iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81569

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 14 Issue IV Apr 2026- Available at www.ijraset.com

Intelligent HR Analytics Platform for Workforce Prediction and Optimization

B.L.C. Chakravarthi[1], Ms. M. Amulya[2], K. Poorna Kumar[3], K. Srinivasulu[4], V. Izak[5]
[2]*Faculty M. Tech, B. Tech Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh – 522510*
[1,3,4,5] *Students, Department of Data Science, Acharya Nagarjuna University, Guntur, Andhra Pradesh – 522510*
Final-Year B. Tech Major Project – 2026

Abstract: ***In today's advanced world of managing people at work HR Management system face problems. They get huge job applications. it's valuable when employees leave the company and it's hard to identify how computer systems make decisions. To Solve these problems, we created a Smart system called Intelligent HR-Analytics Platform. This system integrates different technologies like understanding human language, Machine Learning and it explains how computers think. Our system has four main parts: One it will understands the resume and extract the text from resume using patterns and rank candidates according to the job description and skill match another part explains the work force prediction, In this we introduced to predict who might leave and how well they will perform, a part that explains how the computer makes decisions using Explainable AI(XAI) and a virtual assistant that answers the questions asked by the HR regarding the Employee using Retrieval-Augmented Generation(RAG). We tested our system with fake data that's similar to real data, and it worked much better than old systems that don't work together. It was more accurate and easier to understand, which can help companies make better decisions.***
Keywords: Intelligent HR Analytics Platform, Machine Learning (ML), Explainable AI(XAI), Retrieval-Augmented Generation.

I. INTRODUCTION

Human Resource Management is Crucial to Organizational success. There are many Procedures remain manual leading to inefficiencies and errors. Major Challenges include slow resume screening, Difficulty in identifying in advance attrition signals and limited visibility in AI-driven decisions. Existing Solutions such as NLP based screening, attrition prediction models and HR chatbots are often developed in separation process resulting Unorganized systems. This study simplifies these issues by introducing the Intelligent HR Analytics Platform (IHRAP), An Integrated AI based system. The system performs Resume parsing based on matching skills, experience, education and applies machine learning for attrition prediction and performance analysis includes explainable AI (SHAP) for transparent decision making and includes a chat-oriented Assistant for real time, insights based on data. The platform is designed to be flexible and easy to evaluate.

A. Importance

The proposed Intelligent HR Analytics platform addresses the existing problem and providing the solution using an integrated AI-driven solution that combines Resume parsing, Predictive Analytics, explainable AI, combines into a unified system. In future the assistant enables real-time interaction and insights making the system.

II. LITERATURE SURVEY

Initial Resume Screening Depended on keyword matching, which missed accuracy and faced difficulties with different formats. Our model performs TF-IDF based scoring to evaluate skills and matching along with an explainability module to Guarantee candidate ranking.

A. Employee Attrition Prediction

It is very hard to know when employees will leave a company and predicting employee resignation is difficult. But some machine learning tools like Random Forest can actually works well. Our system can also predict future trends and patterns.

B. Explainable AI in HR Decision Systems

Clarity is required in HR AI systems. Techniques like SHAP help model decisions. But can be Digitally expensive. Our approach uses an Optimized method to deliver fast and reliable explanations.

C. Conversational AI for HR Applications

HR Chatbots today depend on RAG Techniques to provide precise and relevant responses. Our System uses this approach to deliver dependable insights into workforce analytics and HR policies.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. Architectural Overview

IHRAP does not depend too much on another part and the system is divided into four separate parts.

1. Resume Intelligence
2. Workforce Analytics
3. Explainable Decision support
4. Generative HR Assistant

Each Module runs independently through its own API while sharing a common data layer, enabling scalability, Network and flexible deployment. The shared layer stores employee data in JSON format, along with machine learning models.

B. Module 1: Resume Parsing

1) Resume Intelligence Engine:

Imagine a company gets 500 resumes for a job opening. HR has to read all 500, find the right skills, check experience and rank candidates. That takes weeks. But Module I does this automatically in seconds. We have to give a job description and resumes and it gives us ranked list of candidates with scores. In this Module Name Extraction, Email Extraction, Skills Extraction, Experience Extraction According to the patterns followed by the Module I.

