IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

Fake Job Post Prediction Using Deep Learning

[1]I Niharika Naidu, [2]Y Sandhya, [3]Mrs. Shilpa Shesham

[1]Student/ Research Scholar, Department of Artificial Intelligence, Anurag University, Hyderabad, India

[2]Student/ Research Scholar, Department of Artificial Intelligence, Anurag University, Hyderabad, India

[3]Assistant Professor, Department of Artificial Intelligence, Anurag University, Hyderabad, India

ABSTRACT: The study investigates the detection of fraudulent job postings using ML classification models and MLP to address the increasing prevalence of deceptive job offers. The research employs a dataset containing 17,000 job descriptions, utilizing feature extraction techniques to extract valuable insights from the data. To identify fake job postings, the study employs various algorithms, including Logistic Regression, Random Forest, Linear Support Vector Classifier, and K-Nearest Neighbors. Performance metrics and comparisons against the deep learning model are used to assess the effectiveness of the approach. The research aims to provide a robust solution to combat employment fraud, safeguarding job seekers from scams and maintaining the integrity of job postings in the digital realm. The study's comprehensive analysis and methodology demonstrate the potential of ML techniques in accurately predicting and distinguishing between genuine and fraudulent job offers. By presenting a promising avenue for combating fraudulent job postings and bolstering trust in online recruitment processes, the research highlights the value of integrating machine learning-based classification techniques, particularly deep learning, in addressing this critical issue.

KEYWORDS - Fake Job Postings, Machine Learning Classification Models, Multi-Layer Perceptron, Employment Fraud, Job Seekers Protection

I. INTRODUCTION

In the contemporary realm of employment, the confluence of artificial progress and technological invention has expanded and diversified job openings in unknown ways. The digital period has normalized access to job openings, easing global connectivity between associations and implicit campaigners. Social media platforms have surfaced as potent instruments for propagating job bulletins, allowing recruiters to precisely target specific demographics and engage with unresistant job seekers through acclimatized campaigns and interactive content. However, amidst the plenty of licit job bulletins, there lurk deceptive and fraudulent listings that prey on unsuspecting job campaigners. These false announcements allure individualities with pledges of economic positions but frequently lead to scams and exploitation, undermining the trust and confidence of those seeking employment. The frequence of fraudulent job postings has formed a pervasive sense of apprehension among job seekers, compelling them to navigate a geography fraught with uncertainty and threat.

To combat this challenge, innovative results are demanded to effectively discern between genuine job postings and fraudulent listings. One similar result lies in the development of automated systems endowed with prophetic algorithms that can dissect different parameters to descry suspicious patterns or inconsistencies in job postings. By employing the power of machine learning and deep learning, these systems can efficiently sift through vast quantities of data, flagging potentially fraudulent bulletins and mollifying the threat of exploitation for job seekers. The integration of similar slice-edge technologies represents a transformative shift, heralding a new period of translucency and responsibility in the recruitment process. By bolstering the effectiveness and trustability of job announcement platforms, associations can cultivate a further conducive terrain for both recruiters and job seekers, ensuring that licit openings remain accessible to all.

While technology has incontrovertibly revolutionized the recruitment landscape, the proliferation of fraudulent job postings underscores the consummate significance of upholding integrity and trust in the job request. Through combined cooperative efforts among industry stakeholders, policymakers, and technology inventors, we can harness the transformative eventuality of technology to combat fraudulent practices and forge a further indifferent and secure terrain for job seekers worldwide.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

II. RELATED WORK

Statistical features-based real-time detection of drifted Twitter spam by C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min: The Lfun scheme was developed to address the dynamic nature of spam tweets over time, termed "Twitter Spam Drift." This system identifies changed spam tweets from unlabeled data and integrates them into the classifier's training process, enhancing spam detection accuracy in real-world scripts.

Automatically identifying fake news in popular Twitter threads by C. Buntain and J. Golbeck: A system for automating fake news detection on Twitter by predicting accuracy assessments using two credibility-concentrated datasets, CREDBANK and PHEME, was presented. Models trained on crowdsourced workers outperformed those grounded solely on intelligencers' assessments, pressing the significance of non-expert input in fake news discovery.

A performance evaluation of machine learning-based streaming spam tweets detection by C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian: The study estimated machine learning-based streaming spam detection methods across data, point, and model confines. The findings emphasized the complexity of streaming spam tweet detection and the need for robust ways.

