# Reveal Online Fake Job Advert Detection Application using Machine learning

Dr. A.LAXMIKANTH, Associate Professor

,V. BHANU PRASAD, A. VAMSHISRI, K. GOPI,

S. UDAY

SRI INDU COLLEGE OF ENGINEERING AND TECHNOLOGY

Sheriguda (V), Ibrahimpatnam (M), RangareddyDist – 501 510

## ABSTRACT

To avoid fraudulent post for job in the internet, an automated tool using machine learning basedclassification techniques is proposed in the paper. Different classifiers are used for checkingfraudulent post in the web and the results of those classifiers are compared for identifying thebest employment scam detection model. It helps in detecting fake job posts from an enormousnumber of posts. Two major types of classifiers, such as single classifier and ensembleclassifiers are considered for fraudulent job posts detection. However, experimental resultsindicate that ensemble classifiers are the best classification to detect scams over the singleclassifiers.

**INDEX :** job, internet, machine learning, classifiers, scams, single classifiers

## 1. INTRODUCTION

Employment scam is one of the serious issues in recent times addressed in the domain ofOnline Recruitment Frauds (ORF). In recent days, many companies prefer to post theirvacancies online so that these can be accessed easily and timely by the job-seekers. However,this intention may be one type of scam by the fraud people because they offer employment tojob-seekers in terms of taking money from them. Fraudulent job advertisements can be postedagainst a reputed company for violating their credibility. These fraudulent job post detectiondraws a good attention for obtaining an automated tool for identifying fake jobs and reportingthem to people for avoiding application for such jobs.For this purpose, machine learning approach is applied which employs several classificationalgorithms for recognizing fake posts. In this case, a classification tool isolates fake job postsfrom a larger set of job advertisements and alerts the user. To address the problem of identifyingscams on job posting, supervised learning algorithm as classification techniques are consideredinitially. A classifier maps input variable to target classes by considering training data.

Classifiers addressed in the paper for identifying fake job posts from the others are describedbriefly. These classifiers based prediction may be broadly categorized into -Single Classifierbased Prediction and Ensemble Classifiers based Prediction.

### A. Single Classifier based Prediction

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

#### a) Naive Bayes Classifier-

The Naive Bayes classifier is a supervised classification tool that exploits the concept ofBayes Theorem of Conditional Probability. The decision made by this classifier is quiteeffective in practice even if its probability estimates are inaccurate. This classifierobtains a very promising result in the following scenario- when the features areindependent or features are completely functionally dependent. The accuracy of thisclassifier is not related to feature dependencies rather than it is the amount ofinformation loss of the class due to the independence assumption is needed to predictthe1ptimizey.

#### b) Multi-Layer Perceptron Classifier –

Multi-layer perceptron can be used as supervised classification tool by incorporatingoptimized training parameters. For a given problem, the number of hidden layers in amultilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture.

#### c) K-nearest Neighbor Classifier-

K-Nearest Neighbour Classifiers, often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers knumber of objects as the nearest object while determining the class. The main challengeof this classification technique relies on choosing the appropriate value of k.

#### d) Decision Tree Classifier-

A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure. It gains nowledge on classification. Each target class is denoted as a leaf node of DT and non- leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Starting from the beginning at the root this tree are going through it until a leaf node is reached. It is the way of obtaining classification result from a decision tree. Decision tree learning is an approach that has been applied to spam filtering. This can be useful for forecasting the goal based on some criterion by implementing and training this model.

## 2. LITERATURE SURVEY

**TITLE: “An Intelligent Model for Online Recruitment Fraud Detection,”**

**ABSTRACT:** This study research attempts to prohibit privacy and loss of money forindividuals and organization by creating a reliable model which can detect the fraud exposurein the online recruitment environments. This research presents a major contribution representedin a reliable detection model using ensemble approach based on Random forest classifier todetect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud ischaracterized by other types of electronic fraud detection by its modern and the scarcity ofstudies on this concept. The researcher proposed the detection model to achieve the objectivesof this study. For feature selection, support vector machine method is used and for classificationand detection, ensemble classifier using Random Forest is employed. A freely available datasetcalled Employment Scam Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The resultsshowed an obtained accuracy of 97.41%. Further, the findings presented the main features andimportant factors in detection purpose include having a company profile feature, having acompany logo feature and an industry feature.

**TITLE: An Empirical Study of the Naïve Bayes Classifier An empirical study of the naïve Bayes classifier,**

**ABSTRACT:** The naive Bayes classifier greatly simplify learn-ing by assuming that featuresare independent given class. Although independence is generally a poor assumption, in practicenaive Bayes often competes well with more sophisticated classifiers. Our broad goal is tounderstand the data character-istics which affect the performance of naive Bayes. Our approachuses Monte Carlo simulations that al-low a systematic study of classification accuracy forseveral classes of randomly generated prob-lems. We analyze the impact of the distributionentropy on the classification error, showing that low-entropy feature distributions yield goodper-formance of naive Bayes. We also demonstrate that naive Bayes works well for certainnearly-functional feature dependencies, thus reaching its best performance in two oppositecases: completely independent features (as expected) and function-ally dependent features(which is surprising). An-other surprising result is that the accuracy of naive Bayes is notdirectly correlated with the degree of feature dependencies measured as the class-conditionalmutual information between the fea-tures. Instead, a better predictor of naive Bayes ac-curacyis the amount of information about the class that is lost because of the independence assump-tion.

**TITLE: Bayes's Theorem and the Analysis of Binomial Random Variables,**

**ABSTRACT:** A very practical application of Bayes's theorem, for the analysis of binomialrandom variables, is presented. Previous papers (Walters, 1985; Walters, 1986a) have alreadydemonstrated the reliability of the technique for one, or two random variables, and theextension of the approach to several random variables is described. Two biometrical examplesare used to illustrate the method.

**TITLE: Multilayer perceptrons for classification and regression,**

**ABSTRACT:** We review the theory and practice of the multilayer perceptron. We aim ataddressing a range of issues which are important from the point of view of applying thisapproach to practical problems. A number of examples are given, illustrating how themultilayer perceptron compares to alternative, conventional approaches. The application fieldsof classification and regression are especially considered. Questions of implementation, i.e. ofmultilayer perceptron architecture are discussed. Recent studies, which are particularly relevantto the areas of discriminant analysis, and function mapping, are cited.

**TITLE: K -Nearest Neighbour Classifiers,**

**ABSTRACT:** We analyze a Relational Neighbor (RN) classifier, a simple relational predictivemodel that predicts only based on class labels of related neighbors, using no learning and noinherent attributes. We show that it performs surprisingly well by comparing it to morecomplex models such as Probabilistic Relational Models and Relational Probability Trees onthree data sets from published work.

**TITLE: A Survey on Decision Tree Algorithms of Classification in Data Mining,**

**ABSTRACT:** As the computer technology and computer network technology are developing,the amount of data in information industry is getting higher and higher. It is necessary toanalyze this large amount of data and extract useful knowledge from it. Process of extractingthe useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called datamining. Decision tree classification technique is one of the most popular data miningtechniques. In decision tree divide and conquer technique is used as basic learning strategy. Adecision tree is a structure that includes a root node, branches, and leaf nodes. Each internalnode denotes a test on an attribute, each branch denotes the outcome of a test, and each leafnode holds a class label. The topmost node in the tree is the root node. This paper focus on thevarious algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges,advantage and disadvantage.