2) TF-IDF Vectorization and Cosine Similarity

Term Frequency and Inverse Document Frequency defines how many words appears in the document and how rare or unique is that word across all documents.

Example:

Resume has: “Python Python Machine Learning Python”

TF of python = ¾ =0.75 (appears 3 out of 4 words)

IDF of python = high (it’s an important technical word)

TF-IDF of Python = high score

But the word “the” appears everywhere

IDF of the “the” = very low (not unique, not important)

TF-IDF of “the” = very low score

So, TF-IDF gives high scores to important words and low scores to common words.

Cosine Similarity:

After converting text to TF-IDF numbers (vectors), w measure how similar two vectors are.

Similarity = (Resume Vector JD vector) / (Resume * JD)

Result is between 0 and 1

We multiply by 100 get percentage.

C. Module II: Work Force Analytics and Prediction

Once a company hires employees through Module 1, they need to Manage those employees. Module 2 does exactly that. It uses Machine Learning to answer 4 big HR questions:

1) Who is going to leave the company?
2) How will each employee Perform next year?
3) What Skills are employees are missing?
4) How many new people do we need to hire in the next 3 years?

Replacing one employee costs 50 to 200 percent of their annual salary. So, if an employee earning 5 lakhs, when an employee leaves the company, the company spends up to 10 lakhs finding and training a replacement. If HR knows 3 months in advance that someone is likely to Leave the company they can give promotion, increase salary, assign better work that save the company crores of rupees.

we create a fake set of data about 500 employees in this data we have inserted the patterns which are ready to give the accurate results for the above questions. Real HR data is highly confidential companies do not share it publicly. We need full control over the data to demonstrate all features clearly. Our module is flexible with real IBM HR dataset from Kaggle we can plug it in directly.Random Forest plays a key role in attrition prediction it handles imbalanced data and gives feature importance and does not overfit easily.

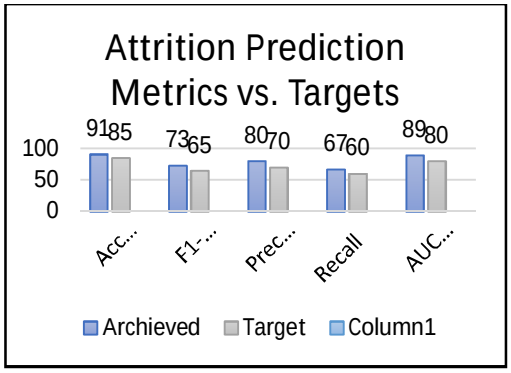


Fig.1 Attrition prediction metrics vs targets

Classification predicts a category and Regression predicts a number it's a continuous value.
Gradient Boosting is the best algorithm for regression on tabular data with mixed feature types like numbers, categories, satisfaction.
Work Demand Forecasting:
Hires Needed = (Current * Attrition rate*year) + (Current * Growth rate * year)
In this Replacement hiring and Growth Hiring will be done

D. Module III: Explainable AI and Decision support

XAI= Explainable Artificial Intelligence
It is the procedure of making AI decisions understandable to humans.
SHAP = Shapely Additive Explanations
When HR Manager clicks on any employee in the table. The system loads the employees 17 feature data created by the company, like Age, Years at company, job satisfaction, overtime, monthly income etc...
It performs base prediction first according to the employee.
After Running SHAP loop on employee features red bars indicates these increase the attrition risk and green bars indicate these decrease the attrition risk. Based on risk factors:
HR decides to conduct 1-on-1 meeting to improve job satisfaction and review compensation salary may be below market rate and reduce overtime workload to improve work life balance.

