Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue S-8 Year 2023 Page 164 -172

Predicting HR Churn with Python and Machine Learning

J. K. Patil[1]*, P. M. Jadhav[2]

[1,2] *Department of Computer Science, Changu Kana Thakur Arts, Commerce and Science College, New Panvel, jadhavprati2013@gmail.com*

***Corresponding Author:** J. K. Patil*
** Department of Computer Science, Changu Kana Thakur Arts, Commerce and Science College, New Panvel, jadhavprati2013@gmail.com*

Abstract

Employee turnover imposes a substantial financial burden, necessitating proactive retention strategies. The aim is to leverage HR analytics, specifically employing a systematic machine learning approach, to predict the likelihood of active employees leaving the company. Using a systematic approach for supervised classification, the study leverages data on former employees to predict the probability of current employees leaving. Factors such as recruitment costs, sign-on bonuses, and onboarding productivity loss are analysed to explain when and why employees are prone to leave. The project aims to empower companies to take pre-emptive measures for retention. Contributing to HR Analytics, it provides a methodological framework applicable to various machine learning problems, optimizing human resource management, and enhancing overall workforce stability. This research contributes not only to predicting turnover but also proposes policies and strategies derived from the model's results. By understanding the root causes and timing of employee departures, companies can proactively implement measures to mitigate turnover, thereby minimizing the associated financial and operational burdens.

Keywords: Employee Turnover, HR Analytics, Churn, Retention.

I. INTRODUCTION

In today's rapidly evolving business landscape, employee retention has become a critical concern for organizations. Given how expensive it is to replace talented workers and the impact it has on ongoing work and productivity, companies are increasingly turning to data-driven approaches for HR churn analysis.

The use of HR analytics has been pivotal in not only gathering data but also predicting how to improve processes to achieve the desired goal. This is particularly important in predicting employee churn and its impact on retention. Telecommunication companies have sought to be more proactive by participating in data mining and machine learning-based models for churn analysis. Additionally, the use of data analytics and artificial intelligence-based methods has proven to be effective in understanding customer churn and developing profitable customer retention programs. Notably, machine learning classification and assembling methods have been incorporated for customer churn analysis, and research has explored support vector machine

methods for predicting employee turnover in the IT industry. This shows a growing trend towards leveraging advanced analytics and machine learning for churn prediction and management.

II. RELATED WORK

This paper examines HR analytics and discovers that, despite proof of its benefits, adoption is minimal. Inadequate IT infrastructure and gap of skills are two issues that HR must deal with. The report highlighted how HR's strategic role is changing and calls for more research to address challenges and hiccups in integrating analytics into HR procedures [1].
This study addressed the significant impact of employee churn on organizations and aimed to identify the best prediction model among naïve Bayes, decision tree, and random forest. Analyzing data from a telecom company in Indonesia, the study concluded that the random forest was the most reliable model, achieving an impressive accuracy of 97.5%. This finding is critical for averting negative organizational consequences [2].
This paper highlighted the significant impact of employee income on organizations, emphasizing the time and cost involved in recruitment and the subsequent period of adjustment for new recruitment. The proposed expert prediction system, utilizing machine learning methods, aimed to address the lack of understanding of employee behavior, helping organizations prevent unnecessary churn and foster mutual growth [3].

This paper explored the transformative role of technology, particularly machine learning, in Human Resource Management during the fifth industrial revolution. Emphasizing the status of employee satisfaction and retention, it highlighted the application of machine learning techniques like classification and clustering to analyze data and make predictions, providing insights to help manage and control employee turnover [4].
This study addressed these issues of employee churn prediction, employing well-established classification methods on HR data. Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes were assessed for accuracy, precision, recall, and F-measure values. The results, along with a feature selection method, offered companies insights to predict and mitigate employee churn, reducing human resource costs[5].
These papers addressed the significant issue of employee attrition, leveraging machine learning, specifically Random Forest, on HR analytics data. The use of this mechanism enhanced model performance, particularly in addressing class imbalance. The studies identified influential factors in attrition, aiding strategic decisions for organizational engagement. The potential for further research to reduce prediction errors was emphasized [6].
In the second research investigation, critical elements contributing to employee attrition in HR analytics were explored, highlighting the impact of factors such as overtime, project involvement, and job level. Utilizing classification algorithms for attrition prediction, the study identified decision trees as the most accurate method. Machine learning, particularly Random Forest, outperformed others, aiding HR in predicting employee attrition effectively [7].

