.31357	0.680104	-0.00881635	0.750494
3	0.52915	-1.59672	0.515238	-0.789956	-0.598851	0.253884	-1.9142	-0.778243	-0.596187
4	0.52915	-0.623483	2.28177	-0.789956	0.311276	0.253884	-0.617048	-0.00881635	-0.596187

Figure 3.2: After Data Preprocessing.

(c) Data Split for Training & Testing

In the final stage of dataset preparation, we split the dataset into a training set and a testing set. For our research, we use 120 individuals' record belonging to 2 classes. We split our data set 80% for training and 20% for testing. So, we have 96 rows for training and 24 rows for testing/ validation.

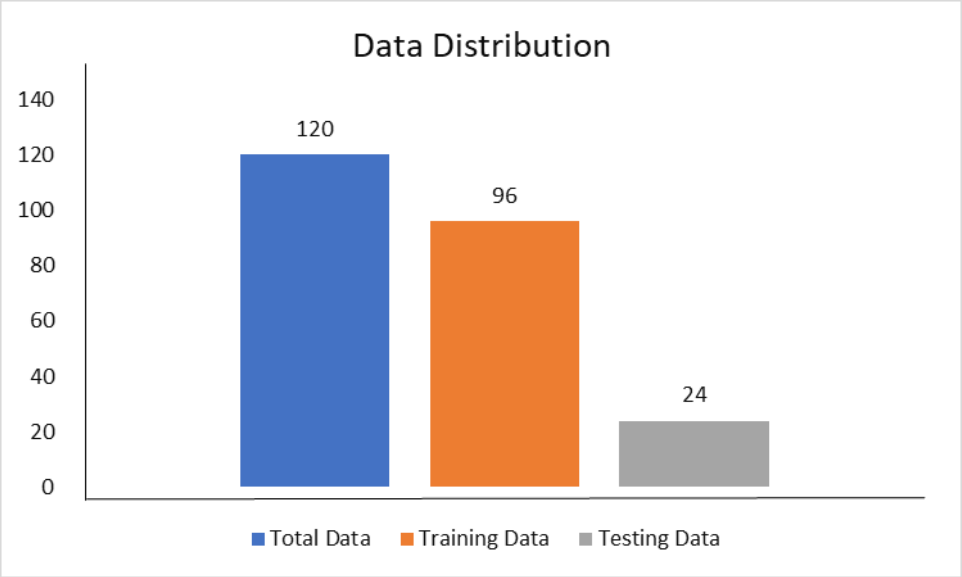


Figure 3.3: Data distribution of training and testing set.

3.3 JOB PREDICTION

For job prediction we use several machine learning models. These models are described briefly in below section.

3.3.1 Logistic Regression

A statistical technique for forecasting binary classes is logistic regression. The result or goal variable has a binary nature. There are just two conceivable classifications when something is dichotomous. It may be used, for instance, to issues with cancer detection. The likelihood of an event happening is calculated.

Given that the target variable is categorical, it is a particular instance of linear regression. As the dependent variable, it employs a log of odds. Using a logit function, logistic regression makes predictions about the likelihood that a binary event will occur.

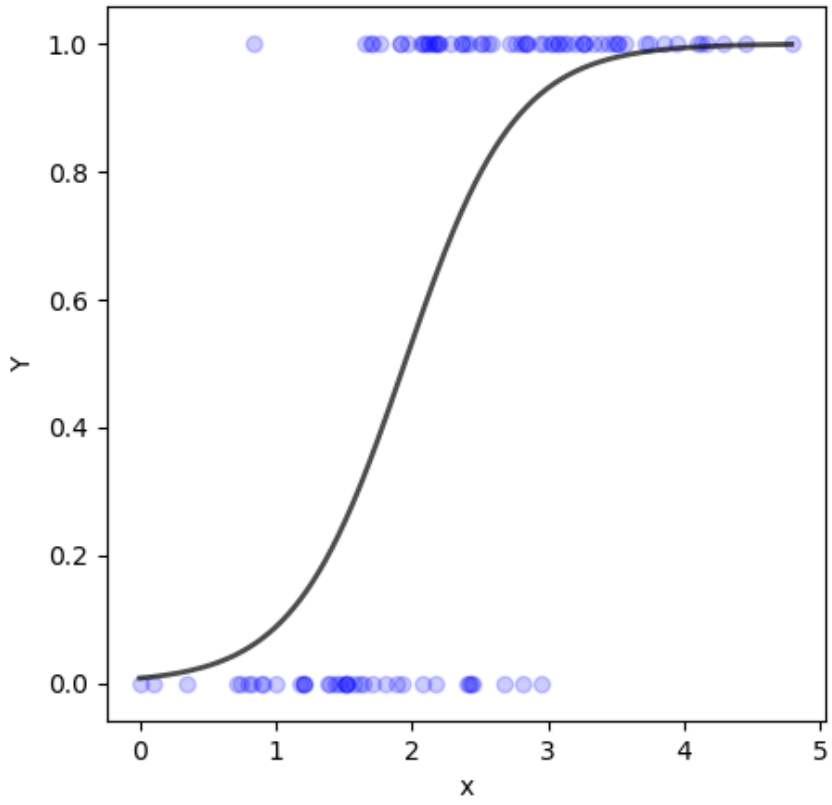


Figure 3.4: Example of logistic regression.

Sigmoid Function:

Any real-valued integer may be mapped into a value between 0 and 1 using the sigmoid function, commonly known as the logistic function. Y anticipated will become 1 if the curve travels to positive infinity, and 0 if the curve goes to negative infinity.

We can categorize the result as 1 or YES if the sigmoid function's output is more than 0.5, and as 0 or NO if it is less than 0.5. The result cannot as an instance, if the result is 0.75, we may state, in terms of probability, that there is a 75% chance that the patient will get cancer.

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{3.1}$$

3.3.2 Gaussian Naive Bayes

A statistical classification method based on the Bayes Theorem is known as Naive Bayes. One of the simplest supervised learning algorithms is this one. The quick, accurate, and trustworthy algorithm is the Naive Bayes classifier. On huge datasets, Naive Bayes classifiers perform quickly and accurately.

Naive The Bayes classifier makes the assumption that the impact of a specific feature on a class is unrelated to the impact of other characteristics. For instance, a loan applicant's suitability depends on factors including income, past loan and transaction history, age, and geography. Even though these qualities are interrelated, they are nonetheless analyzed separately. This assumption makes computing easier, which is why it is viewed as naïve. Class conditional independence is the term used to describe this presumption.

A variation of Naive Bayes known as Gaussian Naive Bayes supports continuous data and adheres to the Gaussian normal distribution. Here we use this Gaussian Naive Bayes.

The formula for Bayes theorem is;

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(B|A)}{P(B)} \tag{3.2}$$

Where,

P(A) = The probability of A occurring

P(B) = The probability of B occurring

P(A|B) = The probability of A given B

P(B|A) = The probability of B given A

P(A∩B) = The probability of both A and B occurring

The assumption that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution is frequently made when working with continuous data. The characteristics' probability is predicated on-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3.3}$$

Sometimes assume variance

- is independent of Y (i.e., σi),

- or independent of Xi (i.e., σk)
- or both (i.e., σ)

Algorithm:

Step 1: determine the prior probability for the specified class labels.

Step 2: Calculate the likelihood probability for each class given each characteristic.

Step 3: Apply the Bayes formula to these values to determine the posterior probability.

Step 4: Assuming the input belongs to the higher probability class, determine which class has a greater probability.

3.3.3 K-Nearest Neighbor

An example of a supervised learning method used for both classification and regression is K-nearest neighbors (KNN). The distance between the test data and all of the training points is calculated by KNN in an effort to forecast the proper class for the test data. Then choose the K spots that are closest to the test data. The KNN method determines which classes of the "K" training data the test data will belong to, and the class with the highest probability is chosen. The value in a regression situation is the average of the 'K' chosen training points.

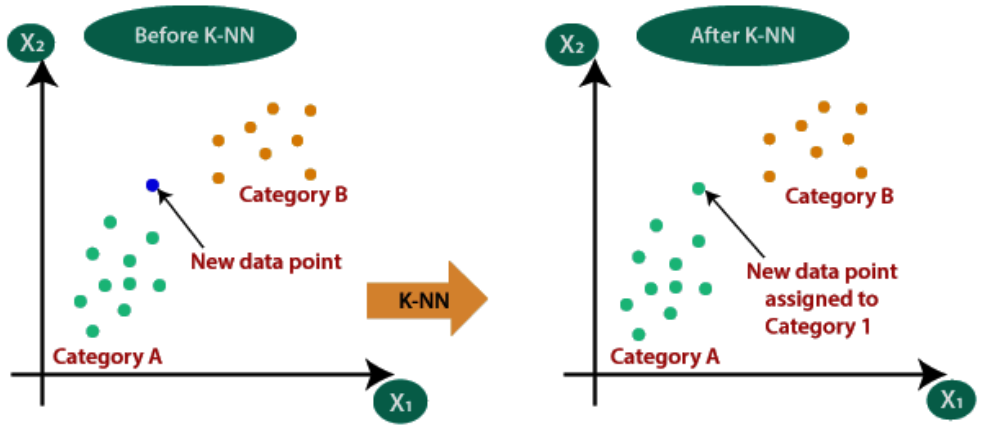


Figure 3.5: Example of K-NN.

If there are two categories, Category A and Category B, and we have a new data point, x1, which category does this data point belong in? We require a KNN method to address this kind of issue. K-NN makes it simple to determine the category or class of a given dataset.

Algorithm:

Step 1: Decide on the neighbors' K-numbers.

Step-2: Calculate the Euclidean distance between K neighbors in step two.

Step 3: Based on the determined Euclidean distance, select the K closest neighbors.

Step 4: Count the number of data points in each category among these k neighbors.

Step 5: Assign the fresh data points to the category where the neighbor count is highest.

Step 6: Our model is complete.

The nearest neighbors' number is shown by K value. Distances between test points and training label points must be calculated. Here we use 4 neighbors for training the model. KNN is a slow learning method because updating distance measures after each iteration are computationally costly.

3.3.4 Support Vector Machine

A supervised machine learning approach called "Support Vector Machine" may be applied to classification or regression problems. However, categorization issues are where it is most frequently utilized. When using the SVM algorithm, each data point is represented as a point in n-dimensional space (where n is the number of features you have), with each feature's value being the value of a certain coordinate. Next, we do classification by identifying the hyper-plane that effectively distinguishes the two classes.

