Machine Learning Powered Employee Attrition Prediction

J. VENKATESWARAMMA[1], V. V. C. H. S. S. SAI CHARAN[2], J. YASWANTH RAM[3], D. PAVAN PRAVEEN[4], R. YASHWANTH VIJAY SAI[5]

[1, 2, 3, 4, 5] *Department of CSE (Data Science), SRK Institute of Technology, Vijayawada, A.P., India.*

Abstract- The project aims to find the Main Factors or Reasons for Employee Attrition to decrease the employee attrition rate by using HR data collected from an organization or a company, for this a dataset has been gathered containing various employee attributes such as ID, Attrition, Years with current manager, over18, Job level, Job Role etc., By analyzing these Features finding the Reasons or factors for Employee Attrition is used to taking steps to control the attrition on a specific Factor. Here we using the Machine Learning Algorithms for finding these Attrition factors, the algorithms using are logistic Regression and Random Forest, before using algorithms few analyses is done on historical data for finding left and stay data and count of Attrition percentage. The ML Models trained with those algorithms and both models working on split data called train and test data, these algorithms are used to predict the top attrition factors from all gathered data, it helps to make solutions for decrease Attrition rate. The Key Components of Project include data collection, data preprocessing, data analysis, data visualization, ML Models training, Model metrics, Top Factors finding, data collected from various sources like historical data and statistical archives and Transformation methods are used to create accurate information for the training model.

Index Terms- Logistic Regression, Random Forest, Recursive Feature Elimination (RFE), Confusion Matrix, Employee Attrition.

I. INTRODUCTION

Employee attrition, a prevalent concern for organizations, is tackled through logistic regression and random forest algorithms. These methods analyze employee demographics and job satisfaction data to predict turnover. Logistic regression computes attrition probabilities, while random forest enhances accuracy by capturing intricate data patterns. Integrating both approaches, the model identifies at-risk employees for tailored retention strategies. This advanced machine learning project furnishes actionable insights to optimize workforce management and mitigate turnover costs, fostering enduring organizational success. Through analysis of diverse attributes such as demographics and performance metrics, the model aids in recruitment, talent development, and employee engagement initiatives. By anticipating turnover beforehand, businesses can implement proactive measures, reducing recruitment expenses and loss of productivity. Ultimately, the goal is to empower organizations with informed decisions for sustained success, nurturing a stable and engaged workforce.

The integration of logistic regression and random forest enables organizations to proactively manage attrition and implement targeted retention strategies, thereby fostering a stable workforce. Through meticulous analysis, this project seeks to empower organizations with actionable insights to mitigate the adverse impacts of employee turnover and optimize workforce management strategies. By developing a predictive model that forecasts attrition, businesses can anticipate turnover before its occurrence and tailor retention strategies to individual employee needs, thereby fostering a more stable and engaged workforce while curbing the costs associated with recruitment and loss of productivity.

Furthermore, the project aims to optimize various aspects of workforce management, including recruitment, talent development, and employee engagement initiatives, by scrutinizing diverse employee attributes such as demographics, job satisfaction, and performance metrics. Ultimately, the goal is to equip organizations with the necessary tools and insights to make informed decisions that promote long-term organizational success by reducing turnover rates and enhancing employee retention. Through proactive measures and targeted interventions, businesses can navigate the challenges posed by employee attrition and cultivate an environment conducive to sustained growth and prosperity.

II. LITERATURE SURVEY

- Understanding and Mitigating Employee Attrition

Aliff, Mohd. "Understanding and Mitigating Employee Attrition." Journal of Human Resource Management (2023).

The study aims to compare the performance of three machine learning classifiers - Decision Tree (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) - in predicting employee attrition. It underscores the significance of employing machine learning techniques to anticipate employee turnover, facilitating timely interventions by HR departments to mitigate attrition's adverse effect.

- Insights into Employee Attrtion

Purpose, Rohit, and Ajith, Pankaj. "Insights into Employee Attrition." Journal of Organizational Behavior (2023).

In their research published in the International Journal of Advanced Research in Artificial Intelligence (IJARAI) in 2016, Rohit Punnoose and Pankaj Ajit address the challenge of accurately predicting employee turnover, which is often under-funded compared to other domains within organizations.

They highlight the prevalence of noise in HRIS data, which can lead to overfitting and inaccurate predictive models. The paper introduces the application of Extreme Gradient Boosting (XG Boost) technique as a novel approach, leveraging its regularization formulation to improve robustness. The study utilizes data from the HRIS of a global retailer to compare XG Boost against six historically used supervised classifiers, demonstrating its significantly higher accuracy in predicting employee turnover.

- Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance

Al-Radaideh, Qasem A., and Eman Al Nagi." Using data mining techniques to build a classification model for predicting employees performance." International Journal of Advanced Computer Science and Applications 3.2(2012).

The study applies data mining techniques to develop a classification model for predicting employee performance. The authors utilize the CRISPDM data mining methodology in their work, focusing on the Decision Tree algorithm as the primary tool for building the classification model. The model generates several classification rules, which are validated through experiments using real data collected from multiple companies. The ultimate goal of the model is to predict the performance of new applicants.

- Using Classification Models for Predicting Employee Attrition

Rahul Yedida, Abhilash, Deepthi Kulkarni." Using classification models for predicting Employee Attrition." (2018).

In this paper presented the effect of voluntary attrition on organizations, and why predicting it is important. It further outlined various classification algorithms based on supervised learning to solve the prediction problem. The results of this research showed the superiority of the KNN classifier in terms of accuracy and predictive effectiveness, by means of the ROC curve. When used with its optimal configuration, it is a robust method that delivers accurate results in spite of the noise in the dataset, which is a major challenge for machine learning algorithms. The authors thus recommend the use of the KNN classifier for accurately predicting employee attrition in an organization, which enables HR to take necessary action for the retention of employees predicted to be at risk of leaving.

III. EXISTING SYSTEM

Present existing systems for addressing employee attrition typically rely on traditional methods such as surveys, exit interviews, and HR analytics tools. These systems often involve manual data collection and analysis, leading to delays in identifying attrition risks and implementing retention strategies. While some organizations may utilize statistical methods like regression analysis, they often lack the ability to capture complex relationships and nonlinear patterns present in employee data. As a result, there's a growing recognition of the need to integrate advanced machine learning techniques into existing systems to enhance predictive accuracy and enable proactive attrition management. These

modern systems leverage algorithms like logistic regression and random forest to analyze diverse employee attributes and provide organizations with actionable insights for optimizing workforce retention efforts.

Disadvantages
- Resource-Intensive Preprocessing
- Lack of Scalability
- Limited Insights
- Inadequate Adaptability
- Limited Predictive Accuracy
- Potential Biases

IV. PROPOSED SYSTEM

The impetus behind the Employee Attrition Prediction Project stems from the pressing challenges organizations face in retaining valuable talent amidst high turnover rates. With turnover rates soaring, the imperative to develop proactive solutions becomes paramount. Studies indicate that traditional methods of attrition prediction often fall short, lacking the sophistication to capture intricate patterns and predictors effectively. Thus, the project seeks to address this gap by harnessing advanced machine learning techniques, namely logistic regression and random forest algorithms, to predict employee attrition accurately. predictors of attrition, and enable organizations to implement practices and foster a stable, engaged workforce.

Advantages
Proactive Intervention: By accurately predicting employee attrition, organizations can implement proactive retention strategies to mitigate turnover before it occurs, fostering a stable workforce.

Comprehensive Insights: The proposed system provides detailed insights into the factors influencing attrition, enabling organizations to understand the root causes and tailor retention efforts accordingly.

Efficiency in Resource Allocation: By using machine learning algorithms such as Logistic Regression and Random Forest, the project can identify the main factors contributing to employee attrition. This allows the organization to understand the root causes behind attrition.

Data-Driven Decision Making: The project enables data-driven decision-making by analyzing HR data collected from the organization. It helps in understanding trends, patterns, and factors contributing to attrition, allowing the organization to take proactive measures.

Efficient HR Management: By analyzing historical data and using machine learning algorithms, the project provides insights into employee behavior and preferences. This helps HR departments in making more informed decisions regarding employee engagement and retention strategies.

Predictive Analysis: The use of machine learning algorithms allows for predictive analysis of employee attrition. By analyzing various employee attributes, the project can predict which employees are most likely to leave the organization, enabling proactive measures to retain them.
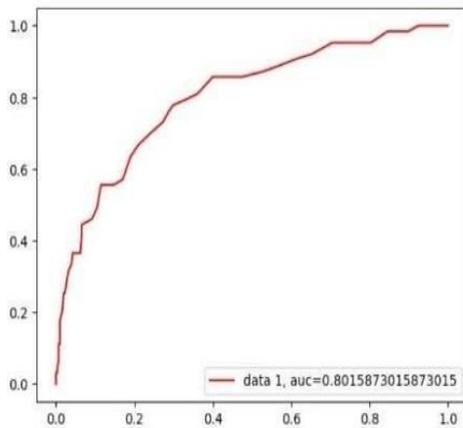
V. RESULTS

