e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

# JOB SCAM DETECTION USING MACHINE LEARNING APPROACH

**Dr. J.B. Jona**[*1], **Manju Bashini Umamaheswaran**[*2], **Swetha Rajaraman**[*3]

[*1]Assistant Professor, Department Of Computer Application- Computer Application, Coimbatore Institute Of Technology, India.

[*2,3]Student, Department Of Computer Application-Computer Application, Coimbatore Institute Of Technology, India.

## ABSTRACT

This paper presents a tool that employs machine learning classification techniques to prevent fraudulent job postings on the internet. Various classifiers, including Random Forest, Naive Bayes, and Support Vector Machine, are employed to scrutinize job posts, and their comparative performance is analysed to identify the most effective employment scam detection model. The proposed tool aids in sifting through a vast number of posts to identify and eliminate fake job listings. The study evaluates multiple classifiers for fraudulent job post detection, revealing that ensemble classifiers outperform individual ones. The results emphasize the effectiveness of an ensemble approach in enhancing the accuracy of job scam detection.

**Keywords:** Fake Job, Online Recruitment, Machine Learning, Ensemble Approach, Job Scam Detection, Random Forest Classifier, Naive Bayes, Support Vector Machine.

## I. INTRODUCTION

Addressing the escalating concern of Employment scams within the realm of Online Recruitment Frauds (ORF), this study delves into the pervasive issue of deceptive job postings. In contemporary times, companies increasingly opt for online platforms to publicize job vacancies, facilitating swift access for job-seekers. Unfortunately, this convenience has given rise to a type of scam wherein fraudulent entities exploit the job-seekers' eagerness by posing as legitimate employers and extracting money under the guise of employment opportunities. Such deceitful practices not only compromise the credibility of reputable companies but also pose a significant threat to unsuspecting applicants.

To counter this challenge, the paper focuses on the crucial task of detecting and exposing fraudulent job postings. The objective is to draw attention to these deceptive practices, enabling job-seekers to discern genuine opportunities from potential scams. Employing a machine learning approach, the study leverages various classification algorithms to discern authentic job advertisements from fraudulent ones. The proposed classification tool acts as a vigilant guardian, singling out fake job postings amidst a sea of legitimate advertisements, thus serving as an invaluable shield for potential job applicants.

To tackle the intricate task of distinguishing scams within job postings, the paper initially explores supervised learning algorithms as classification techniques. These classifiers, mapped with training data, play a pivotal role in identifying and isolating fraudulent job postings from authentic ones. The ensuing sections delve into the specifics of the machine learning algorithms employed, including the Random Forest Classifier, Naive Bayes, and Support Vector Machine, each contributing to the overarching goal of fortifying the online job recruitment landscape against malicious activities. The classifiers-based prediction may be broadly categorized into Single Classifier based Prediction.

### A. Single Classifier Based Prediction

Classifiers are trained for predicting the unknown test cases. The following Algorithm classifiers are used while detecting fake job posts.

#### a) Naïve Bayes

The Naive Bayes algorithm, a supervised classification technique, leverages the principles of Bayes' Theorem of Conditional Probability for the purpose of Job Scam Detection using machine learning. Despite potential inaccuracies in its probability estimates, this classifier demonstrates remarkable effectiveness in practical applications. Particularly, the Naive Bayes classifier yields promising results in scenarios where features exhibit either independence or complete functional dependence. Unlike other classifiers, the accuracy of Naive Bayes is not contingent on feature dependencies; rather, it is influenced by the amount of information loss pertaining to the class due to the assumption of independence. This characteristic makes Naive Bayes well-suited for predicting the accuracy of Job Scam Detection models in scenarios where feature relationships play a crucial role.

**b) Random Forest**

The Random Forest algorithm proves to be a robust and effective tool in the domain of Job Scam Detection using machine learning. As an ensemble learning method, Random Forest builds a multitude of decision trees during training and combines their outputs to enhance overall accuracy and generalizability. In the context of job scams, this algorithm excels at handling complex relationships and interactions among various features extracted from job postings. Each decision tree in the Random Forest contributes to the collective decision-making process, offering resilience against overfitting and increasing the model's ability to discern patterns indicative of fraudulent job postings. Furthermore, the algorithm's ability to assess feature importance aids in identifying key factors contributing to scam detection. The Random Forest's adaptability and superior performance in handling diverse datasets make it a valuable asset in fortifying online job platforms against deceptive practices and safeguarding job seekers from potential scams.