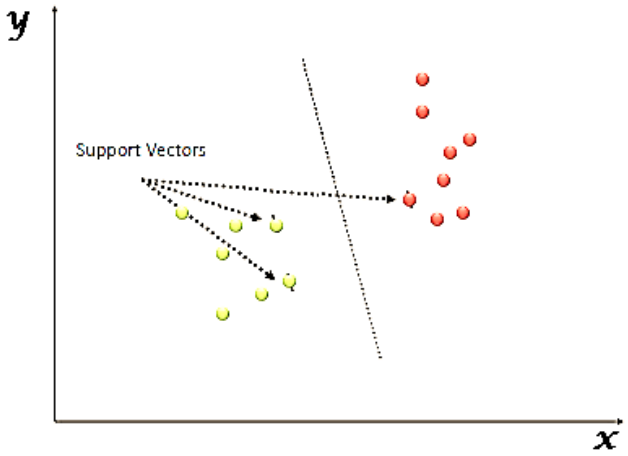


Figure 3.6: Example of Support Vector Machine.

The coordinates of each individual observation are what make up a support vector. A frontier that best separates the two classes (hyper-plane/line) is the SVM classifier.

Hyperplanes and Support Vectors

Decision boundaries known as hyperplanes assist in categorizing the data points. Different classifications can be given to data points that lie on either side of the hyperplane. Additionally, the number of features affects how big the hyperplane is. The hyperplane is essentially a line if there are just two input characteristics. The hyperplane turns into a two-dimensional plane if there are three input characteristics. When there are more than three characteristics, it gets harder to imagine.

SVM Kernels

The SVM algorithm is really implemented with a kernel that modifies the input data space to take the desired form. The kernel trick is a method used by SVM, in which the kernel converts a low-dimensional input space into a higher-dimensional space. To put it simply, the kernel increases the number of dimensions in a problem to make it separable. SVM becomes more potent, adaptable, and accurate as a result. The types of kernels that SVM employs include the following.

- Linear Kernel
- Polynomial Kernel
- Radial Basis Function (RBF) Kernel

Here in this research, we use a polynomial kernel as we are working with nonlinear dataset.

3.3.5 Random Forests Classifiers

An approach to supervised learning is random forests. Both classification and regression may be done with it. Additionally, it is the most user-friendly and adaptable algorithm. Trees are what make up a forest. A forest is supposed to be stronger the more trees it has. On randomly chosen data samples, random forests generate decision trees, obtain predictions from each tree, and vote for the best option. Additionally, it offers a fairly accurate measure of the feature's relevance.

Applications for random forests include feature selection, picture classification, and recommendation engines. It may be used to categorize dependable loan candidates, spot fraud, and forecast sickness. The Boruta method, which chooses significant characteristics in a dataset, is built on it.

Because it requires a shorter training time than other algorithms, we utilize the Random Forest algorithm. It operates effectively even with a big dataset and predicts results with great accuracy. When a significant amount of data is missing, it can still retain accuracy.

Algorithm:

Step 1: choose K data points at random from the training set.

Step 2: Construct the decision trees linked to the chosen data points (Subsets).

Step 3: Select N for the size of the decision trees you wish to construct.

Step-4: Repeat Step 1 & 2.

Step 5: Assign new data points to the category that receives the majority of votes by looking up each decision tree's predictions for the new data points.

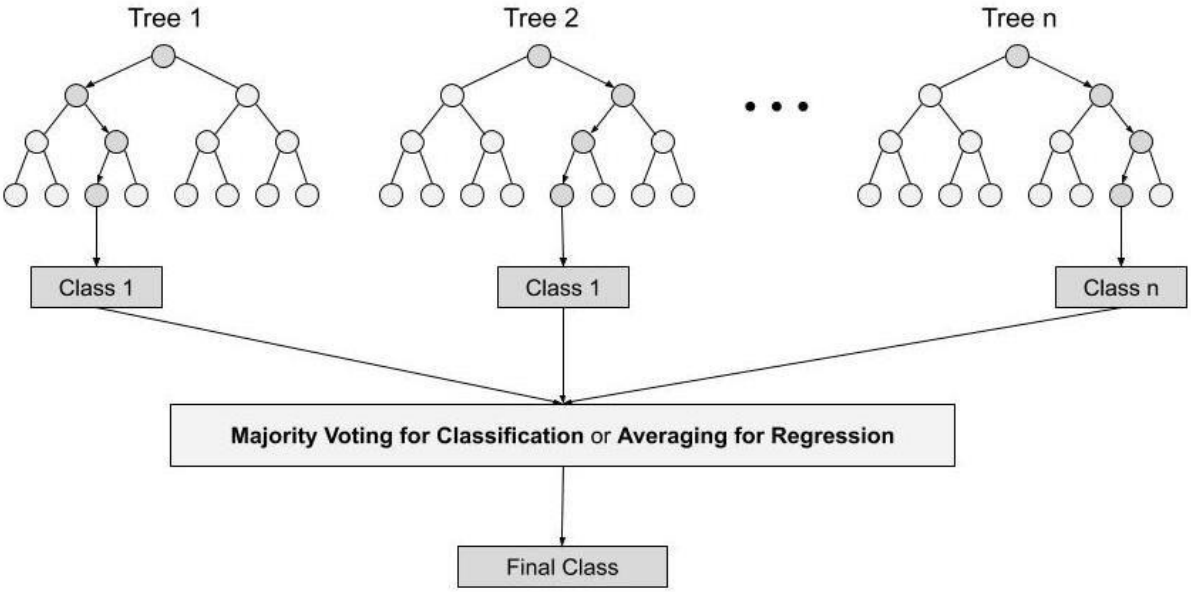


Figure 3.7: Example of Random Forests Classifiers.

3.3.6 Multi Layer Perceptron Neural Network

A complement to feed-forward neural networks is the multi-layer perceptron (MLP). The input layer, output layer, and hidden layer are the three different kinds of layers that make it up. The input layer is where the input signal for processing is received. The output layer completes the necessary task, such as prediction and classification. The actual computational engine of the MLP consists of an arbitrary number of hidden layers positioned between the input and output layers. The data moves from the input to the output layer of an MLP in a manner reminiscent of a feed-forward network.

The back propagation learning technique is used to train the neurons in the MLP. MLPs may resolve issues that are not linearly separable since they are built to approximate any continuous function. Pattern classification, recognition, prediction, and approximation are the main application cases for MLP.

In our research we use MLP with a single hidden layer which consist 20 perceptron neuron, a input layer and a output layer. The input layer consists of the input parameters identified in the previous step. The hidden layer is designed with 20 neurons and to be increased as the training process progresses on demand. Output layer consists of a two parameter which is the prediction of whether one is satisfied with the job or not. For optimization we use adam optimizer and as activation function in hidden layer we use ReLu function.

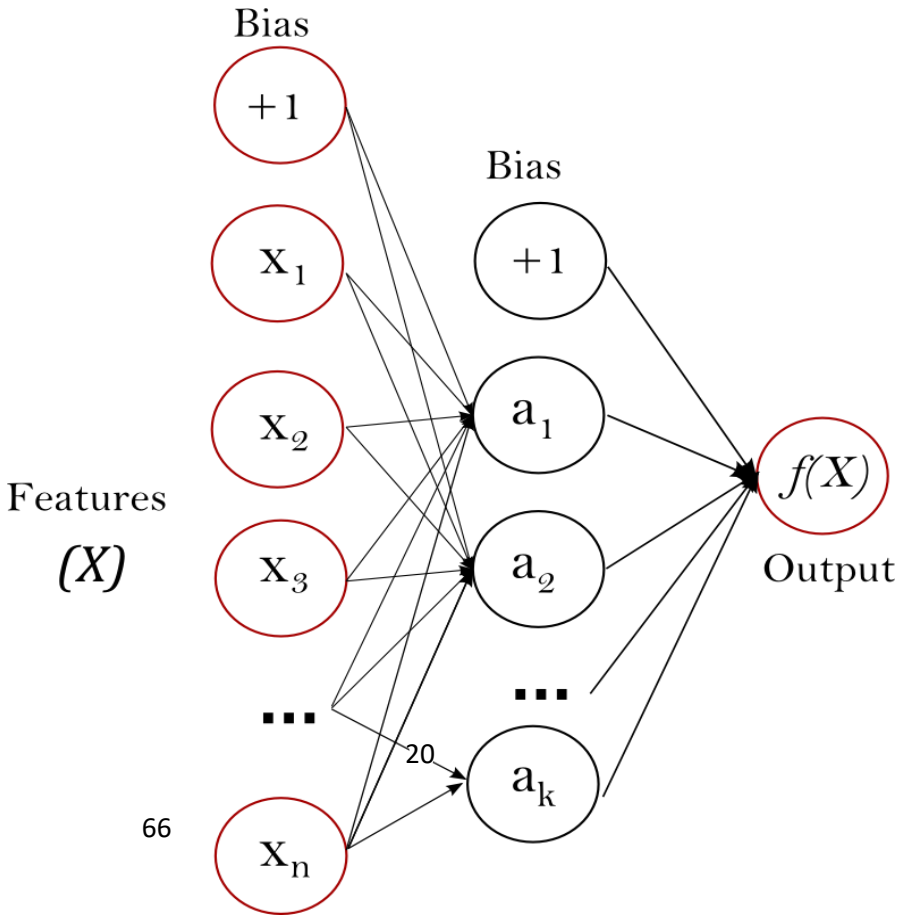


Figure 3.8: MLP structure for Job Satisfaction prediction

3.3.7 Decision Tree Algorithm

Decision Tree is a supervised learning approach that may be used for classification and regression issues; however, it is most commonly employed to solve classification problems. Internal nodes contain dataset attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The Decision Node and the Leaf Node are the two nodes of a decision tree. Leaf nodes are the result of those decisions and do not include any more branches, whereas Choice nodes are used to make any decision and have several branches. Instances are classified using decision trees by sorting them along the tree from the root to a leaf node, which yields the classification. As illustrated in figure 3.3, an

instance is classed by starting at the tree's root node, evaluating the attribute indicated by this node, and then progressing along the tree branch according to the attribute's value. This procedure is then performed for the new node's subtree.