**TITLE: “Machine learning for email spam filtering: review, approaches and open researchproblems,**

**ABSTRACT:** The upsurge in the volume of unwanted emails called spam has created anintense need for the development of more dependable and robust antispam filters. Machinelearning methods of recent are being used to successfully detect and filter spam emails. Wepresent a systematic review of some of the popular machine learning based email spam filteringapproaches. Our review covers survey of the important concepts, attempts, efficiency, and theresearch trend in spam filtering. The preliminary discussion in the study background examinesthe applications of machine learning techniques to the email spam filtering process of theleading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters.Discussion on general email spam filtering process, and the various efforts by differentresearchers in combating spam through the use machine learning techniques was done. Ourreview compares the strengths and drawbacks of existing machine learning approaches and theopen research problems in spam filtering.

**TITLE: ST4_Method_Random_Forest,**

**ABSTRACT:** Several machine-learning algorithms have been proposed for remote sensingimage classification during the past two decades. Among these machine learning algorithms,Random Forest (RF) and Support Vector Machines (SVM) have drawn attention to imageclassification in several remote sensing applications. This paper reviews RF and SVM conceptsrelevant to remote sensing image classification and applies a meta-analysis of 251 peer-reviewed journal papers. A database with more than 40 quantitative and qualitative fields wasconstructed from these reviewed papers. The meta-analysis mainly focuses on: (1) the analysisregarding the general characteristics of the studies, such as geographical distribution, frequencyof the papers considering time, journals, application domains, and remote sensing softwarepackages used in the case studies, and (2) a comparative analysis regarding the performancesof RF and SVM classification against various parameters, such as data type, RS applications,spatial resolution, and the number of extracted features in the feature engineering step. Thechallenges, recommendations, and potential directions for future research are also discussed indetail. Moreover, a summary of the results is provided to aid researchers to customize theirefforts in order to achieve the most accurate results based on their thematic applications.

**TITLE: Bagging classifiers for fighting poisoning attacks in adversarial classification tasks,**

**ABSTRACT:** Pattern recognition systems have been widely used in adversarialclassification tasks like spam filtering and intrusion detection in computer networks. In theseapplications a malicious adversary may successfully mislead a classifier by “poisoning” itstraining data with carefully designed attacks. Bagging is a well-known ensemble constructionmethod, where each classifier in the ensemble is trained on a different bootstrap replicate ofthe training set. Recent work has shown that bagging can reduce the influence of outliers intraining data, especially if the most outlying observations are resampled with a lowerprobability. In this work we argue that poisoning attacks can be viewed as a particular categoryof outliers, and, thus, bagging ensembles may be effectively exploited against them. Weexperimentally assess the effectiveness of bagging on a real, widely used spam filter, and on aweb-based intrusion detection system. Our preliminary results suggest that bagging ensemblescan be a very promising defence strategy against poisoning attacks, and give us valuableinsights for future research work.

## 3. PROBLEM STATEMENT

According to several studies, Review spam detection, Email Spam detection, Fake newsdetection 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 mayguide other purchaser while choosing their products. In this context, spammers canmanipulate reviews for gaining profit and hence it is required to develop techniquesthat detects these spam reviews. This can be implemented by extracting features fromthe reviews by extracting features using Natural Language Processing (NLP). Next,machine learning techniques are applied on these features. Lexicon based approachesmay be one alternative to machine learning techniques that uses dictionary or corpusto eliminate spam reviews.

### 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. Toeradicate this problem, Gmail, Yahoo mail and Outlook service providers incorporatespam filters using Neural Networks. While addressing the problem of email spamdetection, content based filtering, case based filtering, heuristic based filtering,memory or instance based filtering, adaptive spam filtering approaches are taken intoconsideration.

### 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 fakenews is written, how fake news spreads, how a user is related to fake news. Featuresrelated to news content and social context are extracted and a machine learningmodels are imposed to recognize fake news.

## 4. Proposed System & Advantages:

The target of this study is to detect whether a job post is fraudulent or not. Identifying andeliminating these fake job advertisements will help the job seekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle is employed that provides informationregarding a job that may or may not be suspicious.

### A. Implementation of Classifiers

In this framework classifiers are trained using appropriate parameters. For maximizingthe performance of these models, default parameters may not be sufficient enough.Adjustment of these parameters enhances the reliability of this model which may beregarded as the 8ptimized one for identifying as well as isolating the fake job posts fromthe job seekers.

### B. Performance Evaluation Metrics

While evaluating performance skill of a model, it is necessary to employ some metrics tojustify the evaluation. For this purpose, following metrics are taken into considerationin order to identify the best relevant problem-solving approach. Accuracy is a metricthat identifies the ratio of true predictions over the total number of instancesconsidered. However, the accuracy may not be enough metric for evaluating model‘sperformance since it does not consider wrong predicted cases. If a fake post is treatedas a true one, it creates a significant problem. Hence, it is necessary to consider falsepositive and false negative cases that compensate to misclassification. For measuringthis compensation, precision and recall is quite necessary to be considered.

## 5. IMPLEMENTATION

There are 2 modules:

1. Admin

2. User or Candidate

Admin:-

□ Login

□ User Management

□ Pending Users

□ All User

□ Fake job

□ Upload Dataset

□ View Dataset

□ Algorithm

□ SVM Algorithm

□ Decision Tree Algorithm

□ Naïve Bayes Algorithm

□ K-NN Bayes Algorithm

□ Random Forest Algorithm

□ Graph Analysis

□ Comparison Graph

User:-

□ Register

- Login
- Predict

## 6. SYSTEM ARCHITECTURE

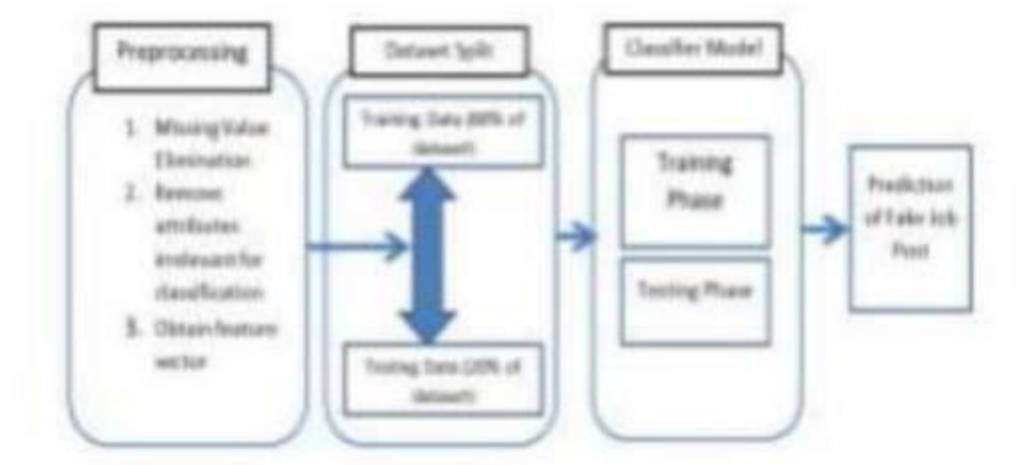
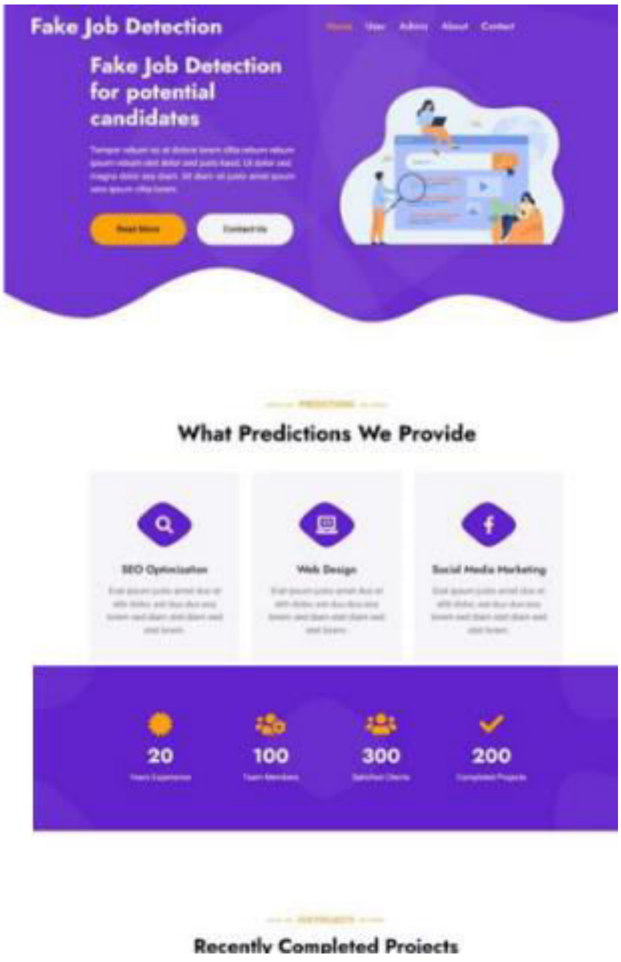

Preprocessing
1. Missing Value Elimination
2. Remove attributes irrelevant for classification
3. Obtain feature vector
Dataset Split
Training Data (80% of dataset)
Testing Data (20% of dataset)
Classifier Model
Training Phase
Testing Phase
Prediction of Fake Job Post