III. PROBLEM DEFINATION

Organizations have a substantial financial cost due to employee turnover, which makes proactive retention strategies creation and execution imperative. The necessity for a predictive strategy to control the probability of active employees leaving a firm is sensitive by the high expenditures linked to recruiting, sign-on incentives, and employee productivity loss during onboarding. This study intends to use HR analytics to tackle this problem by using a methodical machine learning technique for supervised categorization. The main goal is to use data from previous employees to anticipate the likelihood that current employees would leave.
Employee churn, or turnover, refers to the rate at which employees leave a company within a specific time period. High employee churn can be costly for organizations, affecting productivity, morale, and overall business performance. Predicting and understanding the factors that donate to employee churn is crucial for Human Resource (HR) management to take proactive measures and retain valuable talent.

IV. OBJECTIVES

The objective of this project is to develop a machine learning model to predict employee churn based on various features and factors.

- To train machine learning model capable of predicting employee churn.
- To assess and examine metrics for different models.
- To recommendate strategies to mitigate employee churn.

V. METHODOLOGY

The methodology for this research paper involves several steps. First, data collection and preprocessing are conducted to gather relevant HR data and ensure its quality.
Next, exploratory data analysis is performed to gain insights into the dataset and identify any patterns or anomalies.'

Afterwards, feature engineering is carried out to select and transform the necessary attributes which effectively predict employee churn.
This involves techniques such as one-hot encoding, feature scaling, and handling missing data. Once data is prepared, various machine learning algorithms are applied to build predictive models for employee churn. These algorithms include Logistic Regression, Support Vector Machine, and Random Forest.

a. Data Exploration and Preprocessing:
To obtain an understanding of the feature distribution, the first stage of the research is a comprehensive examination of the dataset. This entails being aware of each variable's statistical characteristics and seeing any possible trends or patterns. Preprocessing will be applied on data to handle outliers, missing numbers, and other necessary corrections. This process prepares the dataset for subsequent analysis while also guaranteeing its integrity. Visualizations will be utilized to examine the connections between different characteristics and the desired variable in order to reveal any possible connections and important factors impacting employee turnover.

b. Feature Engineering:
The goal of the feature engineering stage was to find and utilize pertinent features that might have an effect on employee attrition. This involved the careful identification of factors that could be crucial in forecasting attrition. To enhance the model's capacity to identify subtle trends in data, additional characteristics were developed, such as tenure, satisfaction ratings, or other derived metrics. The objective was to add valuable information to the dataset that would improve the accuracy of the prediction models. This stage played a crucial role in enriching the dataset with meaningful features, providing the models with a more comprehensive set of inputs for predicting employee attrition. The features identified and engineered during this phase aimed to capture nuanced patterns within the data, ultimately contributing to the overall effectiveness of the predictive modelling process.

c. Model Selection:
After the dataset had been prepared, appropriate machine learning algorithms for binary classification were selected. Commonly employed algorithms, including logistic regression, decision trees, random forests, and support vector machines, were considered. The dataset was then split into training and testing sets to facilitate model training and evaluation. The selection of suitable models was pivotal to the success of the analysis, and different algorithms were considered to find the most effective one for the specific context of employee churn prediction. This stage marked a critical decision point in determining the approach that would be most adept at capturing patterns within the dataset and accurately predicting employee churn.

d. Model Training:
Using the training dataset, a subset of machine learning models was trained, and performance was maximized by adjusting hyperparameters. To improve the models' predictive power, an incremental refining approach was used. The models' ability to adjust to the distinct features of the data was enhanced by adjusting parameters, which eventually improved the models' predictive accuracy of employee attrition. The trained models were then subjected to a thorough evaluation on the testing dataset using measures including ROC-AUC, accuracy, precision, recall, and F1 score. The goal of this thorough assessment was to measure the models' performance and offer a benchmark for comparison. Considering the advantages and disadvantages of each model, the best option for forecasting employee turnover in the organizational setting was chosen.

VI. IMPLEMENTATION

Employee Churn Prediction:

❖ **Data Analysis:**
Data analysis begin by thoroughly analysing the employee dataset, encompassing both current and past employee records. This step involves gaining insights into the structure of the data, understanding its variables, and identifying potential patterns or trends.

❖ **Data Cleaning and Feature Derivation:**

The dataset for this analysis includes historical information about employees, such as demographics, job-related factors, performance metrics, and any other relevant features. Each record in the dataset represents an employee, and the target variable is a binary indicator (churn or no churn). In the initial phase of analytical approach, the pivotal task revolves around data selection and refinement for the purpose of predicting employee attrition, leveraging the HR Employee Attrition dataset provided by IBM. This dataset encompasses a spectrum of employee details including demographics, experience, skills, nature of work or unit, and position, among others. To streamline; a meticulous curation process was implemented to discern and retain features that hold significance in predicting attrition.

Of the original 34 features, some were deemed extraneous as they either lacked variability or were uniformly applicable across all records; for instance, the age of all employees exceeded 18 years. Additionally, certain attributes such as employee ID or name were recognized as non-contributory to the analysis and were consequently excluded. Following the elimination of these superfluous features, the dataset was refined to encompass 30 pertinent features. Table I delineates the selected features, their respective types, and definitions, providing a comprehensive overview of the refined dataset. This meticulous culling process ensures that subsequent analysis is anchored in a relevant and streamlined dataset conducive to accurate predictions of employee attrition.

Table 1.1 HR Dataset Features

No	Features	Data Type
1	Age	Numeric
2	Attrition	Numeric
3	Business Travel	Categorical
4	Daily Rate	Numeric
5	Department	Categorical
6	Distance from home	Numeric
7	Education	Categorical
8	Education Field	Categorical
9	Employee count	Numeric
10	Employee Number	Numeric
11	Environment Satisfaction	Categorical
12	Gender	Categorical
13	Hourly Rate	Numeric
14	Job Involvement	Categorical
15	Job level	Categorical
16	Job Role	Categorical
17	Job Satisfaction	Categorical
18	Marital status	Categorical
19	Monthly Income	Numeric
20	Monthly rate	Numeric
21	Num Companies Worked	Numeric
22	Age over 18	Numeric
23	Over time	Categorical
24	Percent Salary Hike	Numeric
25	Performance rating	Categorical
26	Relationship Satisfaction	Categorical
27	Standard Hours	Numeric
28	Stock Option Level	Categorical
29	Total Working Years	Numeric
30	Training time Last Year	Numeric
31	Work Life Balance	Categorical
32	Years At Company	Numeric
33	Year Since Last Promotion	Numeric
34	Years With Current Manager	Numeric

Data preprocessing holds pivotal significance in analytical approach, emphasizing its role in yielding robust results even with straightforward algorithms. A pristine dataset is foundational for accurate predictions. To enhance the analysis, The study explored additional attributes derivable from existing employee data. Furthermore, addressing missing data becomes imperative, prompting the application of diverse imputation techniques to ensure data completeness and reliability. This meticulous preprocessing stage ensures that the dataset is not only refined but also enriched, laying a solid foundation for subsequent analyses and predictions in quest to understand and predict employee attrition.

❖ **Feature Selection:**
Not all features contribute equally to predicting churn. Carefully selected features that are most relevant to the churn prediction task. This ensures that the model is focused on the key factors influencing employee turnover.
❖ **Classification Model Exploration:**
Analysis experiment with a variety of classification techniques, ranging from simpler methods like Naive Bayes, linear regression, and nearest neighbours to more sophisticated approaches such as Support Vector Machines (SVM) and Random Forests. Evaluate these models based on performance metrics like accuracy, precision, recall, and F-measure using a dedicated test dataset.
❖ **Feature Selection Refinement:**
To enhance the efficiency of the model, applied feature selection methods. This step involved identifying and retaining the most impactful features for predicting employee churn, further refining the model's predictive capabilities.
❖ **Building the Classification Model:**
With the selected features and refined dataset, constructed a classification model. This model is trained on historical data, learning patterns indicative of employee churn.
❖ **Churn Prediction:**
The trained model is then applied to predict potential churn among current employees. By analysing various indicators, the model helps identify individuals at a higher risk of leaving the company.
❖ **Retention Strategy Decision:**
Based on the model's predictions, provided recommendations for employee retention strategies. These strategies are designed to proactively address potential churn, ultimately aiding in the development of effective HR retention policies and practices.

This comprehensive approach involves a combination of data analysis, feature selection, and the application of diverse classification techniques to build a robust model for predicting and mitigating employee churn. The ultimate goal is to empower organizations with actionable insights to retain their valuable talent and foster a more stable and productive work environment.

❖ **Comparison of Classification Method**
In this evaluation phase, assessed various classification methods to determine their suitability in predicting employee turnover. To avoid overfitting, the study employed a train-test split strategy, using 75% of the dataset for training and 25% for testing. The goal is to gauge the model's accuracy, precision, and F-measure on the test set, providing insights into its predictive performance on unseen data. This approach ensures a robust assessment of the algorithm's generalization capabilities, helping us identify the most effective method for distinguishing between churners and non-churners in the workforce.
Detailed information about the distribution of the datasets after splitting into two parts.

Table 1.2 Train – Test datasets

Dataset	Split Data	percentage
Train	1102	75%
Test	368	25%
Total	1470	100%

Building Machine Learning Models
Initially, employ basic algorithms with default hyperparameters before exploring advanced solutions. The selected algorithms include Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree Classifier, and Gaussian Naive Bayes. Default settings will be utilized for simplicity in this preliminary stage of analysis.

Table 1.3 ML Algorithms

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
0	Logistic Regression	81.64	6.43	76.51	4.36
1	Random Forest	80.03	5.55	85.12	2.86
2	SVM	77.63	5.71	84.30	2.66
4	Gaussian NB	75.11	5.02	66.33	4.99
3	Decision Tree Classifier	65.37	6.30	80.31	4.05

Classification Accuracy, the ratio of correct predictions to total predictions, is a prevalent metric for classification problems. Despite its common usage, it can be misapplied. Accuracy assumes equal class distribution and treats all predictions and errors equally. This assumption may not hold in projects where class imbalances or varying importance of prediction errors exist. In such cases, an alternative scoring metric, accounting for specific project requirements, may be more appropriate for a comprehensive evaluation of model performance, ensuring a nuanced assessment that considers the unique characteristics and goals of the given project.

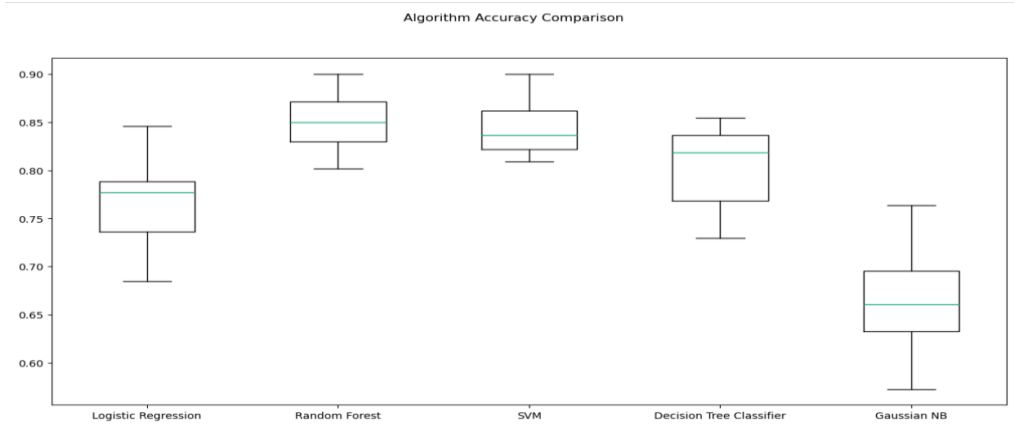

Figure 1: Accuracy Comparison

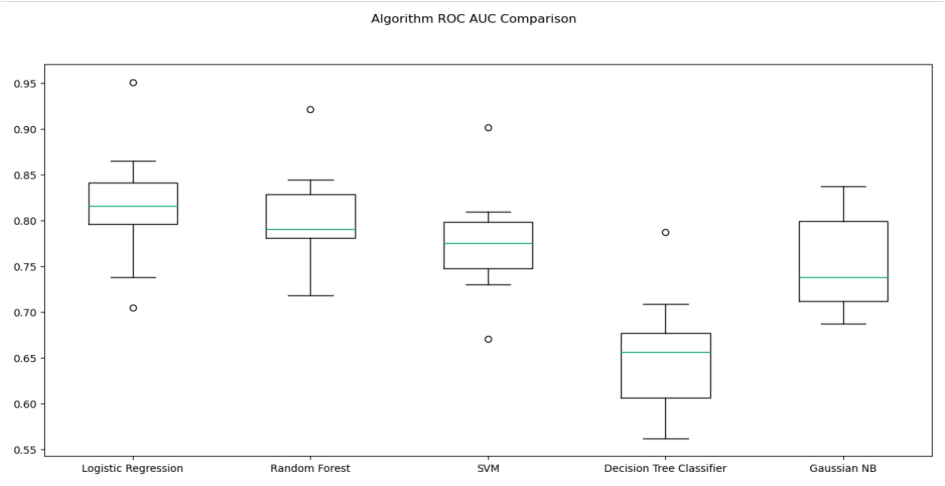

Figure 2: ROC AUC Comparison

ROC AUC is a metric commonly utilized to assess how well binary classification methods work. When you compare the ROC AUC of different algorithms, you are assessing their overall discriminatory power. Here is how the comparison works:

Range of ROC AUC values:
A model with an ROC AUC of 0.5 performs no better than random chance.
A model with an ROC AUC of 1.0 indicates perfect discrimination.

Higher ROC AUC is better:
In general, the algorithm with a higher ROC AUC is considered better at distinguishing between positive and negative instances.

Random baseline:
It is essential to compare the ROC AUC of your models against a random baseline (ROC AUC = 0.5)

Interpretation:
A model with an ROC AUC close to 0.5 might not be very useful for the task.
A model with an ROC AUC between 0.7 and 0.9 is generally considered good.
An ROC AUC above 0.9 suggests excellent discrimination.

ROC Curve Shape:
The shape of the ROC curve can also provide insights. If one algorithm's ROC curve dominates another is across the entire range of thresholds, it is likely to have a higher ROC AUC.
When comparing the AUC of different algorithms, consider the context of the problem and dataset. ROC AUC is a useful metric, but it might not be appropriate for all scenarios. Depending on the application, consider precision, recall, F1 score, or the confusion matrix.
Following ROC AUC comparison, Logistic Regression and Random Forest exhibit the highest mean AUC scores. These top-performing algorithms will undergo further analysis. Details on their characteristics and performance will be explored in the subsequent sections to determine their suitability for the specific project requirements.
Logistic Regression, a machine learning classification algorithm, predicts the probability of a categorical dependent variable. While less sophisticated compared to ensemble methods or boosted decision trees discussed later, it serves as a valuable benchmark. Its simplicity makes it an effective starting point for evaluating and comparing more advanced algorithms in subsequent analyses.
Random Forest, a widely used and versatile machine learning method, adeptly handles both regression and classification tasks. Falling under the umbrella of ensemble learning, it leverages a collection of decision trees for its predictions. Each decision tree within the ensemble contributes to the overall classification or regression output by making a series of decisions based on dataset observations.
Random Forest enables the identification of crucial features in predicting the target feature, such as "attrition" in this project. The following visualization illustrates feature importance rankings.