We got top Attrition factors based on the Algorithms Performance and Its efficiency, which is used to take proactive measures to retain valuable employees.
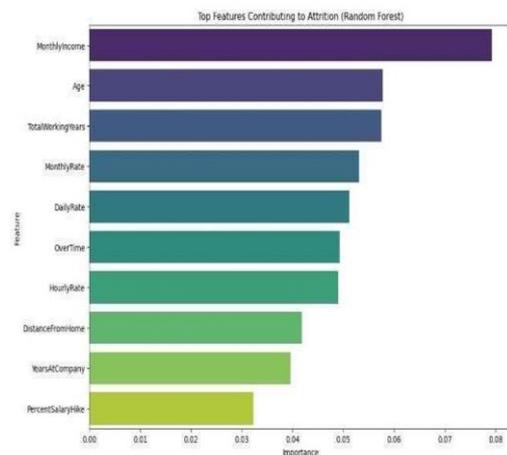
The Outcome from the Random Forest
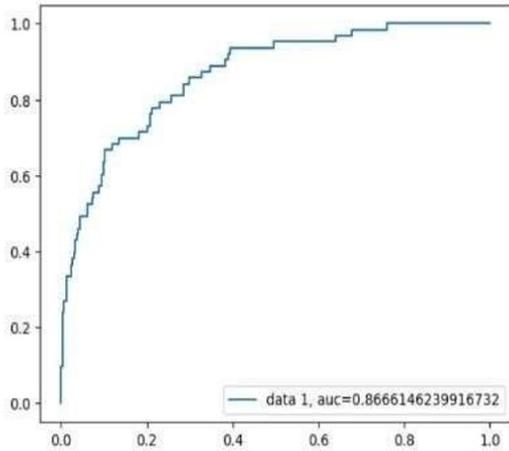
The Accuracy Prediction for the Linear Regression Model

The Accuracy Prediction for the Random Forest Model

CONCLUSION

Improving the strategies for reducing employee attrition based on the results given by our top factors prediction from whole data. According to our project the Logistic Regression gives best accuracy and AUC comparing to Random Forest, accuracy is 0.87 and AUC Percentage is 0.86.

FUTURE SCOPE

Controlling the Employee attrition based on predictions performed by us and taking steps for overcome the current situation and make changes in employment and the attrition percentage in the organizations and companies after the prediction.

REFERENCES

[1] Peng, B. Statistical analysis of employee retention. In Proceedings of the International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), Nanjing, China, 26–28 November 2021; Volume 12163, pp. 7–15. [Google Scholar] [CrossRef]

[2] 19 Employee Retention Statistics That Will Surprise You. 2022. Available online: https://www.apollotechnical.com/employee-retention-statistics/ (accessed on 6 May 2022).

[3] Here's What Your Turnover and Retention Rates Should Look Like. Available online: https://www.ceridian.com/blog/turnover-and-retention-rates-benchmark (accessed on 6 May 2022).

[4] SHRM Survey: Average Cost Per Hire Is $4129. Available online: https://www.businessmanagementdaily.com/46997/shrm-survey-average-cost-per-hire-is-4129/ (accessed on 6 May 2022).

[5] Gandomi, A.H.; Chen, F.; Abualigah, L. Machine Learning Technologies for Big Data Analytics. Electronics 2022, 11, 421. [Google Scholar] [CrossRef]

[6] Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.W.; Zou, B.; Hou, D. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 2021, 270, 116281. [Google Scholar] [CrossRef] [PubMed]

[7] Reshma Ramchandra, N.; Rajabhushanam, C. Machine learning algorithms performance evaluation in traffic flow prediction. Mater. Today Proc. 2022, 51, 1046–1050. [Google Scholar] [CrossRef]

[8] Aljedani, N.; Alotaibi, R.; Taileb, M. HMATC: Hierarchical multi-label Arabic text classification model using machine learning. Egypt. Inform. J. 2021, 22, 225–237. [Google Scholar] [CrossRef]

[9] Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; et al. Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data. Diagnostics 2022, 12, 203. [Google Scholar] [CrossRef] [PubMed]

[10] Aggarwal, S.; Singh, M.; Chauhan, S.; Sharma, M.; Jain, D. Employee Attrition Prediction Using Machine Learning Comparative Study. Smart Innov. Syst. Technol. 2022, 265, 453–466. [Google Scholar] [CrossRef]

[11] Jadhav, A.; Mostafa, S.M.; Elmannai, H.; Karim, F.K. An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. Appl. Sci. 2022, 12, 3928. [Google Scholar] [CrossRef]