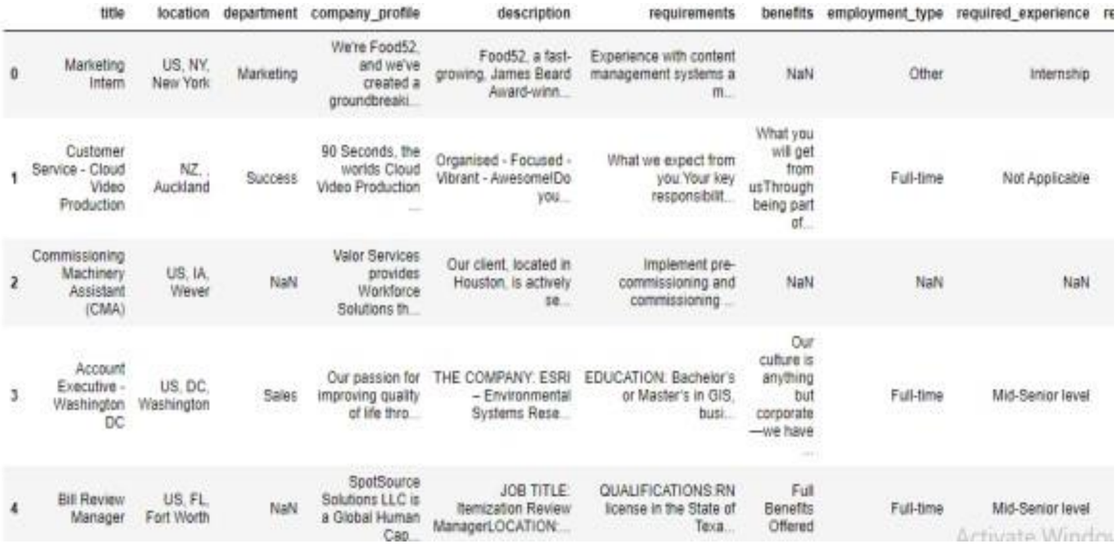
## II. DATASET DESCRIPTITION

The dataset is a collection of job posting records from various websites of with some are fake postings. This dataset contains the attributes ID, Job title, location, department, salary range, employer profile, description, requirements, benefits, and telecommunication, has a company logo, has questions posted, employment type, required experience, required education, industry, function, the target column fraudulent or legit.

**Language Used:** Python

## III. DATA IMPORTING

In the initial phase of the Job Scam Detection project, importing the dataset emerges as a pivotal step, representing a critical component for training and evaluating the machine learning models. The dataset, typically stored in CSV format, encompasses a comprehensive collection of job postings sourced from online platforms. The utilization of machine learning for scam detection demands a meticulous approach to data importing. Employing Python libraries like Pandas, the CSV file is read, and the dataset is loaded into a structured format suitable for analysis. This process establishes the groundwork for subsequent data cleaning and pre-processing steps, ensuring that the machine learning models are trained on a reliable and representative dataset. The imported data encompasses various features extracted from job postings, encompassing textual content, formatting details, and any additional relevant information. With the dataset successfully imported, the workflow advances to the crucial steps of data cleaning, exploration, and, eventually, the application of machine learning algorithms for effective Job Scam Detection.



