A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques

A Report *submitted*
in partial fulfillment of the requirements for the award of degree of

Bachelor of Technology

in

Information Technology

by

JANGITI DWARAKAMAI	18J41A1224
KESIREDDY SAMARA SIMHA REDDY	18J41A1227
MOTHA RAHUL	18J41A1233
CH. YOGESH CHOWDARY	19J45A1201

Under the guidance of
Mr.M.Srinivasa Reddy
Assistant Professor

Pursued in

DEPARTMENT OF INFORMATION TECHNOLOGY

MALLA REDDY ENGINEERING COLLEGE
UGC Autonomous Institution, Approved by AICTE, New Delhi &Affiliated to JNTUH, Hyderabad). Accredited 3rd time by NAAC with ‘A++’ Grade,
Maisammaguda(H), Medchal-Malkajgiri District, Secunderabad, Telangana–500100,
www.mrec.ac.in

JUNE-2022

MALLA REDDY ENGINEERING COLLEGE

UGC Autonomous Institution, Approved by AICTE, New Delhi &Affiliated to
JNTUH, Hyderabad). Accredited 3rd time by NAAC with ‘A++’ Grade,
Maisammaguda (H), Medchal-Malkajgiri District, Secunderabad, Telangana State–
500100, www.mrec.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project work entitled “**A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques**” is submitted by **JANGITI DWARAKAMAI (18J41A1224), KESIREDDY SAMARA SIMHA REDDY (18J41A1227), MOTHA RAHUL (18J41A1233), CH. YOGESH CHOWDARY (19J45A1201) to Malla Reddy Engineering College (Autonomous)**, affiliated to **JNTUH, Hyderabad** in partial fulfillment for the award of the degree of **B.Tech** in (**Information Technology**) is a bonafide record of project work carried out by him/her under my guidance. The contents of this report, in full or in parts, have not been submitted to any other Institution or Universityfor the award of any degree or diploma.

Mr.M.Srinivasa Reddy
Project Guide

External **Counter signature of HOD with seal**

DECLARATION

We hereby declaring that this project work dissertation titled "**A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques**" is original and bonafide work of our own in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology** at **Malla Reddy Engineering College (Autonomous)**, affiliated to **JNTUH, Hyderabad** under the guidance of **Mr.M.Srinivasa Reddy Assistant Professor** and has not been copied from any earlier reports.

Roll Number	Name	Signature
18J41A1224	JANGITI DWARAKAMAI	
18J41A1227	KESIREDDY SAMARA SIMHA REDDY	
18J41A1233	MOTHA RAHUL	
19J45A1201	CH. YOGESH CHOWDARY	

ACKNOWLEDGEMENT

We are extremely thankful to our beloved Chairman and Founder of Malla Reddy Group of Institutions, **Sri. Ch. Malla Reddy,** for providing necessary infrastructure facilities for completing project work successfully.

We express our sincere thanks to our **Principal, Dr. A. Ramaswami Reddy**, who took keen interest and encouraged us in every effort during the project work.

We express our heartfelt thanks to **Dr. Deena Babu Mandru Professor and Head of the Department**, for his kind attention and valuable guidance throughout the project work.

We are extremely thankful to our project guide **Mr.M.Srinivasa Reddy**, **Assistant Professor,** for his constant guidance and support to complete the project work.

We also thank all the teaching and non-teaching staff of department for their cooperation during the project work.

18J41A1224	**JANGITI DWARAKAMAI**
18J41A1227	**KESIREDDY SAMARA SIMHA REDDY**
18J41A1233	**MOTHA RAHUL**
19J45A1201	**Ch YOGESH CHOWDARY**

ABSTRACT

Electronic textual documents are among the most popular teaching content accessible through e-learning platforms. Teachers or learners with different levels of knowledge can access the platform and highlight portions of textual content which are deemed as particularly relevant. The highlighted documents can be shared with the learning community in support of oral lessons or individual learning. However, highlights are often incomplete or unsuitable for learners with different levels of knowledge. This paper addresses the problem of predicting new highlights of partly highlighted electronic learning documents.

In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

TABLE OF CONTENTS

DESCRIPTION **PAGE NUMBERS**

LIST OF FIGURES

ABBREVATIONS/NOMENCLATURE

EMSCAD	Near Field Communication
CNN	convolutional neural network
HTML	Hyper Text Mark-up Language
LAN	Local Area Network
CSS	Cascading Style Sheets
SQL	Simple Query Language
SVM	Support vector machines
KNN	K-Nearest Neighbors
MLP	multilayer perceptron
DNN	Deep Neural Network
NLP	Natural Language Processing

CHAPTER 1

INTRODUCTION

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lack in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information. Thus, the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's belief and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. . Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

A. Fake Job Posting: Job Scam

Online job advertisements which are fake and mostly willing to steal personal and professional information of job seekers instead of giving right jobs to them is known as job scam. Sometimes fraudulent people try to gather money illegally from job seekers. A recent survey by Action Fraud from UK has shown that more than 67% people are at great risk who look for jobs through online advertisements but unaware of fake job posts or job scam [2]. In UK, almost 700000 job seekers complained to lose over $500000 being a victim of job scam. The report showed almost 300% increase over the last two years in UK [2]. Students, fresh graduates are being mostly targeted by the frauds as they usually try to get a secured job for which they are willing to pay extra money. Cybercrime avoidance or protection techniques fail to decrease this offence since frauds change their way of job scam very frequently.

B. Common types of Job Scam

Fraudsters who want to gain other people’s personal information like insurance details, bank details, income tax details, date of birth, national id create fake job advertisements. Advance fee scams occur when frauds ask for money showing reasons like admin charges, information security checking cost, management cost etc. Sometimes fraudsters act them- selves as employers and ask people about passport details, bank statements, driving license etc. as pre-employment check.

Illegal money mulling scams occur when they convince students to pay money into their accounts and then transfer it back [2]. This ’cash in hand’ technique causes to work cash in hand without paying any tax. Scammers usually create fake company websites, clone bank websites, clone official looking documents etc. to trap job seekers. Most of the job scammers try to trap people through email rather than face to face communication. They usually target social media like LinkedIn to prove themselves as recruitment agencies or headhunters. They usually try to represent their company profile or websites to the job seeker as realistic as possible. Whatever the type of job scam they use, the always target the job seeker to fall in their trap, collecting information and making benefit either earning money or any other things

CHAPTER 2
LITERATURE SURVEY

RELATED WORK

According to several studies, Review spam detection, Email Spam detection, Fake news detection have drawn special attention in the domain of Online Fraud Detection.

A. Review Spam Detection

People often post their reviews online forum regarding the products they purchase. It may guide other purchaser while choosing their products. In this context, spammers can manipulate reviews for gaining profit and hence it is required to develop techniques that detects these spam reviews. This can be implemented by extracting features from the reviews by extracting features using Natural Language Processing (NLP). Next, machine learning techniques are applied on these features. Lexicon based approaches may be one alternative to machine learning techniques that uses dictionary or corpus to eliminate spam reviews[11].

B. Email Spam Detection

Unwanted bulk mails, belong to the category of spam emails, often arrive to user mailbox. This may lead to unavoidable storage crisis as well as bandwidth consumption. To eradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporate spam filters using Neural Networks. While addressing the problem of email spam detection, content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, adaptive spam filtering approaches are taken into consideration [7].

C. Fake News Detection

Fake news in social media characterizes malicious user accounts, echo chamber effects. The fundamental study of fake news detection relies on three perspectives- how fake news is written, how fake news spreads, how a user is related to fake news. Features related to news content and social context are extracted and a machine learning model are imposed to recognize fake new.

CHAPTER 3
SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Many researches occurred to predict if a job post is real or fake. A good number of research works are to check online fraud job advertiser. Vidros [1] et al. identified job scammers as fake online job advertiser. They found statistics about many real and renowned companies and enterprises who produced fake job advertisements or vacancy posts with ill-motive. They experimented on EMSCAD dataset using several classification algorithms like naive bayes classifier, random forest classifier, Zero R, One R etc. Random Forest Classifier showed the best performance on the dataset with 89.5% classification accuracy. They found logistic regression performing very poor on the dataset. One R classifier performed well when they balanced the dataset and experimented on that. They tried in their work to find out the problems in ORF model (Online Recruitment Fraud) and to solve those problems using various dominant classifiers.

Alghamdi [2] et al. proposed a model to detect fraud exposure in an online recruitment system. They experimented on EMSCAD dataset using machine learning algorithm. They worked on this dataset in three steps- data pre-processing, feature selection and fraud detection using classifier. In the preprocessing step, they removed noise and html tags from the data so that the general text pattern remained preserved. They applied feature selection technique to reduce the number of attributes effectively and efficiently. Support Vector Machine was used for feature selection and ensemble classifier using random forest was used to detect fake job posts from the test data. Random forest classifier seemed a tree structured classifier which worked as ensemble classifier with the help of majority voting technique. This classifier showed 97.4% classification accuracy to detect fake job posts.

Huynh [3] et al. proposed to use different deep neural network models like Text CNN, Bi-GRU-LSTM CNN and Bi- GRU CNN which are pre-trained with text dataset. They worked on classifying IT job dataset. They trained IT job dataset on Text CNN model consisting of convolution layer, pooling layer and fully connected layer. This model trained data through convolution and pooling layers. Then the trained weights were flattened and passed to the fully connected layer. This model used SoftMax function for classification technique. They also used ensemble classifier (Bi-GRU CNN, Bi-GRULSTM CNN) using majority voting technique to increase classification accuracy. They found 66% classification accuracy using Text CNN and

70% accuracy for Bi-GRU- LSTM CNN individually. This classification task performed best with ensemble classifier having an accuracy of 72.4%.