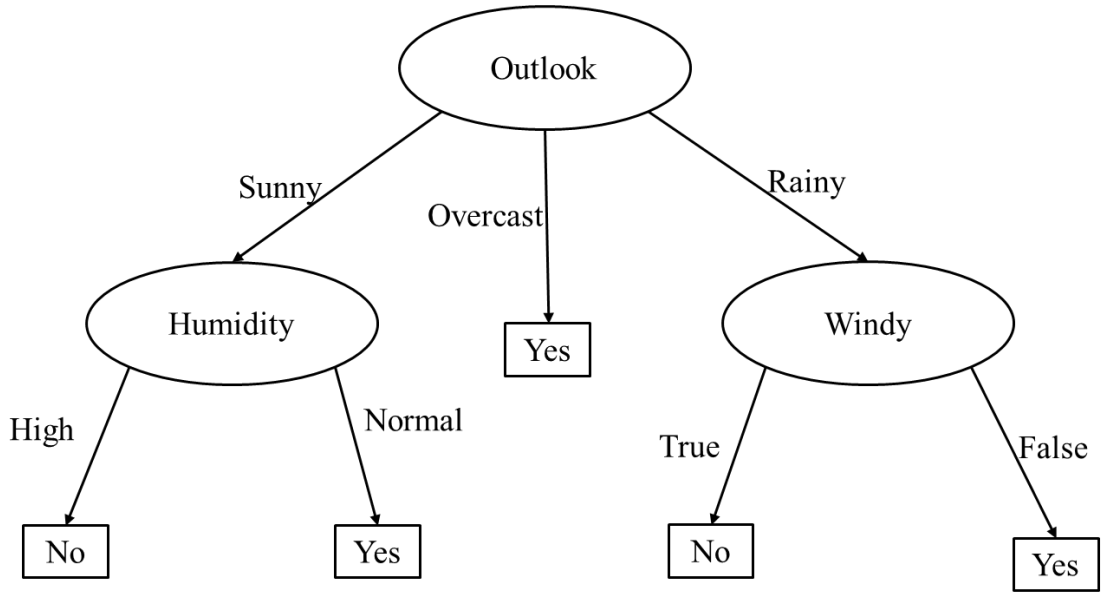


Figure 3.9: Example of a decision tree.

The decision tree in the above graphic categorizes a given morning based on whether it is suitable for tennis play and returns the classification associated with that leaf. (In this situation, the answer is Yes or No).

For example, the instance

(Outlook = Rain, Temperature = Hot, Humidity = High, Wind = Strong) would be sorted down the leftmost branch of this decision tree and would therefore be classified as a negative instance.

To put it another way, a decision tree is a disjunction of conjunctions of constraints on attribute values of instances.

(Outlook = Sunny ^ Humidity = Normal) v (Outlook = Overcast) v (Outlook = Rain ^ Wind = Weak)

Algorithm:

Step 1: According to S, start with the root node, which holds the whole dataset.

Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).

Step 3: Split the S into subgroups based on the best qualities' potential values.

Step 4: Make the optimal attribute decision tree node.

Step 5: Using the subsets of the dataset obtained in step 3, construct new decision trees recursively.

Continue this procedure until the nodes can no longer be classified, at which point the final node is referred to as a leaf node.

3.3.8 GINI Index

Decision Trees are supervised machine learning algorithms that thrive in classification and regression. These algorithms are built by dividing down the training data into subsets of output variables of the same class and applying the specific splitting criteria at each node. This classification procedure starts with the decision tree's root node and grows by applying splitting conditions to each non-leaf node, dividing datasets into a homogenous subset. A decision tree's 'knowledge' is immediately articulated into a hierarchical structure after it has been trained. This framework organizes and presents information in such a way that even non-experts may understand it."

However, because pure homogeneous subsets are impossible to obtain, each node in the decision tree concentrates on selecting a characteristic and a split condition on that feature that minimizes the mixing of class labels, resulting in reasonably pure subsets.

Various splitting indices have been offered to examine "the quality of splitting criterion" or to evaluate how well the splitting is. Gini index and information gain are two of them.

Here in our research, we use the Gini index. The Gini index, also known as the Gini coefficient or Gini impurity, is a version of the Gini coefficient that calculates the risk of a certain variable being incorrectly categorized when picked at random.

It only does binary splitting since it works with categorical data and delivers results of "success" or "failure."

The Gini index ranges from 0 to 1, with 0 being the lowest and 1 being the highest. Where 0 denotes that all of the items belong to the same class, or that only one class exists. The value of 1 in the Gini index indicates that all items are spread randomly among various classes, and a score of 0.5 indicates that the items are evenly distributed across many classes. Mathematically, The Gini Index is represented by;

$$\boldsymbol{Gini} = 1 - \sum_j p_j^2 \qquad (3.4)$$

3.4 MODEL TRAINING

Our used models are supervised machine learning models. That means it learns for all weights and biases from labeled data. This is where our system knows the parameters of the satisfied job by extracting the features from given data. The primary computational part of our system is done in this section. We use a decision tree machine learning algorithm and several other machine learning models. We get 3 phases of the process.

Training: We apply 80% of the data for training phase.
Testing: We apply 20% of the data for testing phase.
Comparison: We compare both phases.

3.5 PERFORMANCE ANALYSIS

Here we describe confusion matrix, AUC-ROC Curve, Precision, Accuracy, Recall, F1 score for analyzing the performance.

(i) Confusion Matrix

A confusion matrix is one of the most common and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problems where the output can be of two or more types of classes. The confusion matrix is a table with two dimensions "Actual class" and "Predicted class" in both dimensions. Actual classifications are Rows and Predicted ones are columns. In the dataset there are two classes one is Class 1 and another is Class 2. A confusion matrix has created as follows:

Table 3.4: Confusion Matrix

		Predicted	
		Positive (Class 1)	Negative (Class 2)
Actual	Positive (Class 1)	TP	FN
	Negative (Class 2)	FP	TN

(ii) AUC – ROC Curve

AUC - ROC is a graphical representation for performance measurement for classification problems. ROC is a probability curve and AUC represents degree or measure of area. It expresses how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the model is more

effective in differentiating between patients with and without a disease when the AUC is higher. TPR and FPR are shown on the y-axis and x-axis, respectively, to represent the ROC curve.

True Positives (TP): True positives are the cases when the actual class of the data point was True and the predicted is also True.
True Negatives (TN): True negatives are the cases when the actual class of the data point was False and the predicted is also false.
False Positives (FP): False positives are the cases when the actual class of the data point was False and the predicted is True.
False Negatives (FN): False negatives are the cases when the actual class of the data point was true and the predicted is False.

(iii) Precision (Positive Predictive Value): It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

$$\boldsymbol{Precision} = \frac{True\ Positive}{True\ positive+False\ Positive} \tag{3.6}$$

(iv) Accuracy: Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. Accuracy comparison is based on the performance among the four classification algorithms.

$$\boldsymbol{Accuracy} = \frac{True\ Positive+True\ Negative}{True\ positive+True\ Negative+False\ Positive+False\ Negative} \tag{3.7}$$

(v) Recall: Recall score gives a value that shows the ability of a model to predict positive classes out of actual positive classes.

$$\boldsymbol{Recall} = \frac{True\ Positive}{True\ positive+False\ Negative} \tag{3.8}$$

(vi) F1-score: F1 measures the precision of the model by a blend of precision and recall. That means it takes false positive and false negative to calculate the F1 score. Often f1 score is calculated to tune the precision and recall value. By the following equation, we can calculate the F1 score;

$$\boldsymbol{F1\ Score} = \frac{2*(Recall*Precision)}{(Recall+Precision)} \tag{3.9}$$

3.6 FLOW CHART

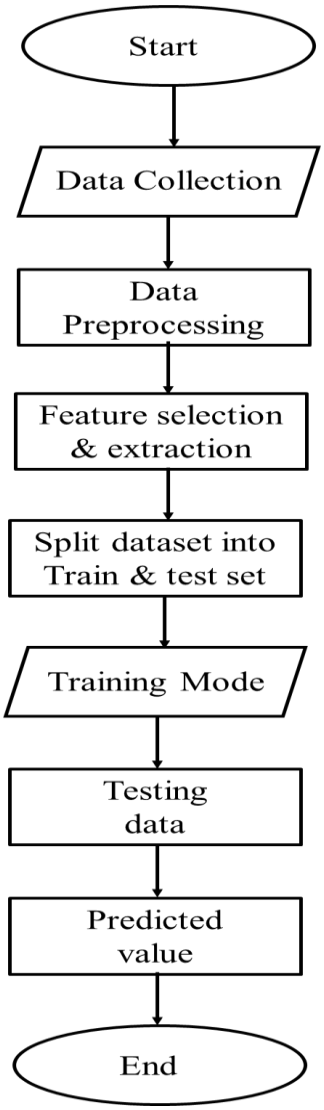


Figure 3.10: Implementation Module Flow Chart.

3.7 SUMMARY

This methodology chapter has discussed about the dataset and how to process dataset and pattern classification algorithms and basic terminologies which are used in this work such as Logistic regression, K-nearest neighbors, Support Vector Machine, Gaussian Naive Bayes, Random Forests Classifiers, Decision Tree and Multi-Layer Perceptron.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 INTRODUCTION

In this chapter, the results of different machine learning classification algorithms are discussed. Performance analysis terminologies such as confusion matrix, roc curve, error calculation etc. also have included in this chapter. Results produced from all the selected algorithms are compared with results from the evaluation process. In this paper we work with different classification algorithm such as Logistic regression, K-nearest neighbors, Support Vector Machine, Gaussian Naive Bayes, Random Forests Classifiers, Decision Tree and Multi-Layer Perceptron. Their performance was evaluated and selects a final model with maximum accuracy.

4.2 PERFORMANCE ANALYSIS OF DISTINCT MODEL

This paper carried out by applying several algorithms from a collected dataset for selecting a suitable job for a candidate. After performing different classification algorithms and values of different terms for each classification algorithm are listed in a table to measure and explore the performance of the classification methods. Comparison is based on different performance metrics like TP rate, FP rate, and Precision, Error, Recall and ROC area are for different classifiers. Model evaluation or Performance Analysis is based on a confusion matrix and all-related measures like- ROC curve, True positives, True negatives, False positives, False negatives, Accuracy, Precision, Recall, F1 Score etc.