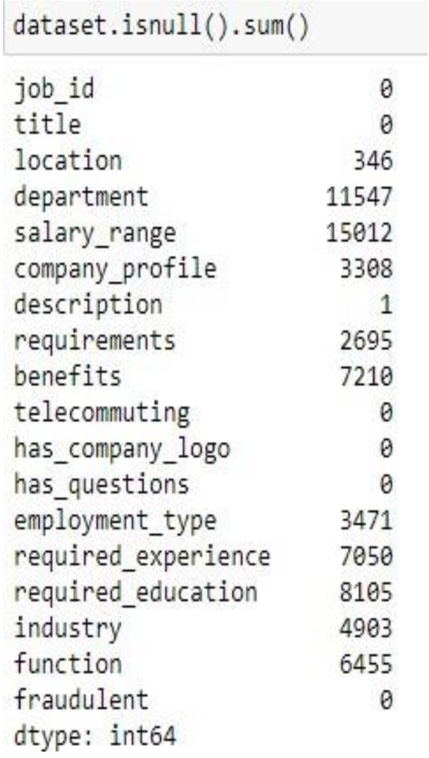
| | title | location | department | company_profile | description | requirements | benefits | employment_type | required_experience | re |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Marketing Intern | US, NY, New York | Marketing | We're Food52, and we've created a groundbreaki... | Food52, a fast-growing, James Beard Award-winn... | Experience with content management systems a m... | NaN | Other | Internship | |
| 1 | Customer Service - Cloud Video Production | NZ, , Auckland | Success | 90 Seconds, the worlds Cloud Video Production ... | Organised - Focused - Vibrant - Awesome!Do you... | What we expect from you:Your key responsibilit... | What you will get from usThrough being part of... | Full-time | Not Applicable | |
| 2 | Commissioning Machinery Assistant (CMA) | US, IA, Wever | NaN | Valor Services provides Workforce Solutions th... | Our client, located in Houston, is actively se... | Implement pre-commissioning and commissioning ... | NaN | NaN | NaN | |
| 3 | Account Executive - Washington DC | US, DC, Washington | Sales | Our passion for improving quality of life thro... | THE COMPANY: ESRI – Environmental Systems Rese... | EDUCATION: Bachelor's or Master's in GIS, busi... | Our culture is anything but corporate—we have ... | Full-time | Mid-Senior level | |
| 4 | Bill Review Manager | US, FL, Fort Worth | NaN | SpotSource Solutions LLC is a Global Human Cap... | JOB TITLE: Itemization Review ManagerLOCATION:... | QUALIFICATIONS:RN license in the State of Texa... | Full Benefits Offered | Full-time | Mid-Senior level | |

**Fig 1:** Data Importing

## IV. DATA CLEANING

After importing the dataset, our Job Scam Detection project focuses on the critical data cleaning phase. Utilizing Python's Pandas library, we meticulously handle missing values, outliers, and inconsistencies to elevate the overall quality of the dataset. Simultaneously, textual content in job postings undergoes pre-processing for uniformity, enabling more meaningful analysis. Exploratory data analysis plays a pivotal role in identifying patterns, outliers, and potential issues that may impact the model's performance. This process guides the removal of redundant features, streamlining the dataset to optimize machine learning model training.

Data cleaning is fundamental for mitigating biases and enhancing the overall effectiveness of our Job Scam Detection system. A refined dataset serves as the foundation for subsequent steps, including data visualization and the application of machine learning algorithms to construct a robust scam detection model.



```
dataset.isnull().sum()
```

```
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
```

**Fig 2:** Data Cleaning

## V. VISUALIZATION

After meticulous data cleaning, our Job Scam Detection project seamlessly moves to data visualization. Using Python's Matplotlib and Seaborn, we create insightful visual representations, including charts and graphs, to understand feature distribution. Visualization uncovers patterns and trends, informing subsequent steps. The objective is a clear depiction of relevant information, aiding in distinguishing features between legitimate job postings and scams. These insights guide decision-making in machine learning model development, actively contributing to constructing a robust Job Scam Detection model.

