Employee Turnover Prediction Using Supervised Machine Learning Models: A Comparative Evaluation Approach

Abstract— **Employee turnover poses a significant threat to organizations, impacting stability and expense. This paper introduces a machine learning model that forecasts whether an employee will quit based on a labeled HR dataset. The data were preprocessed using one-hot encoding and scaling prior to training six models: Logistic Regression, SVM, k-NN, Decision Tree, Random Forest, and Gradient Boosting. On evaluation with accuracy, recall, and precision, Random Forest was the top-performing model. It was then implemented with a prediction pipeline to evaluate new employee records. Results indicate the model's capability to assist HR departments in selecting employees most likely to quit and assist in enhancement of retention practices.**

Keywords— Employee turnover, machine learning, HR analytics, Random Forest, predictive modeling

I. Introduction

Employee turnover poses a serious challenge to companies by negatively impacting productivity and driving up costs of operation. It has always been challenging, through conventional human resource processes, to identify employees who intend to leave prior to their departure. Conventional methods are non-predictive, as they cannot predict turnover. Predictive analytics driven by machine learning presents a potential solution through the use of past worker data to predict turnover risk.

Turnover prediction is generally based on organized data such as demographic and job-related information. This work uses a variety of machine learning algorithms for creating and comparing models to predict employee attrition. Utilizing preprocessing methods to maintain the quality of the data and applying disciplined evaluation metrics like 10-fold cross-validation, the work thoroughly evaluates the performance of classifiers like Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbors, Support Vector Machines, and Naive Bayes.

The comparative study highlights the ability of supervised learning to revolutionize human resource management from being reactive to proactive. The resulting model allows for data-driven retention practices that enhance workforce stability and help drive long-term organizational success.

II. Literature Survey

The research [1] offers a systematic framework to model employee turnover with different machine learning models. The authors experimented with models like Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forest, and Logistic Regression on IBM HR Analytics. Of these, Logistic Regression recorded the maximum accuracy of 87.71% to predict attrition. The analysis revealed gender and promotion as most significant determinants of employee turnover. The research finds that predictive analytics can go a long way in helping HR departments detect impending attrition risks and developing successful retention plans.

The research [2] examines predicting employee turnover within high-stress sectors based on the machine learning methods applied to data related to demographics, job satisfaction, performance, and level of stress. Machine learning algorithms like logistic regression, decision trees, random forests, and neural networks were utilized, with random forests and neural networks providing the best results. The study identifies stress and job satisfaction as robust predictors and underlines ML's role in facilitating early HR interventions to curb turnover in high-demand work environments.

The research [3] inspects machine learning models to estimate turnover intentions among new hires based on the Korea Employment Information Service data. Logistic regression, KNN, and extreme gradient boosting (XGB) were used, of which XGB provided the best accuracy at 78.5%. Job security emerged as the highest predictor, followed by workload and job relevance having lower impacts. The results show that sophisticated ML models, specifically XGB, have the capability to support early turnaround risk detection efficiently, which can enable proactive employee retention and workforce planning.

The research [4] is a systematic review of 52 peer-reviewed articles from 2012 through 2023 on machine learning methods to predict employee turnover. The review determines that supervised learning dominates the literature, being utilized in 96% of the studies, of which the most frequently applied algorithm is Random Forest. Pay and overtime are top predictors throughout the literature. The article emphasizes the growing application of ML in high-risk employee identification, cost reduction in organizations, and improving employee retention. It further suggests existing gaps in research at the moment, e.g., fewer applications in unsupervised techniques and the demand for more diverse and heterogeneous datasets, offering guidance on future industrial and academic research.

The work [5] compares ten supervised machine learning models for predicting employee turnover based on real and simulated HR datasets. Models are Decision Trees, Random Forest, XGBoost, Gradient Boosting, and Neural Networks. Findings indicate that tree-based algorithms, particularly XGBoost and Gradient Boosting Trees (GBT), excel with noisy and imbalanced data. Data quality, amount, and model interpretability in practical applications are emphasized in the research. It suggests XGBoost as the most accurate model for turnover prediction, offering insightful recommendations for researchers and HR professionals on choosing effective prediction models.

The paper [6] studies about employee retention and attrition issues based on primary data gathered through questionnaires. Low wages, bad working conditions, work stress, office politics, and stress related to job roles are pinpointed as primary reasons for turnover. Even though

employees seek job security, dissatisfaction with these issues is the reason for high attrition. The study concerns talent retention problems, especially in the Indian competitive market, and suggests better wages and working conditions for better employee satisfaction and lower turnover.

Research [7] also critiques staff attrition theories within the FMCG sector, with a focus on turnover influencing productivity and profitability. The research outlines four key theories: Herzberg's Two-Factor Theory (job satisfaction and career growth), Employee Equity Model (fairness in the workplace), Expectancy Theory (reward-based motivation), and Job Embeddedness Theory (job matching and social connections). 53 workers from the FMCG sector were interviewed, and the findings revealed that career advancement surpassed remuneration. The findings indicate that companies need to foster stimulating and inclusive work cultures in order to improve retention.

The research [8] examines employee turnover within the IT sector, emphasizing its effects on training and recruitment expenses. The major reasons are improved compensation, occupational or technical changes, and work difficulties. Statistics from IT professionals at all levels reveal that managers as well as employees consider career aspiration one of the most important reasons for resigning. The research emphasizes the need to know the goals of the employees and design efficient retention plans to reduce turnover and its related costs.

The paper [9] predicts employee attrition by applying machine learning techniques to IBM's dataset on employees. The classifiers include K-Nearest Neighbor (KNN), Naive Bayes, Random Forest, and Support Vector Machine (SVM). The study examined the consequences of feature selection; however, using all features resulted in the most favorable outcomes. SVM achieved the most accuracy at 85.3%, coupled with an F1-score of 81% and AUC-ROC of 0.76. Random Forest also reported strong results with an accuracy of 84%. KNN fell short of expectations and was assumed to be a victim of class imbalance. This study emphasizes the effectiveness of SVM while suggesting future refinements, such as implementing oversampling and other model comparisons.

Machine Learning is used for employee attrition prediction in the paper [10] with IBM's HR dataset comprising 35 attributes and roughly 1500 examples. The important features observed are income, age, overtime, and job involvement. Naïve Bayes, Logistic Regression, Random Forest, KNN, and SVM are the experimented-with models using the TDSP methodology. The best result was provided by Gaussian Naive Bayes with a recall of 0.54 and a very low false-negative rate of 4.5%. The findings suggest that ML models can help HR departments better strategize retention.

The study [11] investigates employee turnover prediction using a dataset of 1,470 records from IBM Watson Analytics with 32 features. Several machine learning algorithms—Random Forest, Gradient Boosting, SVM, Logistic Regression, KNN, and Gaussian Naïve Bayes—were applied after preprocessing and train-test splitting. Among them, Random Forest gave the best accuracy with 90.20%, followed by Gradient Booster and Logistic Regression. The study points out that attributes regarding employees such as job role, overtime, and work level significantly influence attrition. The findings justify the usage of ensemble models for predicting turnover in HR analytics.

The paper [12] suggests an XGBoost-based machine learning model to predict employee attrition using IBM's HR data set. Following feature selection and preprocessing, the model is trained to classify employees as either "active," or "likely to leave." There was application of advanced feature engineering methods including years without change, compa ratio, and tenure per job. With 89% accuracy, the model exceeded baseline classifiers including decision trees. Age, job satisfaction, income, years at company, marital status, and job role are identified as the key influencing factors. The study concludes that XGBoost is found to be robust, effective, and quite fit for real-world employee attrition prediction.

The paper [13] is focused on the use of machine learning algorithms to forecast turnover of employees using many different sources of data: demographic information, job satisfaction scores, performance reviews, and engagement metrics and measures. Multiple models were tested, including decision trees, logistic regression, gradient boosting machines (GBM), and ensembles of trees (random forests) - using a variety of performance metrics including F1-score, accuracy, precision, and recall. Random forests and other ensemble models performed the best by predicting the most likely employees to leave. The results show that clearly identified features of turnover are job satisfaction, compensation, and tenure; the models and predictions were supported with overwhelmed interpretability techniques like SHAP values to help explain the models' outputs. The models and work also consider challenges, including data privacy (laws, rights, requests), data objectivity (transparent evaluation), including real-time predictive analytics. Future work and directions include the use of deep learning, including explainable AI, that facilities better prediction outcomes and transparency in HR analytics.

The study [14] investigates the cause of employees' turnover intention in the high-tech industry and establishes a predictive model through machine learning techniques. Employing the use of ridge regression analysis and the XGBoost model, the study identifies workplace interpersonal relationships as the most significant factor predicting employee turnover, followed by work issues, family, and sense of accomplishment. In this study, a ridge regression model was employed to check for collinearity of variables, and the XGBoost model was used to evaluate factor importance and assess accurate predictions of turnover patterns. The results of the study confirmed that the XGBoost model very strong predictive capabilities showed an R^2 of 0.97 and low prediction error. The implications of the findings are that developing interpersonal relations in the workplace can be a good means of preventing employee turnover in high-tech enterprises and contributing knowledge to the areas of human resource management and organizational development within the industry. The development of machine learning techniques provides a powerful methodology for studying and predicting employee's behavior towards organizational commitment, allowing enterprises to engage in early and responsive measures aligned with retention strategies.

III. METHODOLOGY

The data were downloaded from Kaggle and contain employee demographic and job attributes such as level of satisfaction, number of projects, average monthly hours, department, and salary, and the target variable that indicates if an employee has left the company [15]. The data were loaded into a pandas DataFrame for initial inspection and manipulation.

Fig. 1 illustrates the overall framework and preprocessing steps employed in this study. The dataset was initially examined for missing values. Instead of discarding incomplete records, missing data were addressed using imputation techniques to preserve information. Numerical attributes, such as satisfaction level and average monthly hours, were imputed using the mean value and subsequently standardized using z-score normalization. Categorical variables, such as department (labeled as sales in the data) and salary, were filled with the most frequent category and One-Hot Encoded in order to transform them to machine-readable format.

The target variable left was considered to be a binary classification label, 1 if the employee left the organization and 0 if the employee was retained. The data was then divided into training and test subsets in the ratio 80:20, and stratification was used to keep the two sets in class balance.

To prevent overfitting and ensure generalizability, training of the model was done using stratified 10-fold cross-validation. The approach guaranteed that each fold contained a balanced target variable class distribution. Various classification algorithms were attempted, including Logistic Regression, Random Forest, Gradient Boosting, k-Nearest Neighbors, Support Vector Machine with RBF kernel, and Gaussian Naive Bayes. All the models were integrated into a preprocessing pipeline so that the data was processed uniformly during both training and testing phases.

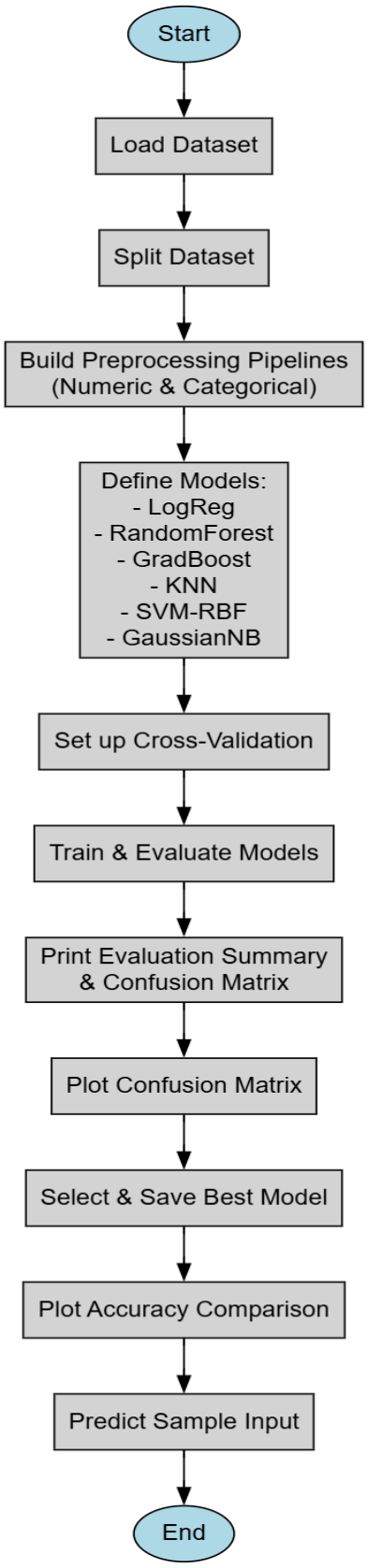

Fig. 1. Flowchart

IV. RESULT

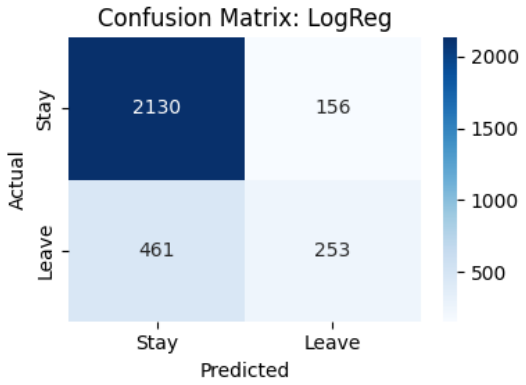

Fig. 2. Confusion Matrix of Logistic Regression.