4.2.1 Logistic Regression

For the training dataset, there was 67 features and 96 input to the logistic regression and for testing, there was 24 data sample. There was two class for output result for whether someone is satisfied with the job or not as a class were “Yes” and “No”. The testing result of logistic regression is given below:

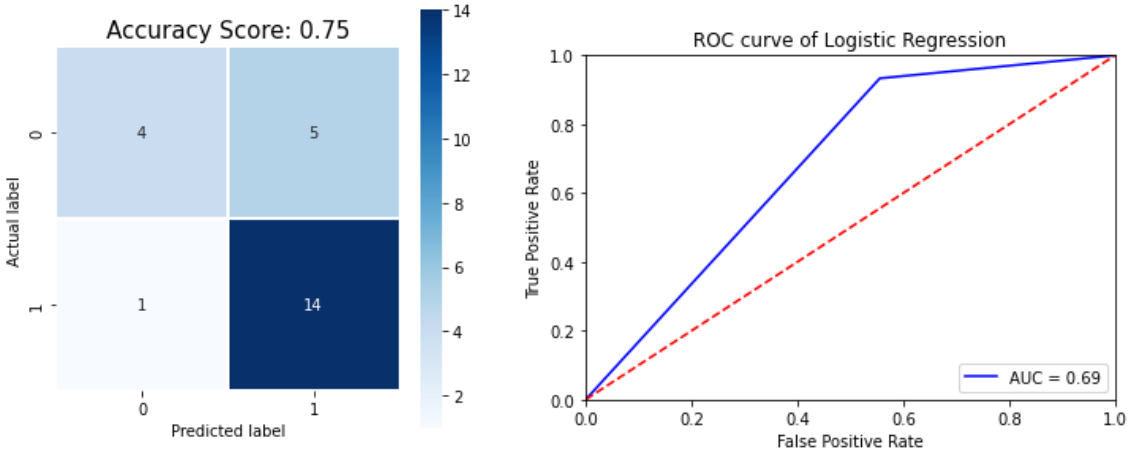
Table 4.1: Performance metrics for logistic regression of individual classes.

Class	precision	recall	f1-score	support
Class 1 (No)	0.80	0.44	0.57	9
Class 2 (Yes)	0.74	0.93	0.82	15

Table 4.2: Average performance metrics and accuracy for logistic regression.

Average	precision	recall	f1-score	Accuracy
Macro avg	0.77	0.69	0.70	0.75
Weighted avg	0.76	0.75	0.73	

Table 4.1 and 4.2 show the results of the logistic regression model. Which shows that average accuracy of the model is about 75%, precision score 77% means positive prediction rate is 77%, the recall value is 69% and the f1-score is 70%. In the case of individual class prediction



(a) (b)

Figure 4.1: Logistic Regression Model: (a) Confusion matrix, (b) ROC curve.

accuracy, it shows precision of 80% for class 1 (Yes) and 74 % precision for class 2 (No). We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class.

The figure 4.1 (a) shows that logistic regression model was able to correctly classify 18 data sample and 6 data sample was misclassified among 24 testing data. Figure 4.1 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Logistic regression is 69% means it is able to distinguish between classes 69% time accurately.

4.2.2 Gaussian Naive Bayes

The testing result of Gaussian Naive Bayes is given below:

Table 4.3: Performance metrics for Gaussian Naive Bayes of individual classes.

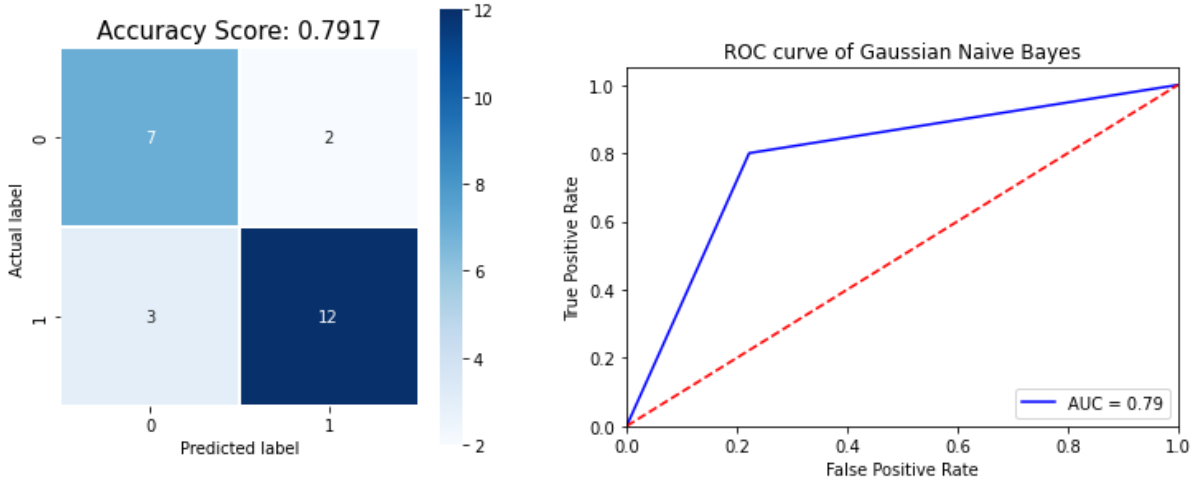
Class	precision	recall	f1-score	support
Class 1 (No)	0.70	0.78	0.74	9
Class 2 (Yes)	0.86	0.80	0.83	15

Table 4.4: Average performance metrics and accuracy for Gaussian Naive Bayes.

Average	precision	recall	f1-score	Accuracy
Macro avg	0.78	0.79	0.78	0.79
Weighted avg	0.80	0.79	0.79	

Table 4.3 and 4.4 show the results of the Gaussian Naive Bayes model. Which shows that average accuracy of the model is about 79%, precision score 80% means positive prediction rate is 80%, the recall value is 79% and the f1-score is 79%. In the case of individual class prediction accuracy, it shows precision of 70% for class 1 (Yes) and 86 % precision for class 2 (No).

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. In below the confusion matrix for Gaussian Naive Bayes model is given;



(a) (b)

Figure 4.2: Gaussian Naive Bayes Model: (a) Confusion matrix, (b) ROC curve.

The figure 4.2 (a) shows that Gaussian Naive Bayes model was able to correctly classify 19 data sample and 5 data sample was misclassified among 24 testing data. The figure 4.2 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Gaussian Naive Bayes is 79% means it is able to distinguish between classes 79% time accurately.

4.2.3 Support Vector Machine

The testing result of Support Vector Machine is given below:

Table 4.5: Performance metrics for Support Vector Machine of individual classes.

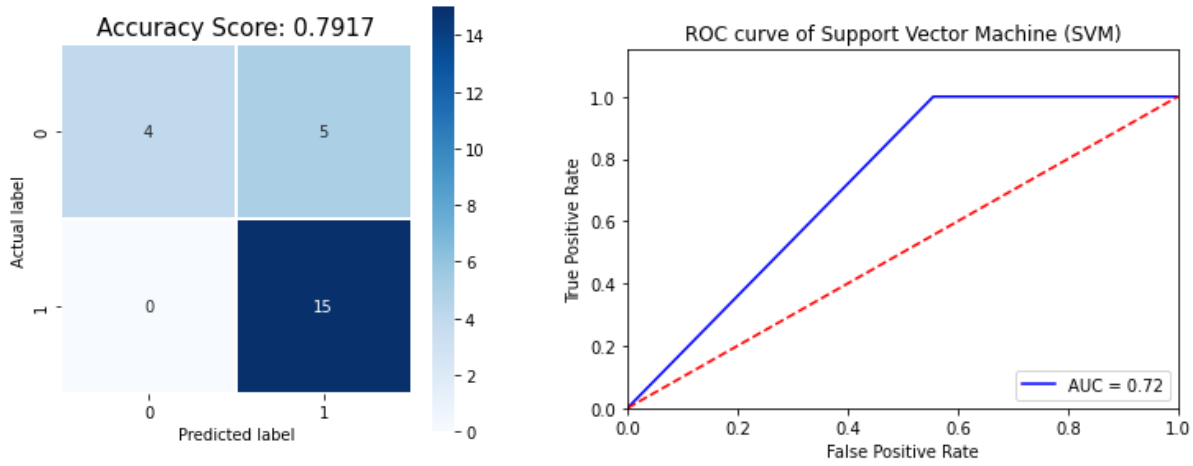
Class	precision	recall	f1-score	support
Class 1 (No)	1.00	0.44	0.62	9
Class 2 (Yes)	0.75	1.00	0.86	15

Table 4.6: Average performance metrics and accuracy for Support Vector Machine.

Average	precision	recall	f1-score	Accuracy
Macro avg	0.88	0.72	0.74	0.79
Weighted avg	0.84	0.79	0.77	

Tables 4.5 and 4.6 show the results of the Support Vector Machine model. This shows that the average accuracy of the model is about 79%, the precision score 88% means a positive prediction rate is 80%, the recall value is 72% and the f1-score is 74%. In the case of individual class prediction accuracy, it shows precision of 100% for class 1 (Yes) and 75 % precision for class 2 (No).

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. In below the confusion matrix for Support Vector Machine model is given;



(a) (b)

Figure 4.3: Support Vector Machine: (a)Confusion matrix, (b) ROC curve.

The figure 4.3(a) shows that Support Vector Machine model was able to correctly classify 19 data sample and 5 data sample was misclassified among 24 testing data. The figure 4.3 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Support Vector Machine is 72% means it is able to distinguish between classes 72% time accurately.

4.2.4 K-Nearest-Neighbor

The testing result of K-Nearest-Neighbor is given below:

Table 4.7: Performance metrics for K-Nearest-Neighbor of individual classes.