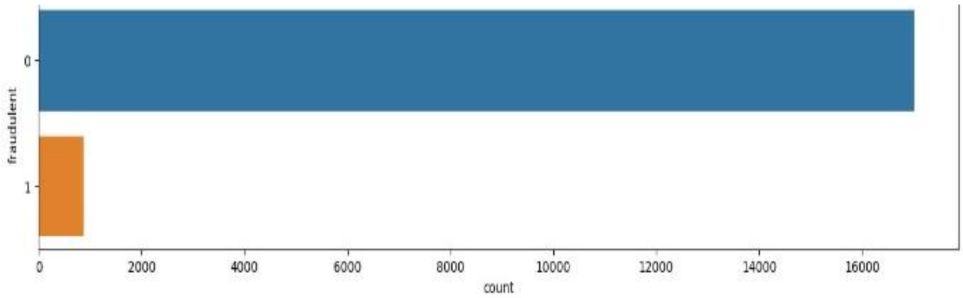


**Fig 3:** Fraudulent Postings in the Dataset

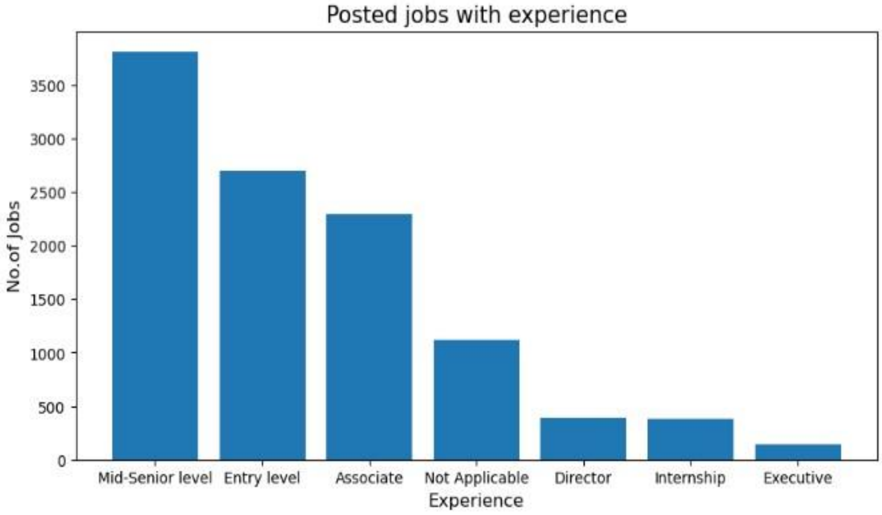


**Fig 4:** Posted Job with Experience

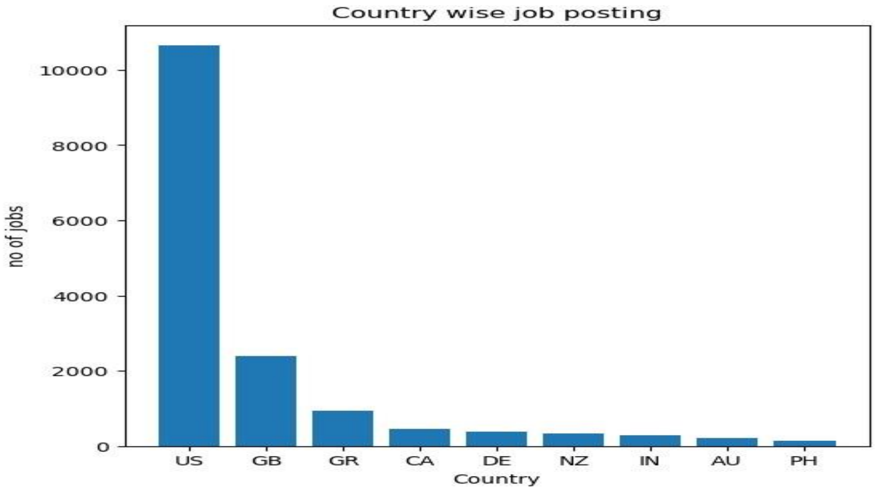
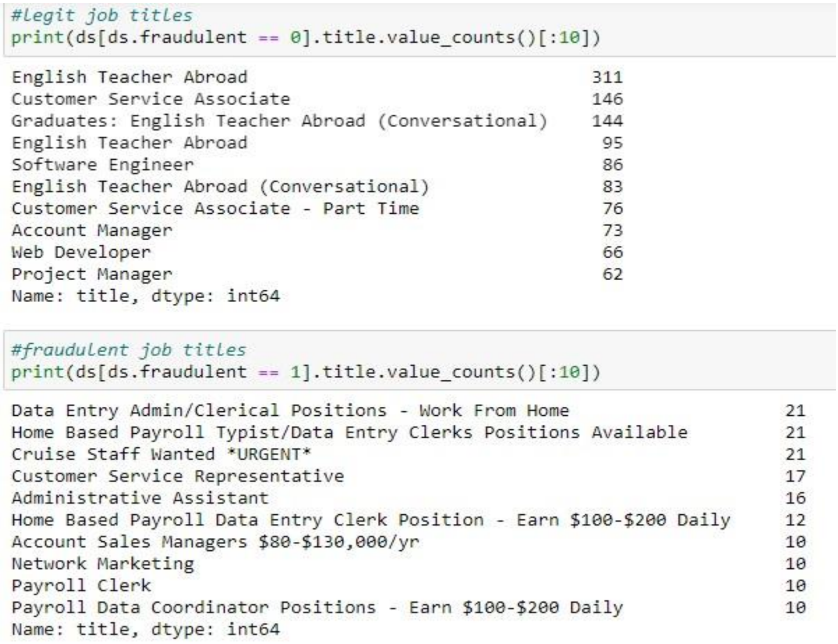