Fig. 2. shows a confusion matrix for Logistic Regression model. In this model, 253 instances were accurately classified as the positive class (TP), while 2130 instances were correctly identified as the negative class (TN). There were 156 instances misclassified as the positive class (FP), and 461 instances incorrectly classified as the negative class (FN).

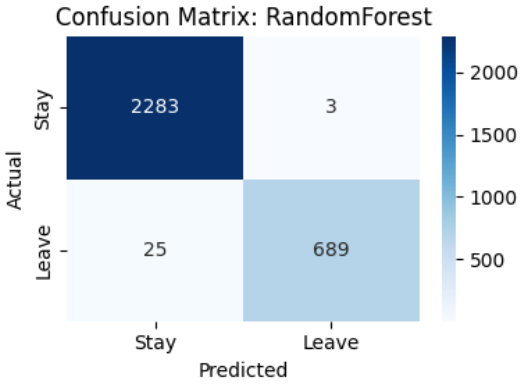

Fig. 3. Confusion Matrix of Random Forest.

Fig. 3. shows a confusion matrix for Random Forest model. In this model, 689 instances were accurately classified as the positive class (TP), while 2283 instances were correctly identified as the negative class (TN). There were 3 instances misclassified as the positive class (FP), and 25 instances incorrectly classified as the negative class (FN).

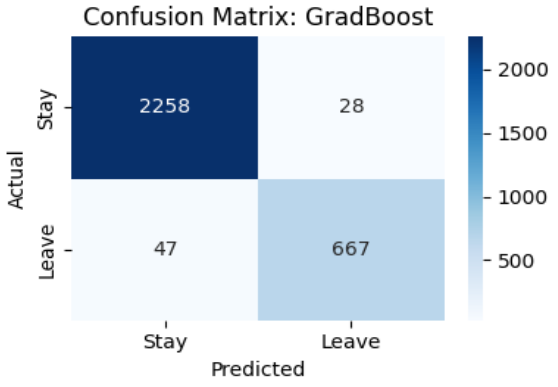

Fig. 4. Confusion Matrix of Gradient Boosting.

Fig. 4. shows a confusion matrix for Gradient Boosting model. In this model, 667 instances were accurately classified as the positive class (TP), while 2258 instances were correctly identified as the negative class (TN). There were 28 instances misclassified as the positive class (FP), and 47 instances incorrectly classified as the negative class (FN).

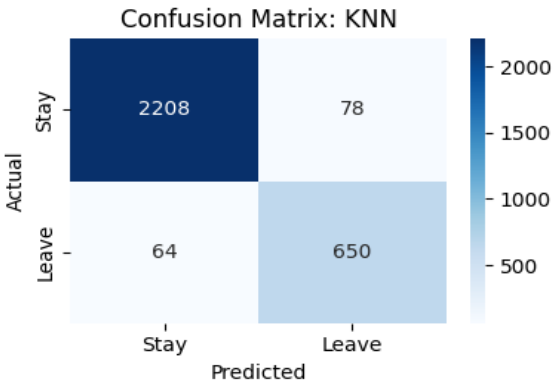

Fig. 5. Confusion Matrix of KNN.

Fig. 5. shows a confusion matrix for KNN model. In this model, 650 instances were accurately classified as the positive class (TP), while 2208 instances were correctly identified as the negative class (TN). There were 78 instances misclassified as the positive class (FP), and 64 instances incorrectly classified as the negative class (FN).

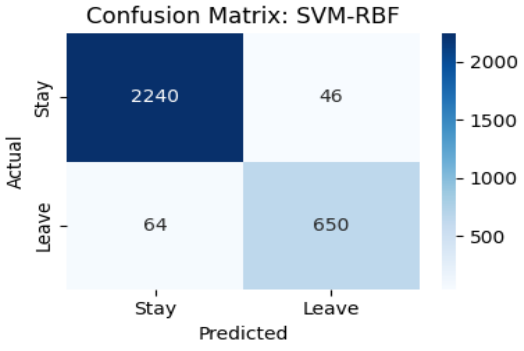

Fig. 6. Confusion Matrix of SVM-RBF.