Class	precision	recall	f1-score	support
Class 1 (No)	0.80	0.89	0.84	9
Class 2 (Yes)	0.93	0.87	0.90	15

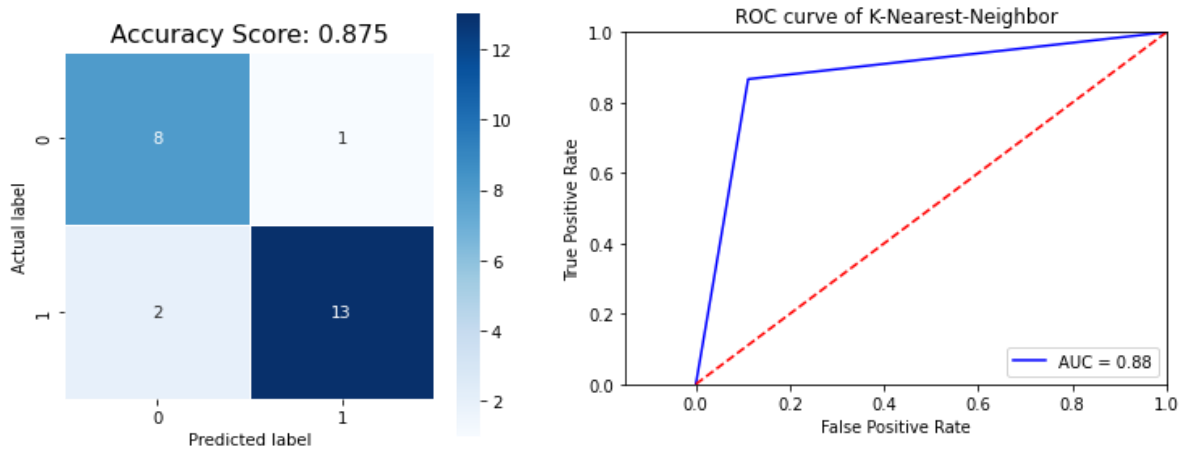
Table 4.8: Average performance metrics and accuracy for K-Nearest-Neighbor

Average	precision	recall	f1-score	Accuracy
Macro avg	0.86	0.88	0.87	0.875
Weighted avg	0.88	0.88	0.88	

Tables 4.7 and 4.8 show the results of the K-Nearest-Neighbor model. This shows that the average accuracy of the model is about 87.5%, the precision score 86% means a positive

prediction rate is 86%, the recall value is 88% and the f1-score is 87%. In the case of individual class prediction accuracy, it shows precision of 80% for class 1 (Yes) and 93 % precision for class 2 (No).

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. In below the confusion matrix for K-Nearest-Neighbor model is given:



(b)

Figure 4.4: K-Nearest-Neighbor Model: (a)Confusion matrix, (b) ROC curve.

The figure 4.4 (a) shows that K-Nearest-Neighbor model was able to correctly classify 19 data sample and 5 data sample was misclassified among 24 testing data. Figure 4.4 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for K-Nearest-Neighbor is 88% means it is able to distinguish between classes 88% time accurately.

4.2.5 Random Forest Tree

The testing result of Random Forest Tree is given below:

Table 4.9: Performance metrics for Random Forest of individual classes.

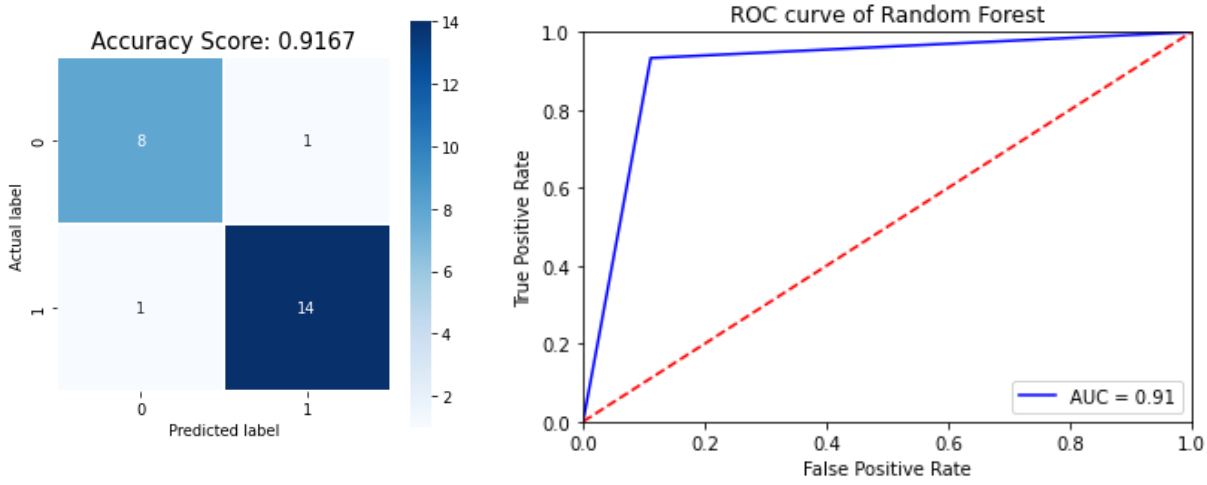
Class	precision	recall	f1-score	support
Class 1 (No)	0.89	0.89	0.89	9
Class 2 (Yes)	0.93	0.83	0.93	15

Table 4.10: Average performance metrics and accuracy for Random Forest

Average	precision	recall	f1-score	Accuracy
Macro avg	0.91	0.91	0.91	0.92
Weighted avg	0.92	0.92	0.92	

Tables 4.9 and 4.10 show the results of the Random Forest model. This shows that the average accuracy of the model is about 92%, the precision score of 91% means a positive prediction rate is 91%, the recall value is 91% and the f1-score is 91%. In the case of individual class prediction accuracy, it shows a precision of 89% for class 1 (Yes) and 93 % precision for class 2 (No).

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. In below the confusion matrix for Random Forest model is given;



(a) (b)

Figure 4.5: Random Forest Model: (a)Confusion matrix, (b) ROC curve.

The figure 4.5 (a) shows that Random Forest model was able to correctly classify 22 data sample and 2 data sample was misclassified among 24 testing data. Figure 4.5 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Random Forest is 91% means it is able to distinguish between classes 91% time accurately.

4.2.6 Multi-Layer Perceptron Neural Network Model

The testing result of Multi-Layer Perceptron Neural Networks is given below:

Table 4.11: Performance metrics for Multi-Layer Perceptron Neural Networks of individual classes.

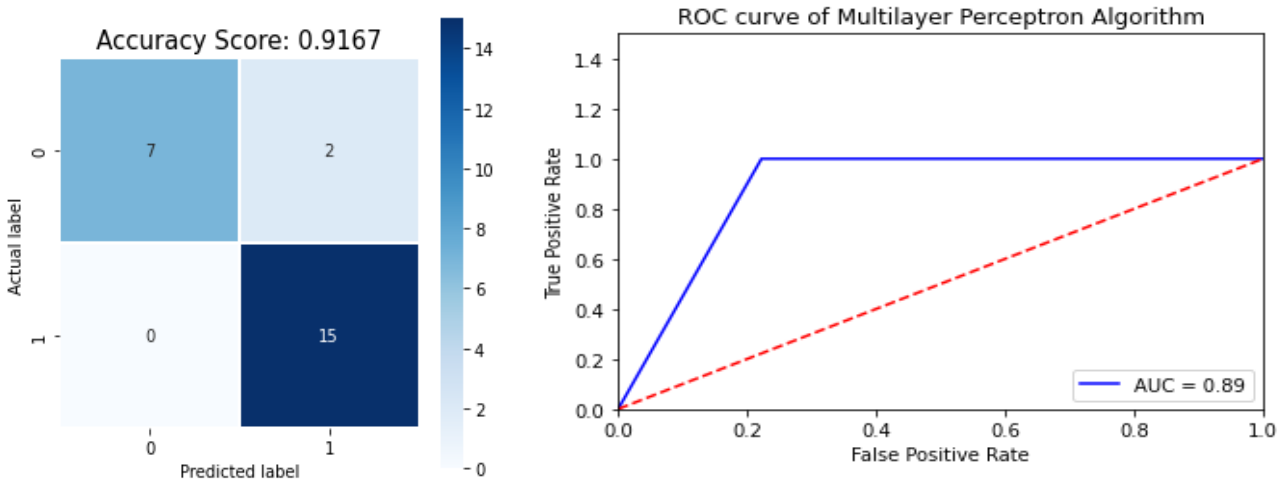
Class	precision	recall	f1-score	support
Class 1 (No)	1.00	0.78	0.88	9
Class 2 (Yes)	0.88	1.00	0.94	15

Table 4.12: Average performance metrics and accuracy for Multi-Layer Perceptron Neural Networks

Average	precision	recall	f1-score	Accuracy
Macro avg	0.94	0.89	0.91	0.916
Weighted avg	0.93	0.92	0.91	

Tables 4.11 and 4.12 show the results of the Multi-Layer Perceptron Neural Networks model. This shows that the average accuracy of the model is about 91.6%, the precision score of 93% means a positive prediction rate is 93%, the recall value is 92% and the f1-score is 92%. In the case of individual class prediction accuracy, it shows a precision of 100% for class 1 (Yes) and 88 % precision for class 2 (No).

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. In below the confusion matrix for Multi-Layer Perceptron Neural Networks model is given;



(a) (b)

Figure 4.6: MLP Model: (a)Confusion matrix, (b) ROC curve.

The figure 4.6 (a) shows that Multi-Layer Perceptron Neural Networks model was able to correctly classify 22 data sample and 2 data sample was misclassified among 24 testing data. The figure 4.6 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Multi-Layer Perceptron Neural Networks is 89% means it is able to distinguish between classes 89% time accurately.

4.2.7 Decision Tree Model

The testing result of the Decision Tree is given below:

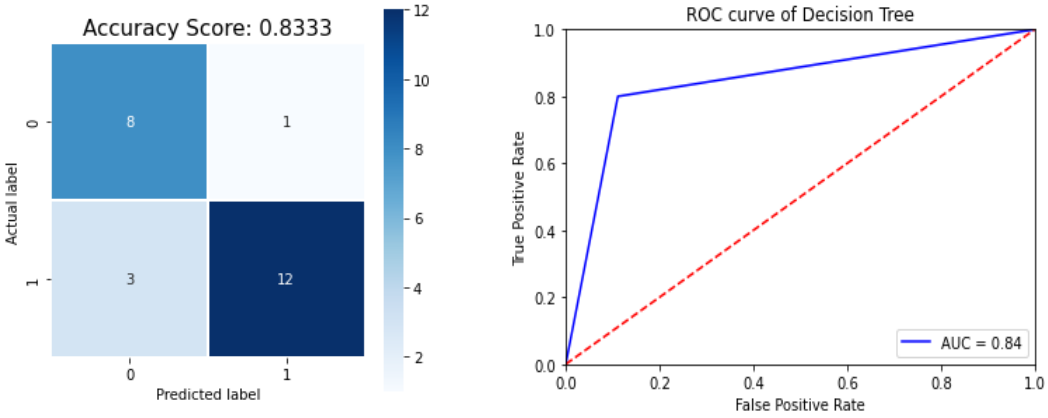
Table 4.13: Performance metrics for Decision Tree of individual classes.

Class	precision	recall	f1-score	support
Class 1 (No)	0.73	0.89	0.80	9
Class 2 (Yes)	0.92	0.80	0.86	15

Table 4.14: Average performance metrics and accuracy for Decision Tree of individual classes

Average	precision	recall	f1-score	Accuracy
Macro avg	0.83	0.84	0.83	0.83
Weighted avg	0.85	0.83	0.84	

Tables 4.13 and 4.14 show the results of the Decision Tree model. This shows that the average accuracy of the model is about 83%, the precision score of 85% means a positive prediction rate is 85%, the recall value is 83% and the f1-score is 84%. In the case of individual class prediction accuracy, it shows a precision of 73% for class 1 (Yes) and 92 % precision for class 2 (No).



(a) (b

Figure 4.7: Decision Tree Model: (a)Confusion matrix, (b) ROC curve.

We can see the summary of the model from confusion matrix which shows the correctness of predicted class as actual class. The figure 4.7 (a) shows that Decision Tree model was able to correctly classify 20 data sample and 4 data sample was misclassified among 24 testing data. Figure 4.7 (b) shows the AUC - ROC curve where ROC is a probability curve and AUC score represents the degree or measure of separability. The AUC score for Decision Tree is 84% means it is able to distinguish between classes 84% time accurately.

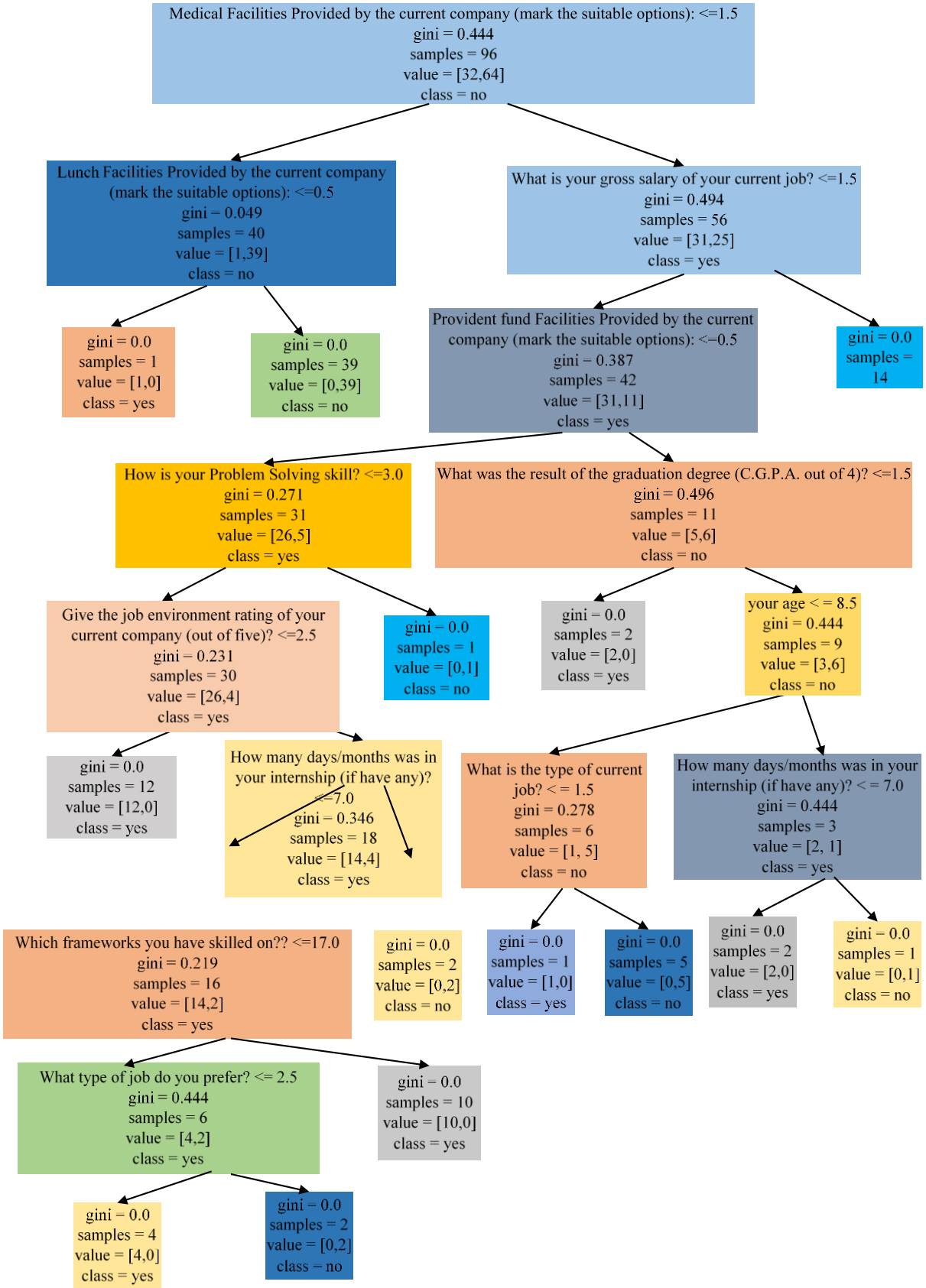


Figure 4.15: Tree construction of Decision Tree.

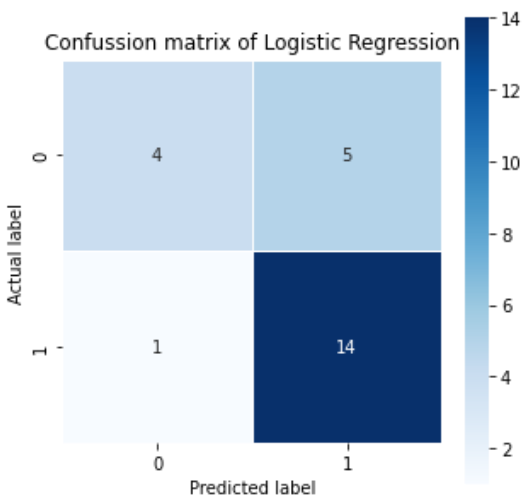
Figure 4.8 shows the constructed tree for decision tree algorithms with some important features. The tree split on the basis of Gini Index value. The leaves node of the tree indicates the decision and the internal nodes represent the features.

4.3 COMPARATIVE ANALYSIS OF DIFFERENT MODEL

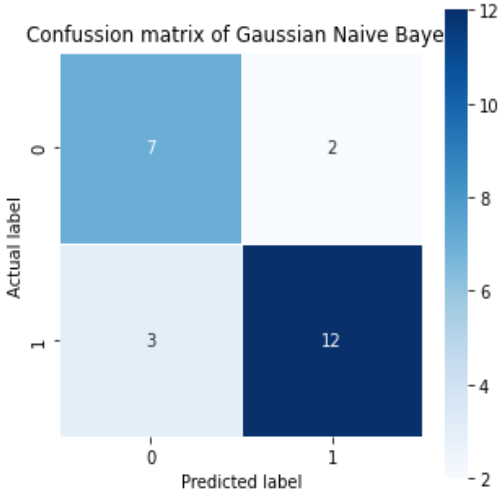
Comparison of different performance terminologies like confusion matrix, ROC, precision score, recall value, f1 score, accuracy etc. of different machine learning model are shown below:

Comparison of confusion matrix between different Machine learning model is shown in figure 4.9

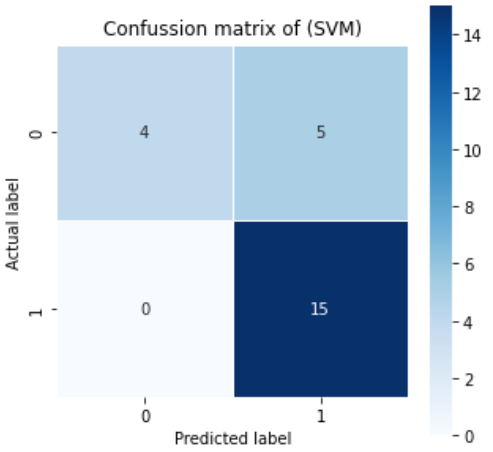
.



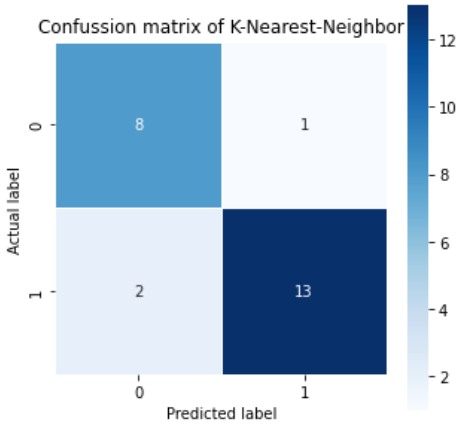
(a)



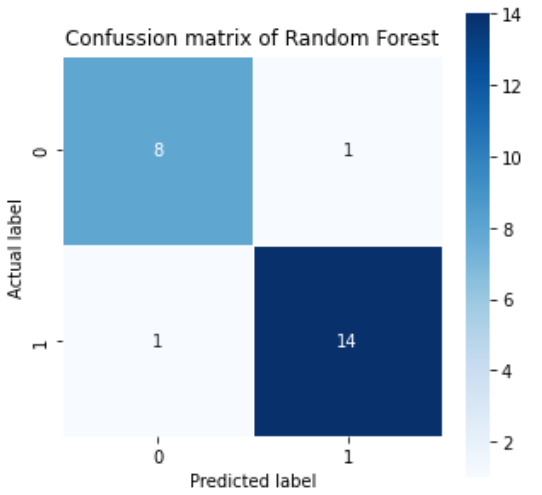
(b)



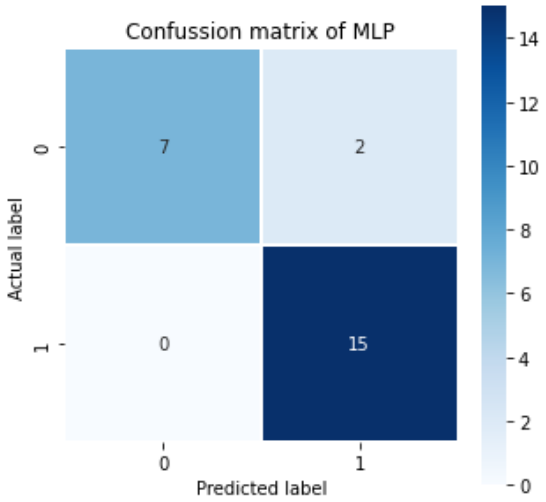
(c)



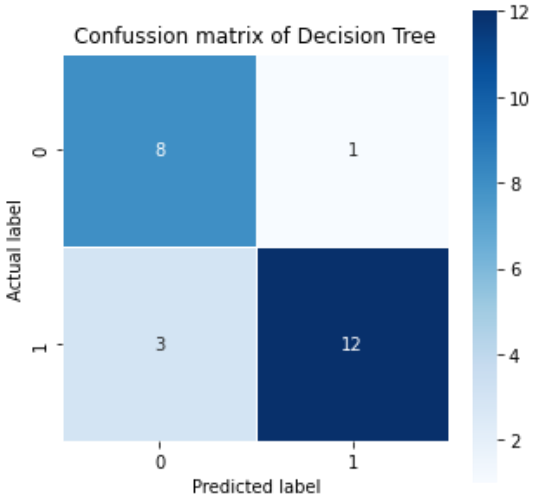
(d)



(e)



(f)



(g)

Figure 4.16: Comparison of confusion matrix between different model (a)Logistic Regression, (b) Gaussian Naive Bayes, (c) Support Vector Machine, (d) K-Nearest-Neighbor, (e) Random Forest Tree, (f) Multilayer Perceptron Algorithm, (g) Decision Tree

Table 4.15: Comparison of Performance metrics for different models

Machine Learning Model Name	Precision Score	Recall Value	F1-Score	Accuracy
Logistic Regression	0.76	0.75	0.73	0.75
Gaussian Naive Bayes	0.80	0.79	0.79	0.79
K-Nearest-Neighbor	0.88	0.88	0.88	0.875
Support Vector Machine	0.84	0.79	0.77	0.79
Random Forest Tree	0.92	0.92	0.92	0.92
Multilayer Perceptron Algorithm	0.93	0.92	0.91	0.916
Decision Tree	0.85	0.83	0.84	0.83

Table 4.15 shows the comparison of different machine learning algorithms. Here we see Random Forest tree gives best accuracy among other models.

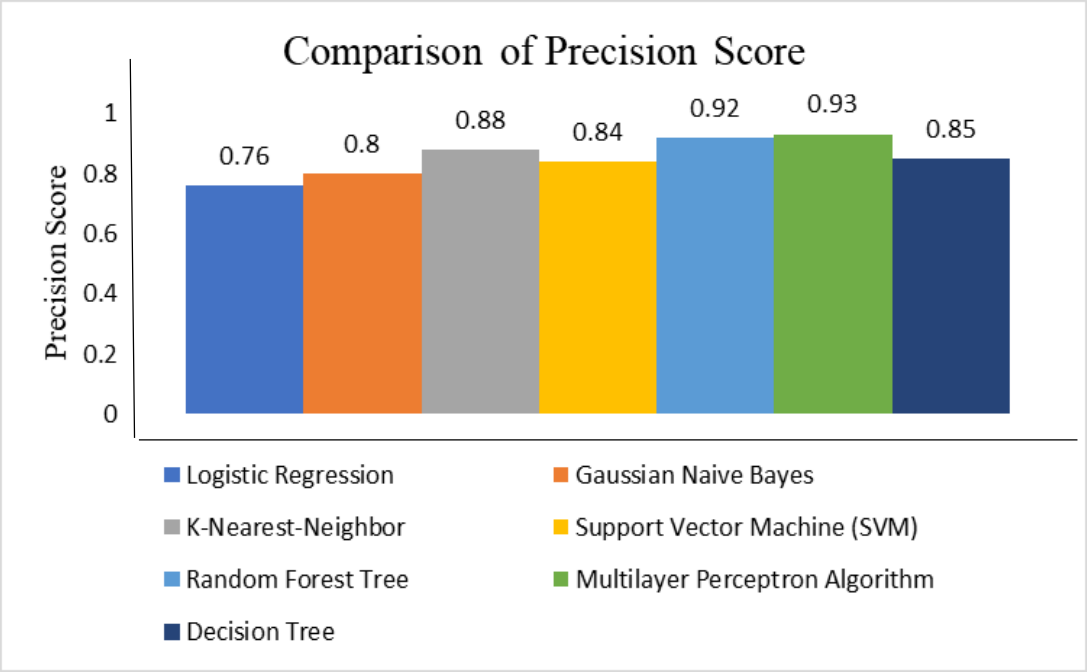


Figure 4.17: Comparison of Precision Score for different Models.

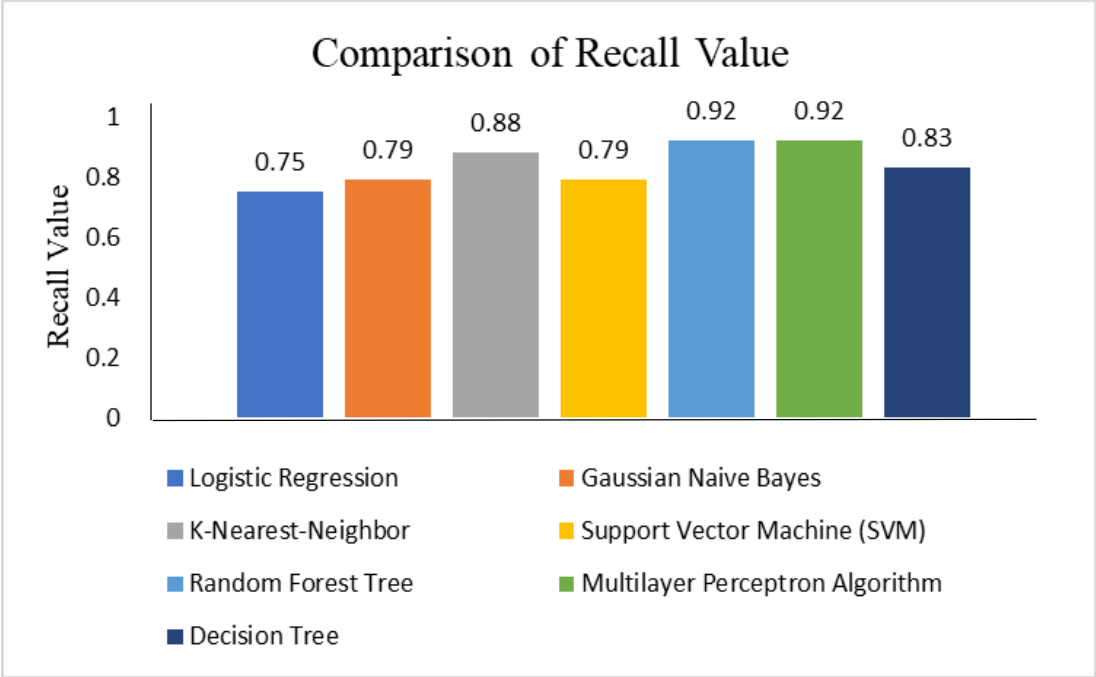


Figure 4.18: Comparison of Recall Value for different Models

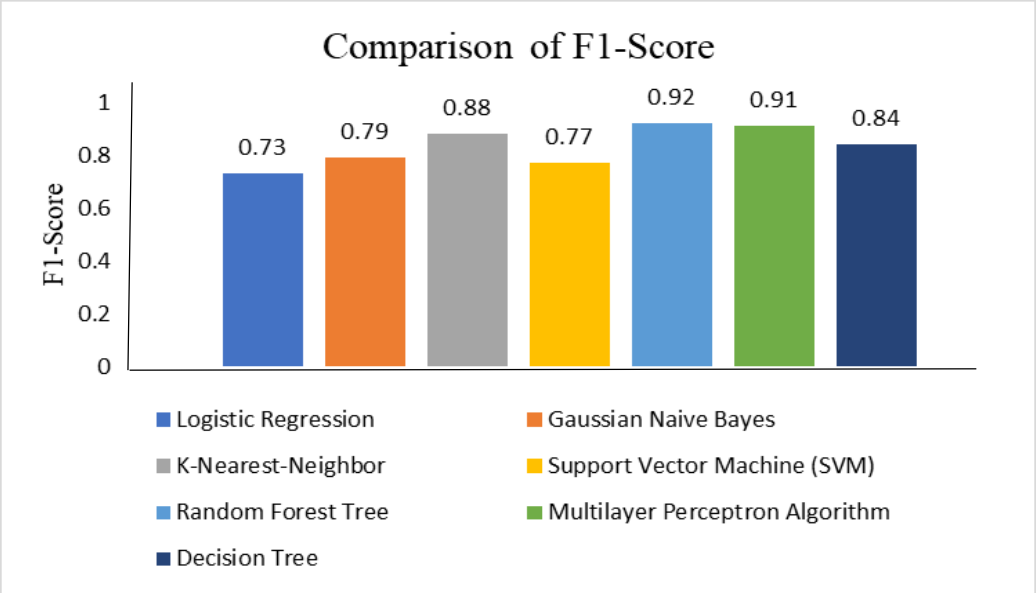


Figure 4.19: Comparison of F1-Score for different Models

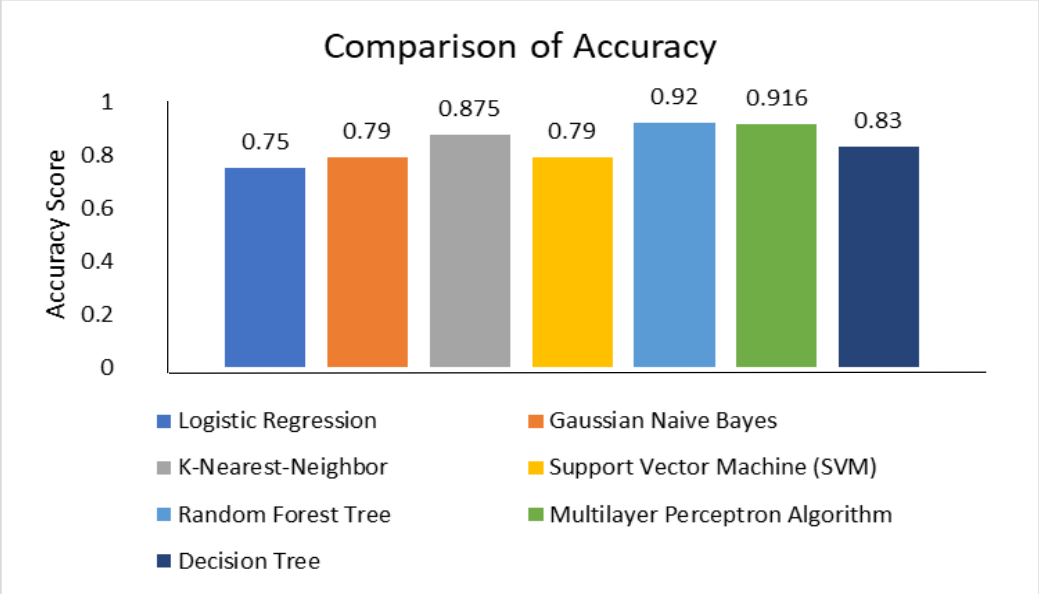


Figure 4.20: Comparison of Accuracy Score for different Models.

Figure 4.10, 4.11, 4.12, and 4.13 plot the bar chart to show the Comparison of Precision Score, Recall value, F1-Score, and accuracy of different machine learning model. From the comparison, we can easily see that the random forest tree model shows promising performance in job satisfying prediction.

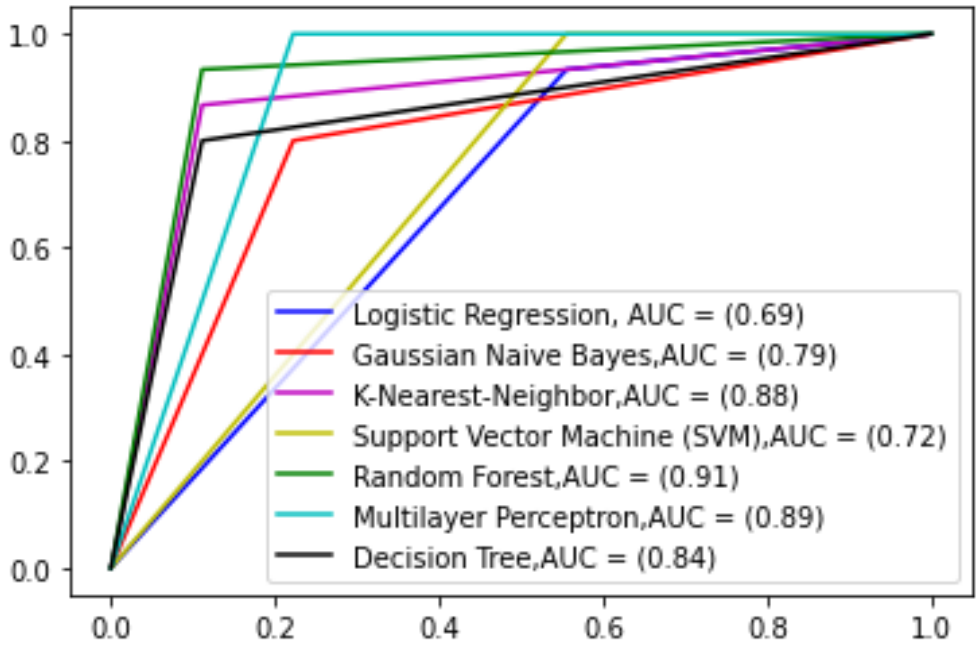


Figure 4.21: Comparison of ROC curve and AUC score for different Models.

From figure 4.14 we also notice that the Random Forest Tree model has a higher AUC score which is 0.91 that means this algorithm predicts class more accurately than other algorithms.

4.4 TESTING

Here we will see the real data set value for 24 entries and also the predicts result of the different models. We compare the real value with the predicted value and get result such as is the real value and the model value same or not. If any model gives better result than other models, then we will reach the final decision for this problem.

Table 4.16: Model Prediction for the Dataset

Sample No.	Dataset Result	Predicted Results						
		Logistic Regression	Gaussian Naive Bayes	K-Nearest-Neighbor	Support Vector Machine	Random Forest Tree	Multilayer Perceptron Algorithm	Decision Tree
i	No	Yes	No	No	Yes	No	Yes	No
ii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
iii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
iv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
v	No	No	No	No	No	No	No	No
vi	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
vii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
viii	Yes	No	No	Yes	Yes	Yes	Yes	No
ix	Yes	Yes	No	No	Yes	No	Yes	No
x	No	No	No	No	No	No	No	No
xi	No	Yes	Yes	No	Yes	No	No	No
xii	No	No	No	No	No	No	No	No
xiii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
xiv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
xv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
xvi	No	Yes	No	Yes	Yes	No	No	No
xvii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
.xviii	No	No	No	No	No	No	No	No
xix	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
xx	Yes	Yes	No	No	Yes	Yes	Yes	Yes
xxi	No	Yes	Yes	No	Yes	Yes	Yes	Yes
xxii	No	Yes	No	No	Yes	No	No	No
xxiii	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
xxiv	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 4.15 shows the prediction of different machine learning algorithms.

The performance of different machine learning algorithms which is obtained under Section 3.2 to 3.8 is described below:

Table 4.17: Comparison of Correct & Incorrect Prediction

Machine Learning Model Name	Correct Prediction	Incorrect Prediction	Accuracy (%)
Logistic Regression	18	6	70%
Gaussian Naive Bayes	19	5	62%
K-Nearest-Neighbor	21	3	85.8%
Support Vector Machine	19	5	62%
Random Forest Tree	22	2	90%
Multilayer Perceptron Algorithm	22	2	90%
Decision Tree	20	4	80%

Table 4.16 shows the prediction comparison score of different machine learning algorithms. Here we see Random Forest tree and multilayer perception model gives best accuracy among other models.

4.5 FINAL SELECTION

From our research, we have reached in this point that Random Forest Tree model is better among all models in this job selection problem.

4.6 SUMMARY

In this chapter, analyze the parameters, results of all algorithms are discussed in details and select the final model among different machine learning model which is considered for this work. And show the comparison of performance metrics of different model

CHAPTER 5

CONCLUSION

5.1 INTRODUCTION

This chapter describes the conclusion and future research scope of this thesis work. First the overall summary of this thesis work is presented in the conclusion section. Then the potential future research directions are provided in the future research scope section

5.2 CONCLUSION

This research focuses on the use of different supervised classification techniques to enhance the Job selection process for job-seeking candidates. Job selection for a fresh graduate is a very complicated task. Here used the real-time dataset of job selection problems which are collected from individual job holder's person and LinkedIn. Parameters are analyzed to find parameters for job selection or job satisfaction in this paper. In the proposed model, all the experiment is based on a real dataset, and jupyter notebook software is used for experiments. This research work used a multi-layer perceptron neural networks algorithm and six more machine learning algorithms Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, K-Nearest-Neighbor, Random Forest Tree, and decision tree. 66 parameters are used as an input in the input layer, one hidden layer with 20 neurons is used for training the multi-layer perceptron neural network algorithm. Each algorithm's outcome is used to determine the better classifiers among them. Comparisons of these selected algorithms are based on performance factors.

In the testing state, the Logistic Regression Algorithm gives a 80% precision score,75% accuracy and AUC value of 0.69. Gaussian Naive Bayes gives an 80% precision score, 79% accuracy, and AUC value of 0.79. K-Nearest-Neighbor gives 88% precision score, 87.5% accuracy and AUC value of 0.88, Support Vector Machine gives 84% precision score, 79% accuracy, and AUC value of 0.72, Random Forest Tree gives 92% precision score, 92% accuracy and AUC value 0.91, Multilayer Perceptron Algorithm gives 93% precision score, 91.6% accuracy and AUC value 0.89, Decision Tree gives 85% precision score, 83% accuracy and AUC value 0.84.

This research paper's goal is to provide better performance and prediction.

5.3 FUTURE RESEARCH SCOPE

There is a possibility to generate a new algorithm or a hybrid one which will provide better accuracy than the existing algorithms. Data from several districts of Bangladesh can be used for wide research in this topic in future. Domain based research can be conduct in near future.

REFERENCES

[1] A. A. Biswas, A. Majumder, M. Jueal Mia, R. Basri, and M. Sabab Zulfiker, "Career prediction with analysis of influential factors using data mining in the context of bangladesh," in *Advances in Intelligent Systems and Computing*, 2021, vol. 1309, pp. 441–451. doi: 10.1007/978-981-33-4673-4_35.

[2] Veneri and Carolyn M., "Here Today, Jobs of Tomorrow: Opportunities in Information Technology," 1998.

[3] P. Bolander and J. Sandberg, "How Employee Selection Decisions are Made in Practice," *Organization Studies*, vol. 34, no. 3, pp. 285–311, Mar. 2013, doi: 10.1177/0170840612464757.

[4] N. Magesh *et al.*, "Evaluating The Performance Of An Employee Using Decision Tree Algorithm." [Online]. Available: www.ijert.org

[5] L. Matos Pombo, "Landing on the right job: a Machine Learning approach to match candidates with jobs applying semantic embeddings."

[6] John F. Magee, "11. Decision Trees for Decision Making," *Harvard Business Review*, 1964.

[7] H. Patel, J. Sanghavi, S. Shah, S. Thacker, and S. Mishra, "CAREER GUIDANCE SYSTEM USING MACHINE LEARNING," 2021. [Online]. Available: www.ijcrt.org

[8] P. Kiselev, B. Kiselev, V. Matsuta, A. Feshchenko, I. Bogdanovskaya, and A. Kosheleva, "Career guidance based on machine learning: Social networks in professional identity construction," in *Procedia Computer Science*, 2020, vol. 169, pp. 158–163. doi: 10.1016/j.procs.2020.02.128.

[9] N. Vidya Shreeram and Dr. A. Muthu kumar avel, "Student Career Prediction Using Machine Learning Approaches," Jun. 2021. doi: 10.4108/eai.7-6-2021.2308642.

[10] H. Al-Dossari, Z. Al-Qahtani, F. A. Nughaymish, M. Alkahlifah, and A. Alqahtani, "CareerRec: A Machine Learning Approach to Career Path Choice for Information Technology Graduates," 2020. [Online]. Available: www.etasr.com

[11] K. S. Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, and J. Priyanka, "Student Career Prediction Using Advanced Machine Learning Techniques," 2018.

[12] Y. Lou, R. Ren, and Y. Zhao, “A Machine Learning Approach for Future Career Planning.”

[13] A. Nigam, A. Roy, H. Singh, and H. Waila, “Job Recommendation through Progression of Job Selection,” 2019.

[14] U. K. Sah, M. A. Singh, and M. Tech Scholar, “Student Career Prediction Using Machine Learning,” *ijsdr.org International Journal of Scientific Development and Research*, vol. 7, p. 343, 2022, [Online]. Available: www.ijsdr.org

[15] O. Awujoola, Philip O Odion, Martins E Irhebhude, and Halima Aminu, “Performance Evaluation of Machine Learning Predictive Analytical Model for Determining the Job Applicants Employment Status,” *Malaysian Journal of Applied Sciences*, vol. 6, no. 1, pp. 67–79, Apr. 2021, doi: 10.37231/myjas.2021.6.1.276.