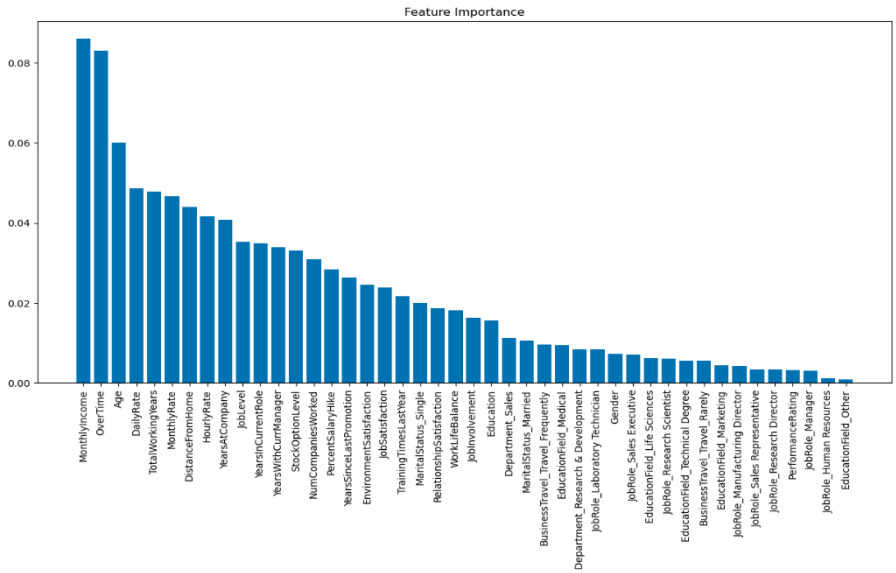

Figure 3: Feature Importance

Classification report for the optimised Random Forest Regression
The metrics provided are the evaluation results for a binary classification model, specifically applied to Human Resources (HR) churn analysis using the Random Forest algorithm in Python. Here is a more detailed explanation with a focus on HR churn analysis:

	precision	**recall**	**f1-score**	**support**
0	0.88	0.97	0.93	309
1	0.7	0.32	0.44	59
accuracy			0.87	368
macro avg	0.79	0.65	0.68	368
weighted avg	0.85	0.87	0.85	368

- Precision (for class 0.0, representing non-churn):
  Precision is 0.88, indicating that 88% of the employees predicted as not churning(class 0.0) by the model were not churning.
- Recall (Sensitivity, for class 0.0):
  Recall is 0.97, suggesting that the model correctly identified 97% of the employees who were not churning.
- F1-score (for class 0.0):
  F1-score is 0.93, which is the mean of precision and recall. It provides a balance between correctly identifying non-churning employees and avoiding false positives.
- Support (for class 0.0):
  There are 309 instances of employees who are not churning in the dataset.
- Precision (for class 1.0, representing churn):
  Precision is 0.70, indicating that 70% of the employees predicted as churning (class 1.0) by the model were churning.
- Recall (Sensitivity, for class 1.0):
  Recall is 0.32, suggesting that the model correctly identified only 32% of the employees who were churning. This is a lower value and indicates that the model is not as effective at capturing all instances of actual churn.
- F1-score (for class 1.0):
  F1-score is 0.44, which is lower than for class 0.0. This indicates that the model's performance on predicting churn (class 1.0) is not as balanced as for non-churn.
- Support (for class 1.0):
  There are 59 instances of employees who are churning in the dataset.
- Accuracy:
  The overall accuracy of the model is 0.87 or 87%, which is the ratio of correctly predicted observations to the total observations. This represents the overall performance of the model.
- Macro Avg:
  The macro-averaged precision, recall, and F1-score are calculated independently for each class and then averaged. In this case, the macro average precision is 0.79, recall is 0.65, and F1-score is 0.68.
- Weighted Avg:
  The weighted average accounts for the class imbalance by considering the class support. The weighted-averaged precision, recall, and F1-score are 0.85, 0.87, and 0.85, respectively.

In HR churn analysis, high recall for class 0.0 (non-churn) is crucial to identify employees who are likely to stay. However, the lower recall for class 1.0 (churn) suggests that the model is not capturing all instances of employees who are churning, which could be a concern for HR teams aiming to proactively address employee attrition. Fine-tuning the model or exploring other algorithms may help improve its performance, especially in identifying employees at risk of churn.

VII. CONCLUSION:

The HR Churn Analysis project is pivotal for organization's sustained success, focusing on leveraging data-driven insights to uncover and address factors influencing employee turnover. Recognizing the critical link between employee satisfaction, organizational stability, and overall productivity, this initiative aims to systematically collect and analyse relevant data on employee behaviour, performance, and engagement. By implementing advanced analytics and predictive modelling, identified key indicators of potential churn, allow

for targeted interventions. The anticipated impact includes improved employee satisfaction, enhanced organizational stability, and increased overall productivity. This strategic investment underscores commitment to cultivating a positive workplace environment, ensuring talent retention, and fostering long-term success. Through continuous monitoring, feedback loops, and adaptive strategies, we aspire to create a workplace where employees thrive, contributing to the sustained growth and prosperity of the organization.

REFERENCES

1. Zeidan, S. and Itani, N. (2020). HR Analytics and Organizational Effectiveness. International Journal on Emerging Technologies, 11(2): 683–688.
2. Andry Alamsyah , Nisrina Salma. (2022). A Comparative Study of Employee Churn Prediction Model, Research Gate
3. Aniket Tambde, Dilip Motwani. (2019) Employee Churn Rate Prediction and Performance Using Machine Learning. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019
4. Rodrigo Miranda Cabrera. Impact Of Machine Learning in Human Resource Management: Towards the Modernization of Leadership. Journal of Positive School Psychology http://journalppw.com 2022, Vol. 6, No. 2S, 290-299
5. İbrahim Onuralp Yiğit. An Approach for Predicting Employee Churn by Using Data Mining, September 2017 DOI:10.1109/IDAP.2017.8090324
6. Shobhanam Krishna and Sumati Sidharth. HR Analytics: Employee Attrition Analysis using Random Forest vol. 18, no. 4, April 2022, pp. 275-281 DOI: 10.23940/ijpe.22.04.p5.275281
7. Elham Mohammed Thabit A. Alsaadi "Identification of human resource analytics using machine learning algorithms ". Vol. 20, No. 5, October 2022, pp. 1004~1015 ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v20i5.21818
8. Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. Journal of Organizational Effectiveness: People and Performance, 4(2), 127-148.
9. Davenport, T.H., Harris, J.G. and Morison, R. (2010). Analytics at Work: Smarter Decisions, Better Results, Harvard Business School Press, Boston, MA.
10. Boudreau, J. W., & Ramstad, P. M. (2005). Talentship, talent segmentation, and sustainability:
11. a new HR decision science paradigm for a new strategy definition. Human Resource
12. Management, 44(2), 129- 136.
13. Jones, K. (2014). Conquering HR Analytics: Do you need a rocket scientist or a crystal ball?
14. Workforce Solutions Review, 5(1), 43–44.
15. Conboy, K., Dennehy, D., & O'Connor, M. (2020). 'Big time': An examination of temporal complexity and business value in analytics. Information & Management, 57(1), 1-13.
16. Dulebohn, J., & Johnson, R. (2013). Human resource metrics and decision support: A
17. classification framework. Human Resource Management Review, 23(1), 71–83.
18. Hota, J. & Ghosh, D. (2013). Workforce Analytics Approach: An emerging Trend of Workforce Management. AIMS International Journal of Management, 7(3), 167-179.
19. Chattopadhyay, D., Biswas, B. D., & Mukherjee, S. (2017). A new look at HR analytics. GMJ, 11(1), 41-51.
20. Sousa, M. J. (2018). HR analytics models for effective decision making. 14th European Conference on Management, Leadership and Governance, ECMLG, 256-263.
21. Reddy, P. R., & Lakshmikeerthi, P. (2017). HR Analytics' - An Effective Evidence Based HRM Tool. International Journal of Business and Management Invention, 6(7), 23-34.