A model-based approach for identifying spammers in social networks by F. Fathaliani and M. Bouguessa: A model-grounded approach to identify spammers in social networks using admixture modeling principles was proposed. This system requires no mortal intervention for setting threshold parameters, demonstrating its connection across various online social platforms.

Spam detection of Twitter traffic: A framework based on random forests and non-uniform feature sampling by C. Meda, E. Ragusa, C. Gianoglio, R. Zunino, A. Ottaviano, E. Scillia, and R. Surlinelli: A framework using a variant of the Random Forests Algorithm and non-uniform feature sampling to identify spammers within Twitter traffic was proposed. The experimental results validated the effectiveness of the fortified feature sampling method in distinguishing between spammers and licit users.

III. PROPOSED WORK

The main aim of this system is to give job seekers a dependable way to separate between real job openings and fraudulent ones, therefore allowing them to concentrate their efforts on genuine prospects. To achieve this, the system utilizes Multi-Layer Perceptron Method with a dataset sourced from Kaggle, which contains various details about job postings similar as job ID, title, location, and department. Before feeding this data into the classification model, the system undergoes a preprocessing stage where it cleans and refines the dataset. This preprocessing involves several tasks, including removing any gratuitous spaces, eliminating entries with missing values, and filtering out stop words that carry little or no meaningful information. Once the dataset is cleaned and prepared, it's also ready for classification. The classifier, trained on preprocessed data, works to make predictions about the authenticity of each job advertisement. By applying machine learning algorithms, the system can directly identify genuine job openings while filtering out fraudulent ones. This ensures that only licit job postings are presented to job seekers, thereby saving them time and effort that would otherwise be spent sifting through potentially deceptive listings. Eventually, this approach enhances the responsibility of the job search process and provides job seekers with a more dependable platform to pursue employment openings.

3.1 Algorithms

Linear Support Vector Classifier (SVC):

Linear SVC finds optimal hyperplanes to separate data classes, accommodating high-dimensional data like text and images. While versatile, it faces challenges with non-linear or imbalanced datasets, prompting ongoing research for scalability and efficiency improvements.

Logistic Regression:

Logistic regression models the probability of binary outcomes based on predictor variables, offering interpretability and handling noise well. However, it struggles with complex nonlinear relationships, addressed through techniques like regularization and feature engineering.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

Random Forest:
Random Forest constructs decision trees and effectively handles complex datasets with high dimensionality. Despite its strengths, overfitting and imbalanced classes pose challenges, mitigated by hyperparameter tuning and resampling techniques.

K-Nearest Neighbors (KNN):
KNN classifies data points based on their nearest neighbors and suits complex, nonlinear datasets. However, its scalability is hindered by computational complexity, and optimal performance requires careful parameter tuning and feature selection.

Multilayer Perceptron (MLP):
MLP captures intricate data relationships through multiple layers of neurons, suitable for various data types. Ongoing research focuses on improving efficiency and scalability through hyperparameter tuning and architectural enhancements.

3.2. Methodology

A. Data Collection:
Gather job posting data from diverse online platforms, ensuring a balanced representation of legitimate and fraudulent postings. Curate a comprehensive dataset containing textual information such as job descriptions, titles, locations, and company details.

B. Data Preprocessing:
Clean the raw text data by removing HTML tags, punctuation, and non alphanumeric characters to ensure uniformity. Tokenize the text into individual words or tokens and regularize the text by converting it to lowercase. Exclude stopwords(generally being words with little semantic value) and perform lemmatization or stemming to reduce words to their base forms.

C. Feature Engineering:
Transform the preprocessed text data into numerical features suitable for machine learning algorithms. Use ways like CountVectorizer or TF- IDF to convert textual data into numerical representations. Explore additional features similar as job titles, locales, and company information to enhance the prophetic power of the model.

D. Model Training:
Split the dataset into training and testing sets to assess model performance accurately. Train colorful classification models similar as Support Vector Machines(SVM), Random Forests, and Neural Networks using the training data. Finetune hyperparameters using ways like grid search or arbitrary search to optimize model performance.

E. Model Evaluation:
Estimate the trained models using multiple performance metrics, including accuracy, precision, recall, and F1score. Employ cross validation to ensure the robustness and generalizability of the models across different subsets of the data. Analyze confusion matrices to gain perceptivity into model errors, particularly false positives and false negatives.

F. Visualization:
Visualize important features and performance metrics using libraries like Matplotlib and Seaborn to gain insights into model behavior and effectiveness. Generate word clouds to visualize the most common terms associated with legitimate and fraudulent job postings, aiding in interpretability.