Formula:

Φ_i = P (all 117 features) – P (17 features with feature i = 0)
Where:
Φ_i = contribution of feature i
P () = prediction probability from the model
Positive Φ = feature is pushing toward "will leave"
Negative Φ = feature is protecting against leaving.
AutoML Workflow
Without Manual Work selecting which algorithm to use, AutoML tries Multiple Algorithms and it tests each one picks the best automatically.
Step 1: Load all 500 employee records
Step 2: set up 4 candidate algorithms
Candidate 1: Logistic Regression
Candidate 2: Decision Tree (max_depth = 6)

Candidate 3: Random Forest (100 trees)
Candidate 4: Gradient Boosting (100 trees)
It splits the 500 employees into 5 groups (100 each)
Each Algorithm gets 5 different scores
Final score = Average of all 5 scores
Step 4: Compare F1 scores
Step 5: Pick the Winner

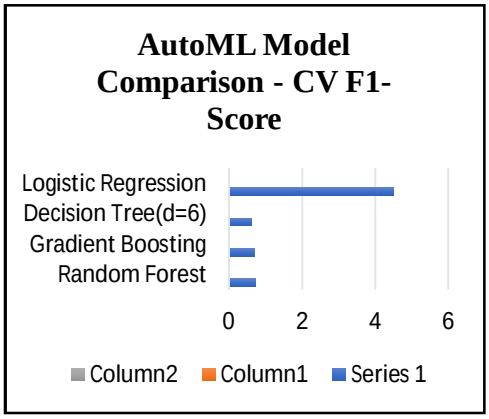


Fig. 2 AutoML Model

Bias Detection
Bias means the AI is unfair to certain groups of people.
If the model predicts higher attrition risk for women simply because there are fewer women in the training data that is bias not real risk.

E. MODULE IV: Generative AI powered HR Assistant
Without this module HR Manager would have to open Module 2 dashboard to know the highest attrition.
Using this Module HR Manager work becomes efficient. HR will take to Assistant like person-to-person Communication.
In this HR Manager simply types question in plain English, System understands the question, System retrieves relevant HR policy documents, System combines policies and live ML predictions, System generates a clear, accurate answer with sources.
RAG = Retrieval Augmented Generation
It is a technique that makes AI answers correct and reliable.
Retrieval – Find relevant documents from a knowledge base.
Augmented – Add extra context to those documents.
Generation – Create a clear answer based on retrieved + augmented info.
How RAG Works:
Step 1: User types a Question
Q: Which employees are at highest risk of leaving?
Step 2: Tokenize the Query
Query tokens = {which, employees, highest, risk, leaving}
Stop words removed = {are, at, of}
Step 3: The system figure outs what kind of question is this is:
According to the keywords it finds the categories of matching question.
Step 4: Document Retrieval
The system counts how many query words appear in the document and finds the top 3 documents retrieved those who are having highest score.
Step 5: Live ML
This makes module 4 special compared to normal assistants.
Instead of answering from the old documents, it pulls live data.

Step 6: Response Generation

This is based on retrieved documents and live ML data.

We have Generated Some static questions that would use generally. We have created 8 questions when HR Managers click any of these questions instead of typing. It Makes HR Managers work more efficient and reliable.

Feature	Normal Chatbot	Module 4 (RAG)
Data source	Static training data	LiveML Predictions + fresh docs
Hallucination	High risk	Very low grounded in facts
HR Policy accuracy	May be wrong	Pulled from actual policy docs
Attrition stats	Outdated or guessed	Real-time from module 2 models
Source attribution	None	Policy document chips shown

Table 1. RAG Representation

IV. IMPLEMENTATION DETAILS

A. Technology Stack

IHRAP is implemented in python 3 using and freely available tools. Machine Learning part is built using scikit-learn which helps in creating models like Random Forest and Gradient Boosting and also it allows testing the model using correct cross validation. The backend is created using flask it is a lightweight framework used to create the APIs.

For formatting data and performing calculations NumPy, Pandas are used. Here, saved models are stored using a method called pickle. So, they can be retrieved later without training the model again. The frontend is created using web technologies like HTML, CSS and JavaScript without using any frameworks. This makes model looks simple and easy to use and deploy.

B. Dataset Generation Protocol

A Dataset is created by our own using a static random value so that same data can be created again if needed. Employees were assigned to different departments across seven units. Numerical details like age, salary and distance from work were generated within practical boundaries using random values.

Whatever the ratings and predictions were based on probabilities to match real world HR data patterns.

C. Model Training Protocol

In general Model training is done using the dataset and that dataset is divided into two parts that one part is 80% of the data inn dataset is used for training and 20 % of the data is used for testing to maintain model predictions efficient.

Various algorithms are trained like this, the algorithms are Random Forest Classifier, Fivefold cross validation.

V. EXPERIMENTAL RESULTS

Module 1 Analysis was conducted using a Data Science job description against a sample five candidate profiles with changing qualification levels. Each employee was placed into one of seven departments and we tried to separate them accordingly. A candidate ranked third with skill = 60%, In the end about 15 out of 100 employees leave which looks like realistic.

Random Forest Model performed very well. It correctly predicted outcomes 91% of the time.

The model was good at finding employees who might leave the company. It got a best balance parameter between correctly finding employees who will leave and not wrongly marking too many people.

Metric	Test set	CV (5-fold)	Target	Status
Accuracy	91.0%	----	≥85%	Met
F1-Score	0.730	0.728 ± 0.104	≥ 0.65%	Met
Precision	0.800	-----	≥ 0.70	Met
Recall	0.670	-----	≥ 0.60	Met
AUC-ROC	0.891	----	≥ 0.80	Met

Table 2. Prediction results

VI. EVALUATION METRICS

The model we created had found that some factors are more important than other predicting if an employee will leave. There are mainly two factors: An employee years at company is the most important factor. Employees who are new (0-2 years) having more chances to leave. And Second factor is somewhat interesting that an employee Number of companies Worked is the second most important. People who changed jobs many times before are having more chances to leave again. And also, our model also predicts how well employees will perform the prediction is very small in the sense model is fairly accurate.

The model explains about 63% of employee performance which is good. The remaining part depends on things that are hard to measure, like teamwork personal situations or project difficulty. We have not used static algorithms, instead of this we have used AutoML comparative Evaluation it means using different models which will give best results like Random Forest, Gradient Boosting, Logistic Regression. This shows that predicting employee attrition is not simple and basic models don't work well.

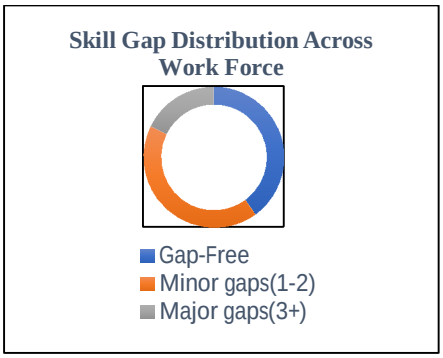


Fig. 3 Skill gap Distribution

This system also checks what skills employees are missing, on average each employee is missing 1-2 skills. The most common missing skills are: SQL, python, Docker, AWS.

In our employee dataset 40% have all required skills, 42% have small gaps (1-2 skills), 18% have big gaps.

Result Category	Metric	Value	Target	Status
Resume Ranking	Ranking speed	< 1sec/ 10 resumes	< 5sec	Exceeded
Resume Ranking	Verdict accuracy	Correctly ordered	Expert-aligned	Validated
Attrition	Test accuracy	91.0%	>=85%	Exceeded
Attrition	F1 Score	0.730	>=0.65	Exceeded
Attrition	CV F1	0.728	>=0.65	Exceeded
Attrition	Precision	0.800	>=0.70	Exceeded
Attrition	Recall	0.670	>=0.60	Exceeded
Attrition	AUC-ROC	0.891	>=0.80	Exceeded
Performance	MAE	0.356	<=0.50	Exceeded
Performance	R^2 Score	0.627	>=0.60	Exceeded
XAI	SHAP speed	<1 sec/ employee	< 3sec	Exceeded
Auto ML	Best F1	0.730	>=0.65	Exceeded
System	API response time	< 0.5 sec avg	< 2 sec	Exceeded

Table 3. Model results

The Current model has three disadvantages. First, the synthetic is aligned with only some patterns, it cannot fully perform real world organizational pattern complexity. While training model would need to train on real HR data.

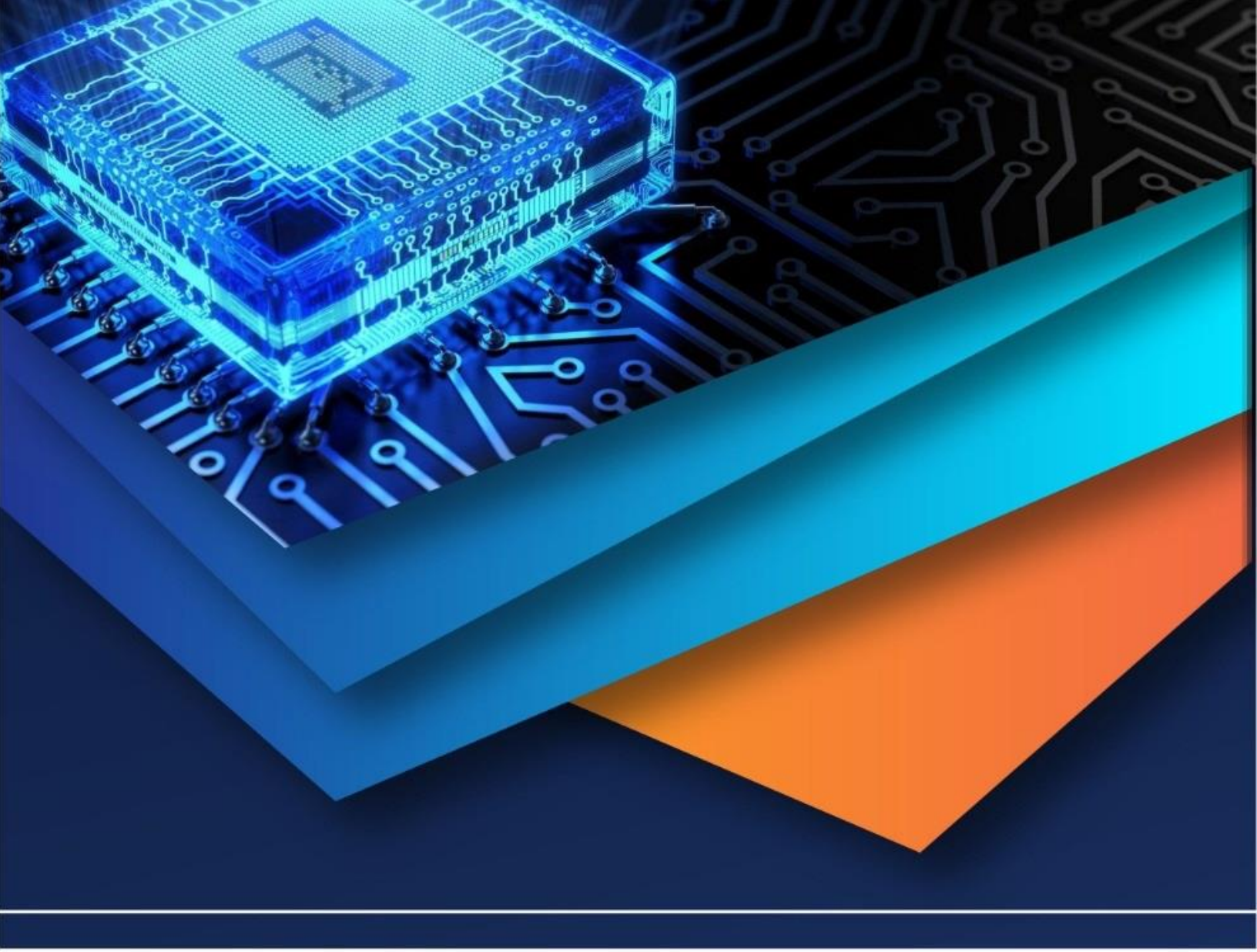
VII. CONCLUSION AND FUTURE WORK

This paper introduced Intelligent HR Analytics Platform framework designed to make the HR Management work more efficient. By using NLP-based resume analysis and Machine Learning for workforce analytics, SHAP based explainability, AutoML Algorithms model selection, RAG based conversational assistant. The system provides one common approach to HR decision making. Experiment results represent strong performance with 91% accuracy and F1 score of 0.73 for attrition prediction. By depending on individual metrics, the main contribution of IHRAP lies in its ability to combine multiple HR functions into a single architecture. This study shows that insights can be achieved with limited data offering both research foundation and practical framework that organizations can adapt to their own HR systems.

Future work will focus on improving both model capability and real-world suitability. First the retrieval component can be improved using dense vector-based methods (e.g. Sentence Transformers) to better handle in terms of meaning similar queries and improves response accuracy. Second, this model includes time-based patterns that will allow the system to identify the changes over time, enabling more accurate prediction of employee performance particularly for identifying lack of involvement. Finally deploying and evaluating the system in a real organizational setting using HR data will be a critical next step. This will help validate is effectiveness in practice, find problems during implementation and improve the system to better support decision making and employee outcomes.

REFERENCES

[1] R. E. Miles and C. C. Snow, "Organizational Strategy, Structure, and Process," McGraw-Hill, 1978.
[2] P. Nandhini and R. Bhatnagar, "Automated Resume Screening Using Natural Language Processing Techniques," Int. J. Comput. Sci. Mob. Comput., vol. 9, no. 4, pp. 21–29, 2020.
[3] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Comput. Sci., vol. 2, art. 160, 2021.
[4] M. Tuffaha, B. Pandya, and M. R. Perello-Marin, "AI-Powered Chatbots in Human Resource Management," J. Contemp. Issues Bus. Gov., vol. 28, no. 3, 2022.
[5] A. Holm, "Cognition Constructs in Automated Resume Screening: A Review," Comput. Human Behav., vol. 30, pp. 493 503, 2014.
[6] P. Sriram and R. Bhatnagar, "An NLP Framework for Automated Candidate Shortlisting," in Proc. ICTAI, 2020, pp. 45 52.
[7] E. Saatci et al., "Resume Screening Using Natural Language Processing," AI Appl. J., vol. 12, no. 2, pp. 77–89, 2024.
[8] V. Jagwani, S. Meghani, K. Pai, and S. Dhage, "Resume Evaluation through NLP and Machine Learning for Effective Candidate Selection," arXiv:2301.09756, 2023.
[9] H. Kaur and R. Sharma, "Employee Attrition Prediction Using Machine Learning Techniques," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 45–58, 2022.
[10] Z. Zhao, Y. Liu, and S. Chen, "A Systematic Meta-Analysis of Machine Learning Methods for Employee Attrition Prediction," Expert Syst. Appl., vol. 211, art. 118662, 2023.
[11] [M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Hiring," in Proc. ACM FAccT, 2020, pp. 469–481.
[12] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," in Proc. FAT* Conf., 2018, pp. 149–159.
[13] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.
[14] I. Covert and S. I. Lee, "Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression," in Proc. AISTATS, 2021, pp. 3457–3465.
[15] T. B. Brown et al., "Language Models are Few-Shot Learners," Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020.
[16] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Adv. Neural Inf. Process. Syst., vol. 33, pp. 9459–9474, 2020.
[17] IBM HR Analytics Employee Attrition & Performance Dataset. [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
[18] D. G. Allen, P. C. Bryant, and J. M. Vardaman, "Retaining Talent: Replacing Misconceptions with Evidence-Based Strategies," Acad. Manage. Perspect., vol. 24, no. 2, pp. 48–64, 2010.
[19] C. D. Crossley, R. J. Bennett, S. M. Jex, and J. L. Burnfield, "Development of a Global Measure of Job Embeddedness," J. Appl. Psychol., vol. 92, no. 4, pp. 1031–1042, 2007. [20] World Economic Forum, "The Future of Jobs Report 2023," WEF, Geneva, Switzerland, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY