iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39580

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Classification of Genuinity in Job Posting Using Machine Learning

Charan Lokku[1], Kesava Naga Sivaram Kolli[2], Santhosh Puganuru[3]
[1,2,3]*Student, Department of Computer Science and Engineering, SRM Institute of Science and Technology*

Abstract*: *To avoid fraudulent Job postings on the internet, we target to minimize the number of such frauds through the Machine Learning approach to predict the chances of a job being fake so that the candidate can stay alert and make informed decisions if required. The model will use NLP to analyze the sentiments and pattern in the job posting and TF-IDF vectorizer for feature extraction. In this model, we are going to use Synthetic Minority Oversampling Technique (SMOTE) to balance the data and for classification, we used Random Forest to predict output with high accuracy, even for the large dataset it runs efficiently, and it enhances the accuracy of the model and prevents the overfitting issue. The final model will take in any relevant job posting data and produce a result determining whether the job is real or fake.

Keywords*: *Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF), Synthetic Minority Oversampling Technique (SMOTE), Random Forest.

I. INTRODUCTION

We are living in an unprecedented time due to COVID-19 Pandemic, hurting economics in every continent. Unemployment rates are increasing every single day. In these desperate times, when thousands and millions of people are on the lookout for a job, it provides a perfect opportunity for online scammers to take advantage of their desperation. We see a daily rise in these fake job postings where the posting seems reasonable, often these companies will have a website as well, and they will have a recruitment process that is like other companies in the industry. If one looks hard enough, one can spot the differences between these fake postings and genuine ones. Most of the time these postings don't have a company logo on these postings, the initial response from the company is from an unofficial email account, or during an interview they might ask you for personal confidential information such as your credit card details by saying they need it for personnel verification. In normal economic conditions, all these are evident hints that there is something suspicious about the company, but these are not normal economic conditions. These are the worst times we all have seen in our lifetimes, and at this time, desperate individuals just need a job, and by this, these individuals are directly playing into the hands of these scammers. The current market situation has led to high unemployment. Economic stress and the coronavirus's impact have significantly reduced job availability and job loss for many individuals. A case like this presents an appropriate opportunity for scammers. Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. Most scammers do this to get personal information from the person they are scamming. Personal information can contain addresses, bank account details, social security numbers, etc. In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the jobseekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. For this purpose, a machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user.

II. RELATED WORK

FHA. Shibly, Uzzal Sharma and HMM [1]. Used two-class decision boosted tree and two-class decision forest algorithms for classification of data to determine which of these two algorithms perform well. They concluded that the efficiency of the two-class boosted decision tree is more than the two-class decision forest algorithm. The limitation of this paper is it takes more time to train the model and it has many hyperparameters so it can overfit.

Prof. R. S. Shishupal, Varsha, Supriya Mane, Vinita Singh, Damini Wasekar [2]. Implemented android based application to compare various classifiers results for prediction of fake job profiles. In this paper, they used Multinomial Naive Bayes, Android, Flask API, Blender, NLP. This model takes input as text data and gives results in the form of speech. But it cannot classify numeric data as it takes text data as input.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Shawni Dutta and Prof.Samir Kumar Bandyopadhyay [3]. Concluded that ensembling classification gives better results than the single classifier. In this paper, they used Naive Bayes, Decision Tree, Multi-Layer Perceptron Classifier, K-nearest Neighbor, AdaBoost, Gradient Boost, and Random Forest Classifiers to compare the best model among the given classifiers. It determines the best classification algorithm among others. But it is computationally expensive and lacks interpretability.

Ibrahim M. Nasser and Amjad H. Alzaanin [4]. Used Multinomial Naive Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbors, Random Forest algorithms for classification which determines the best classification algorithm among others. And uses TF-IDF vectorizer for feature extraction. It is a simple model with effective accuracy. It does not have balanced data so there will be a problem of underfitting and data loss to do the classification.

Bandar Alghamdi, Fahad Alharby [5]. Used Support vector machine, random forest classifier, and data mining tools. This research is an empirical study based on observation, testing, and evaluation. Weka tool is used to implement and evaluate the performance of the proposed model. Gives efficient accuracy and enhances the model by using data mining techniques. It couldn't analyze the company profile, company logo, and required main string attributes.