G. Deployment:
Select the best performing model based on evaluation results and deploy it for real time prediction. Implement the model in a web application or API, ensuring seamless integration with existing platforms. Prioritize security measures to safeguard sensitive user data during model deployment and usage.

H. Monitoring and Maintenance:
Continuously monitor the deployed model's performance over time, retraining it periodically with streamlined data to adapt to evolving patterns of fraudulent activity. Handle model updates, bug fixes, and advancements to maintain its effectiveness and trustability in detecting fake job postings.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

3.3. Technology Used

A. Programming Language:

Python: Widely used for its simplicity and readability, Python dominates in machine learning and NLP due to its extensive libraries and rapid development capabilities.

B. Machine Learning Libraries:

Scikitlearn: Comprehensive ML library in Python, offering diverse algorithms and an intuitive API for easy model implementation and deployment.

TensorFlow: Google's open-source framework for building and training deep learning models, providing scalability and efficient execution on CPUs and GPUs.

Keras: High-level neural networks API atop TensorFlow, simplifying model development with a user-friendly interface for rapid prototyping.

C. Natural Language Processing (NLP) Libraries:

NLTK: Python's go-to for NLP tasks, offering a range of functionalities and access to corpora and pretrained models for text analysis.

D. Data Visualization Libraries:

Matplotlib: Versatile plotting library in Python for creating static, interactive, and publication-quality visualizations with customizable options.

Seaborn: Statistical data visualization library extending Matplotlib, providing high-level functions for visually appealing graphics and seamless integration with Pandas.

E. Development Frameworks:

Flask: Lightweight web framework for Python, ideal for building RESTful APIs and web services with simplicity and scalability.

3.4. Implementation

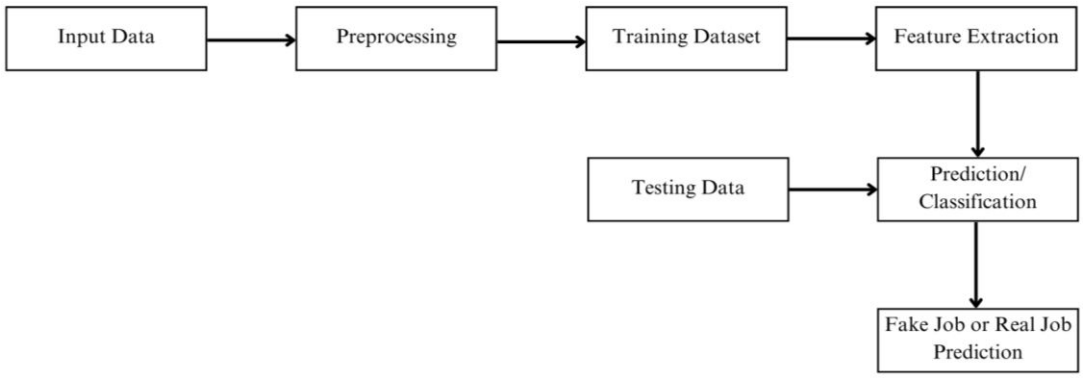


Figure 1. Data Flow Diagram

Modules

A. Data Collection:

The dataset “EMSCAD”, sourced from www.Kaggle.com, comprises diverse samples of job postings, encompassing various parameters such as title, location, department, salary range, description, and required education.

B. Pre-Processing:

Data is converted into scalar format, followed by the creation of new features, which are then fed into the algorithm. These features are stored as 'x', while labels are stored as 'y'.

C. Train and Test:

The dataset is split into training and test sets, utilizing 70% of the data for training and the remaining 30% for testing. This split is achieved using a parameter that divides the data frame in a 70-30 ratio.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

D. Model Training:

Split the dataset for accurate model assessment. Train classification models like Linear Support Vector Classifier, Random Forest, Logistic Regression, K-Nearest Neighbors, and Multi-Layer Perceptron. Optimize model performance through hyperparameter fine tuning using techniques like grid or random search.

E. Fake Job Post Prediction:

Input details in the form of a CSV file containing various job profiles are fed into the system for prediction. The system predicts the authenticity of the job postings based on these details and stores the results in a new CSV file along with the prediction outcomes. Ultimately, the culmination of this process is the model's output, wherein it furnishes predictions for each job listing in the testing data, signaling whether it perceives the job as fake or real based on the knowledge acquired during training.

```
Data columns (total 18 columns):
job_id                 17880 non-null int64
title                  17880 non-null object
location               17534 non-null object
department             6333 non-null object
salary_range           2868 non-null object
company_profile        14572 non-null object
description            17879 non-null object
requirements           15185 non-null object
benefits               10670 non-null object
telecommuting          17880 non-null int64
has_company_logo       17880 non-null int64
has_questions          17880 non-null int64
employment_type        14409 non-null object
required_experience    10830 non-null object
required_education     9775 non-null object
industry               12977 non-null object
function               11425 non-null object
fraudulent             17880 non-null int64
```

Figure 2. Dataset

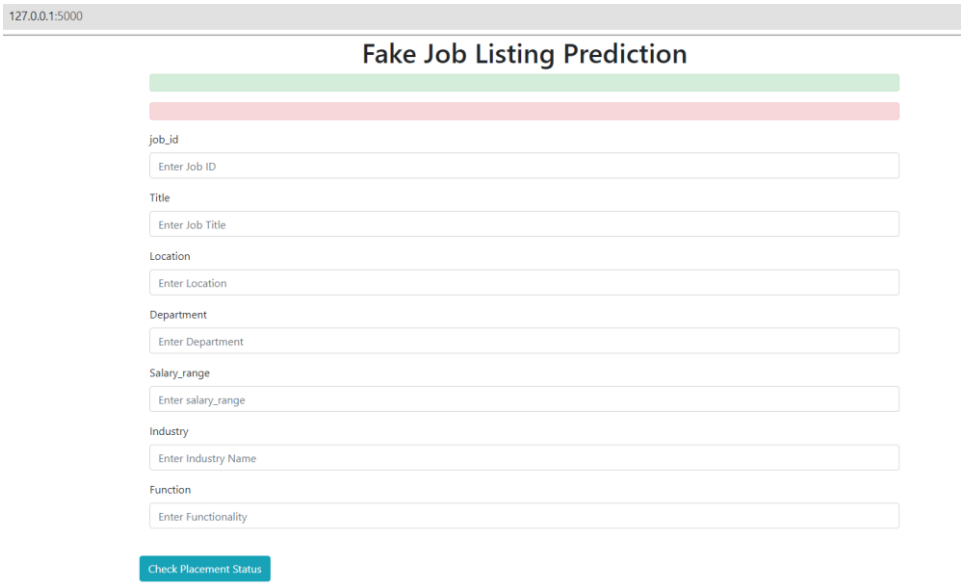
Figure 3. Target Distribution in Dataset

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

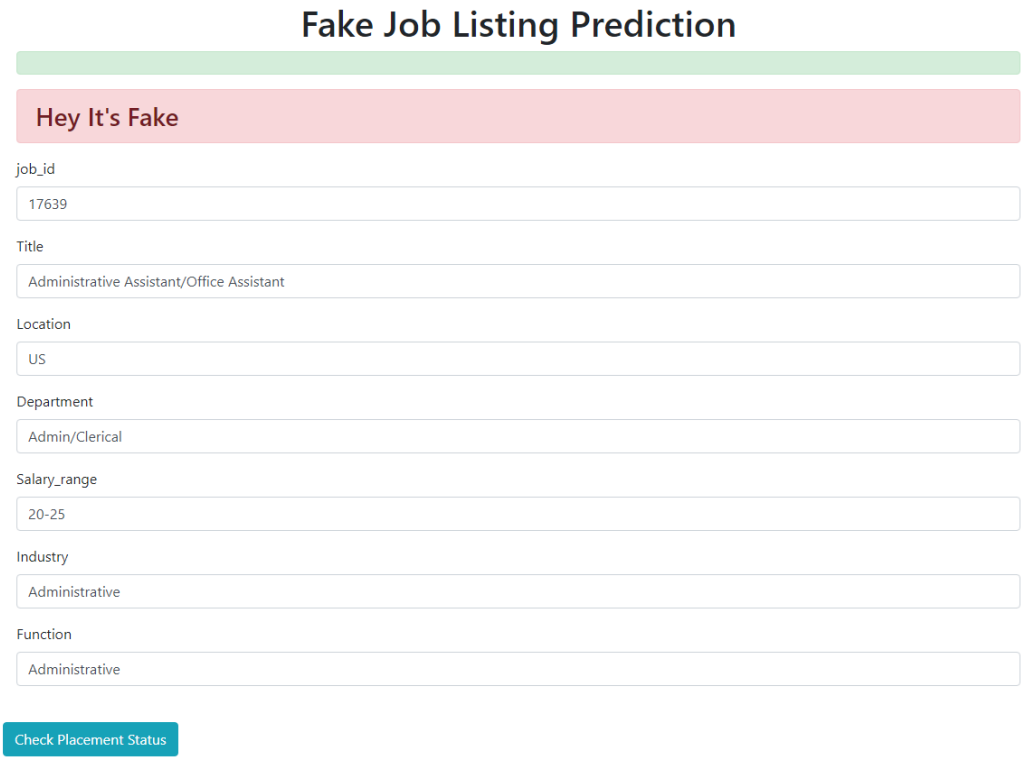
|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

IV. RESULT ANALYSIS



Screen 1. Web Page

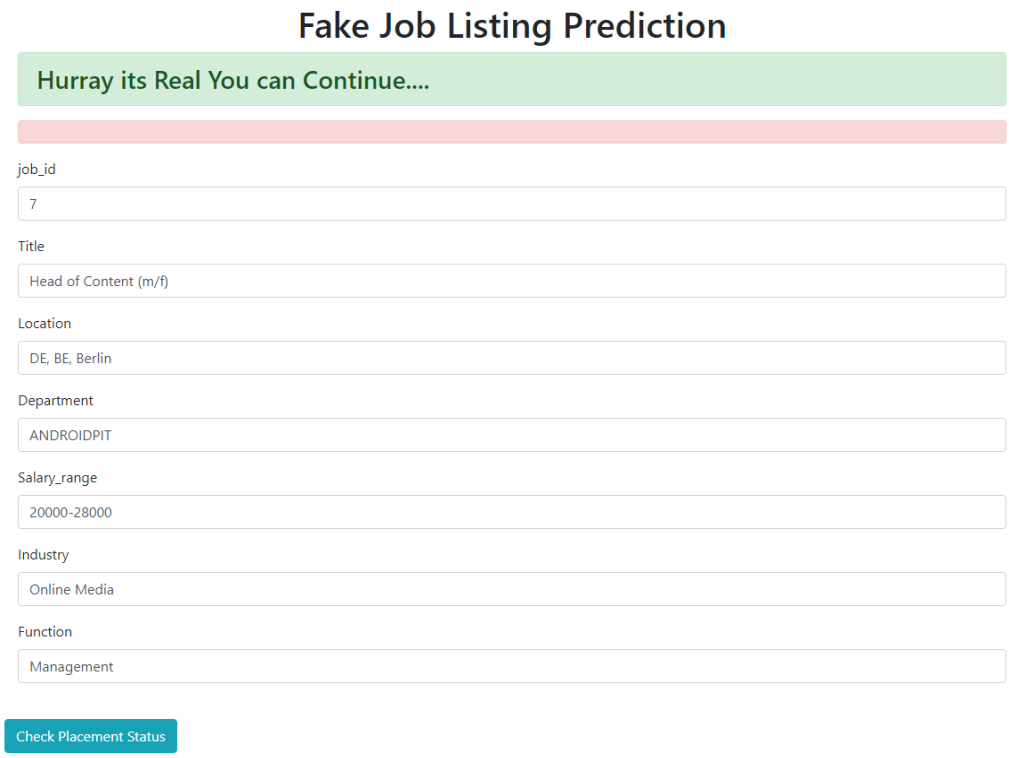


Screen 2. Prediction is: Fake Job Post

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |



Screen 3. Prediction is: Real Job Post

V. CONCLUSION

Job scam detection is a pressing issue globally, prompting our study using the EMSCAD dataset to analyze fraudulent job postings. We explore traditional machine learning and deep learning methods, finding the Random Forest Classifier effective, while the Deep Neural Network shows promise with 0.96 accuracy. Our study offers insights into improving fraud detection mechanisms for safeguarding job seekers in the digital landscape.

VI. FUTURE WORK

The real and fake job prediction with cloud connectivity:
It offers scalability and efficiency benefits, utilizing computational resources for processing large datasets. Cloud deployment enables seamless integration with job search platforms for real-time authenticity assessments, but stringent data privacy measures are essential, including robust encryption and access controls.

Enhancing workflow optimization for implementation within an Artificial Intelligence framework:
It includes fine-tuning algorithms and enabling continuous learning to adapt to emerging trends. Collaborative filtering personalizes job recommendations, enhancing user engagement, while Explainable methodologies ensure transparency in system decisions, fostering trust.

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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304408 |

[5] P. Cunningham and S. J. Delany, K -Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.
[6] H. Sharma and S. Kumar, A Survey on Decision Tree Algorithms of Classification in Data Mining,‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.
[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems,‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.
[8] L. Breiman, ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.
[9] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details