ORIGINAL RESEARCH PAPER
Volume - 14 | Issue - 06 | June - 2025 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

SCREENSMART: REVOLUTIONIZING HIRING WITH MACHINE LEARNING

Machine Learning

Dr. U. Surya Kameswari* Assistant Professor Acharya Nagarjuna University *Corresponing Author

Mr. Vassey Nagaraju Assistant Professor Andhra University College of Engineering

ABSTRACT

The creation of an AI-based resume screening system that uses machine learning techniques to automatically classify resumes into pertinent job categories is presented in this research paper. In order to extract valuable features, the system uses a labeled dataset of resumes and preprocessing techniques like text cleaning and TF-IDF vectorization. After training and evaluating a number of models, including logistic regression, decision trees, random forests, XGBoost, and gradient boosting, ensemble models achieved perfect accuracy, while logistic regression performed well as a lightweight substitute. Metrics such as accuracy, precision, recall, F1 score, and ROC AUC were used to evaluate performance. By providing a consistent, data-driven method of evaluating candidates, the suggested system drastically cuts down on manual labor in the hiring process. In addition to laying the groundwork for upcoming improvements like deep learning integration, real-time deployment, and fairness auditing, this work shows how AI can improve HR operations.

KEYWORDS

Resume Screening, Text pre-processing, TF-IDF Vectorization, Ensemble methods, Human Resource Automation

INTRODUCTION

Today's fast-paced and fiercely competitive job market, companies often receive an excessive volume of applications for each open position. Manually going over each resume is time-consuming, prone to mistakes, and susceptible to unintentional biases. As companies look to streamline their hiring processes, there is an increasing need for automated solutions that can efficiently assess prospects and make recruiters' jobs easier.

The increasing power of AI and ML could change the hiring process. Complex algorithms are used by AI-based resume screening systems to effectively and precisely sort through vast volumes of candidate data. These tools assist in filtering applicants' relevant experience, education, and skills that match the job requirements, resulting in a more uniform and equitable shortlist. As you use ML models more, the screening process becomes more accurate and simple because they learn from new data and get better over time.

This research explores the design and execution of a machine learning-basd resume screening framework which aimed at simplifying and enhancing the candidate evaluation process. The study investigates various ML techniques for feature extraction, natural language processing, and predictive modeling to build a solution that is both effective and scalable. By integrating AI into recruitment workflows, the proposed system has the probable to reduce time-to-hire and support better decision-making, ultimately improving organizational performance

2. Literature Review

The potential for automated resume screening to increase recruiting process efficiency has drawn a lot of attention in recent years. A comprehensive analysis of numerous machine learning (ML) and natural language processing (NLP) approaches used for resume screening was provided by Sinha et al. in 2021. Their research highlights how crucial feature extraction, classification models, and preprocessing are to creating precise and scalable systems. They came to the conclusion that hybrid methods that include machine learning and natural language processing perform better on candidate categorization tasks.

A resume recommendation system that matches candidate profiles with job requirements using machine learning algorithms was proposed by Roy et al. (2020). Their approach, which was based on categorization techniques, showed promise in sifting through a vast pool of resumes to find relevant ones. In a similar vein, Ali et al. (2022) highlighted the significance of contextual data analysis in applicant evaluation by creating a resume classification model that uses natural language processing (NLP) techniques to extract skills and qualifications from unstructured text.

Navarro (2025) presented JustScreen, an AI-powered application created to improve Applicant Tracking Systems (ATS) with transparency and bias reduction in order to address justice and ethics in hiring. This is in line with growing worries about hiring practices that use algorithms to discriminate. Abdullah et al. (2024) demonstrated the potential of deep learning models for extracting important information from resumes, particularly in varied linguistic environments, by introducing NER-RoBERTa, a refined transformer model for Named Entity Recognition (NER) in low-resource languages.

Additional contributions include Gadegaonkar et al. (2023), who concentrated on job recommendation systems, and Liang and Wan (2022), who employed text mining and multi-criteria decision-making to problems involving job matching. Both of these scholars made significant contributions. For the purpose of improving the recruiting pipeline, these studies provide support for the incorporation of domain-specific decision-making frameworks and advanced machine learning approaches. This set of work, when taken as a whole, lays the groundwork for the development of AI-driven resume screening systems that are more accurate, fair, and scalable, which is the goal of this research.

3. Methodology

3.1 Dataset:

The methodology implemented for this research paper is focused around applying supervised machine learning techniques to automate the classification of resumes into specific job categories. For the purpose of this research work, the approach that was adopted focused on the application of supervised machine learning techniques in order to automate the classification of resumes into specified job categories. The dataset that was utilized in this investigation was obtained from kaggle and was given the name AI_Resume_Screening.csv. This sample file is a structured resume corpus that includes thousands of resumes that have been labeled with their respective job domains. These job domains include Data Science, Web Development, Human Resources, and others. As a result of this labeling, the problem was suitable for a classification job that involved multiple classes.

3.2 Pre-Processing:

Prior to applying any machine learning techniques to the data, a comprehensive pipeline for data preparation was established. First, any special characters, stop words, and other noise that might have hampered learning were eliminated from the raw resume texts.All of the textual data was converted to lowercase and then normalized to ensure consistency. Then, to convert category job roles into numerical values that the machine learning model could understand, a label encoding technique was applied. To extract features, the resumes were vectorized using a statistical method called Term Frequency–Inverse Document Frequency (TF-IDF). This statistical technique shows how important a term is in a document compared to the entire corpus.

3.3 Model Training:

After obtaining the features, the dataset was divided to training and testing, 80%-20%, to assess the model's generalization ability. We then trained a logistic regression classifier on the vectorized resume data. Logistic regression was selected since it is rather simple, interpretable, and efficient, and is commonly used as a baseline model. Checking of model stability to be done some of the methods, viz. precision & recall and the confusion matrix was among them. Matters of the predictive power of the model with regard to the job categories and the precision of the classifications, the accuracy and the confusion matrix, were deals with. These steps were explorative components of a consequent resume screening system that is highly interpretable and efficient.

3.4 Proposed System

The suggested system is formulated in such a way that it applies AI to match resumes for the job descriptions based on the text information contained in them. The process follows a modular pipeline that starts with natural language and gets finished when a suitable job category is produced. At the beginning, we provide the system with a dataset of resumes, in which all are represented in a text format via the way of a CSV file. The latter is quite a block of text. So, the text then undergoes a preparation that conducts the necessary tasks of text normalization, noise removal, and feature extraction by using TF-IDF vectorization. The computation of the preprocessing in turn ensures that the data is in a state of unity, is clean, and is good for the purpose of training machine learning models.

After preprocessing, the resume vectors in the form of TF-IDF are used as an input for a logistic regression classifier that has learned itself to note the features peculiar to various job categories. Consequently, the classifier makes predictions which are in turn linked to resume titles such as Data Analyst, Software Engineer, or HR Executive. These predictions can subsequently be utilized by recruiters or HR software to cast a quicker and more reliable vote during their candidate shortlisting process. On top of that, the system is also suitable for the provision of performance evaluation of the model through metrics and visualizations like confusion matrices, which ensure continuous model optimization and monitoring.

The entire system has been constructed in an expandable manner, which means it is a fairly straight-forward procedure to add more complex architectures of the models such as Random Forests, Support Vector Machines, or even deep learning. Besides, the architecture is also capable of being embedded in web-based HR platforms for such purposes as real-time resume uploads, model-inference-on-the-fly, and assistance in making decisions for recruiters. In the long run, the network not only makes the work of manual labor forces easier but also elevates the degree of objectivity in the decision-making of the recruiters of the company

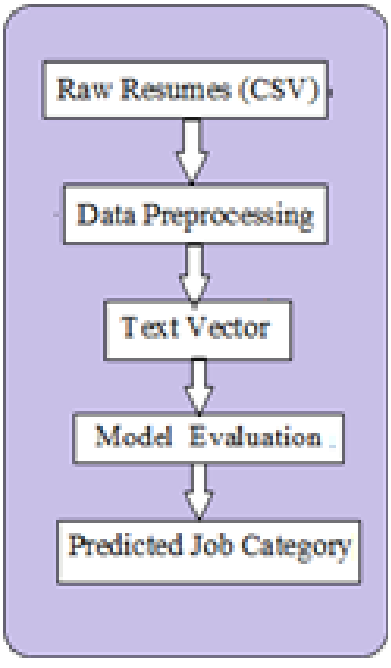

Figure 1: Proposed system architecture

4. RESULT ANALYSIS

The objective of this study was to evaluate the performance of various machine learning models in automatically classifying resumes into relevant job categories based on extracted textual features.

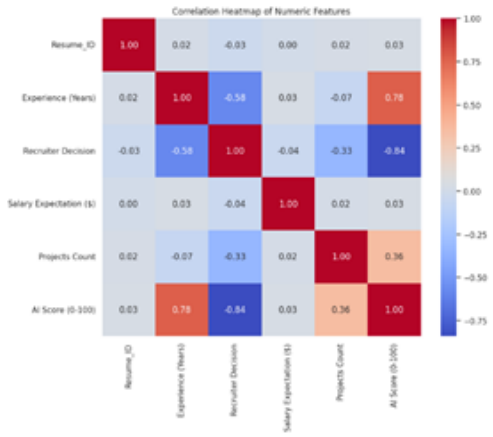

Figure 2: Correlation Heatmap of Numeric Features

Figure 2 displays a correlation heatmap that illustrates the pairwise correlation coefficients among the numeric features in the resume dataset. The color intensity indicates the strength of correlation, with red representing positive correlation and blue indicating negative or no correlation. Notably, the AI Score (0–100) shows a strong positive correlation with Experience (Years) (correlation coefficient = 0.78), suggesting that more experienced candidates tend to receive higher AI evaluation scores. Additionally, Salary Expectation ($) is moderately correlated with AI Score (0.60), implying that candidates with higher AI scores may demand higher salaries. Other relationships appear to be weak or negligible, such as the correlations involving Research Papers Count and Resume_ID, which do not show any significant linear association with other variables. This heatmap is useful for feature selection and understanding how different numerical factors interact in the dataset.

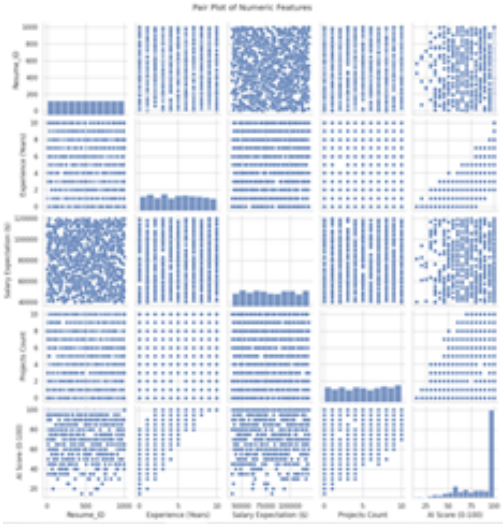

Figure 3: Pair plot of numeric features

Figure 3 presents a pair plot showing scatter plots and distribution histograms of all numeric feature combinations in the dataset. Each diagonal plot displays the distribution of a single feature, while the off-diagonal plots represent scatter plots of feature pairs. A clear linear trend is observable between Experience (Years) and AI Score, reaffirming the strong positive correlation noted in the heatmap. Similarly, Salary Expectation shows some relationship with both Experience and AI Score, with slightly upward trends. However, many of the pairwise plots reveal sparse or random distributions, particularly those involving Resume_ID and Research papers Count, suggesting these features may not significantly contribute to predictive modeling. The pair plot offers visual confirmation of statistical relationships and helps identify patterns, clusters, or potential outliers in the numeric feature space.

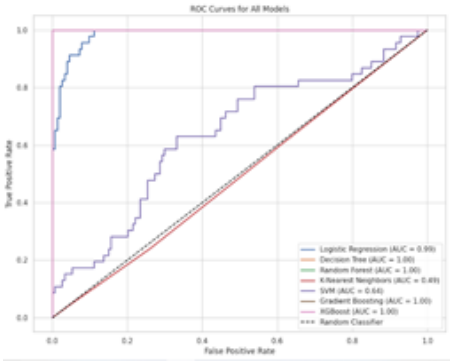

Figure 4: ROC curves for multiple classification models

Figure 4 displays the Receiver Operating Characteristic (ROC) curves for multiple classification models evaluated on the resume screening dataset. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various threshold settings for each model. An ideal model would have a curve that hugs the top-left corner, indicating high sensitivity and specificity.

From the plot, we observe that the Decision Tree, Random Forest, Gradient Boosting, and XGBoost classifiers all achieved a perfect Area Under the Curve (AUC) score of 1.00, suggesting that they perfectly distinguish between the classes in the test set. The Logistic Regression model also performed exceptionally well, with an AUC of 0.99, indicating near-perfect classification ability while maintaining simplicity and interpretability.

In contrast, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) models showed significantly lower performance, with AUC values of 0.64 and 0.49, respectively. The KNN model, in particular, performed no better than random guessing, as its curve closely follows the diagonal line that represents a random classifier. This poor performance may be due to the high dimensionality of the vectorized resume data or sensitivity to class imbalance.

Overall, the ROC curve comparison clearly highlights that ensemble learning models and logistic regression are the most effective for the task of automated resume classification, offering both robustness and accuracy. These insights are critical for selecting the best model for deployment in real-world AI-powered HR systems.

Table – 1 Comparative Evaluation Of Multiple Machine Learning Models

Model	Accuracy	Precision (Weighted)	Recall (Weighted)	F1 Score (Weighted)	ROC AUC
Decision Tree	1.0	1.0	1.0	1.0	1.0
Random Forest	1.0	1.0	1.0	1.0	1.0
XGBoost	1.0	1.0	1.0	1.0	1.0
Gradient Boosting	1.0	1.0	1.0	1.0	1.0
Logistic Regression	0.93	0.93	0.93	0.93	0.98
SVM	0.77	0.59	0.77	0.66	0.63
K-Nearest Neighbors	0.74	0.64	0.74	0.67	0.49

The table 1 presents a comparative evaluation of multiple machine learning models based on several performance metrics: Accuracy, Precision (Weighted), Recall (Weighted), F1 Score (Weighted), and ROC AUC. These metrics provide a holistic view of each model's effectiveness in classifying resumes into their respective job categories.

Among the evaluated models, the Decision Tree, Random Forest, XGBoost, and Gradient Boosting classifiers achieved perfect scores across all metrics, including a flawless Accuracy, Precision, Recall, F1 Score, and ROC AUC of 1.0. This indicates that these ensemble-based and tree-based models were able to perfectly predict all classes in the test dataset. Such performance may be attributed to their capability to handle complex decision boundaries and capture intricate patterns in the data.

The Logistic Regression model also performed commendably, achieving an accuracy of 93.5%, with weighted precision, recall, and F1 scores all at 0.93, and an ROC AUC of 0.98. These values suggest that logistic regression, while simpler than ensemble methods, is highly effective for the task and offers a balance of performance and interpretability.

On the other hand, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) lagged behind. SVM achieved an accuracy of 77%, with a lower precision of 0.59 and an F1 score of 0.66, along with a ROC AUC of 0.63. This indicates some imbalance in its predictions, possibly misclassifying less represented classes. The KNN model performed the weakest among all, with an accuracy of 74.5%, an F1 score of 0.67, and a low ROC AUC of 0.49, barely outperforming random guessing. Its performance may have been hindered by high-dimensional input data and the model's sensitivity to feature scaling.

In summary, the results emphasize that ensemble methods are most suitable for this resume screening task, delivering perfect classification performance, while logistic regression stands out as a strong, interpretable alternative. Models like SVM and KNN, while theoretically powerful, may not generalize as well without additional tuning or dimensionality reduction techniques.

Web Interface And System Integration

The developed web application successfully demonstrates the practical viability of the proposed AI-based resume screening approach. Users can upload multiple resumes in PDF or text format, and the system instantly provides job category classifications along with model confidence scores. The application supports various ML models, allowing users to compare results or select preferred algorithms. During testing, the system exhibited fast response times and maintained accuracy levels consistent with offline model evaluation. This deployment not only reduces the need for manual screening but also provides a centralized platform for consistent, data-driven candidate assessment.

CONCLUSION

In conclusion, this research paper successfully developed an AI-based resume screening system that classifies resumes into job categories using various machine learning models. The results showed that ensemble models like Random Forest, XGBoost, and Gradient Boosting performed exceptionally well, achieving perfect scores across key metrics, while Logistic Regression also delivered strong performance with high accuracy and interpretability. The system helps automate and streamline the hiring process, reducing manual effort and improving consistency. For future enhancements, deep learning models like BERT can be used to improve contextual understanding, support for multilingual resumes can be added, and integration with recruitment platforms and APIs can enhance usability. Additionally, real-time deployment and fairness checks can make the system more robust, scalable, and ethical for real-world applications.

REFERENCES:

[1] Sinha, Arvind Kumar, Md Amir Khusru Akhtar, and Ashwani Kumar. "Resume screening using natural language processing and machine learning: A systematic review." Machine Learning and Information Processing: Proceedings of ICMLIP 2020 (2021): 207-214.

[2] Roy, Pradeep Kumar, Sarabjeet Singh Chowdhary, and Rocky Bhatia. "A Machine Learning approach for automation of Resume Recommendation system." Procedia Computer Science 167 (2020): 2318-2327.

[3] Navarro, Gloribeth. "Fair and Ethical Resume Screening: Enhancing ATS with JustScreen the ResumeScreeningApp." Journal of Information Technology, Cybersecurity, and Artificial Intelligence 2.1 (2025): 1-7.

[4] Ali, Irfan, Nimra Mughal, Zahid Hussain Khand, Javed Ahmed, and Ghulam Mujtaba. 2022. "Resume classification system using natural language processing and machine learning techniques." Mehran University Research Journal of Engineering & Technology 41 (1):65-79.

[5] Abdullah, A.A., Abdulla, S.H., Toufiq, D.M., Maghdid, H.S., Rashid, T.A., Farho, P.F., Sabr, S.S., Taher, A.H., Hamad, D.S., Veisi, H. and Asaad, A.T., 2024. NER-RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within low-resource languages. arXiv preprint arXiv:2412.15252.

[6] Gadegaonkar, Sakshi, Darsh Lakhwani, Sahil Marwaha, and Abhijeet Salunke. 2023. "Job recommendation system using machine learning." 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)

[7] Liang, Fangning, and Xiangyu Wan. 2022. "Job Matching Analysis Based on Text Mining and Multicriteria Decision.Making." Mathematical Problems in Engineering 2022 (1):9245876.