Zhang [4] et al. proposed an automatic fake detector model to distinguish between true and fake news (including articles, creators, subjects) using text processing. They had used a custom dataset of news or articles posted by PolitiFact website twitter account. This dataset was used to train the proposed GDU diffusive unit model. Receiving input from multiple sources simultaneously, this trained model performed well as an automatic fake detector model.

DISADVANTAGES OF EXISTING SYSTEM:

- The system is implemented by Conventional Machine Learning.
- The system doesn't implement for analyzing large data sets.

3.2 PROPOSED SYSTEM

The system has used EMSCAD to detect fake job post. This dataset contains 18000 samples and each row of the data has 18 attributes including the class label. The attributes are job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, telecommunication, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent (class label). Among these 18 attributes, we have used only 7 attributes which are converted into categorical attribute. Telecommuting, has_company_logo, has_questions, employment_type, required experience, required_education and fraudulent are changed into categorical value from text value. For example, "employment_type" values are replaced like this- 0 for "none", 1 for 'full-time", 2 for "part-time" and 3 for "others", 4 for "contract' and 5 for "temporary". The main goal to convert these attributes into categorical form is to classify fraudulent job advertisements without doing any text processing and natural language processing. In this work, we have used only those categorical attributes.

ADVANTAGES OF PROPOSED SYSTEM:

- The proposed has been implemented EMSCAD technique which is very accurate and fast.
- The system is very effective due to accurate detection of Fake job posts which creates inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time.

3.3 MODULE IMPLEMENTATION

In this we have implemented three modules. They are:

I. Service Provider

II. View and Authorize Users

III. Remote User

Different Modules used:

3.3.1 Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Predict Job Post Type Details, Find Job Post Type Prediction Ratio, Download Trained Data Sets, View Job Post Type Prediction Ratio Results, View All Remote Users.

3.3.2 View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorize the users.

3.3.3 Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, POST JOB POST DATA SETS, PREDICT JOB POST PREDICTION, VIEW YOUR PROFILE.

In this module, the admin has to login by using valid user name and password. After login successful he can do some operations, such as View all users and authorize, View E – Commerce Website users and authorize, add bank with its details such as ,View Credit card request and Process with ,generate card transaction Bill for a period, Show all Bank Fraud detection in Random-tree-based random forest, View all Normal and Abnormal CC Users for cash limit and give link to show in chart, View all

3.4 PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts fromthe thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine, address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, i.e., preliminary investigation begins.

The activity has three parts:

I. Request Clarification

II. Feasibility Study

III. Request Approval

3.4.1 REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network (LAN). In today's busy schedule man need everything should be provided in a readymade manner. So, taking into consideration of the vastly use of the net in day-to-day life, the corresponding development ofthe portal came into existence.

3.4.2 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3.4.3 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. Aftera project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

CHAPTER 4
SYSTEM REQUIREMENT SPECIFICATIONS

Software Environment

4.1.1 PYTHON

Python is a **high-level, interpreted**, **interactive** and **object-oriented scripting language**. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted:** Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive:** You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented:** Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language:** Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

4.1.2 History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

4.1.3 Python Features

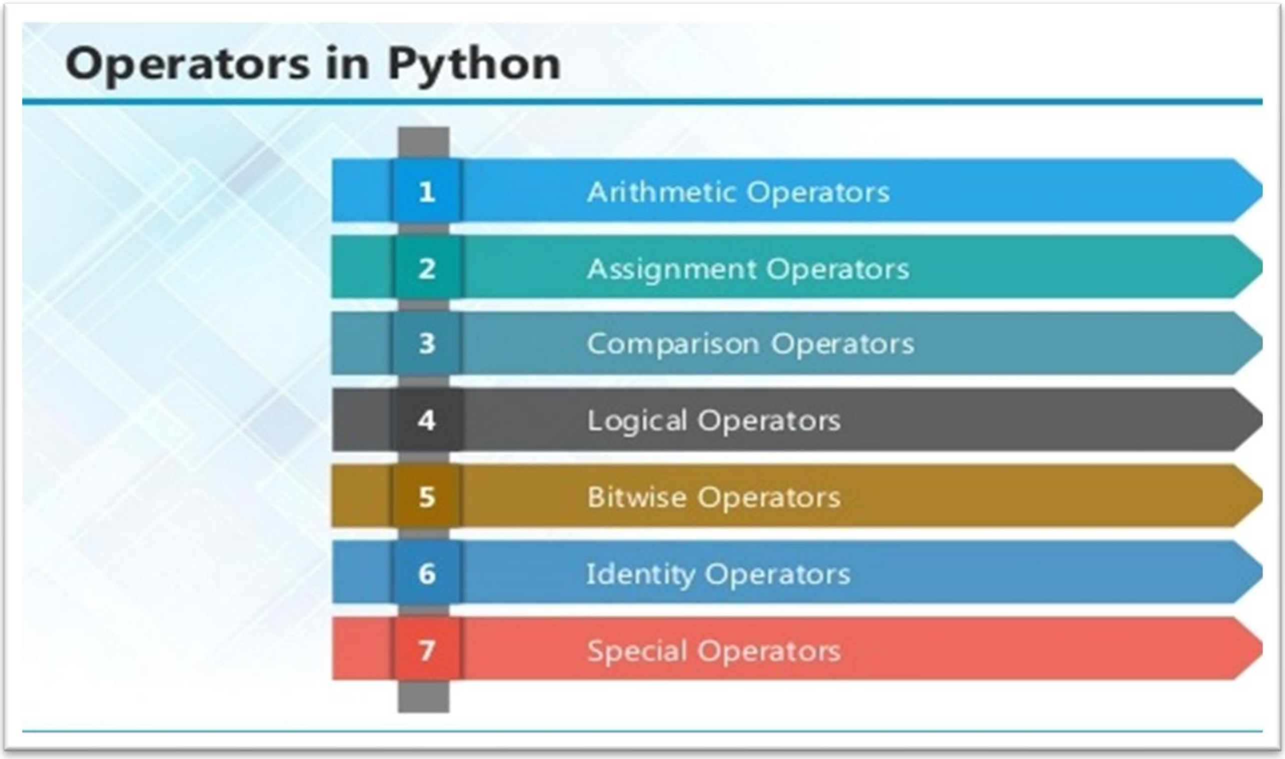
Python's features include:

- **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.
- **A broad standard library:** Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode:** Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable:** You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases:** Python provides interfaces to all major commercial databases.
- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

Python has a big list of good features:

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.

- IT supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.



Operators in Python

1 Arithmetic Operators

2 Assignment Operators

3 Comparison Operators

4 Logical Operators

5 Bitwise Operators

6 Identity Operators

7 Special Operators

ARITHMETIC OPERATORS

Operator	Description	Example
+ Addition	Adds values on either side of the operator.	a + b = 30
- Subtraction	Subtracts right hand operand from left hand operand.	a – b = -10

* Multiplication	Multiplies values on either side of the operator	a * b = 200
/ Division	Divides left hand operand by right hand operand	b / a = 2
% Modulus	Divides left hand operand by right hand operand and returns remainder	b % a = 0
** Exponent	Performs exponential (power) calculation on operators	a**b =10 to the power 20
//	Floor Division - The division of operands where the result is the quotient in which the digits after the decimal point are removed. But if one of the operands is negative, the result is floored, i.e., rounded away from zero (towards negative infinity):	9//2 = 4 and 9.0//2.0 = 4.0, -11//3 = -4, -11.0//3 = -4.0

ASSIGNMENT OPERATOR

Operator	Description	Example
=	Assigns values from right side operands to	c = a + b assigns

	left side operand	value of a + b into c
+= Add AND	It adds right operand to the left operand and assign the result to left operand	c += a is equivalent to c = c + a
-= Subtract AND	It subtracts right operand from the left operand and assign the result to left operand	c -= a is equivalent to c = c - a
*= Multiply AND	It multiplies right operand with the left operand and assign the result to left operand	c *= a is equivalent to c = c * a
/= Divide AND	It divides left operand with the right operand and assign the result to left operand	c /= a is equivalent to c = c / ac /= a is equivalent to c = c / a

%= Modulus AND	It takes modulus using two operands and assign the result to left operand	c %= a is equivalent to c = c % a
**= Exponent AND	Performs exponential (power) calculation on operators and assign value to the left operand	c **= a is equivalent to c = c ** a
//= Floor Division	It performs floor division on operators and assign value to the left operand	c //= a is equivalent to c = c // a

IDENTITY OPERATOR

Operator	Description	Example
is	Evaluates to true if the variables on either side of the operator point to the same object and false otherwise.	x is y, here **is** results in 1 if id(x) equals id(y).

is not	Evaluates to false if the variables on either side of the operator point to the same object and true otherwise.	x is not y, here **is not** results in 1 if id(x) is not equal to id(y

COMPARISON OPERATOR

Operator	Description	Example
& Binary AND	Operator copies a bit to the result if it exists in both operands	(a & b) (means 0000 1100)
| Binary OR	It copies a bit if it exists in either operand.	(a | b) = 61 (means 0011 1101)
^ Binary XOR	It copies the bit if it is set in one operand but not both.	(a ^ b) = 49 (means 0011 0001)
~ Binary Ones Complement	It is unary and has the effect of 'flipping' bits.	(~a) = -61 (means 1100 0011 in 2's complement form due to a

		signed binary number.
<< Binary Left Shift	The left operands value is moved left by the number of bits specified by the right operand.	a << 2 = 240 (means 1111 0000)
>> Binary Right Shift	The left operands value is moved right by the number of bits specified by the right operand.	a >> 2 = 15 (means 0000 1111)

LOGICAL OPERATOR

Operator	Description	Example
and Logical AND	If both the operands are true then condition becomes true.	(a and b) is true.
or Logical OR	If any of the two operands are non-zero then condition becomes true.	(a or b) is true.
not Logical	Used to reverse the logical	Not(a and b) is

NOT	state of its operand.	false.

Membership Operators

Operator	Description	Example
in	Evaluates to true if it finds a variable in the specified sequence and false otherwise.	x in y, here in results in a 1 if x is a member of sequence y.
not in	Evaluates to true if it does not finds a variable in the specified sequence and false otherwise.	x not in y, here not in results in a 1 if x is not a member of sequence y.

Python Operators Precedence

Operator	Description
**	Exponentiation (raise to the power)
~ + -	Complement, unary plus and minus (method names for

	the last two are +@ and -@)
* / % //	Multiply, divide, modulo and floor division
+ -	Addition and subtraction
>> <<	Right and left bitwise shift
&	Bitwise 'AND'
^ |	Bitwise exclusive `OR' and regular `OR'
<= < > >=	Comparison operators
<> == !=	Equality operators
= %= /= //= -= += *= **=	Assignment operators
is is not	Identity operators
in not in	Membership operators
not or and	Logical operators

4.2 SYSTEM REQUIREMENTS

Hardware System Configuration:-

Processor	-	Pentium –IV
RAM	-	4 GB (min)
Hard Disk	-	20 GB
Key Board	-	Standard Windows Keyboard
Mouse	-	Two or Three Button Mouse
Monitor	-	SVGA

SOFTWARE REQUIREMENTS:

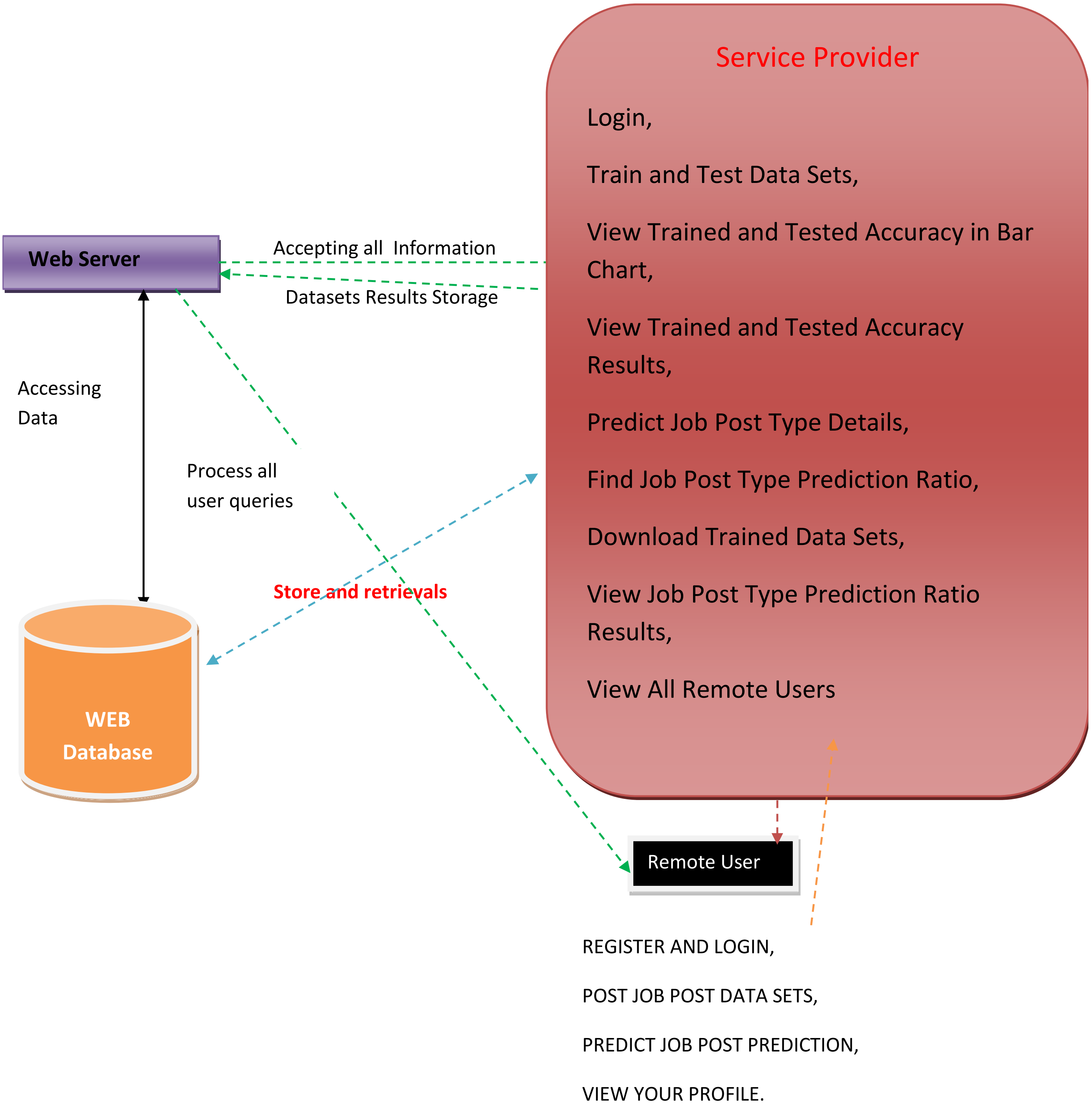
Operating system	:	Windows 7 Ultimate.
Coding Language	:	Python.
Front-End	:	Python.
Back-End	:	Django-ORM
Designing	:	Html, css, javascript.
Data Base	:	MySQL (WAMP Server).

CHAPTER 5

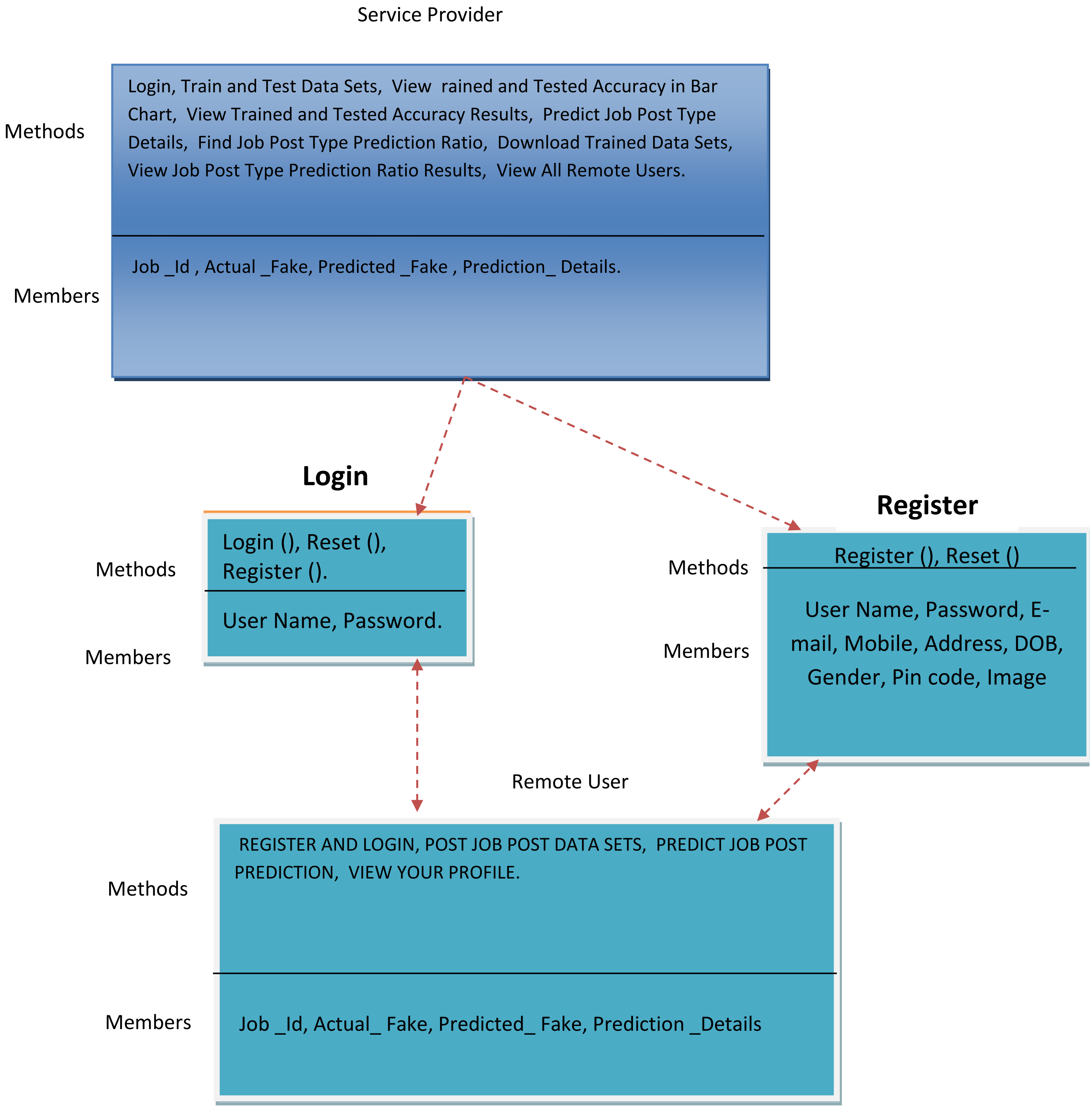
SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

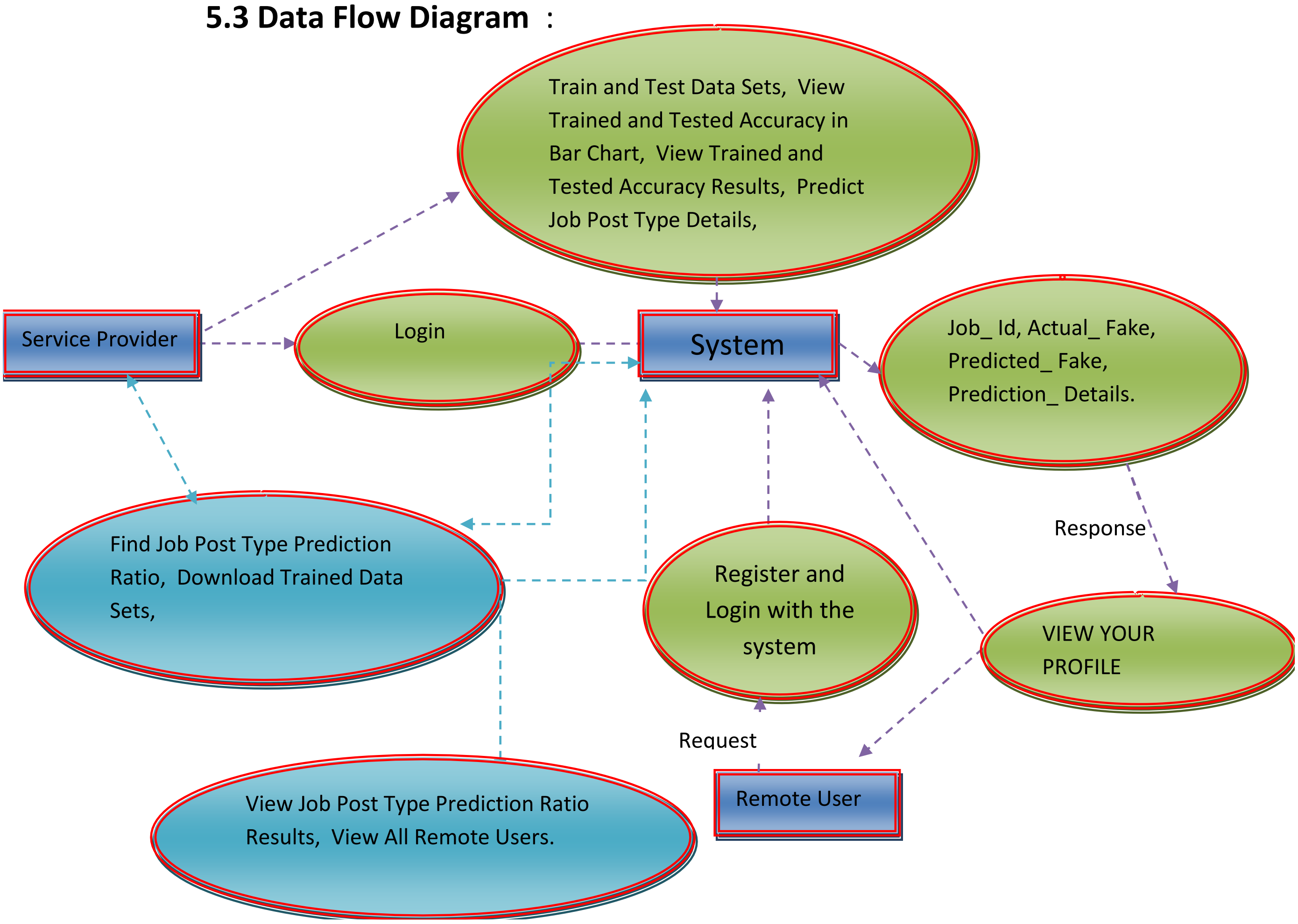
Architecture Diagram


Service Provider
Login,
Train and Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
POST JOB POST DATA SETS,
PREDICT JOB POST PREDICTION,
VIEW YOUR PROFILE.

5.2 Class Diagram :

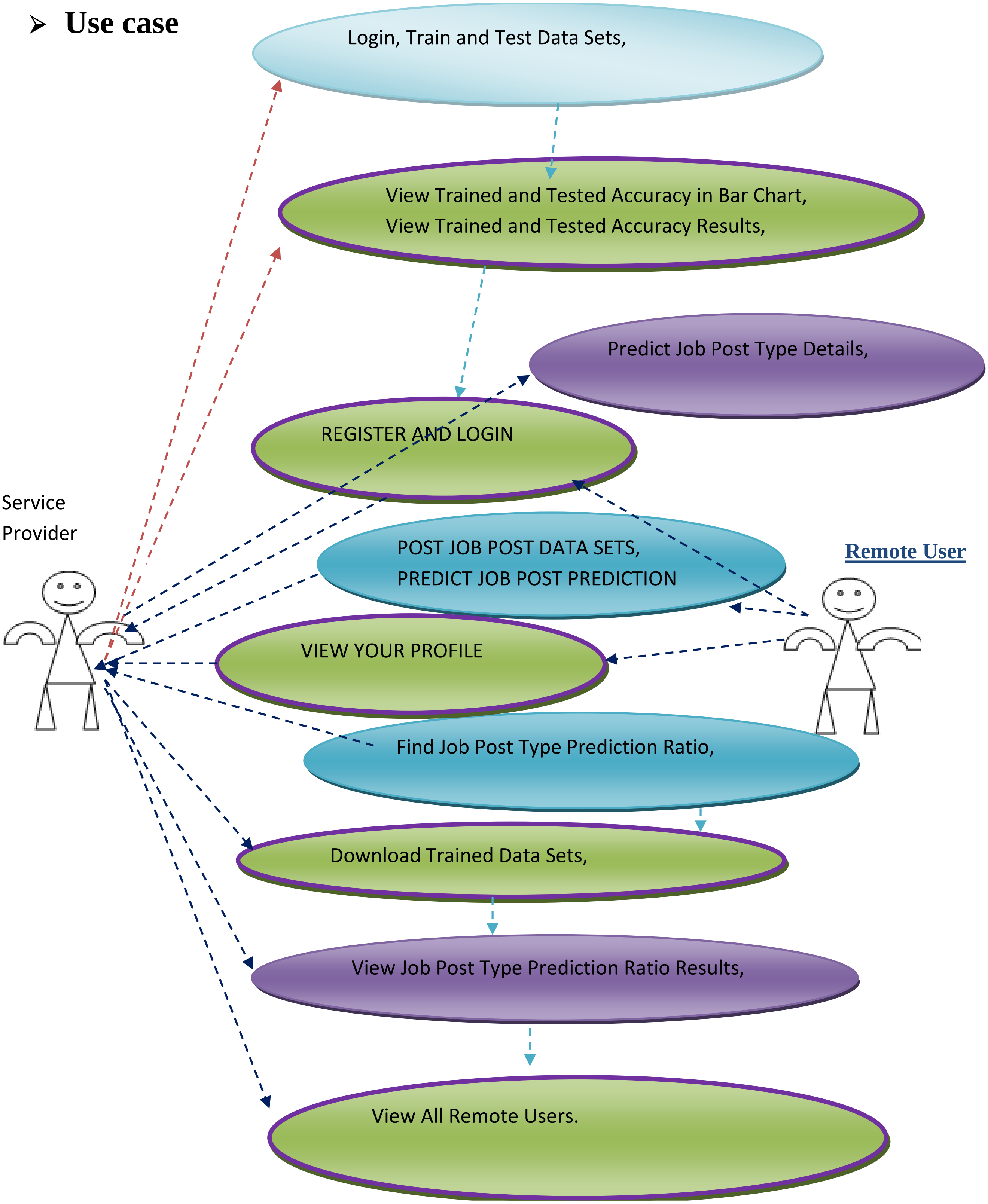

Service Provider
Methods
Login, Train and Test Data Sets, View rained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Predict Job Post Type Details, Find Job Post Type Prediction Ratio, Download Trained Data Sets, View Job Post Type Prediction Ratio Results, View All Remote Users.
Members
Job _Id , Actual _Fake, Predicted _Fake , Prediction_ Details.
Login
Methods
Login (), Reset (), Register ().
Members
User Name, Password.
Register
Methods
Register (), Reset ()
Members
User Name, Password, E-mail, Mobile, Address, DOB, Gender, Pin code, Image
Remote User
Methods
REGISTER AND LOGIN, POST JOB POST DATA SETS, PREDICT JOB POST PREDICTION, VIEW YOUR PROFILE.
Members
Job _Id, Actual_ Fake, Predicted_ Fake, Prediction _Details

5.3 Data Flow Diagram :


Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Predict Job Post Type Details,
Service Provider
Login
System
Job_ Id, Actual_ Fake, Predicted_ Fake, Prediction_ Details.
Response
Find Job Post Type Prediction Ratio, Download Trained Data Sets,
Register and Login with the system
VIEW YOUR PROFILE
Request
Remote User
View Job Post Type Prediction Ratio Results, View All Remote Users.

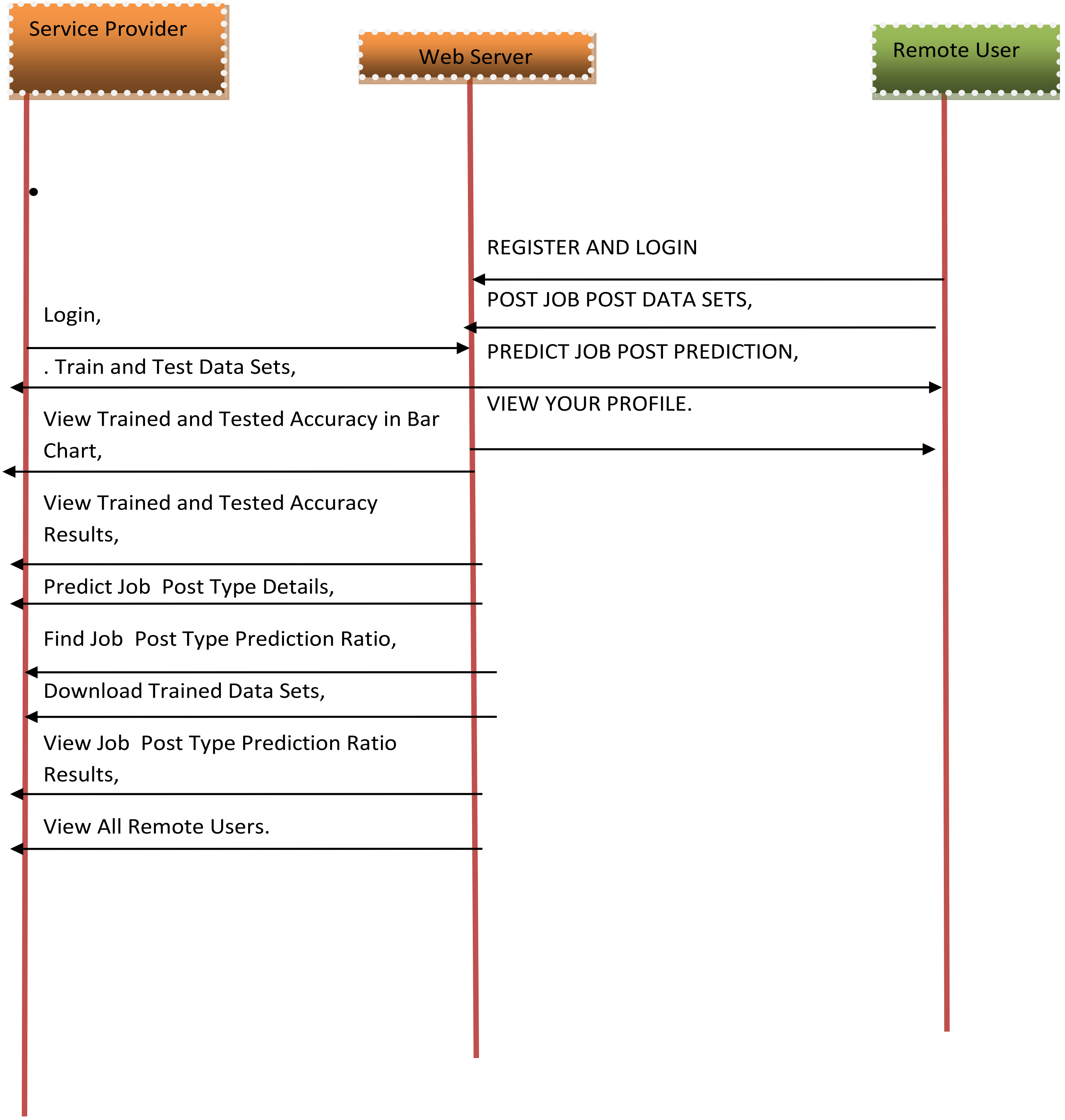
5.4 USE CASE DIAGRAM

- **Use case**

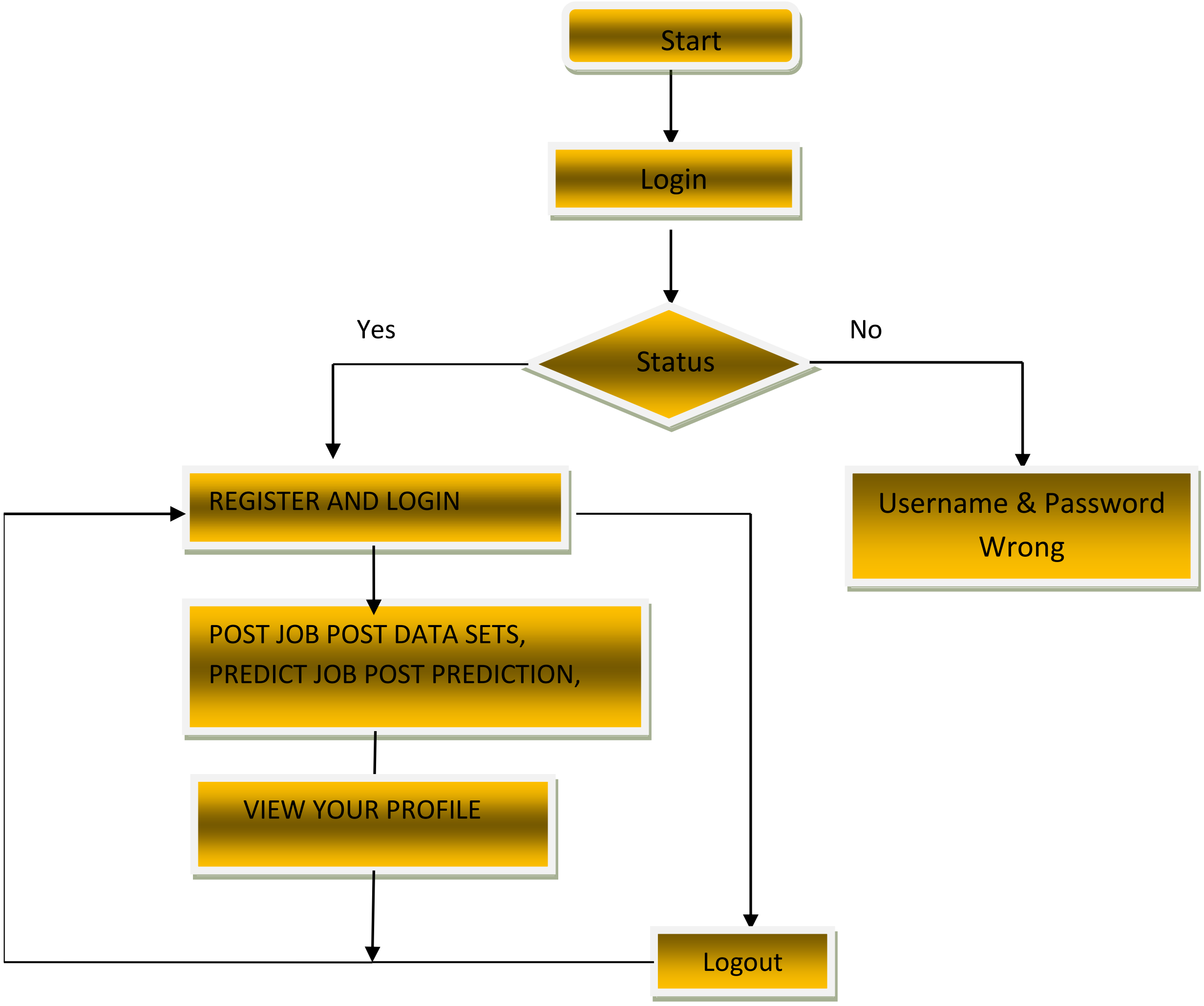

Login, Train and Test Data Sets,
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
REGISTER AND LOGIN
Service
Provider
POST JOB POST DATA SETS,
PREDICT JOB POST PREDICTION
Remote User
VIEW YOUR PROFILE
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users.

5.5 Sequence Diagram

Sequence Diagram


Service Provider
Web Server
Remote User
REGISTER AND LOGIN
POST JOB POST DATA SETS,
Login,
PREDICT JOB POST PREDICTION,
. Train and Test Data Sets,
VIEW YOUR PROFILE.
View Trained and Tested Accuracy in Bar Chart,
View Trained and Tested Accuracy Results,
Predict Job Post Type Details,
Find Job Post Type Prediction Ratio,
Download Trained Data Sets,
View Job Post Type Prediction Ratio Results,
View All Remote Users.

5.6 Flow Chart : Remote User


Start
Login
Yes
Status
No
REGISTER AND LOGIN
Username & Password Wrong
POST JOB POST DATA SETS, PREDICT JOB POST PREDICTION,
VIEW YOUR PROFILE
Logout

5.7 Input and Output Design

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So, inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rest with the administrator only.

Chapter 6
Implementation and Algorithm

ALGORITHM:

Decision tree classifiers

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision-making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2, Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi, we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

Simple, but a very powerful classification algorithm
Classifies based on a similarity measure
Non-parametric
Lazy learning
Does not "learn" until the test example is given
Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

Training dataset consists of k-closest examples in feature space

Feature space means, space with categorization variables (non-metric variables)
Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naïve Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .
Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they do not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "Blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms (GAs) or perceptrons, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement

CHAPTER 7

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and doesnot fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

7.1 TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.
Functional testing is centered on the following items:

Valid Input	: identified classes of valid input must be accepted.
Invalid Input	: identified classes of invalid input must be rejected.
Functions	: identified functions must be exercised.
Output	: identified classes of application outputs must be exercised.
Systems/Procedures:	interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

7.2 TESTING METHODOLOGIES

The following are the Testing Methodologies:

- Unit Testing.
- Integration Testing.
- User Acceptance Testing.
- Output Testing.
- Validation Testing.

Unit Testing-

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing. During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All-important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing-

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

The following are the types of Integration Testing:

1) Top-Down Integration -This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner. In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

2) Bottom-up Integration -This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the

need for stubs is eliminated.

User Acceptance Testing

User Acceptance of a system is the key factor for the success of any system. The system under consideration is tested for user acceptance by constantly keeping in touch with the prospective system users at the time of developing and making changes wherever required. The system developed provides a friendly user interface that can easily be understood even by a person who is new to the system.

Output Testing

After performing the validation testing, the next step is output testing of the proposed system, since no system could be useful if it does not produce the required output in the specified format. Asking the users about the format required by them tests the outputs generated or displayed by the system under consideration. Hence the output format is considered in 2 ways – one is on screen and another in printed format.

Validation Checking

Validation checks are performed on the following fields.

Text Field:

The text field can contain only the number of characters lesser than or equal to its size. The text fields are alphanumeric in some tables and alphabetic in other tables. Incorrect entry always flashes and error message.

Numeric Field:

The numeric field can contain only numbers from 0 to 9. An entry of any character flashes an error message. The individual modules are checked for accuracy and what it has to perform. Each module is subjected to test run along with sample data. The individually tested modules are integrated into a single system. Testing involves executing the real data information is used in the program the existence of any program defect is inferred from the output. The testing should be planned so that all the requirements are individually tested.

A successful test is one that gives out the defects for the inappropriate data and produces and output revealing the errors in the system.

Preparation of Test Data

Taking various kinds of test data does the above testing. Preparation of test data plays a vital role in the system testing. After preparing the test data the system under study is tested using that test data. While testing the system by using test data errors are again uncovered and corrected by using above testing steps and corrections are also noted for future use.

Using Live Test Data:

Live test data are those that are actually extracted from organization files. After a system is partially constructed, programmers or analysts often ask users to key in a set of data from their normal activities. Then, the systems person uses this data as a way to partially test the system. In other instances, programmers or analysts extract a set of live data from the files and have them entered themselves.

It is difficult to obtain live data in sufficient amounts to conduct extensive testing. And, although it is realistic data that will show how the system will perform for the typical processing requirement, assuming that the live data entered are in fact typical, such data generally will not test all combinations or formats that can enter the system. This bias toward typical values then does not provide a true system test and in fact ignores the cases most likely to cause system failure.

Using Artificial Test Data:

Artificial test data are created solely for test purposes, since they can be generated to test all combinations of formats and values. In other words, the artificial data, which can quickly be prepared by a data generating utility program in the information systems department, make possible the testing of all login and control paths through the program.

The most effective test programs use artificial test data generated by persons other than those who wrote the programs. Often, an independent team of testers formulates a testing plan, using the systems specifications.

The package “Virtual Private Network” has satisfied all the requirements specified as per software requirement specification and was accepted.

7.3 USER TRAINING

Whenever a new system is developed, user training is required to educate them about the working of the system so that it can be put to efficient use by those for whom the system has been primarily designed. For this purpose, the normal working of the project was demonstrated to the prospective users. Its working is easily understandable and since the expected users are people who have good knowledge of computers, the use of this system is very easy.

7.4 MAINTAINENCE

This covers a wide range of activities including correcting code and design errors. To reduce the need for maintenance in the long run, we have more accurately defined the user’s requirements during the process of system development. Depending on the requirements, this system has been developed to satisfy the needs to the largest possible extent. With development in technology, it may be possible to add many more features based on the requirements in future. The coding and designing is simple and easy to understand which will make maintenance easier.

7.5 TESTING STRATEGY :

A strategy for system testing integrates system test cases and design techniques into a well-planned series of steps that results in the successful construction of software. The testing strategy must co-operate test planning, test case design, test execution, and the resultant data collection and evaluation .A strategy for software testing must accommodate low-level tests that are necessary to verify that a small source code segment has been correctly implemented as well as high level tests that validate major system functions against user requirements.
Software testing is a critical element of software quality assurance and represents the ultimate review of specification design and coding. Testing represents an interesting anomaly for the software. Thus, a series of testing are performed for the proposed system before the system is ready for user acceptance testing.

SYSTEM TESTING:

Software once validated must be combined with other system elements (e.g., Hardware, people, database). System testing verifies that all the elements are proper and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

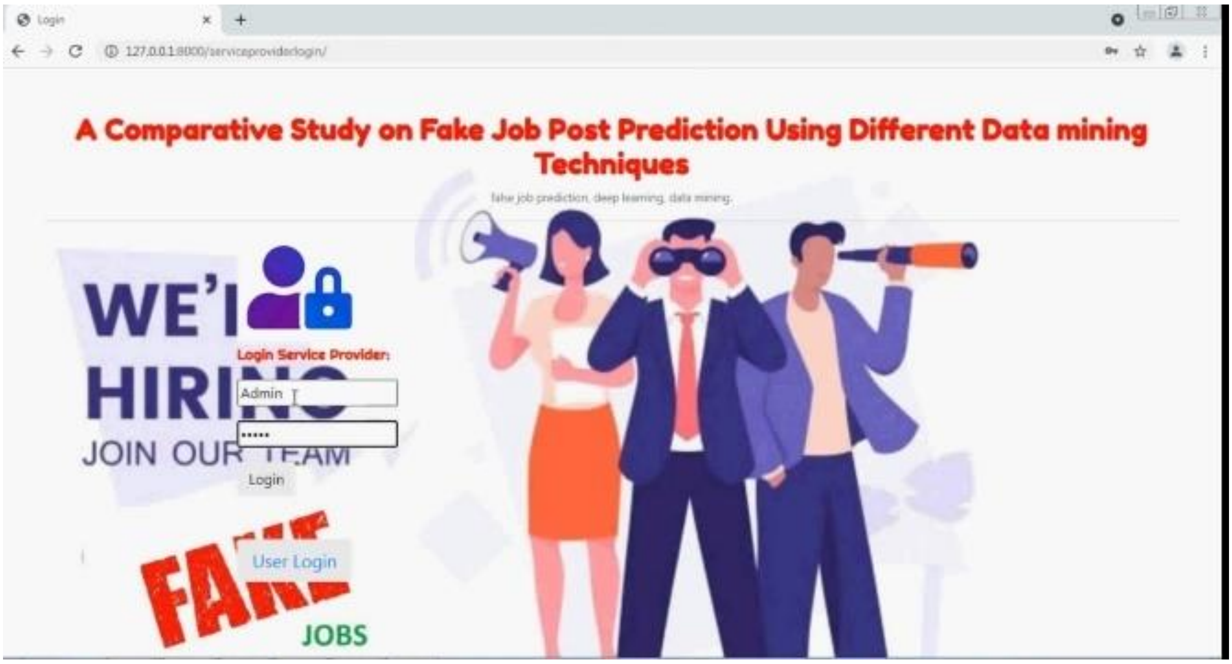
UNIT TESTING:

In unit testing different are modules are tested against the specifications produced during the design for the modules. Unit testing is essential for verification of the code produced during the coding phase, and hence the goals to test the internal logic of the modules. Using the detailed design description as a guide, important Conrail paths are tested to uncover errors within the boundary of the modules. This testing is carried out during the programming stage itself. In this type of testing step, each module was found to be working satisfactorily as regards to the expected output from the module. In Due Course, latest technology advancements will be taken into consideration. As part of technical build-up many components of the networking system will be generic in nature so that future projects can either use or interact with this. The future holds a lot to offer to the development and refinement of this project.

CHAPTER 8

RESULT

To run this application, go to any browser and type http://127.0.0.1.8000/



Login

127.0.0.1:8000/serviceproviderlogin/

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

fake job prediction, deep learning, data mining.

Login Service Provider:

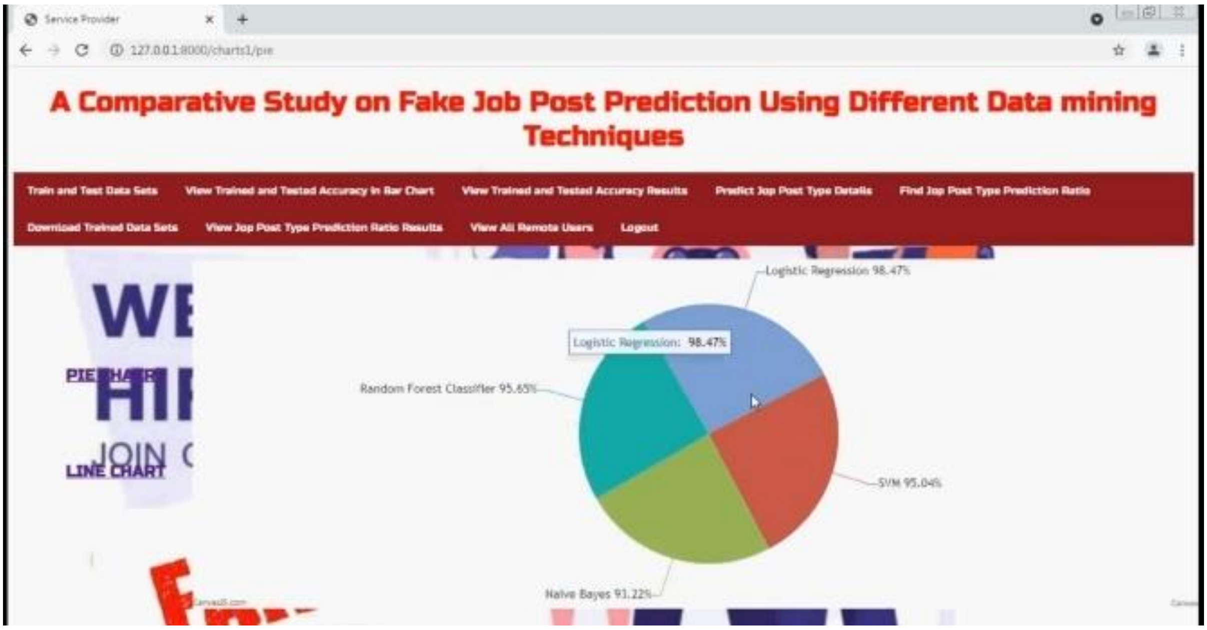
Admin

•••••

Login

User Login

Fig 8.1 Admin login



Service Provider

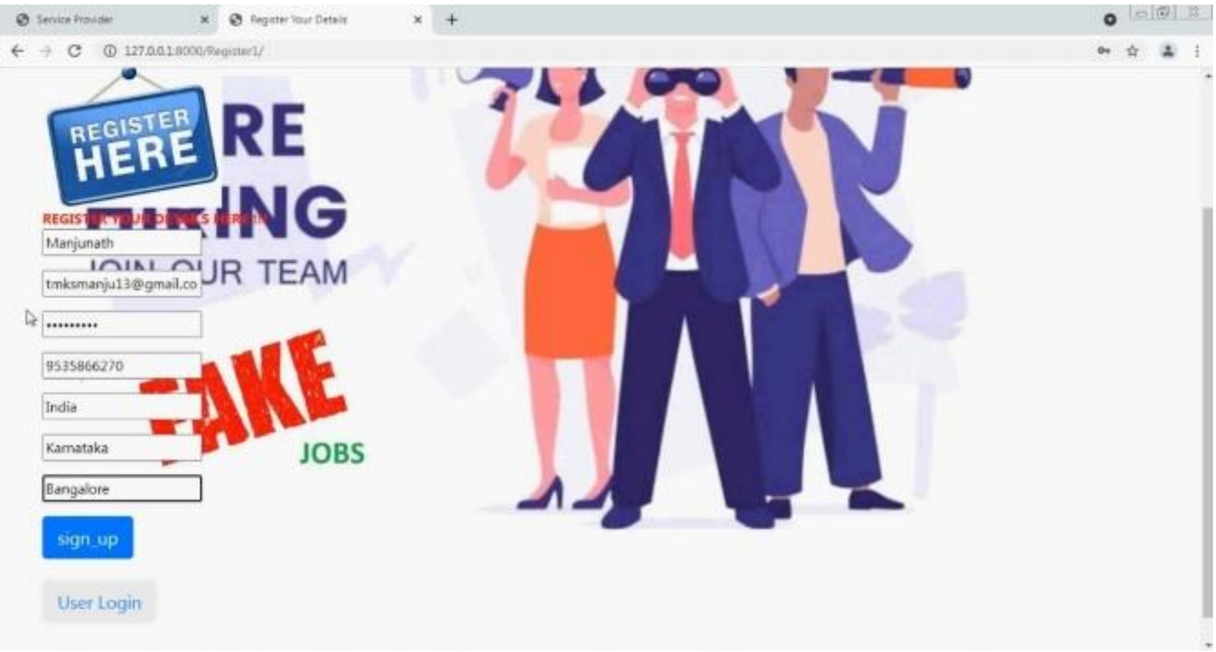
127.0.0.1:8000/charts1/pie

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Train and Test Data Sets | View Trained and Tested Accuracy in Bar Chart | View Trained and Tested Accuracy Results | Predict Jop Post Type Details | Find Jop Post Type Prediction Ratio | Download Trained Data Sets | View Jop Post Type Prediction Ratio Results | View All Remote Users | Logout

PIE CHART
LINE CHART
Logistic Regression 98.47%
Logistic Regression: 98.47%
Random Forest Classifier 95.65%
SVM 95.04%
Naive Bayes 91.22%
CanvasJS.com

Fig 8.2 Pie chart of EMDSCAN data set.



Manjunath

tmksmanju13@gmail.co

•••••••••

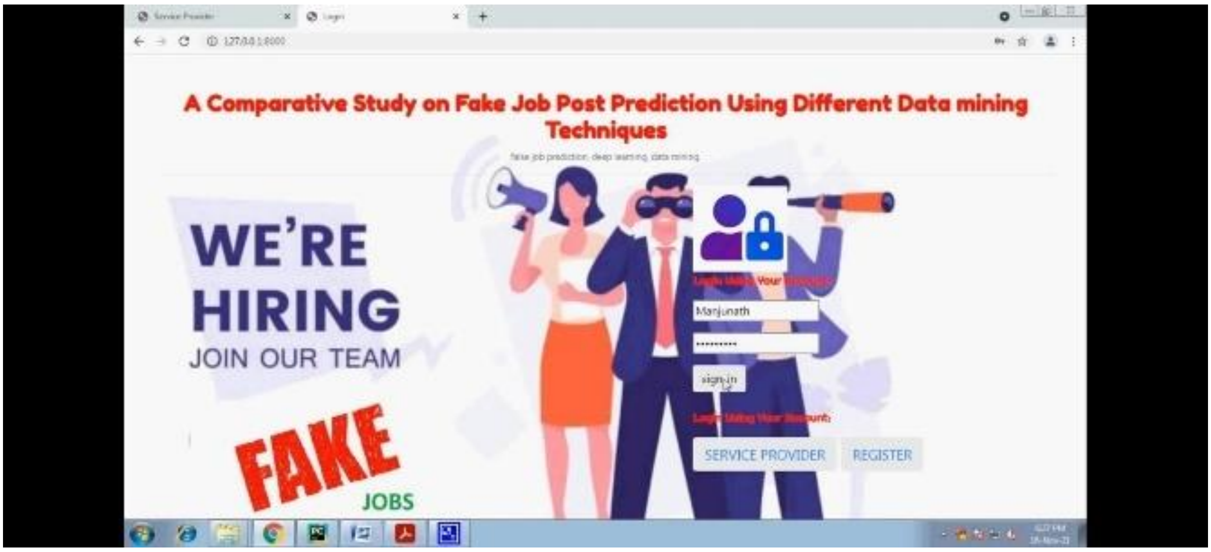
9535866270

India

Karnataka

Bangalore

Fig 8.3 User registration Page

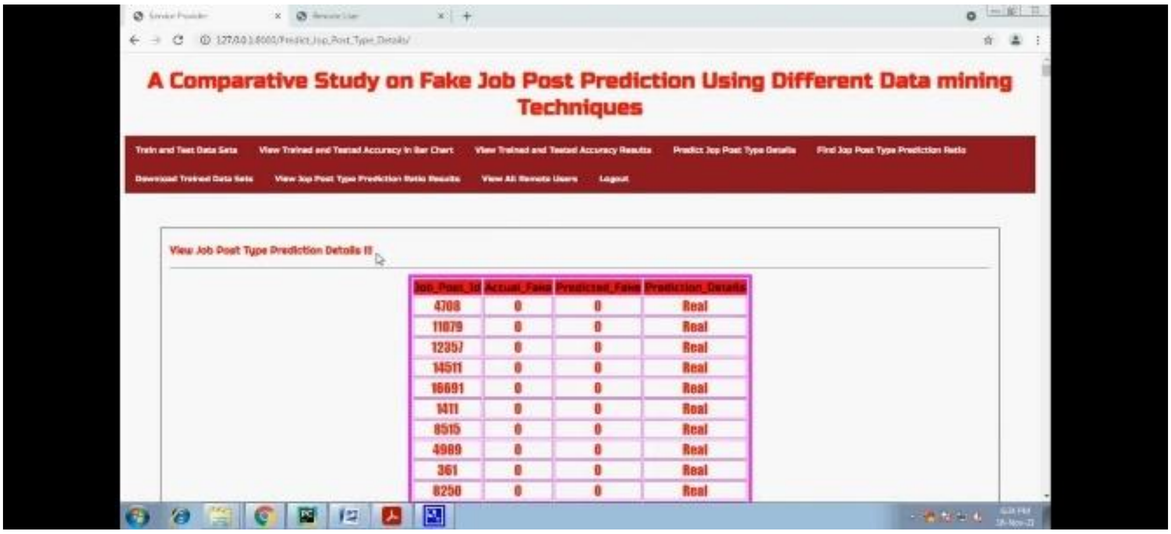


A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Manjunath

•••••••••

Fig 8.4 User login

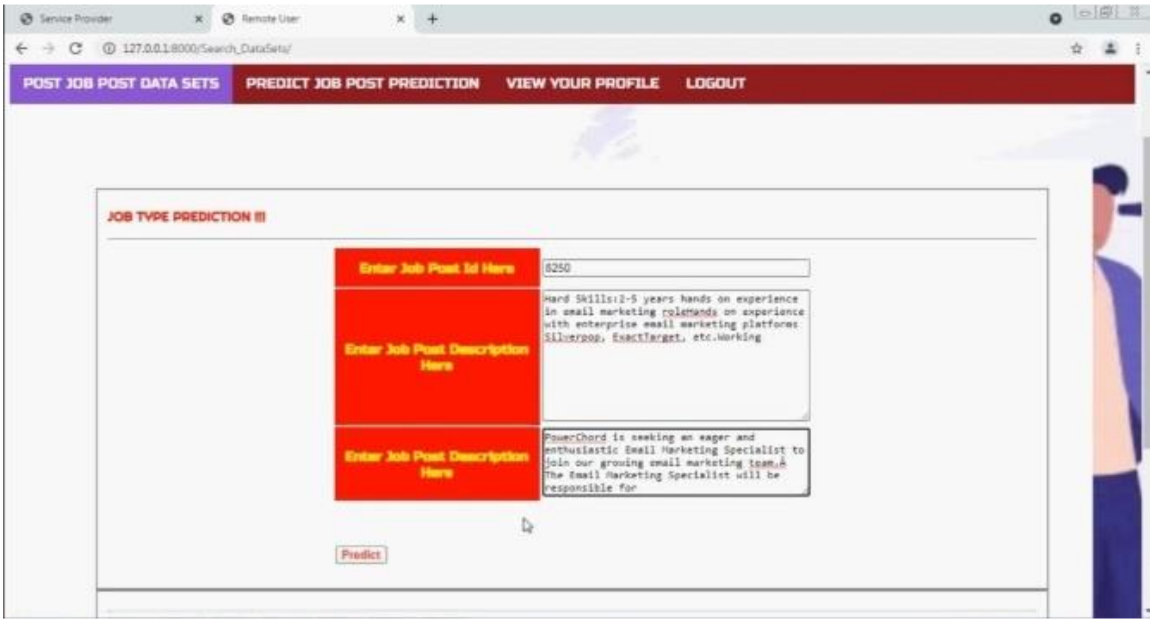

Service Provider
Remote User
127.0.0.1:8000/Predict_Job_Post_Type_Details/
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Job Post Type Details
Find Job Post Type Prediction Ratio
Download Trained Data Sets
View Job Post Type Prediction Ratio Results
View All Remote Users
Logout

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

View Job Post Type Prediction Details !!!

Job_Post_Id	Actual_Fake	Predicted_Fake	Prediction_Details
4708	0	0	Real
11079	0	0	Real
12357	0	0	Real
14511	0	0	Real
16691	0	0	Real
1411	0	0	Real
8515	0	0	Real
4989	0	0	Real
361	0	0	Real
8250	0	0	Real

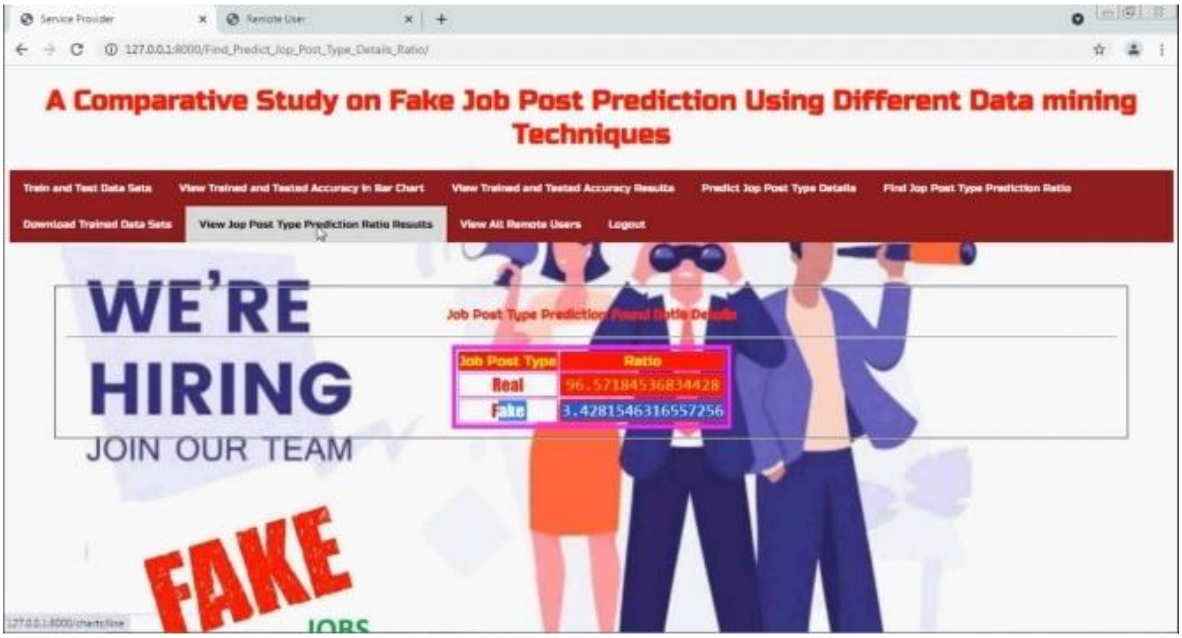
Fig 8.5 After data cleaning true and fake jobs posts



JOB TYPE PREDICTION !!!

Enter Job Post Id Here	8250
Enter Job Post Description Here	Hard Skills:2-5 years hands on experience in email marketing roleHands on experience with enterprise email marketing platforms Silverpop, ExactTarget, etc.Working
Enter Job Post Description Here	PowerChord is seeking an eager and enthusiastic Email Marketing Specialist to join our growing email marketing team.Â The Email Marketing Specialist will be responsible for

Fig 8.6 . checking the job post

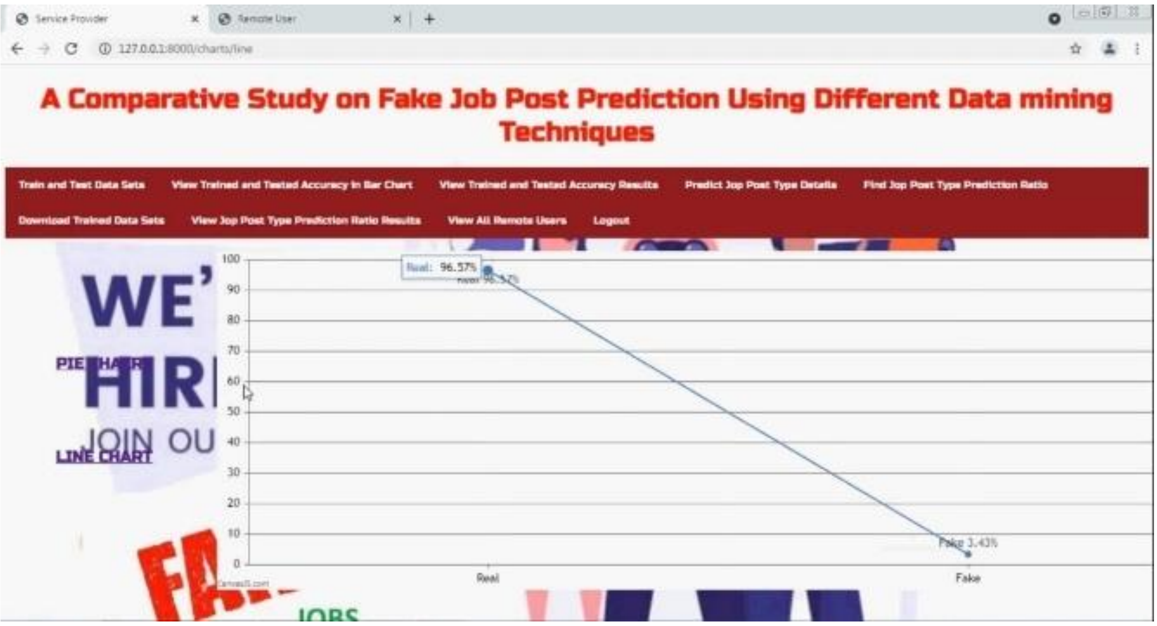

Service Provider
Remote User
127.0.0.1:8000/Find_Predict_Jop_Post_Type_Details_Ratio/
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
WE'RE
HIRING
JOIN OUR TEAM
FAKE
JOBS
127.0.0.1:8000/charts/line

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Job Post Type Prediction Found Ratio Details

Job Post Type	Ratio
Real	96.57184536834428
Fake	3.4281546316557256

Fig 8.7 Distinguishing between true and fake job


Service Provider
Remote User
127.0.0.1:8000/charts/line
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
WE'
HIR
JOIN OU
PIE CHART
LINE CHART
100
90
80
70
60
50
40
30
20
10
0
Real: 96.57%
Real 96.57%
Fake 3.43%
Real
Fake
CanvasJS.com
JOBS

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Fig 8.8 Graphical representation of true and fake jobs

CHAPTER 9
CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, KNN, Naïve Bayes , Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

CHAPTER 10

REFERENCES

[1] B. Alghamdi and F.Alharby, An Intelligent Model for Online Recruitment Fraud Detection," J.Inf.Secur., vol.10, no.03, pp.155–176, 2019, doi:10.4236/jis.2019.103009.

[2] I. Rish, An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no.January 2001, pp.41–46, 2014.

[3] D. E. Walters, Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol.30, no.7, pp.817–825, 1988, doi:10.1002/bimj.4710300710.

[4] F. Murtagh, Multilayer perceptron for classification and regression,‖ Neurocomputing, vol.2, no.5–6, pp.183–197, 1991, doi:10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S.J.Delany, K-Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] H. Sharma and S.Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int.J.Sci.Res., vol.5, no.4, pp.2094–2097, 2016, doi:10.21275/v5i4.nov162954.

[7] E. G.Dada, J.S.Bassi, H.Chiroma, S.M.Abdulhamid, A.O.Adetunmbi, and O.E.Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol.5, no.6, 2019, doi:10.1016/j.heliyon.2019.e01802.

[8] L. Breiman, ST4_Method_Random_Forest,‖ Mach.Learn.,vol.45, no.1, pp.5–32, 2001, doi:10.1017/CBO9781107415324.004.

[9] B. Biggio, I.Corona, G.Fumera, G.Giacinto, and F.Roli, Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol.6713 LNCS, pp. 350–359, 2011, doi:10.1007/978-3-642-21557-5_37.

[10] A. Natekin and A. Knoll, Gradient boosting machines, atutorial,‖ Front. Neurorobot., vol. 7, no. DEC, 2013, doi:10.3389/fnbot.2013.00021.

[11] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M.Kaleem, ―Spam review detection techniques: A systematic literature review,‖ Appl. Sci., vol. 9, no. 5, pp. 1–26, 2019, doi: 10.3390/app9050987.

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, ―Fake News Detection on social media,‖ ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017,

doi:10.1145/3137597.3137600.

[13] Shivam Bansal (2020, February). [Real or Fake] Fake JobPosting Prediction,Version 1 Retrieved March 29,2020 from https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

[14] H. M and S. M.N, ―A Review on Evaluation Metrics for Data Classification Evaluations,‖ Int. J. Data Min. Knowl.Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi:10.5121/ijdkp.2015.5201.

[15] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, Cohen's kappa coefficient as a performance measure for feature selection," 2010 IEEE World Congr. Comput. Intell. WCCI2010, no. May 2016, 2010, doi:10.1109/FUZZY.2010.5584447.