Okti Nindyati, I Gusti Bagus, Baskara Nugraha [6]. This research explores the issue of employment scam detection. The contribution of this research is to create a new dataset namely IESD and propose behavioral context-based features to determine whether a job advertisement is legitimate or fraudulent. As mentioned in the result, behavioral features can improve the performance of employment scam detection. It reached 90% inaccuracy. They used Naive Bayes, K-nearest Neighbor, Logistic regression, decision tree, neural networks, support vector machine. Balanced data and effective identification of job vacancies. The limitations of this paper they classified a very small dataset containing data from different fields.

Sangeeta Lal, Rishabh Jaiswal, Neethu Sardana, Ayushi Verma, Amanpreeth Kaur, Rahul Mourya [7]. In this paper, they proposed an ensemble-based model ORF Detector for ORF(Online Fraud Detection) detection. We have taken three baselines classifies, J48, Logistic Regression (LR), and Random Forest (RF). We applied three ensemble techniques Average Vote (AV), Majority Vote (MV), and Maximum Vote (MXV) on these baseline classifiers to build the ORF Detector framework and evaluated the proposed ORF Detector model on a publicly available dataset. The proposed model is found to be effective and gives an average f1-score and accuracy of 94% and 95.4, respectively. But model suffers from a lack of interpretability and is usually computationally expensive.

Elsevier B V [8]. Explained the performance of various machine learning techniques in the detection of financial frauds using Classification and Regression Tree, Naïve Bayes, K-Nearest Neighbor. In this paper, it was found that hybrid fraud detection techniques are the most used, as they combine the strengths of several traditional detection methods and the studies do not smother all types of fraud, and each type of fraud has constraints specific to it; response required in real-time, text analysis. In this model, It can handle billions of transactions and respond at lightning speed with absolute accuracy. As there are massive amounts of data involved, businesses also need to invest in data storage and management.

III. METHODOLOGY

A. Dataset and Data Preprocessing

The data for this project is available at Kaggle. The dataset consists of 17,880 observations and 18 features. The structure of the dataset is shown in Fig.1

```
job_id                 int64
title                 object
location              object
department            object
salary_range          object
company_profile       object
description           object
requirements          object
benefits              object
telecommuting          int64
has_company_logo       int64
has_questions          int64
employment_type       object
required_experience   object
required_education    object
industry              object
function              object
fraudulent             int64
```

Fig 1: Structure of the Dataset

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

We aim to find the fake job postings from the advertisements of the given dataset. First, we go through the data preprocessing as shown in Figure.2. It is the text cleaning process necessary to highlight attributes that are required in the model so, we remove the unnecessary data like missing values and other techniques like stop word removal, splitting of words, and analyzing the accurate meaning of the data using nltk(Natural Language Toolkit) libraries

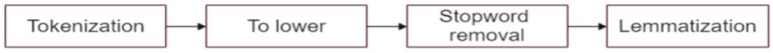


Fig 2: Data Preprocessing

1) *Tokenization:* The textual data is split into smaller units. In this case, the data is split into words. For example, Plata o Plomo-> 'Plata','o','Plomo'.
2) *To Lower:* The split words are converted to lowercase
3) *Stop word Removal:* Stop words are words that do not add much meaning to sentences. For example, the, a, an, he, have, etc. These words are removed.
4) *Lemmatization:* The process of lemmatization groups in which inflected forms of words are used together.

B. TF-IDF (Term Frequency-Inverse Document Frequency)

It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. The TF-IDF value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods,

1) *Term Frequency (TF):* It is a measure of the frequency of a word (w) in a document (d). TF is defined as the ratio of a word's occurrence in a document to the total number of words in a document. The denominator term in the formula is to normalize since all the corpus documents are of different lengths.

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

2) *Inverse Document Frequency (IDF):* It is the measure of the importance of a word. Term frequency (TF) does not consider the importance of words. Some words such as' of', 'and', etc. can be most frequently present but are of little significance. IDF provides weightage to each word based on its frequency in corpus D. It is the product of TF and IDF. TFIDF gives more weightage to the word that is rare in the corpus (all the documents). TFIDF provides more importance to the word that is more frequent in the document as shown in the below figure 3. The table shows the sample data which contains the higher term importance on different data labels and its weightage in the particular labeled data.

$$IDF(w,D) = \ln\left(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w}\right)$$

	abil	account	also	amp	applic	base	benefit	best	build	busi	...	year	telecommuting	has_company_logo	has_qu
0	0.000000	0.000000	0.116434	0.102818	0.000000	0.000000	0.000000	0.119893	0.000000	0.091077	...	0.000000	0	1	
1	0.000000	0.041469	0.033522	0.059203	0.000000	0.061404	0.000000	0.000000	0.000000	0.104886	...	0.000000	0	1	
2	0.000000	0.000000	0.000000	0.092825	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0	1	
3	0.039152	0.566401	0.000000	0.000000	0.000000	0.000000	0.089800	0.000000	0.118249	0.293025	...	0.058842	0	1	
4	0.000000	0.109922	0.000000	0.000000	0.086042	0.000000	0.095852	0.182991	0.000000	0.069505	...	0.062808	0	1	

Fig 3: Term Frequency Vector Model

C. SMOTE (Synthetic Minority Oversampling Technique) – Oversampling

In Machine Learning and Data Science, we often come across a term called Imbalanced Data Distribution, which generally happens when observations in one of the classes are much higher or lower than the other classes. As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution. This problem is prevalent in examples such as Fraud Detection, Anomaly Detection, Facial recognition, etc. Standard ML techniques such as Decision Tree and Logistic Regression have a bias towards the majority class, and they tend to ignore the minority class. They tend only to predict the majority class, hence, having major misclassification of the minority class in comparison with the majority class. In more technical words, if we have imbalanced data distribution in our dataset then our model becomes more prone to the case when the minority class has a negligible or very lesser recall.

SMOTE Algorithm

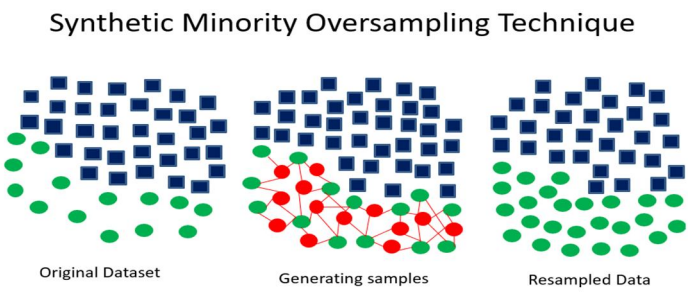


Fig 4: Data Resampling

It aims to balance class distribution by randomly increasing minority class examples by replicating them as shown in the above Fig 4.

SMOTE synthesizes new minority instances between existing minority instances. It generates the virtual training records by linear interpolation for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed, and several classification models can be applied to the processed data as it undergoes the following steps as given below.

1) *Step 1:* Setting the minority class for original dataset, for each, the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and every other sample in set.
2) *Step 2:* The sampling rate N is set according to the imbalanced proportion. For each N examples (i.e x1, x2, …xn) are randomly selected from its k-nearest neighbors, and they construct the set.
3) *Step 3:* For each example (k=1, 2, 3…N), the following formula is used to generate a new example: in which rand (0, 1) represents the random number between 0 and 1

$$x' = x + rand(0,1) * | x - x_k |$$

D. Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and predicts the final output. As shown the random forest algorithm follows the steps given below

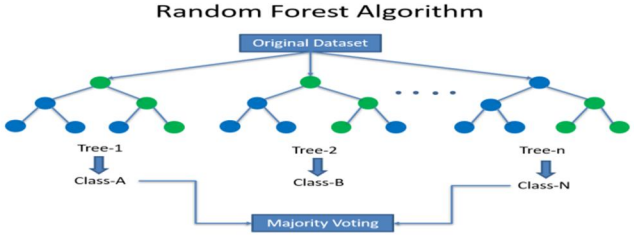


Fig 5: Random Forest Algorithm

1) *Step-1:* Select random K data points from the training set.
2) *Step-2:* Build the decision trees associated with the selected data points (Subsets).
3) *Step-3:* Choose the number N for decision trees that you want to build.
4) *Step-4:* Repeat Step 1 & 2.
5) *Step-5:* For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

IV. SYSTEM ARCHITECTURE

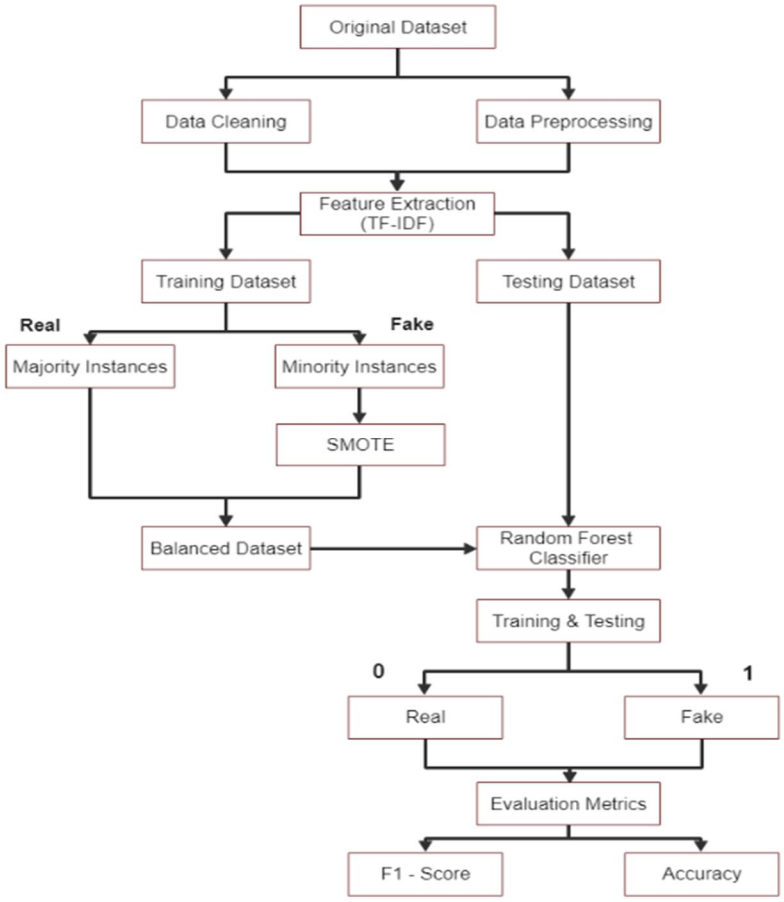


Fig 6: System Architecture

In the above system architecture diagram, we have taken the original dataset then we have done feature extraction after cleaning and preprocessing the data. We split the extracted data into training and testing datasets. The training dataset consists of majority and minority instances, we apply the SMOTE (Synthetic Minority Oversampling Technique) to the minority instances to balance the dataset. we apply the random forest algorithm for the classification of the train and test split consisting of real and fake data. Finally, we get the evaluation metrics F1-score and accuracy of the dataset.

V. IMPLEMENTATION

A. Training & Testing Datasets

The dataset we had taken for the classification consists of 17014 real data and 866 fraudulent data. After feature extraction, we split the data into both training and testing datasets. We consider the training dataset as 75% and the testing dataset as 25% for the classification process.

B. Data Balancing

The dataset consists of 94% real data and 6% fake data which is highly imbalanced, to get better results we need to balance the data. Consider the real data as majority samples and the fake data as minority samples, in order to balance the data, we use SMOTE to the minority sample data then the data will be balanced by replicating the virtual training samples and synthesizing them by selecting randomly generated samples. As we can see in Fig.7 below

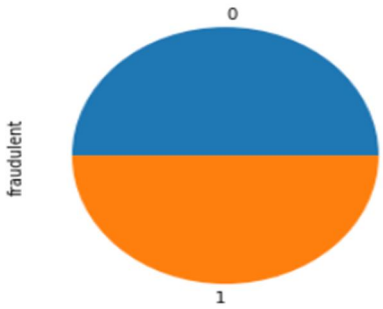


Fig 7: Balanced Dataset

C. Evaluation Metrics

The models will be evaluated based on two metrics:

1) Accuracy:

2) This metric is defined by this formula

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

3) F1-Score: F1 score is a measure of a model's accuracy on a dataset. The formula for this metric is

$$F_1 = \frac{True\ Positive}{True\ Positive + \frac{1}{2}(False\ Positive + False\ Negative)}$$

A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class.