**Fig 5:** Country wise Job Posting



```
#legit job titles
print(ds[ds.fraudulent == 0].title.value_counts()[:10])
```

```
English Teacher Abroad                                  311
Customer Service Associate                              146
Graduates: English Teacher Abroad (Conversational)    144
English Teacher Abroad                                   95
Software Engineer                                        86
English Teacher Abroad (Conversational)                  83
Customer Service Associate - Part Time                   76
Account Manager                                          73
Web Developer                                            66
Project Manager                                          62
Name: title, dtype: int64
```

```
#fraudulent job titles
print(ds[ds.fraudulent == 1].title.value_counts()[:10])
```

```
Data Entry Admin/Clerical Positions - Work From Home                    21
Home Based Payroll Typist/Data Entry Clerks Positions Available         21
Cruise Staff Wanted *URGENT*                                            21
Customer Service Representative                                         17
Administrative Assistant                                                16
Home Based Payroll Data Entry Clerk Position - Earn $100-$200 Daily     12
Account Sales Managers $80-$130,000/yr                                  10
Network Marketing                                                       10
Payroll Clerk                                                           10
Payroll Data Coordinator Positions - Earn $100-$200 Daily               10
Name: title, dtype: int64
```

**Fig 6:** Legit and Fraudulent Job Titles

## VI. WORD CLOUD

In Job Scam Detection, word clouds visually capture essential information from job postings, unveiling patterns and potential scam indicators post data cleaning. This graphical representation swiftly identifies common keywords, distinguishing between legitimate opportunities and scams. By emphasizing frequent words, word clouds assist analysts in focusing on essential elements, facilitating the identification of distinguishing features between authentic and fraudulent opportunities. Integrating word clouds enhances interpretability, contributing to the overall effectiveness of Job Scam Detection systems. This includes two types of word clouds as shown below in fig 7 and fig 8.

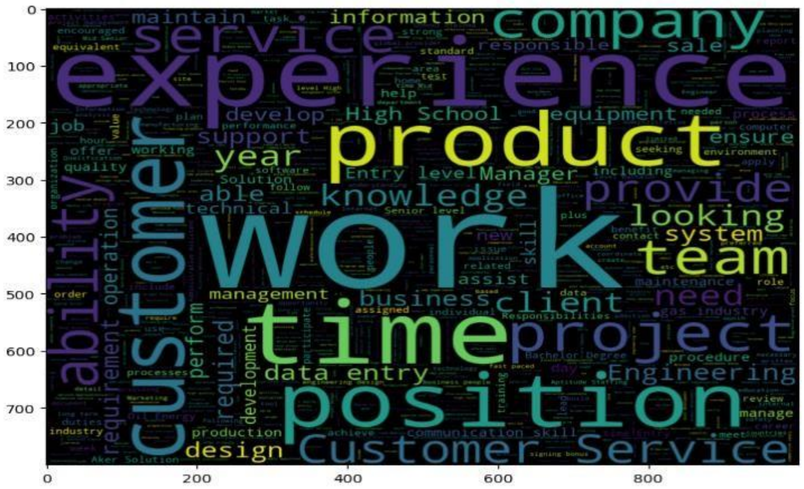


**Fig 7:** Word Cloud of Fraudulent Jobs

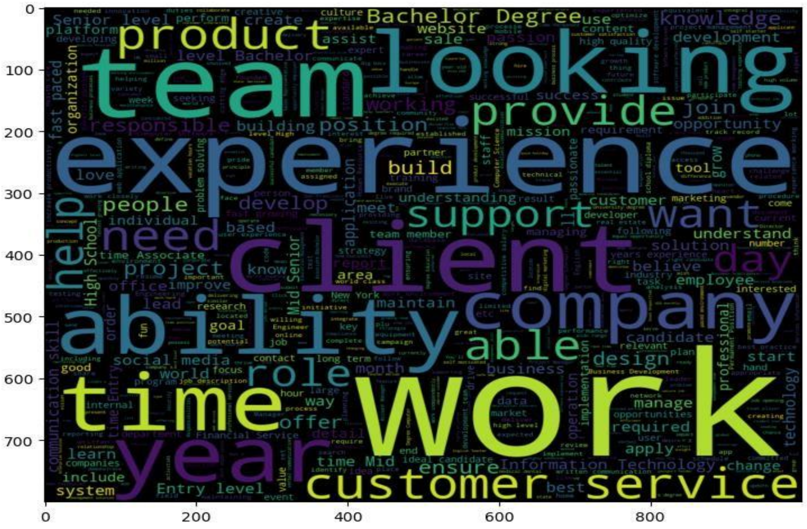
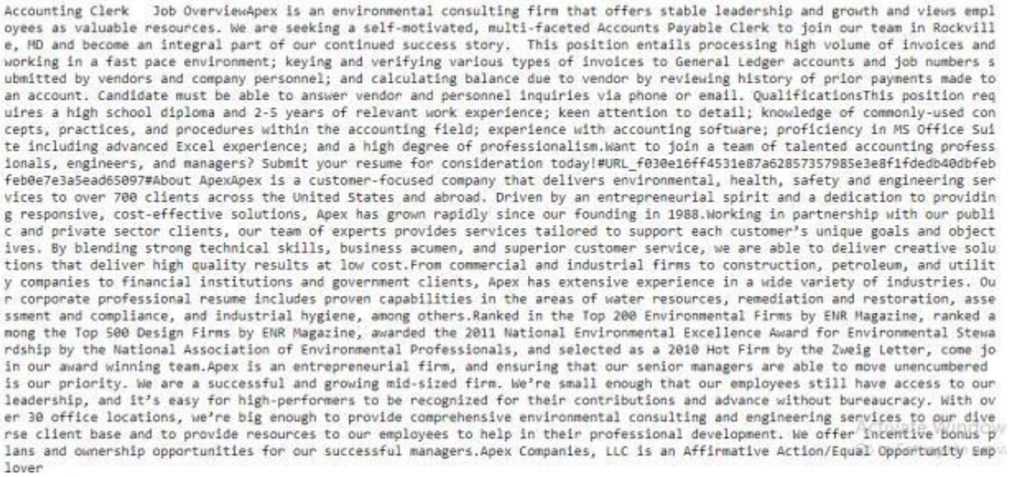


**Fig 8:** Word Cloud of Real Jobs

## VII. PRE-PROCESSING

In Job Scam Detection using machine learning, pre-processing is a critical step to refine the dataset and optimize it for effective model training. This involves handling missing values, addressing outliers, and encoding categorical variables to ensure a comprehensive and balanced dataset. Textual data from job postings undergoes normalization and stemming for uniformity, aiding in the extraction of meaningful features. Feature engineering is employed to enhance the model's ability to discern patterns, and data scaling ensures that features are on a consistent scale. The dataset is then split into training and testing sets for model evaluation. Additionally, methods to address class imbalance are applied to prevent bias. This meticulous pre-processing lays the groundwork for robust machine learning models, enhancing their accuracy and reliability in identifying potential job scams.

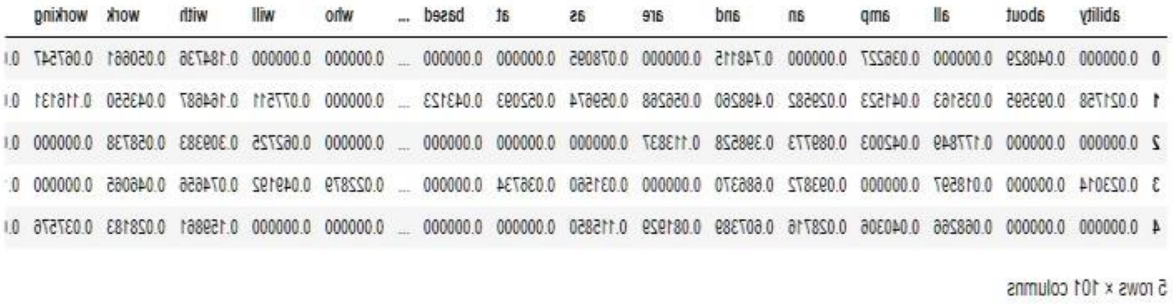


```
Accounting Clerk   Job OverviewApex is an environmental consulting firm that offers stable leadership and growth and views empl
oyees as valuable resources. We are seeking a self-motivated, multi-faceted Accounts Payable Clerk to join our team in Rockvill
e, MD and become an integral part of our continued success story.  This position entails processing high volume of invoices and
working in a fast pace environment; keying and verifying various types of invoices to General Ledger accounts and job numbers s
ubmitted by vendors and company personnel; and calculating balance due to vendor by reviewing history of prior payments made to
an account. Candidate must be able to answer vendor and personnel inquiries via phone or email. QualificationsThis position req
uires a high school diploma and 2-5 years of relevant work experience; keen attention to detail; knowledge of commonly-used con
cepts, practices, and procedures within the accounting field; experience with accounting software; proficiency in MS Office Sui
te including advanced Excel experience; and a high degree of professionalism.Want to join a team of talented accounting profess
ionals, engineers, and managers? Submit your resume for consideration today!#URL_f030e16ff4531e87a62857357985e3e8f1fdedb40dbfeb
feb0e7e3a5ead65097#About ApexApex is a customer-focused company that delivers environmental, health, safety and engineering ser
vices to over 700 clients across the United States and abroad. Driven by an entrepreneurial spirit and a dedication to providin
g responsive, cost-effective solutions, Apex has grown rapidly since our founding in 1988.Working in partnership with our publi
c and private sector clients, our team of experts provides services tailored to support each customer's unique goals and object
ives. By blending strong technical skills, business acumen, and superior customer service, we are able to deliver creative solu
tions that deliver high quality results at low cost.From commercial and industrial firms to construction, petroleum, and utilit
y companies to financial institutions and government clients, Apex has extensive experience in a wide variety of industries. Ou
r corporate professional resume includes proven capabilities in the areas of water resources, remediation and restoration, asse
ssment and compliance, and industrial hygiene, among others.Ranked in the Top 200 Environmental Firms by ENR Magazine, ranked a
mong the Top 500 Design Firms by ENR Magazine, awarded the 2011 National Environmental Excellence Award for Environmental Stewa
rdship by the National Association of Environmental Professionals, and selected as a 2010 Hot Firm by the Zweig Letter, come jo
in our award winning team.Apex is an entrepreneurial firm, and ensuring that our senior managers are able to move unencumbered
is our priority. We are a successful and growing mid-sized firm. We're small enough that our employees still have access to our
leadership, and it's easy for high-performers to be recognized for their contributions and advance without bureaucracy. With ov
er 30 office locations, we're big enough to provide comprehensive environmental consulting and engineering services to our dive
rse client base and to provide resources to our employees to help in their professional development. We offer Incentive bonus p
lans and ownership opportunities for our successful managers.Apex Companies, LLC is an Affirmative Action/Equal Opportunity Emp
loyer
```

**Fig 9:** Pre-Processing

## VIII. MODELLING

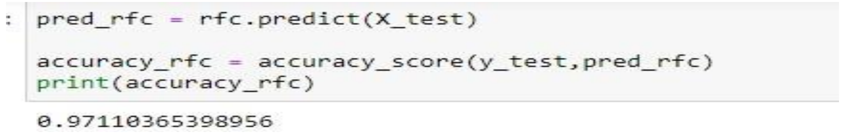
In the modelling phase of Job Scam Detection, machine learning algorithms are deployed to analyze the pre-processed dataset and identify patterns indicative of fraudulent job postings. Commonly used algorithms such as Random Forest Classifier, Naive Bayes, and Support Vector Machine are applied to learn from the training data and make predictions on new instances. Ensemble methods, like Random Forest, enhance model robustness by combining multiple classifiers. The choice of algorithms depends on the nature of the dataset and the specific characteristics of job scam patterns. Model hyper parameters are fine-tuned to optimize performance, and cross-validation is often employed to assess generalization to new data. The ultimate goal is to develop a predictive model capable of accurately distinguishing between legitimate and potentially fraudulent job opportunities, contributing to a more secure online recruitment environment.



| | ability | about | all | amp | an | and | are | as | at | based | ... | who | will | with | work | working | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.040829 | 0.000000 | 0.036227 | 0.000000 | 0.748115 | 0.000000 | 0.078095 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.184736 | 0.050661 | 0.067547 | 0.1 |
| 1 | 0.021758 | 0.093595 | 0.035163 | 0.041523 | 0.029582 | 0.498260 | 0.056268 | 0.059674 | 0.052093 | 0.043123 | ... | 0.000000 | 0.077511 | 0.164687 | 0.043550 | 0.116131 | 0.1 |
| 2 | 0.000000 | 0.000000 | 0.177849 | 0.042003 | 0.089773 | 0.398528 | 0.113837 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.062725 | 0.309383 | 0.058738 | 0.000000 | 0.1 |
| 3 | 0.023014 | 0.000000 | 0.018597 | 0.000000 | 0.093872 | 0.688370 | 0.000000 | 0.031560 | 0.036734 | 0.000000 | ... | 0.022879 | 0.049192 | 0.074656 | 0.046065 | 0.000000 | 0. |
| 4 | 0.000000 | 0.000000 | 0.068256 | 0.040306 | 0.028716 | 0.607389 | 0.081929 | 0.115850 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.159861 | 0.028183 | 0.037576 | 0.1 |

5 rows × 101 columns

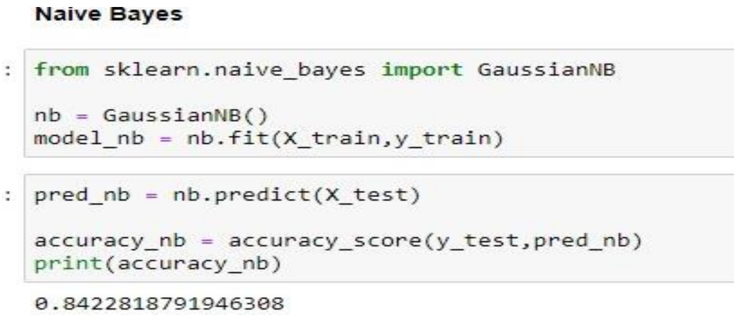
**Fig 10:** Modelling



```
pred_rfc = rfc.predict(X_test)

accuracy_rfc = accuracy_score(y_test,pred_rfc)
print(accuracy_rfc)
```

```
0.97110365398956
```

**Fig 11:** Final Model Random Forest Accuracy Report



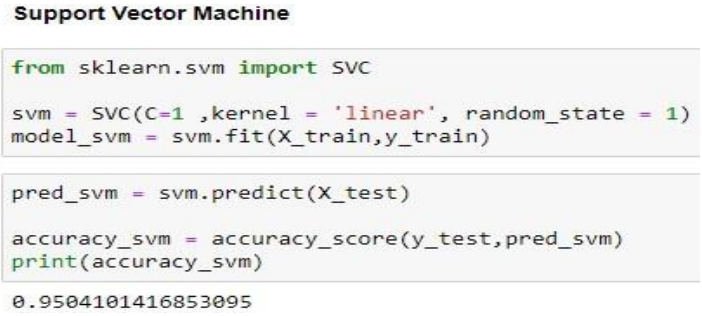
Naive Bayes

```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()
model_nb = nb.fit(X_train,y_train)
```

```
pred_nb = nb.predict(X_test)

accuracy_nb = accuracy_score(y_test,pred_nb)
print(accuracy_nb)
```

```
0.8422818791946308
```

**Fig 12:** Final Model Naive Bayes Accuracy Report



Support Vector Machine

```
from sklearn.svm import SVC

svm = SVC(C=1 ,kernel = 'linear', random_state = 1)
model_svm = svm.fit(X_train,y_train)
```

```
pred_svm = svm.predict(X_test)

accuracy_svm = accuracy_score(y_test,pred_svm)
print(accuracy_svm)
```

```
0.9504101416853095
```

**Fig 13:** Final Model Support Vector Machine Accuracy Report

## IX. CONCLUSION

In conclusion, this study tackles the Urgent issue of fraudulent job postings in Online Recruitment Frauds (ORF) by deploying machine learning classifiers, including Random Forest, Naive Bayes, and Support Vector Machine. The proposed Job Scam Detection system, relying on ensemble classifiers, effectively identifies scams, serving as a vigilant guardian in sifting through job posts to discern authentic advertisements and eliminate fake listings. The meticulously imported and pre-processed dataset establishes a robust foundation for machine learning models, incorporating crucial techniques such as data cleaning, visualization, and word cloud analysis. These approaches significantly contribute to the interpretability and effectiveness of the Job Scam Detection system, aiding in the distinction between legitimate and potentially fraudulent job opportunities. The final modelling phase underscores the accuracy of algorithms, particularly the resilience of the ensemble approach. This research plays a substantial role in fortifying the online job recruitment landscape, highlighting the pivotal contribution of machine learning to create a secure online recruitment environment.

## X. REFERENCE

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

[2] 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] 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.

[5] S. M. Vieira, U. Kaymak, and J. M. C. Sousa, ―Cohen's kappa coefficient as a s2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447.

[6] 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. lNotes Bioinformatics), vol. 6713 LNCS, pp. 350– 359, 2011, doi: 10.1007/978-3-642-21557-5_37.

[7] 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.