Fig. 2.Detailed description for working of Classifiers

## 7. OUTCOME RESULTS

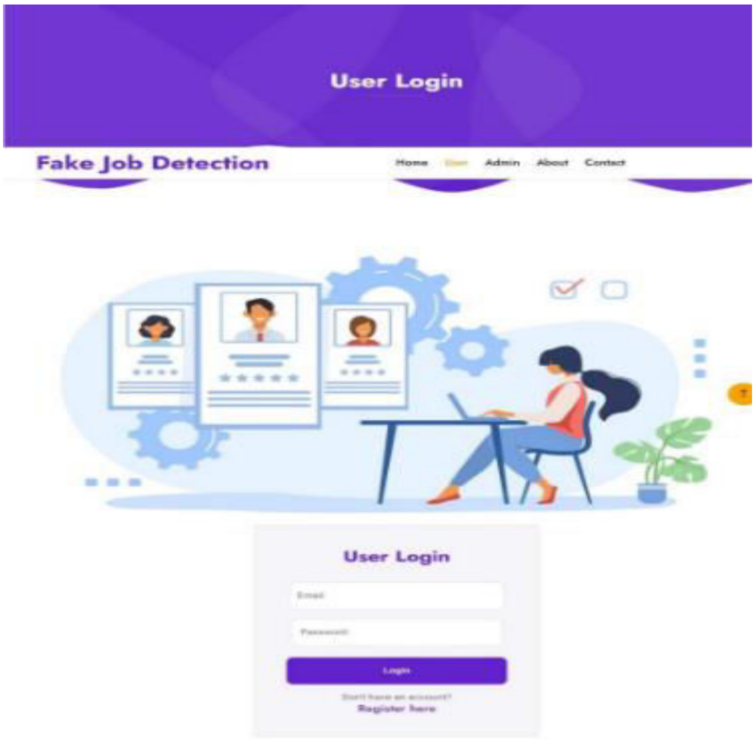


Fake Job Detection

Fake Job Detection for potential candidates

What Predictions We Provide

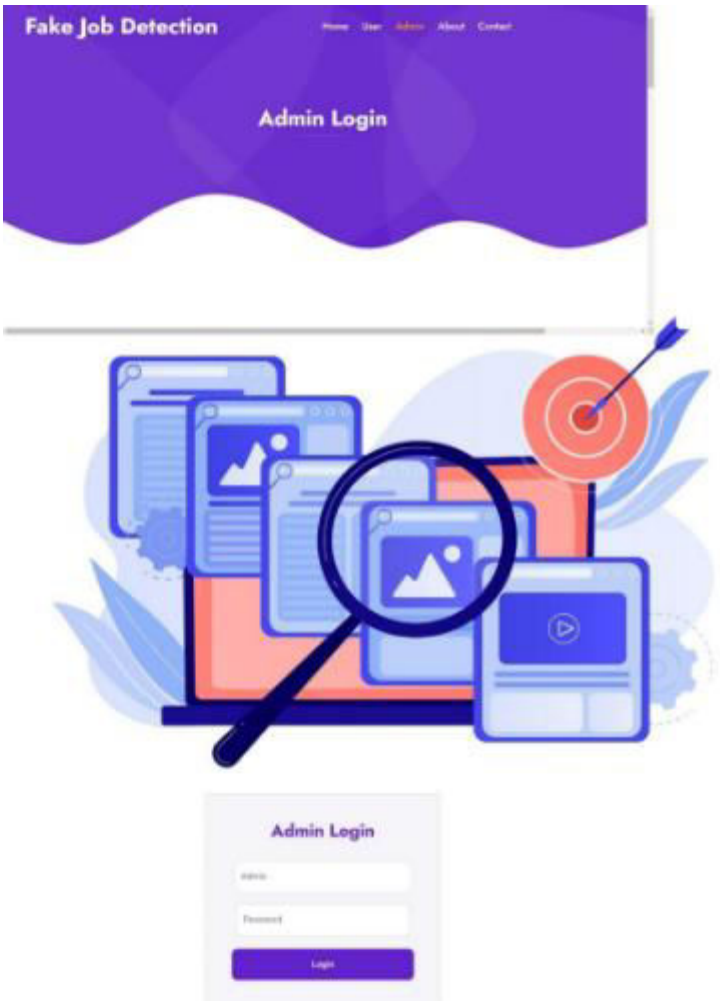
Recently Completed Projects



User Login

Fake Job Detection

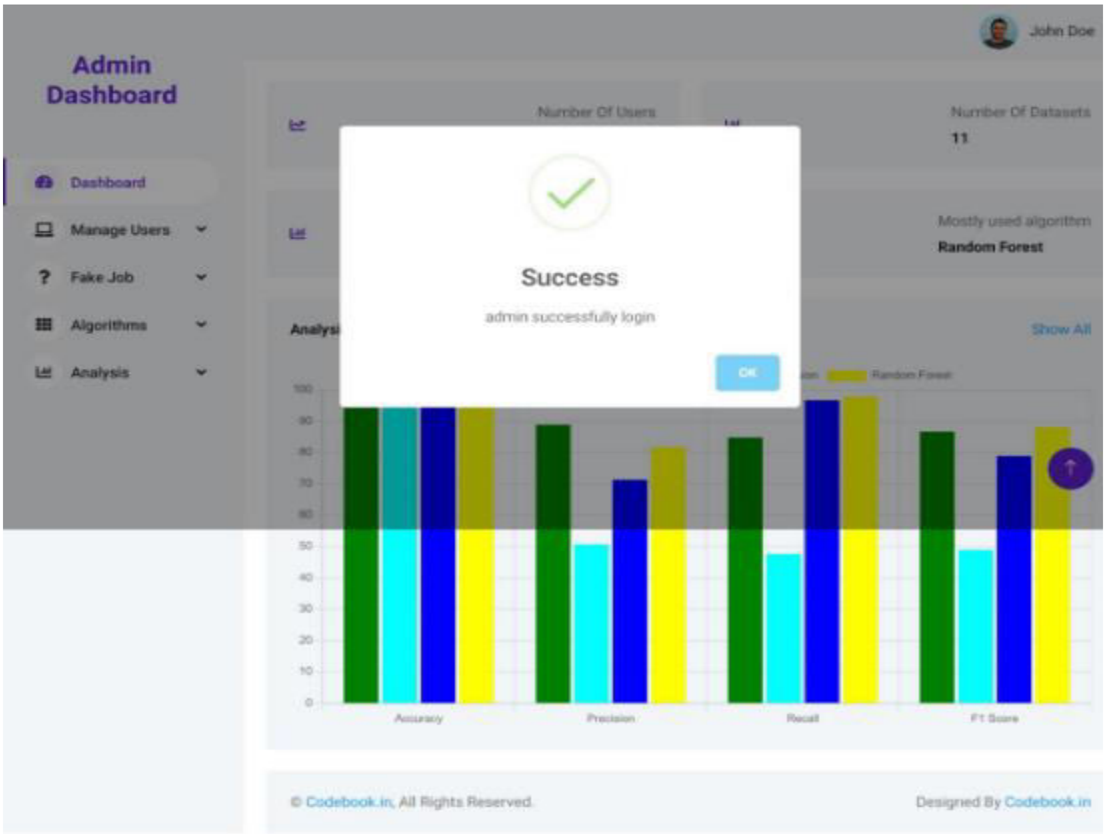
User Login



Fake Job Detection

Admin Login

Admin Login



John Doe

Admin Dashboard

Number Of Users

Number Of Datasets 11

Mostly used algorithm Random Forest

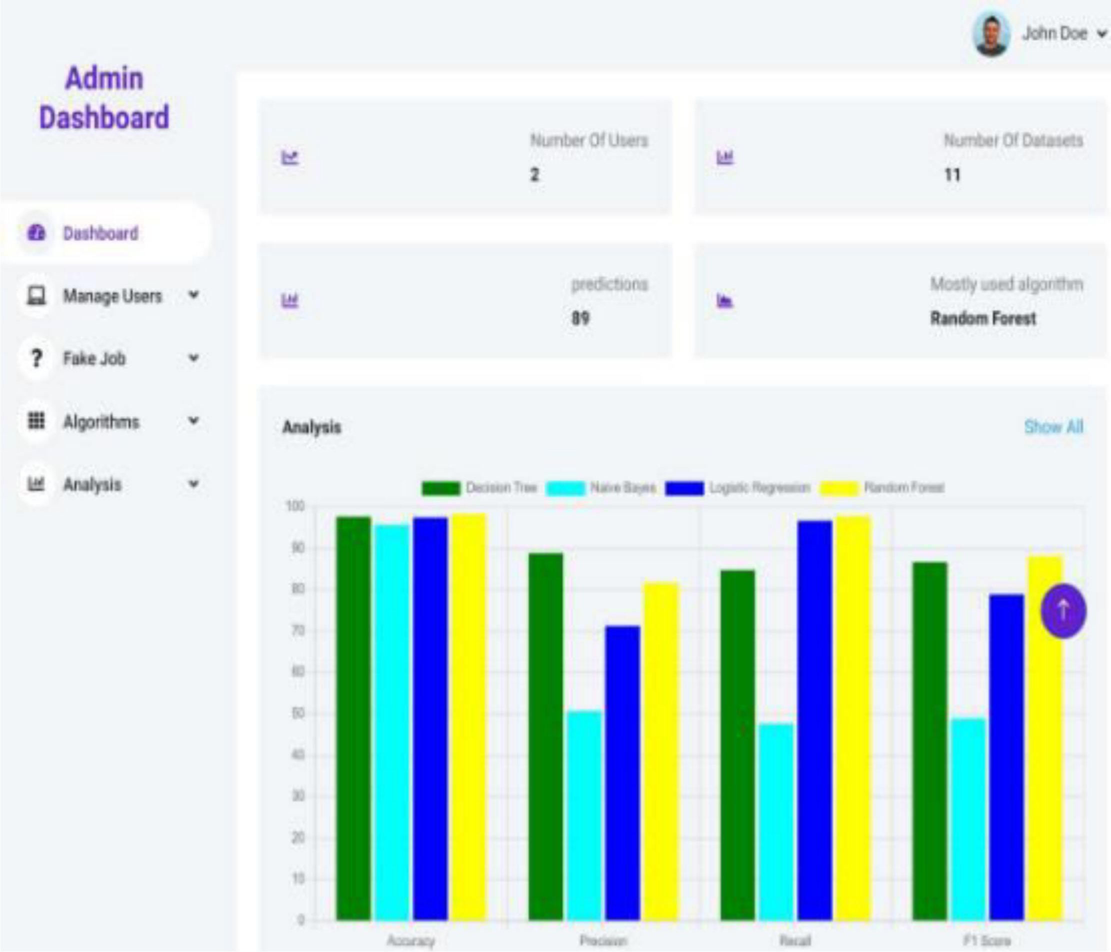
Success

admin successfully login

© Codebook.in, All Rights Reserved.

Designed By Codebook.in

Fig 3.6



John Doe

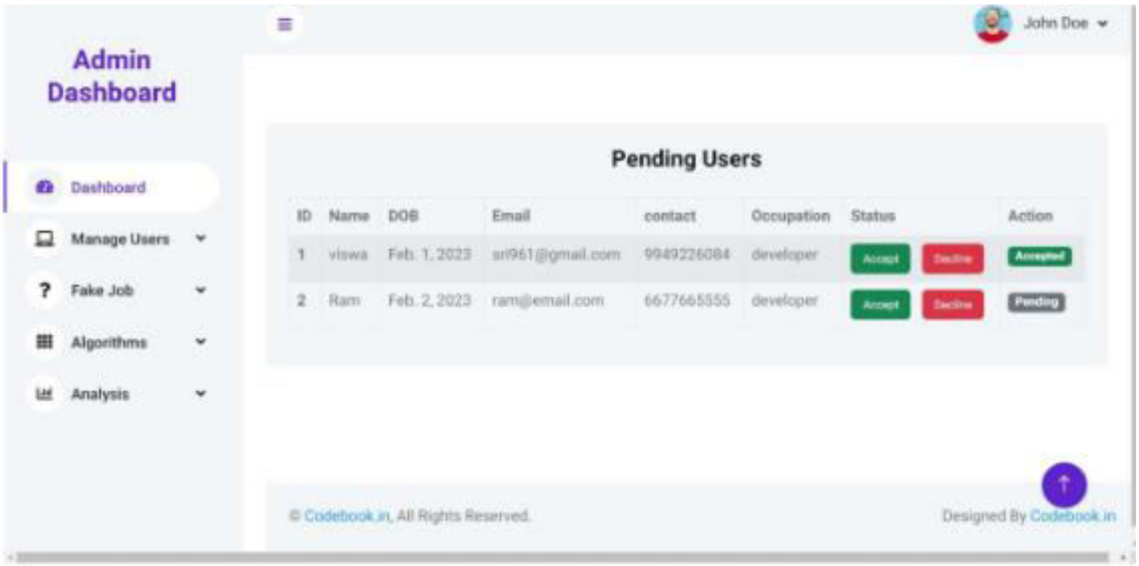
Admin Dashboard

Number Of Users 2

Number Of Datasets 11

predictions 89

Mostly used algorithm Random Forest

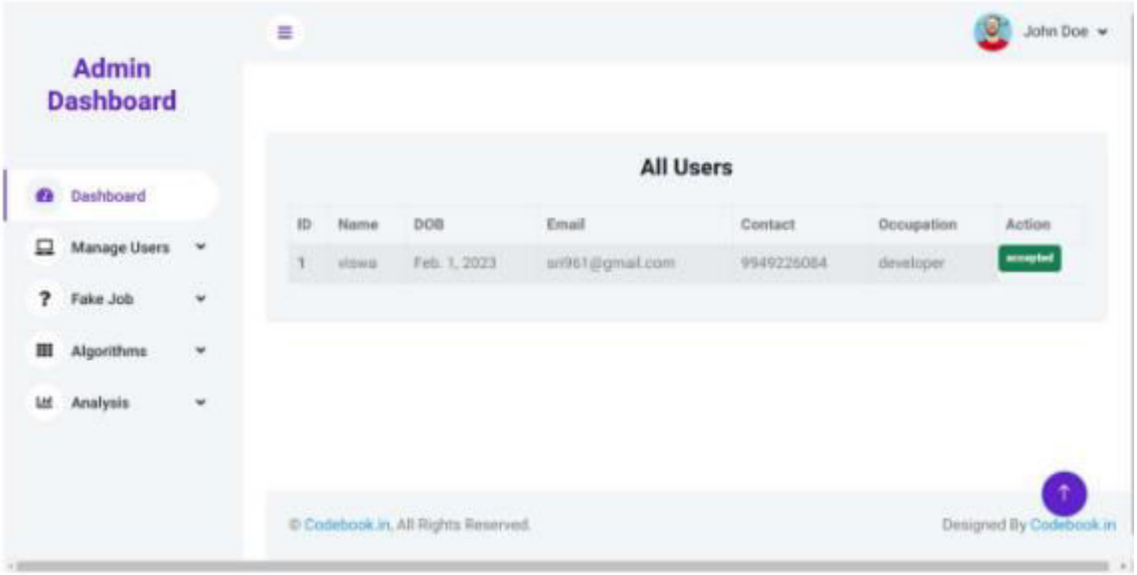


Pending Users

| ID | Name | DOB | Email | contact | Occupation | Status | Action |
|---|---|---|---|---|---|---|---|
| 1 | viswa | Feb. 1, 2023 | sri961@gmail.com | 9949226084 | developer | Accept Decline | Accepted |
| 2 | Ram | Feb. 2, 2023 | ram@email.com | 6677665555 | developer | Accept Decline | Pending |

© Codebook.in, All Rights Reserved.

Designed By Codebook.in

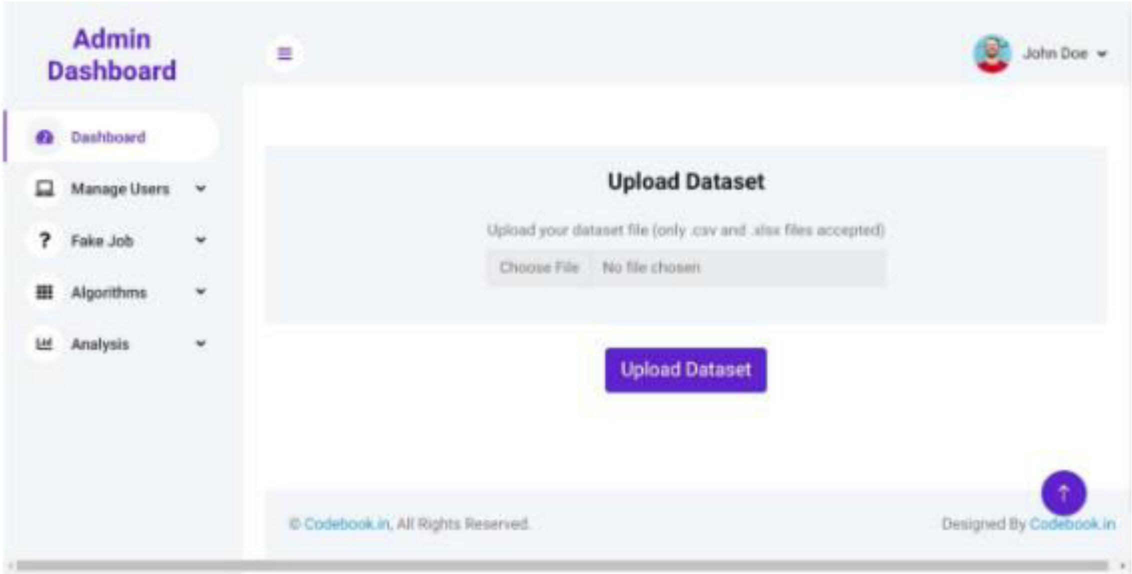


All Users

| ID | Name | DOB | Email | Contact | Occupation | Action |
|---|---|---|---|---|---|---|
| 1 | viswa | Feb. 1, 2023 | sri961@gmail.com | 9949226084 | developer | accepted |

© Codebook.in, All Rights Reserved.

Designed By Codebook.in

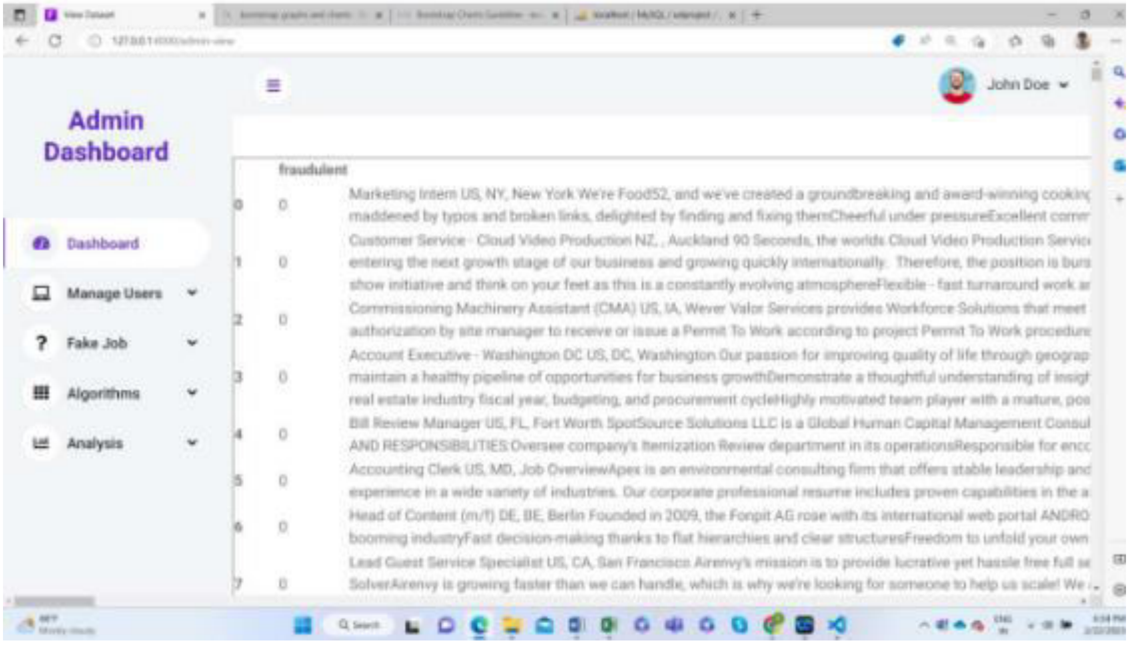


Upload Dataset

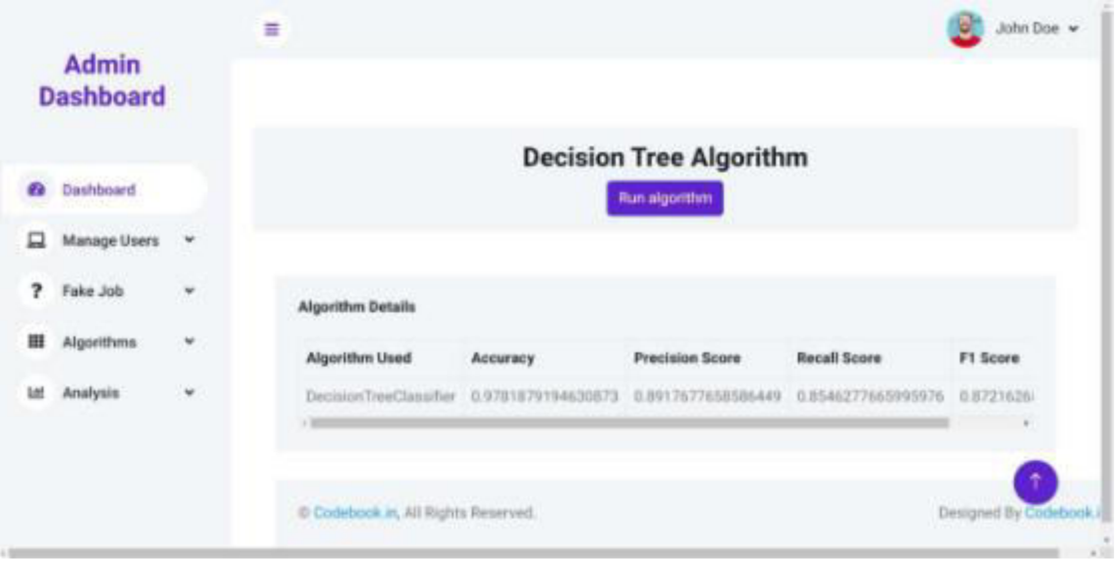
Upload your dataset file (only .csv and .xlsx files accepted)

© Codebook.in, All Rights Reserved.

Designed By Codebook.in



| | fraudulent | |
|---|---|---|
| 0 | 0 | Marketing Intern US, NY, New York We're Food52, and we've created a groundbreaking and award-winning cooking<br>maddened by typos and broken links, delighted by finding and fixing themCheerful under pressureExcellent comm |
| 1 | 0 | Customer Service - Cloud Video Production NZ, , Auckland 90 Seconds, the worlds Cloud Video Production Servic<br>entering the next growth stage of our business and growing quickly internationally. Therefore, the position is burs<br>show initiative and think on your feet as this is a constantly evolving atmosphereFlexible - fast turnaround work ar |
| 2 | 0 | Commissioning Machinery Assistant (CMA) US, IA, Wever Valor Services provides Workforce Solutions that meet<br>authorization by site manager to receive or issue a Permit To Work according to project Permit To Work procedure |
| 3 | 0 | Account Executive - Washington DC US, DC, Washington Our passion for improving quality of life through geograp<br>maintain a healthy pipeline of opportunities for business growthDemonstrate a thoughtful understanding of insigt<br>real estate industry fiscal year, budgeting, and procurement cycleHighly motivated team player with a mature, pos |
| 4 | 0 | Bill Review Manager US, FL, Fort Worth SpotSource Solutions LLC is a Global Human Capital Management Consul<br>AND RESPONSIBILITIES:Oversee company's Itemization Review department in its operationsResponsible for encc |
| 5 | 0 | Accounting Clerk US, MD, Job OverviewApex is an environmental consulting firm that offers stable leadership and<br>experience in a wide variety of industries. Our corporate professional resume includes proven capabilities in the a |
| 6 | 0 | Head of Content (m/f) DE, BE, Berlin Founded in 2009, the Fonpit AG rose with its international web portal ANDRO<br>booming industryFast decision-making thanks to flat hierarchies and clear structuresFreedom to unfold your own |
| 7 | 0 | Lead Guest Service Specialist US, CA, San Francisco Airenvy's mission is to provide lucrative yet hassle free full se<br>SolverAirenvy is growing faster than we can handle, which is why we're looking for someone to help us scale! We . |

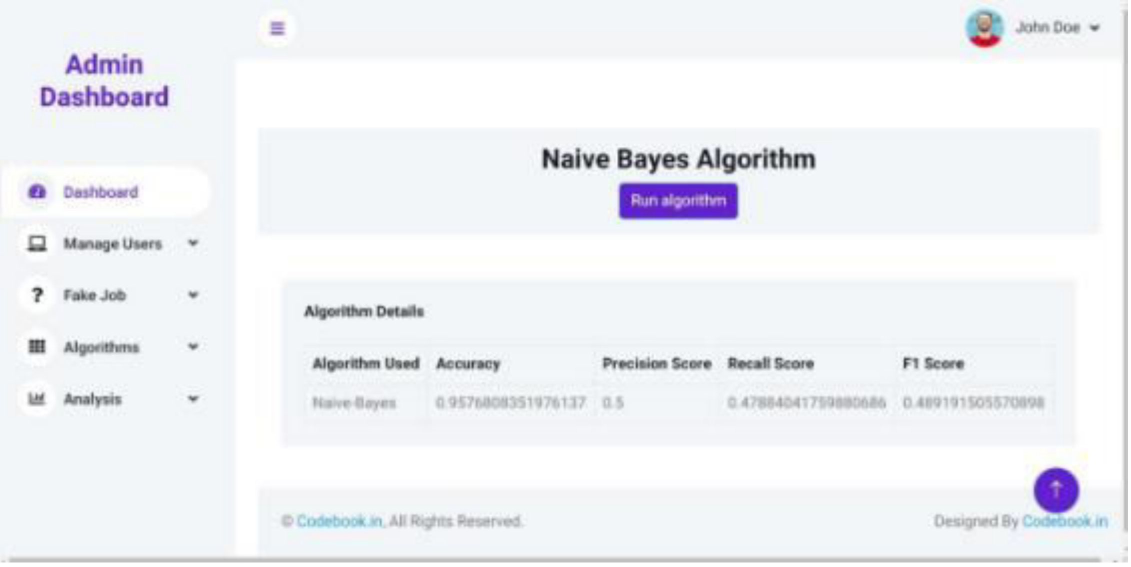


Decision Tree Algorithm

Algorithm Details

| Algorithm Used | Accuracy | Precision Score | Recall Score | F1 Score |
|---|---|---|---|---|
| DecisionTreeClassifier | 0.9781879194630873 | 0.8917677658586449 | 0.8546277665995976 | 0.8721626 |

© Codebook.in, All Rights Reserved. Designed By Codebook.i

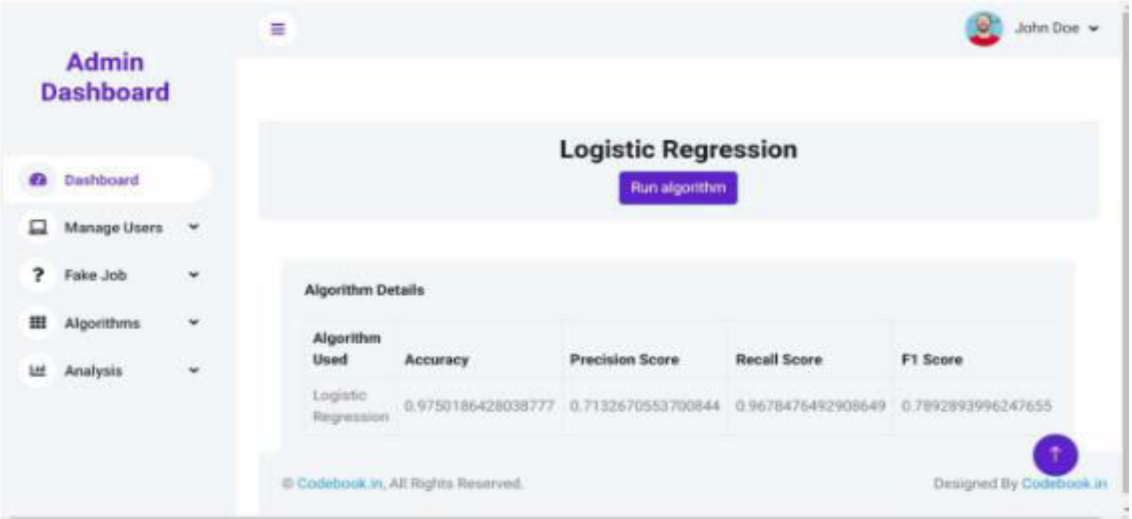


Naive Bayes Algorithm

Algorithm Details

| Algorithm Used | Accuracy | Precision Score | Recall Score | F1 Score |
|---|---|---|---|---|
| Naive-Bayes | 0.9576808351976137 | 0.5 | 0.4788404175988068 | 0.4891915055870898 |

© Codebook.in, All Rights Reserved. Designed By Codebook.in

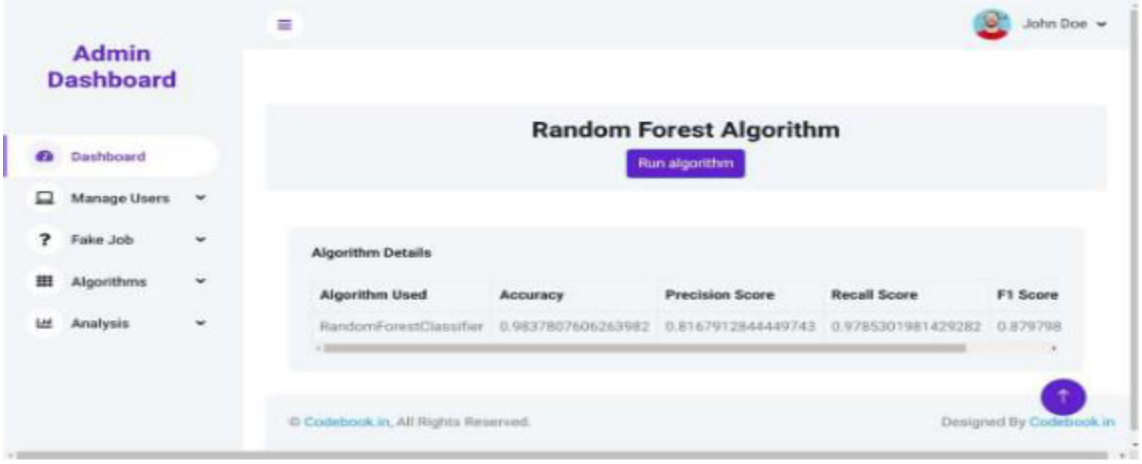


Logistic Regression

Algorithm Details

| Algorithm Used | Accuracy | Precision Score | Recall Score | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.9750186428038777 | 0.7132670553700844 | 0.9678476492908649 | 0.7892893996247655 |

© Codebook.in, All Rights Reserved. Designed By Codebook.in

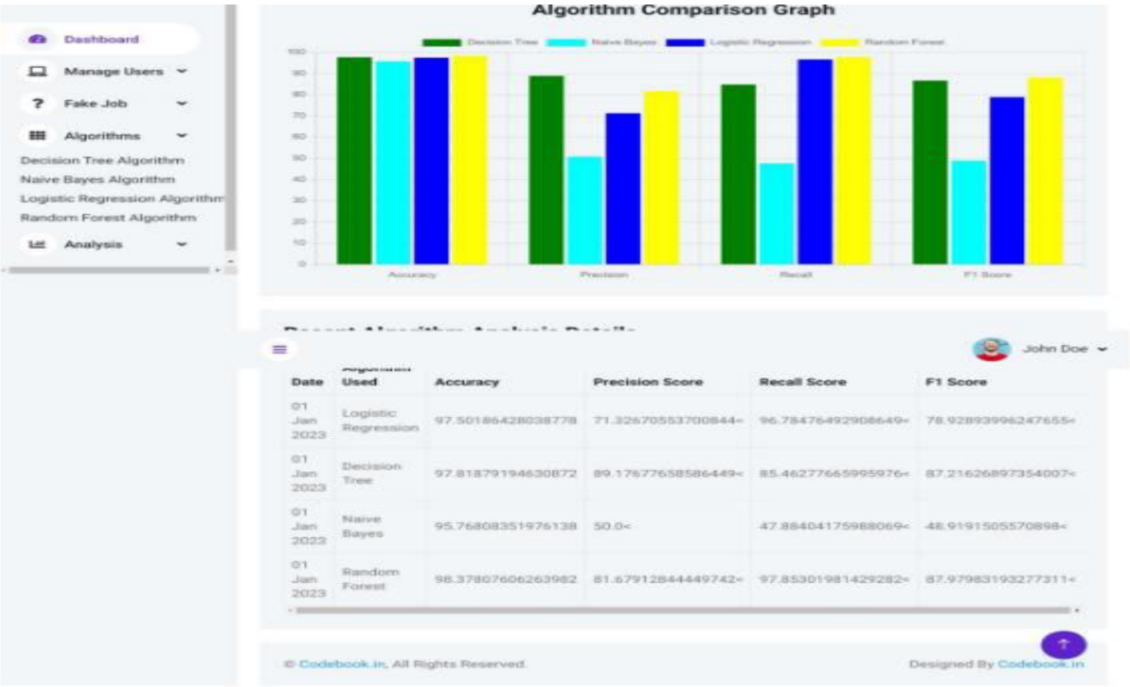


Random Forest Algorithm

Algorithm Details

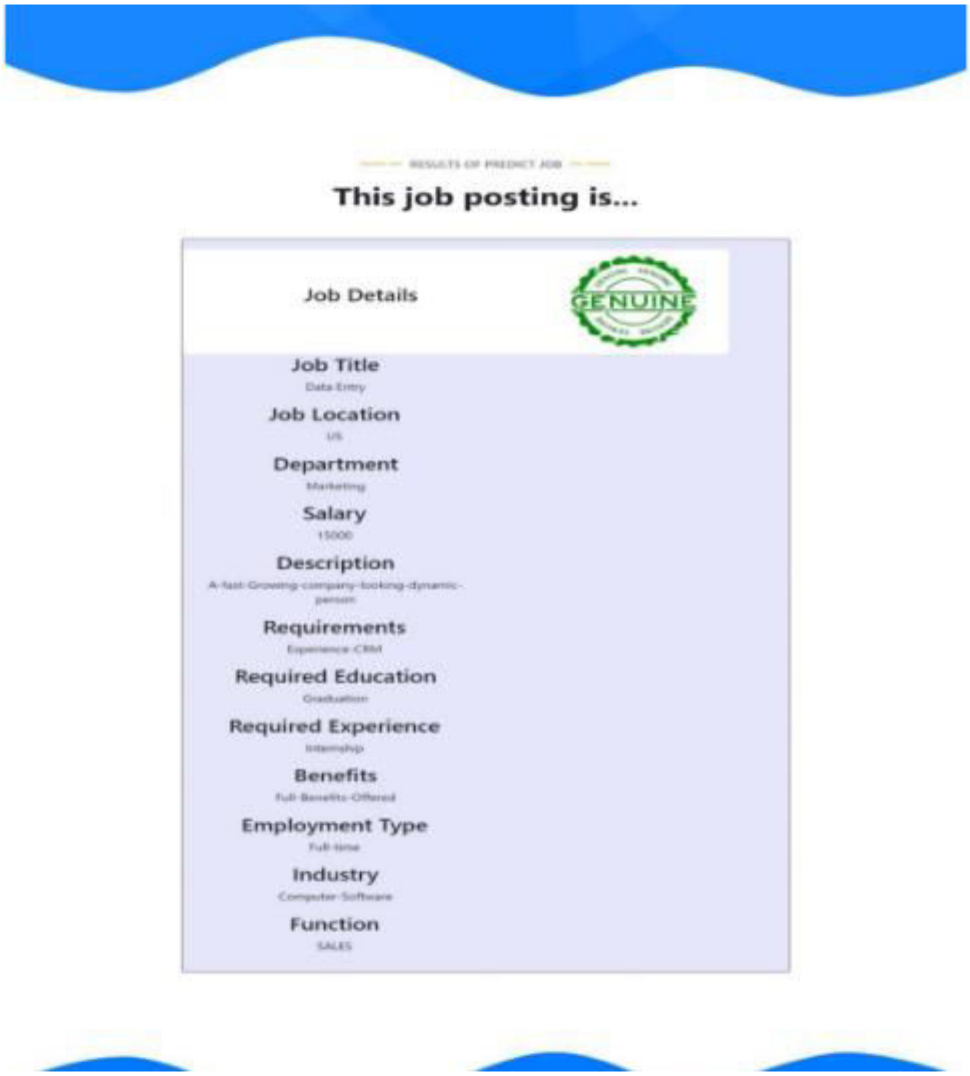
| Algorithm Used | Accuracy | Precision Score | Recall Score | F1 Score |
|---|---|---|---|---|
| RandomForestClassifier | 0.9837807606263982 | 0.8167912844449743 | 0.9785301981429282 | 0.879798 |

© Codebook.in, All Rights Reserved. Designed By Codebook.in



| Date | Algorithm Used | Accuracy | Precision Score | Recall Score | F1 Score |
|---|---|---|---|---|---|
| 01 Jan 2023 | Logistic Regression | 97.50186428038778 | 71.32670553700844< | 96.78476492908649< | 78.92893996247655< |
| 01 Jan 2023 | Decision Tree | 97.81879194630872 | 89.17677658586449< | 85.46277665995976< | 87.21626897354007< |
| 01 Jan 2023 | Naive Bayes | 95.76808351976138 | 50.0< | 47.88404175988069< | 48.9191505570898< |
| 01 Jan 2023 | Random Forest | 98.37807606263982 | 81.67912844449742< | 97.85301981429282< | 87.97983193277311< |

© Codebook.in, All Rights Reserved. Designed By Codebook.in



RESULTS OF PREDICT JOB

This job posting is...

Job Details

GENUINE

Job Title

Data Entry

Job Location

US

Department

Marketing

Salary

15000

Description

A-fast-Growing-company-looking-dynamic-person

Requirements

Experience-CRM

Required Education

Graduation

Required Experience

Internship

Benefits

Full-Benefits-Offered

Employment Type

Full-time

Industry

Computer-Software

Function

SALES

## 8. CONCLUSIONS

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

## 9. 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 perceptrons for classification and regression,‖Neurocomputing,

vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] C. Anglano, M. Canonico, and M. Guazzone, "Forensic analysis of Telegram Messenger on Android smartphones," Digit. Investig., vol. 23, pp. 31–49, 2017.

[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] "Kali Linux | Penetration Testing and Ethical Hacking Linux Distribution." [Online]. Available: https://www.kali.org/. [Accessed: 21- Aug-2018].

[8] L. Breiman, ―ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001,

doi: 10.1017/CBO9781107415324.004.

[9] "Welcome to Python.org." [Online]. Available: https://www.python.org/. [Accessed: 21-Aug-2018].

.

[10] A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot.,

vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[11]. "SWGDE." [Online]. Available: https://www.swgde.org/. [Accessed: 30- Aug-2018].

[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] ShivamBansal (2020, February). [Real or Fake] Fake JobPosting Prediction,Version1.Retrieved March 29,2020 from https://www.kaggle.com/shivamb/real-or-fake-fake

jobposting-prediction

[Gobierno del Ecuador, “Ley Orgánica de Educación Intrcultural.” 2012.

[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. WCCI

2010, no. May 2016, 2010, doi: 10.1109/FUZZY.2010.5584447