A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class.

D. Results

According to Fig.2 we used a Random Forest classifier to train and test the model for detecting the real and fake job postings for the given dataset. The following Table.1 shows the confusion matrix of the model with respect to Fig.8.

Table- I: Confusion Matrix

RF Confusion Matrix		
	Real (Predicted)	Fake (Predicted)
Real (Actual)	4215 (TP)	26 (FN)
Fake (Actual)	31 (FP)	4232 (TN)

	precision	recall	f1-score	support
0	0.99	0.99	0.99	4241
1	0.99	0.99	0.99	4263
accuracy			0.99	8504
macro avg	0.99	0.99	0.99	8504
weighted avg	0.99	0.99	0.99	8504

Fig 8: Confusion Matrix

Table.2 shows the results of the classifier based on Fig.9 with respect to the evaluation metrics in terms of precision, recall, f1-score, and overall accuracy of the model.

Table- II: Evaluation Metrics Results

	Precision	Recall	F1- Score	Accuracy
RF	0.99	0.99	0.99	0.99

```
confusion_matrix(y_test, rfc_predict)

array([[4215,   26],
       [  31, 4232]], dtype=int64)
```

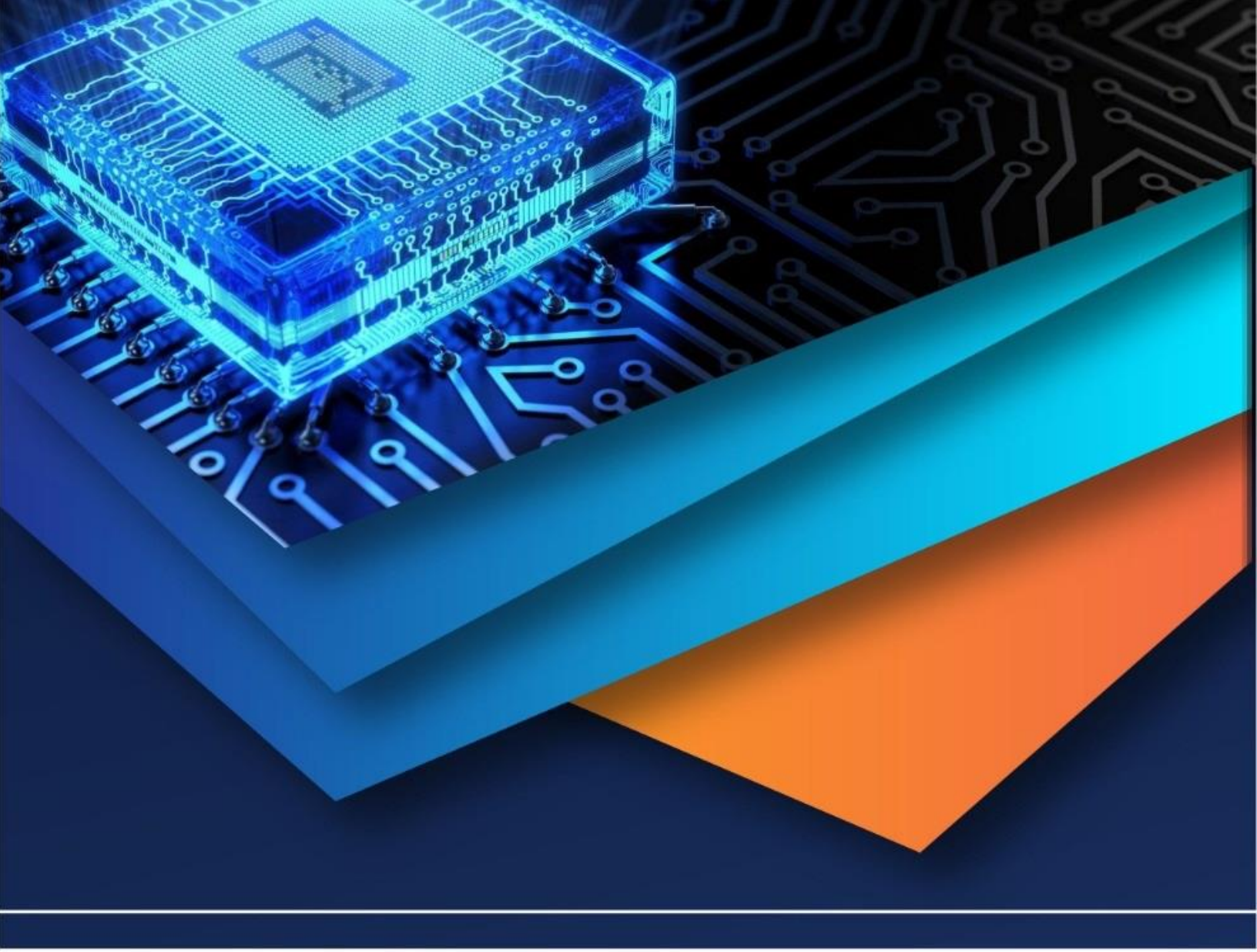
Fig 9: Evaluation metrics

VI. CONCLUSION

Based on our research and work it shows that the machine learning approach is the best way to guide job seekers from the traps of scammers. There are many supervised algorithms are used for classification but results show that Random Forest has produced the best results in classifying the real and fake job postings, as it works more efficiently for large datasets. Another important factor is the oversampling technique improved the overall accuracy by balancing the dataset we got the best performance in every evaluation metric. Finally, we got 99% overall accuracy and precision which is higher than the existing works.

REFERENCES

[1] Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.
[2] Nasser, Ibrahim & Alzaanin, Amjad. (2020). Machine Learning and Job Posting Classification: A Comparative Study. 4. 6-14.
[3] FHA. Shibly, Uzzal Sharma, HMM. Naleer, "Performance Comparison of Two-Class Boosted Decision Tree snd Two-Class Decision Forest Algorithms in Predicting Fake Job Postings", *Annals of RSCB*, pp. 2462 –, Apr. 2021.
[4] Alghamdi, Bandar & Alharby, Fahad. (2019). An Intelligent Model for Online Recruitment Fraud Detection. Journal of Information Security. 10. 155-176. 10.4236/jis.2019.103009.
[5] Nindyati, Okti & Nugraha, I. (2019). Detecting Scam in Online Job Vacancy Using Behavioral Features Extraction. 1-4. 10.1109/ICISS48059.2019.8969842.
[6] Shishupal, Prof & Varsha, & Mane, Supriya & Singh, Vinita & Wasekar, Damini. (2021). Efficient Implementation using Multinomial Naive Bayes for Prediction of Fake Job Profile. International Journal of Advanced Research in Science, Communication and Technology. 286-291. 10.48175/IJARSCT-1241.
[7] Reis, Julio & Correia, Andre & Murai, Fabricio & Veloso, Adriano & Benevenuto, Fabrício & Cambria, Erik. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems. 34. 76-81. 10.1109/MIS.2019.2899143.
[8] Athira Das | Swati Ashok Desale "**Techniques to Analyse, Identify & Verify the Online Job Offers by Fake Companies Worldwide**" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-2 | Issue-4, June 2018, pp.2660-2663.
[9] Hemamou, Leo & Felhi, Ghazi & Vandenbussche, Vincent & Martin, Jean-Claude & Clavel, Chloé. (2019). HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 573-581. 10.1609/aaai.v33i01.3301573.
[10] Sangeeta, Sangeeta & Jiaswal, Rishabh & Sardana, Neetu & Verma, Ayushi & Kaur, Amanpreet & Mourya, Rahul. (2019). ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection. 1-5. 10.1109/IC3.2019.8844879.
[11] Sadgali, Imane & Sael, Nawal & Benabbou, Faouzia. (2019). Performance of machine learning techniques in the detection of financial frauds.
[12] Vidros, Sokratis & Kolias, Constantinos & Kambourakis, Georgios & Akoglu, Leman. (2017). Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset. Future Internet. 9. 6. 10.3390/fi9010006.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY