IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# PREDICTING FAKE JOB POSTS USING MACHINELEARNING MODELS

[#1] Mr. B.J.M Ravi Kumar, [#2] Mr. Melkamu Boka Eba, [#3] Mr. Ridwan Mohammed

[#1] Assistant Professor, [#2] B. Tech Student, [#3]B. Tech Student

Department of Information Technology and Computer Application, Engineering College A, Andhra University, Visakhapatnam, India - 530003

***Abstract:*** Due to recent advancements in social media and modern technologies, posting job openings online has become more commonplace. Therefore, predicting fake job ads will be of great importance to everyone. Predicting fake job listings presents several issues similar to other sorting jobs. Using machine learning-based categorisation approaches, the research proposes an automated solution to thwart fake online job adverts.

In this research, we propose applying the *Apriori Algorithm* for data mining to determine the confidence value. Following this, weuse various classification algorithms, such as decision trees, logistic regression, support vector machines, naïve Bayes classifiers, random forest classifiers, and multilayer perceptron, to determine whether a job post is genuine or fraudulent. Various classifiers are employed to verify fraudulent posts on the internet, and the output of those classifiers is compared to determine the most effective model for detecting job scams. From an extensive range of notifications, it assists in identifying fake job postings.

***Key Words*** **- False Job Prediction, Machine Learning, Data Mining.**

## I. INTRODUCTION

Picking their dream job is today's greatest obstacle for any graduate. Unluckily, most of them fall for fake job postings and lose money and time during their search. The suggested system uses a deep learning-based system and a web page to assist non-technicalusers in analysing these fake scams and landing their dream jobs. The prevalence of phoney employment in the modern era underscores the risks associated with their consequences and the difficulties distinguishing them from genuine employment.

However, technological advancements and news spread via various social media platforms have exacerbated the number of fraudulent jobs today. Due to this, the effects of fake jobs have grown dramatically in recent years, and action needs to be taken tosafeguard them from arising in the future. Our objective is to utilise machine learning to distinguish between genuine and fraudulentjobs as least as accurately as people.

To do this, a method based on machine learning is used that utilises a variety of classification algorithms to detect fake posts. Here,a classification tool warns users when it detects phoney job advertisements amid a larger set of job postings. To solve the problemof identifying scammers on job advertising, supervised learning algorithms are first regarded as classification techniques. By considering training data, a classifier links an input variable to a target class. A brief description of the classifiers discussed in the study is given to distinguish fake job posts from real ones. These predictions based on classifiers tend to be divided into two categories: predictions based on ensemble classifiers and predictions based on single classifiers.

One can distinguish between fake and genuine postings if one looks carefully enough. Nearly all of the moment, the company's response to these postings comes through an unofficial email account, or they may ask for private information, like your credit card number, during an interview, claiming that they require it for personnel verification. All of these would be clear indicators that there is something fishy about the organisation under regular economic circumstances, but these are not normal circumstances. These are the worst circumstances any of us has ever experienced. Desperate people currently need work, and by doing so, they are giving these con artists direct leverage over them. The classifier uses the training data to determine which desired class the input variables belong to. To distinguish the fraudulent classified ads from the rest, let's briefly examine the classifiers used in this investigation. This classifier-based forecast is categorised broadly using an ensemble and single classifier-based prediction.

Several people bring up the issue of fake job post detection, and businesses frequently post false job vacancies online to mislead job seekers. These fictitious job postings may be used to sell goods, services, or training courses or to gather personal data. Fake job postings may occasionally entice job seekers into dubious schemes like pyramid schemes or investment fraud.

In general, the issue of detecting false job postings is complicated and multifaceted, necessitating a combination of technical, legal, and educational solutions. Some possible answers are advanced fraud detection algorithms, more education and awareness campaigns for job seekers, and tougher legislation and enforcement mechanisms for companies and job sites.

### 1.1. Single Classifier-based Prediction

To anticipate unknown test instances, classifiers undergo training. To identify phony job postings, the following classifiers are employed.

1. ***Naive Bayes Classifier***

A statistical classification method named Naive Bayes is predicated on the Bayes Theorem. One of the most basic supervised learning strategies is this one. One quick, accurate, and reliable method is the naive Bayes classifier. When applied to large datasets, naive Bayes classifiers are fast and accurate. According to the Naive Bayes classifier, the impact of one feature on a class is unaffected by the influence of other characteristics.