Fig. 6. shows a confusion matrix for SVM-RBF model. In this model, 650 instances were accurately classified as the positive class (TP), while 2240 instances were correctly identified as the negative class (TN). There were 46 instances misclassified as the positive class (FP), and 64 instances incorrectly classified as the negative class (FN).

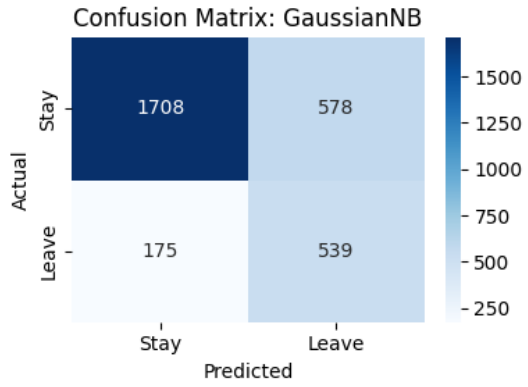

Fig. 7. Confusion Matrix of Gaussian NB.

Fig. 7. shows a confusion matrix for Gaussian NB model. In this model, 539 instances were accurately classified as the positive class (TP), while 1708 instances were correctly identified as the negative class (TN). There were 578 instances misclassified as the positive class (FP), and 175 instances incorrectly classified as the negative class (FN).

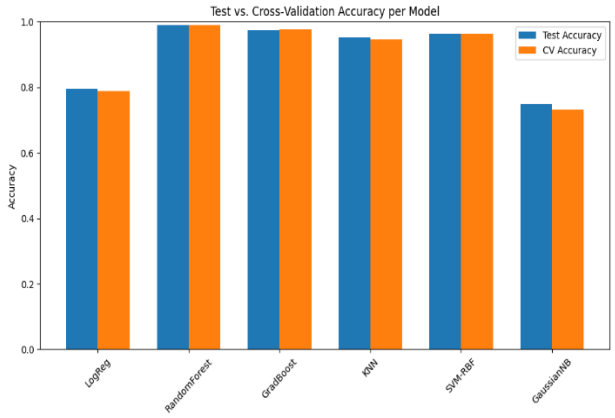

Fig. 8. Accuracy Comparison Graph.

Fig. 8. shows the test and cross-validation accuracy comparison of machine learning models employed. It can be noted that Random Forest was the most accurate, followed by KNN, SVM-RBF, and Gradient Boosting. Logistic Regression and Gaussian Naive Bayes had relatively lower accuracy, with the lowest accuracy among the models employed being that of GaussianNB.

Table.1. Comparison of Classification Models Based on Evaluation Metrics

Model	CV_Accuracy	Test_Accuracy	Recall	Precision	F1 Score
LogReg	78.91	79.43	35.43	61.86	45.06
Random Forest	99.07	99.07	96.50	99.57	98.01
Grad Boost	97.67	97.50	93.42	95.97	94.68
KNN	94.70	95.27	91.04	89.29	90.15
SVM-RBF	96.25	96.33	91.04	93.39	92.20
Gaussian NB	73.16	74.90	75.49	48.25	58.87

Table. 1 gives a comparative evaluation of some of the models based on cross-validation accuracy, test accuracy, recall, precision, and F1 score. The Random Forest model leads with the highest accuracy scores for both cross-validation (99.07%) and test data (99.07%) and has high recall, precision, and F1 score, indicating robust and reliable performance. Gradient Boosting also performs exceptionally well, with high accuracy and well-balanced recall and precision, reflecting a strong F1 score. The SVM with RBF kernel and K-Nearest Neighbors models reflect good test accuracies above 95%, with competitive recall and precision readings, marking them as reliable choice.

Logistic Regression reflects moderate accuracy but with very low recall, reflecting that many positive cases are missed despite good precision. Gaussian Naïve Bayes, reflecting relatively low accuracy, has a recall higher than Logistic Regression, but precision and F1 score are comparatively weak, reflecting more false positives.

Random Forest and Gradient Boosting reflect overall superior and well-balanced performance for all measures, reflecting the most effective models for the classification task.

V. Conclusion

In comparative analysis, various models for predicting employee turnover were compared based on significant performance metrics such as accuracy, recall, precision, and F1-score. All of the models had both strengths and weaknesses. Gaussian Naive Bayes showed relatively higher recall but low precision and was therefore less ideal where minimizing false positives is a priority. Logistic Regression was characterized by high precision but low recall and is therefore ideal where false positives must be minimized. KNN and SVM-RBF balanced performance to some extent, with SVM-RBF performing slightly better in precision. However, the Random Forest model consistently exhibited high performance in all the performance measures, which is a characteristic of excellent generalization and minimal trade-off. Its balanced and robust performance makes it the best and most ideal for real-world utilizations in turnover prediction tasks.

REFERENCES

[1] P. Kumar, S. B. Gaikwad, S. T. Ramya, T. Tiwari, M. Tiwari, and B. Kumar, "Predicting Employee Turnover: A Systematic Machine Learning Approach for Resource Conservation and Workforce Stability," Engineering Proceedings, vol. 59, no. 1, p. 117, Dec. 2023. https://doi.org/10.3390/engproc2023059117

[2] K. B. Adeusi, P. Amajuoyi, and L. B. Benjamin, "Marketing, communication, banking, and Fintech: personalization in Fintech marketing, enhancing customer communication for financial inclusion," International Journal of Management & Entrepreneurship Research, vol. 6, no. 5, pp. 1687–1701, May 2024. [Online]. Available: https://www.fepbl.com/index.php/ijmer/article/view/1142

[3] J. Park, Y. Feng, and S. P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques," Scientific Reports, vol. 14, Article 1221, Jan. 2024. [Online]. Available: https://doi.org/10.1038/s41598-023-50593-4

[4] M. Al Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review," Expert Systems with Applications, vol. 238, Article 121794, Mar. 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.121794

[5] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee Turnover Prediction with Machine Learning: A Reliable Approach," in Advances in Intelligent Systems and Computing, vol. 869, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham: Springer, 2019, pp. 737–758. doi: 10.1007/978-3-030-01057-7_56. [Online]. Available: https://doi.org/10.1007/978-3-030-01057-7_56

[6] S. Srilatha and V. Divya, "A Study on Employee Attrition and Retention Analysis – Indiabulls," International Journal of Innovative Research in Technology, vol. 9, no. 7, pp. 449–456, Nov. 2023. [Online].Available:https://www.researchgate.net/publication/375800378_A_STUDY_ON_EMPLOYEE_ATTRITION_AND_RETENTION_ANALYSIS_-INDIABULLS

[7] V. Mangal and S. Dhamija, "Analysing Theoretical Models for Predicting Employee Attrition: A Comparative Study in the FMCG Sector," Journal of Advanced Zoology, vol. 44, no. S-3, pp. 1179–1191, Oct. 2023. [Online]. Available: https://doi.org/10.17762/jaz.v44iS-3.1295

[8] V. Saraf and M. A. Peshave, "An Analysis on Employee-Attrition in IT Industry," Mukt Shabd Journal, vol. 9, no. 7, pp. 2751–2758, July 2020. [Online]. Available: https://hmct.dypvp.edu.in/Documents/research papers-publication/Resarch-publications/60.pdf

[9] I. F. Alsuahim, F. A. Alotaibi, M. A. AlAsiri, M. S. Alkharji, and S. A. Alharthi, "Predicting employee attrition using machine learning," in Proc. Int. Conf. Ind. Eng. Oper. Manag., Riyadh, Saudi Arabia, Nov. 2019, pp. 1007–1013. [Online]. Available: https://www.ieomsociety.org/gcc2019/papers/124.pdf

[10] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, Nov. 2020. [Online]. Available: https://doi.org/10.3390/computers9040086

[11] R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches," *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Kuala Lumpur, Malaysia, 2021, pp. 1-6, doi: https://doi.org/10.1109/GUCON50781.2021.9573759

[12] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2018, pp. 113-120, doi: https://doi.org/10.1109/SYSMART.2018.8746940

[13] G. Manoharan, V. Pushpa, A. V. Deshpande, M. Lourens, M. K. Sharma and A. Jain, "Machine Learning for Employee Turnover Prediction," 2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2024, pp. 1-5, doi: https://doi.org/10.1109/ICSES63760.2024.10910479

[14] S. Zhang and Y. -C. Chang, "High-Tech Industry Employees Turnover Intention Prediction Model Based on Machine Learning," 2023 International Conference on Computer Science and Automation Technology (CSAT), Shanghai, China, 2023, pp. 364-368, doi: https://doi.org/10.1109/CSAT61646.2023.00100

[15] G. Srikant, "HR Employee Retention," GitHub, Kaggle, 2024. 2021. [Online].https://www.kaggle.com/datasets/gummulasrikanth/hr-employee-retention/versions/1.