The choice of this classifier performs well in practice, despite its erroneous probability predictions. In the following cases, the classifier produces a promising result: either the features are functionally independent or coupled to something else. There is no relationship between the dependency of this classifier's features and its accuracy. Instead, because of the assumption of independence, the total amount of data lost in the class is required to predict accuracy.

1. **Multi-Layer Perceptron Classifier.**

A multiple-layer perceptron with proper training parameters can be used as a supervised classification tool. The number of nodes in each layer and hidden layers in a multi-layer perceptron may differ for a specific task. The structure of the network and the training data are considered while choosing the parameters.

### *1.2. Ensemble Approach-based Classifiers*

The ensemble approach allows multiple machine learning algorithms to work together to increase the system's overall accuracy. Regarding classification difficulties, Random Forest (RF) uses the regression technique and ensemble learning approach. This classifier assimilates several tree-like classifiers and applies them to different subsamples of the dataset. Each tree then votes for the class that best fits the input.

Mainly, Naïve Bayes, SVM, Random Forest, and Logistic Regression are the supervised learning algorithms employed in this project's classification. Let us examine each one of them carefully.

***Supervised Learning Algorithms***

When a model is trained on a "Labeled Dataset," it is called a supervised learning algorithm. Datasets with labels have parameters for both input and output. Algorithms that use supervised learning learn to map points between inputs and accurate outputs. They contain labelled datasets for both training and validation. The following lists the two major types of supervised learning: Classification and Regression.

**Classification:** Predicting categorical target variables, which stand for discrete classes or labels, is the task of classification. For example, determining whether an email is spam or a patient is at high risk for heart disease. Classification algorithms acquire the ability to associate the input features with one of the pre-established classes. Examples of classification algorithms are K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Support Vector Machine, Logistic Regression, and Decision Tree.

For classification jobs where the objective is to predict the likelihood that an instance belongs to a specific class or not, supervisedmachine learning algorithms like logistic regression are employed. Logistic regression is a statistical method for examining the relationship between two data components. The article discusses the kinds, applications, and foundations of logistic regression. In this binary classification, a sigmoid function, which accepts input as independent variables and outputs a probability value between0 and 1, is used.

The outcome of a categorical dependent variable is predicted using logistic regression. The outcome must therefore be a discrete orcategory value. Instead of fitting a regression line (0 or 1), logistic regression uses a "S" shaped logistic function to represent the two maximum values. We imported the logistic regression algorithm from the sklearn_linear_model module in order to apply it forthis project.

Assumptions for Logistic Regression:

- The dependent variable in a logistic regression analysis must be categorical.
- Multicollinearity should not exist in the independent variable.

The type of logistic regression utilised in this paper is called binomial. There are only two conceivable forms of dependent variablesin binomial logistic regression, such as 0 or 1, Pass or Fail, etc.

## II. EXISTING SYSTEM

One of the major problems in online recruitment fraud (ORF) that has drawn attention recently is employment scams. Nowadays, many companies have chosen to list their openings online so that job seekers can access them quickly. However, as they promise jobs to applicants in exchange for their money being taken, this may be one of the fraudsters' intentions. Fake job postings can be made against a reputable company to damage its reputation. The discovery of fraudulent job posts brings attention to the need for an automated program to recognise phoney job postings and alert users so they don't apply.

Numerous researchers were able to determine whether a job posting was genuine or fake. The EMSCAD dataset has been evaluatedusing various classification techniques, including KNN, naive Bayes, random forest, Zero R, One R, and others. The Random ForestClassifier performed the best with 95.5% classification accuracy.

## III. LIMITATION OF PRIMITIVE SYSTEM

1. Low Accuracy is the consequence of training the data with fewer features taken into account.
2. The proposed system has a higher processing accuracy than the support vector machine model.
3. It's also a time-consuming and complicated process.
4. Random-Forest and SVM may not capture complex non-linear correlations in the data as well as they do with othermethods, such as logistic regression.

## IV. PROPOSED SYSTEM AND ITS ADVANTAGES

The algorithm has identified fake job postings using EMSCAD. This dataset contains 18,000 samples and 18 attributes per row, including the class label. The attributes include the following: employment type, necessary experience, education, industry, function,fraud (class label), job ID, title, location, department, salary range, company profile, description, requirements and benefits, telecommunication, business logo, and questions. Only seven of these eighteen features have been converted into categories of content.

Textual features are converted into categorical forms for easy classification. The method reduces computing complexity and improves efficiency by doing away with complicated text processing and natural language techniques by breaking down the feature space into categorical properties. This strategy also enhances interpretability by clearly classifying criteria like work type, needed experience,and education and making the decision-making process explicit. By considering training data, a classifier translates input variablesto target classes. A brief description of the classifiers discussed in the study is given to distinguish fake job posts from real ones. These classifier-based predictions can be broadly divided into two categories: ensemble classifier-

based predictions and single classifier-based predictions. For this research, the logistic regression algorithm achieved the best results.

***Abbreviations and Acronyms***

1. ML – Machine Learning
2. RF – Random Forest
3. ORF – Online Requirement Frauds
4. EMSCAD – Employment Scam Aegean Dataset
5. KNN – K Nearest Neighbor

## V. METHODOLOGY

After preprocessing and cleaning the EMSCAD dataset, which contains over 17880 job posts, we have used it to train varioussupervised machine learning models.

The suggested method employs the Logistic Regression technique to distinguish between genuine and fraudulent job postings. Themodel is trained to be as accurate as possible while considering the different ways that jobs are posted on professional and non- professional websites. The dataset comes from a double-masked study.

Clients may feel more at ease looking for job online as a result of the increased success of job hunting compared to the past. The dataset that was used is quite helpful because it has been thoroughly studied. The front end can be used by users to anticipate job descriptions. The proposed method uses Django and Python to create an easy-to-use web interface for non-technical people. In addition, we would like to develop an accurate way to distinguish between legitimate and fraudulent job listings.

To prevent overfitting, the dataset was compiled from a variety of reliable sources and viewpoints, which adds

to its integrity.Before this project is implemented, the following procedures are completed.

**Import the Libraries:** We must first import the relevant libraries to implement the algorithm in Python. The NumPy libraries willbe imported for scientific computation.

**Fetch the Data:** Using 'pandas_datareader,' we will retrieve the data from a CSV file and save it in a data frame.

**Split the Dataset:** The dataset will be divided into two categories: training and test datasets. 30% of our data will be used for testingand 70% for training. To accomplish this, we shall divide the data frame by half, 70%.

**Create Machine Learning Classifier Model:** First, we will divide the dataset into training and test datasets. Next, we will use the'fit' function to create several classifiers that fit the train data. Then, we will keep the classifier as a model.

**Prediction**: As input prediction is carried out, a new CSV file with distinct job profile details is provided, and information is savedin a new CSV with the prediction outcomes.

**Deployment**: A responsive website is deployed using Python as a programming language and Django as a framework.

### 5.1. Diagram of Proposed system

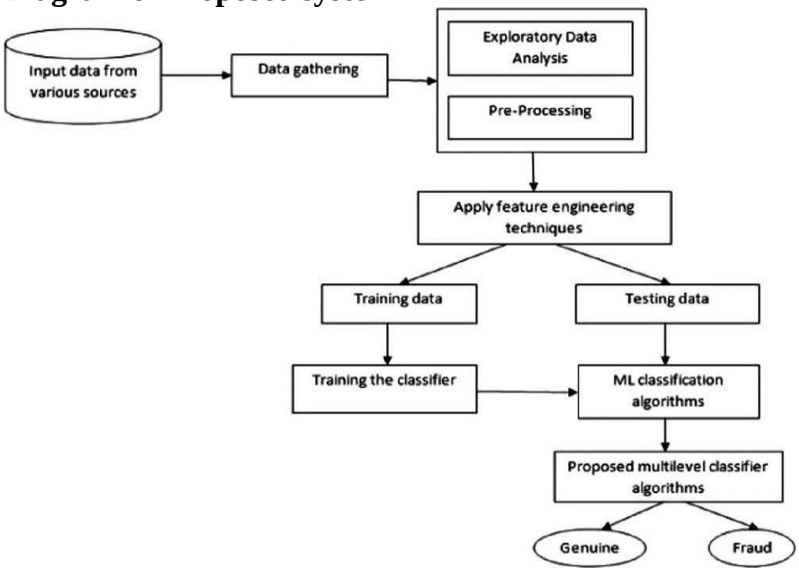



Fig 5.1. Metrics of the proposed system

### 5.2. Implementation Phase

The process of transforming the theoretical design into a programmable format is called implementation. The program will be divided into several modules, and the deployment code will be developed. Here, the current application is implemented utilising Python as a programming language. The following seven modules make up the majority of the application. These are listed in the following order:

1. Load Dataset Module
2. Generate Test and Train Data
3. Run Several Algorithms
4. Detect fake profile recruitment identification from the Test Dataset
5. Comparative analysis
6. Predict the trained Dataset
7. Identify the genuine and fake job

**1. Load Dataset Module**

The project's data can be accessed via Kaggle - Real / Fake Job Posting Prediction (kaggle.com). Eighteen characteristics and 17,880observations make up the dataset. Integer, binary, and text datatypes are all combined in the data.

**2. Generate Test and Train Data**

Here, we attempt to separate the data into test and train datasets, partitioning the entire dataset into several segments using a 70:30 %ratio. In this case, 70% of the data records are used to train the system, and 30% are utilised to test the model.

**3. Run Several Algorithms**

Here, we attempt to run many algorithms on the training dataset to determine the probability of every attribute present in that particular record. After processing all the documents, we try to decide which ones include fraud activity and which have regular activity. We can determine the accuracy of each technique once we apply Gaussian Mixture and Isolation Forest to the training dataset. Ultimately, it is evident that the Gaussian mixture yields better results than all other techniques.

**4. Detect fake Job**

In this case, we attempt to use multiple techniques and validate the model using test data. After input, the test data can be divided into two categories: the number of records that are real recruitments and those that are not.

**5. Compare analysis**

In the current application, we evaluated the dataset using Random Forest, SVM, Naïve Bayes, and Logistic Regression. We ultimately determined that Logistic Regression yields the best results considering the accuracy, which is 97.37%.

# VI. DATA ANALYSIS

## 6.1. *Exploratory Analysis*

Making a correlation matrix to examine the relationship between numerical data is the first step in visualizing the dataset for this project. There aren't any particularly strong positive or negative correlations between the numerical data in the correlation matrix. The information is composed of text, binary, and integer data types. The provided dataset has great value because it can be utilisedto address the following inquiries:

1. Build a classification model to identify which job descriptions are genuine or fraudulent based on text data characteristicsand meta-features.
2. Determine which essential characteristics (words, entities, and phrases) from job descriptions are fake.
3. To find the job descriptions that are the most comparable, run a contextual embedding model.
4. Conduct exploratory data analysis to find intriguing insights from this dataset.A brief definition of the variables is given in below table:

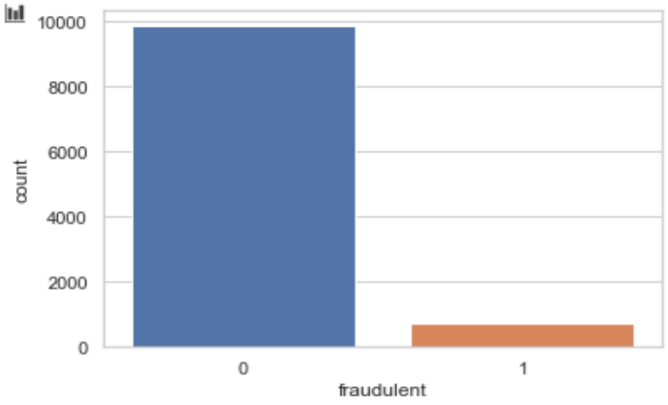
| # | Variable | Datatype | Description |
|---|---|---|---|
| 1 | job_id | int | Identification number given to each job posting |
| 2 | title | text | A name that describes the position or job |
| 3 | location | text | Information about where the job is located |
| 4 | department | text | Information about the department this job is offered by |
| 5 | salary_range | text | Expected salary range |
| 6 | company_profile | text | Information about the company |
| 7 | description | text | A brief description about the position offered |
| 8 | requirements | text | Pre-requisites to qualify for the job |
| 9 | benefits | text | Benefits provided by the job |
| 10 | telecommuting | boolean | Is work from home or remote work allowed |
| 11 | has_company_logo | boolean | Does the job posting have a company logo |
| 12 | has_questions | boolean | Does the job posting have any questions |
| 13 | employment_type | text | 5 categories – Full-time, part-time, contract, temporary and other |
| 14 | required_experience | text | Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable |
| 15 | required_education | text | Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational |
| 16 | Industry | text | The industry the job posting is relevant to |
| 17 | Function | text | The umbrella term to determining a job's functionality |
| 18 | Fraudulent | boolean | The target variable → 0: Real, 1: Fake |

This situation does not require a summary statistic because most data types are text or Boolean. The only integer that matters for this analysis and the letters that represent the requirements and description are the Job ID. We investigate the dataset further to findnull values.

```
job_id                   0
title                    0
location               346
department           11547
salary_range         15012
company_profile       3308
description              1
requirements          2695
benefits              7210
telecommuting            0
has_company_logo         0
has_questions            0
employment_type       3471
required_experience   7050
required_education    8105
industry              4903
function              6455
fraudulent               0
dtype: int64
```

This dataset contains 17,880 job posts used in the proposed methods for testing the strategy's overall performance. A multistep approach is used to produce a balanced dataset, which helps to comprehend the target as a baseline better. Some pre-processing methods are used on this dataset before fitting it to any classifier. Pre-processing techniques include removing superfluous space, stop words, irrelevant attributes, and missing values. Variables like department and pay range have a lot of missing data, so additional analysis is not performed on these columns.

The dataset is highly unbalanced, with 9868 (93% of the jobs) being real and only 725 or 7% of the jobs being fraudulent. A countplot of the same can clearly show the disparity, as seen in the following graph.




## VII. EXPERIMENTAL RESULTS

In this section, we attempt to formulate our present model in Python, utilising Django as the application's framework. We may nowassess our suggested application's performance as follows:

### *6.1. Importing Libraries*

```
from django.db.models import  Count, Avg
from django.shortcuts import render, redirect
from django.db.models import Count
from django.db.models import Q
import datetime
import xlwt
from django.http import HttpResponse
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from sklearn.ensemble import VotingClassifier
import sklearn as sk
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
#Model from SciKit-Learn
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn import feature_selection
from sklearn.impute import SimpleImputer

# Model Evaluations from SciKit Learn
from sklearn.metrics import
accuracy_score,classification_report,confusion_matrix,precision_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import precision_score
```

### *6.2. Modules Service Provider*

In this module, the Service Provider must log in using a valid username and password. After logging in successfully, he can performoperations such as Train and Test Data Sets, View Trained and Tested Accuracy in a Bar Chart, View Trained and Tested AccuracyResults, Predict Job Post Type Details, Find Job Post Type Prediction Ratio, Download Trained Data Sets, View Job Post Type Prediction Ratio Results, and View All Remote Users.

***View and Authorize Users***

In this module, the admin can view the list of registered users, view the users' details, such as username, email, and address, andauthorise the users.

***Remote User***

There are no users in this module. Users should register before performing any operations. Once users register, their details arestored in the database. After successful registration, they must log in using an authorised username and password.

The list of several modules utilised in our application is displayed in the window above.

### *6.2. Testing and Training Dataset*

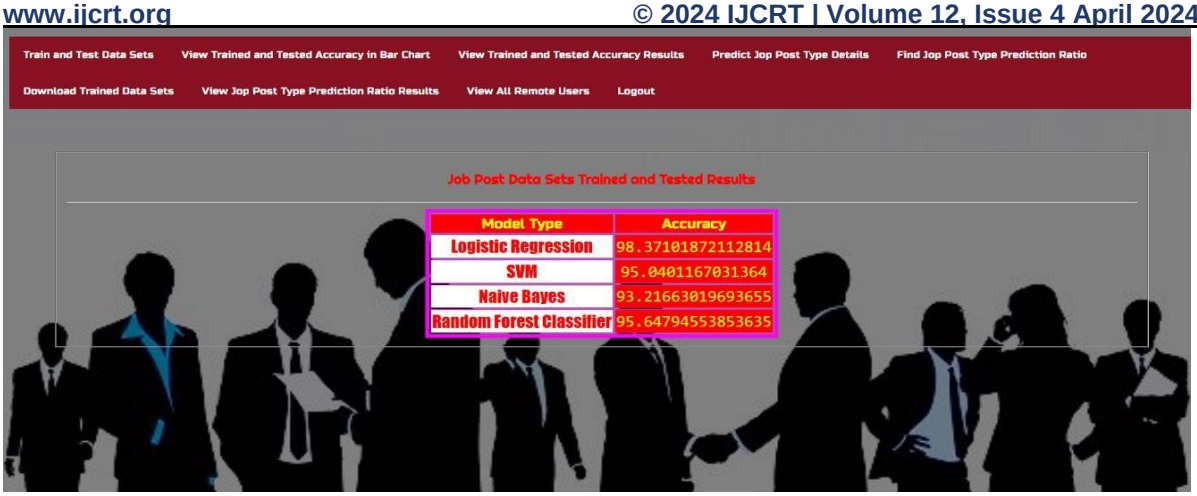
```
[09/Mar/2024 20:26:03] "GET /favicon.ico HTTP/1.1" 404 4353
X Trans
                                        title            location   department salary_range  ... job_id telecommuting has_company_logo has_questions
0                                 Marketing Intern      US, NY, New York    Marketing      Missing  ...      1            0                1             0
1              Customer Service - Cloud Video Production       NZ, , Auckland      Success      Missing  ...      2            0                1             0
2                Commissioning Machinery Assistant (CMA)        US, IA, Wever      Missing      Missing  ...      3            0                1             0
3                     Account Executive - Washington DC    US, DC, Washington        Sales      Missing  ...      4            0                1             0
4                               Bill Review Manager    US, FL, Fort Worth      Missing      Missing  ...      5            0                1             1
...                                             ...                   ...          ...          ...  ...    ...          ...              ...           ...
17875               Account Director - Distribution       CA, ON, Toronto        Sales      Missing  ...  17876            0                1             1
17876                           Payroll Accountant  US, PA, Philadelphia   Accounting      Missing  ...  17877            0                1             1
17877  Project Cost Control Staff Engineer - Cost Con...       US, TX, Houston      Missing      Missing  ...  17878            0                0             0
17878                           Graphic Designer        NG, LA, Lagos      Missing      Missing  ...  17879            0                0             1
17879                  Web Application Developers     NZ, N, Wellington  Engineering      Missing  ...  17880            0                1             1

[17880 rows x 17 columns]
```

```
Y Trans
0        0
1        0
2        0
3        0
4        0
        ..
17875    0
17876    0
17877    0
17878    0
17879    0
```

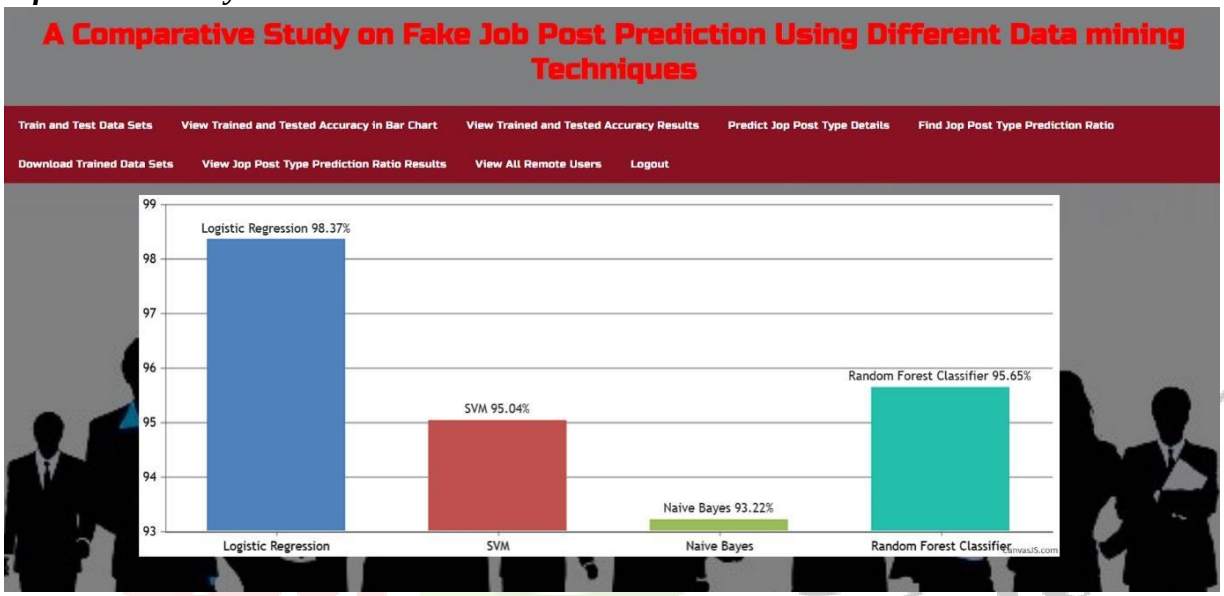
The above area shows that the data has been preprocessed and divided into test and train.

### *6.3. Comparing the algorithms used in the analysis*




The output above shows that, for our study strategy, the Logistic Regression model predicts fraudulent job posts with the highest accuracy.

### *6.4. Performance analysis*




The graph above demonstrates that the Logistic Regression model in this project obtained an impressive accuracy of 98.37%, whichmay be attributed to numerous advantages inherent in this technique. First, linear decision limits in the feature space can be roughlyapproximated by Logistic Regression, which assumes a linear relationship between input features and the log-likelihood of the outcome.
This makes it especially appropriate for situations where the decision border between classes is primarily linear. Furthermore, LogisticRegression models are well known for being interpreted easily, offering insights into the significance and direction of each feature'sinfluence on the result. This interpretability makes comprehending the elements that lead to fake job postings easier.

Furthermore, logistic regression can handle big Datasets with comparatively little computer power since it is scalable and computationally efficient. Its regularisation algorithms guarantee improved generalisation performance on unobserved data by preventing overfitting. Moreover, producing well-calibrated probabilities using Logistic Regression provides dependable estimationsof the likelihood that examples belong to distinct classes, which is essential for fraud detection jobs.

Logistic Regression distinguishes between legitimate and fraudulent job posts exceptionally well because of its interpretability, regularisation capabilities, efficiency, simplicity, and well-calibrated probability.

### *6.5. Predicted Job Post Ratio*

As can be seen from the graph below, 96.57% of the job listings examined in the dataset were found to be real, while 3.42% were found to be fraudulent. This revised language highlights the substantial margin that real job posts outnumber fake ones. It also strengthens confidence in the reliability of the dataset by reaffirming that most job advertisements are real.


## VIII. CONCLUSION

Job seekers might be guided by employment fraud detection to receive only authentic offers from businesses. Several Machine learning methods, including Random Forest, SVM, Naïve Bayes, and Logistic Regression, are suggested in this research as countermeasures to address the issue of job scam detection. Several classifiers for job fraud detection are demonstrated using a supervised technique. According to trial results, the Logistic Regression classifier works better than its peer classification tool. Compared to the current methods, the suggested approach's accuracy rate of 98.37% is significantly greater.

## IX. ACKNOWLEDGMENT

We deeply thank **Mr B.J.M. RAVI KUMAR, Assistant Professor of Computer Science and Systems Engineering, for his excellent guidance throughout the project "Predicting Fake Job Posts: Comparative Analysis of ML Models**". His suggestions have been extremely useful. His encouragement and motivation have gone a long way toward the successful completion of the project. We feel glad to express our sincere thanks and acknowledge our indebtedness to our belovedHead of the Department of Information Technology and Computer. **Prof. Kunjam Nageswara Rao**, for his spontaneous expressionof knowledge, helped us bring up this project throughout the academic year. We thank all those who contributed directly and indirectly to completing this project.

**MELKAMU BOKA EBARIDWAN** **320106411057**
**MOHAMMED** **320106411058**

## REFERENCES

**[1]** S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics,Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.
**[2]** B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", Journal of Information Security,2019, Vol 10, pp. 155 176, https://doi.org/10.4236/iis.2019.103009.
**[3]** Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep NeuralNetwork Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.
**[4]** Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.
**[5]** Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3,5, 2014, https://doi.org/10.1186/s13388-014-0005-5.
**[6]** Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882, 2014.
**[7]** T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, "Hate Speech

Detection on VietnameseSocial Media Text using the Bi-GRU-LSTM-CNN Model," arXiv Prepr. arXiv1911.03644, 2019. Types of Machine Learning - Futural Net. https://futural.net/ai/types-of-machine-learning/

**[8]** P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolu/tionalneural network for improving short text classification," Neurocomputing, vol. 174, pp. 806 814, 2016.