[12] Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.E.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. Heliyon 2018, 4, e00938. [Google Scholar] [CrossRef] [Green Version]

[13] Ganthi, L.S.; Nallapaneni, Y.; Perumalsamy, D.; Mahalingam, K. Employee Attrition Prediction Using Machine Learning Algorithms. Lect. Notes Netw. Syst. 2022, 288, 577–596. [Google Scholar] [CrossRef]

[14] Jiang, Z.; Gao, B.; He, Y.; Han, Y.; Doyle, P.; Zhu, Q. Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports. Math. Probl. Eng. 2021, 2021, 6619088. [Google Scholar] [CrossRef]

[15] Qutub, A.; Al-Mehmadi, A.; Al-Hssan, M.; Aljohani, R.; Alghamdi, H.S. Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. Int. J. Mach. Learn. Comput. 2021, 11, 110–114. [Google Scholar] [CrossRef]

[16] Habous, A.; Nfaoui, E.H.; Oubenaalla, Y. Predicting Employee Attrition using Supervised Learning Classification Models. In Proceedings of the 2021 Fifth International Conference on Intelligent Computing in Data Sciences (ICDS), Fez, Morocco, 20–22 October 2021. [Google Scholar] [CrossRef]

[17] Najafi-Zangeneh, S.; Shams-Gharneh, N.; Arjomandi-Nezhad, A.; Zolfani, S.H. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. Mathematics 2021, 9, 1226. [Google Scholar] [CrossRef]

[18] Pratt, M.; Boudhane, M.; Cakula, S. Employee attrition estimation using random forest algorithm. Balt. J. Mod. Comput. 2021, 9, 49–66. [Google Scholar] [CrossRef]

[19] Sadana, P.; Munnuru, D. Machine Learning Model to Predict Work Force Attrition. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Pune, India, 2–4 April 2021. [Google Scholar] [CrossRef]

[20] Kaya, İ.E.; Korkmaz, O. Machine Learning Approach for Predicting Employee Attrition and Factors Leading to Attrition. Cukurova Univ. J. Fac. Eng. 2021, 36, 913–928. [Google Scholar] [CrossRef]

[21] Mazumder, R.K.; Salman, A.M.; Li, Y. Failure risk analysis of pipelines using data-driven machine learning algorithms. Struct. Saf. 2021, 89, 102047. [Google Scholar] [CrossRef]

[22] Mate, Y.; Potdar, A.; Priya, R.L. Ensemble Methods with Bidirectional Feature Elimination for Prediction and Analysis of Employee Attrition Rate During COVID-19 Pandemic. Lect. Notes Electr. Eng. 2022, 806, 89–101. [Google Scholar] [CrossRef]

[23] Joseph, R.; Udupa, S.; Jangale, S.; Kotkar, K.; Pawar, P. Employee attrition using machine learning and depression analysis. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 1000–1005. [Google Scholar] [CrossRef]

[24] Bhartiya, N.; Jannu, S.; Shukla, P.; Chapaneri, R. Employee Attrition Prediction Using Classification Models. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019. [Google Scholar] [CrossRef]

[25] Lai, H.; Hossin, M.A.; Li, J.; Wang, R.; Hosain, M.S. Examining the Relationship between COVID-19 Related Job Stress and Employees' Turnover Intention with the Moderating Role of Perceived Organizational Support: Evidence from SMEs in China. Int. J. Environ. Res. Public Health 2022, 19, 3719. [Google Scholar] [CrossRef]

[26] HR-Employee-Attrition-Dataset by Aaizemberg|Data.World. Available online: https://data.world/aaizemberg/hr-employee-attrition (accessed on 6 May 2022).

[27] Karthiga, R.; Usha, G.; Raju, N.; Narasimhan, K. Transfer Learning Based Breast cancer Classification using One-Hot Encoding

Technique. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 115–120. [Google Scholar] [CrossRef]

[28] Shobhanam, K.; Sumati, S. HR Analytics: Employee Attrition Analysis using Random Forest. Int. J. Perform. Eng. 2022, 18, 275. [Google Scholar] [CrossRef]

[29] Baldomero-Naranjo, M.; Martínez-Merino, L.I.; Rodríguez-Chía, A.M. A robust SVM-based approach with feature selection and outliers detection for classification problems. Expert Syst. Appl. 2021, 178, 115017. [Google Scholar] [CrossRef]

[30] Dong, S. Multi class SVM algorithm with active learning for network traffic classification. Expert Syst. Appl. 2021, 176, 114885. [Google Scholar] [CrossRef]

[31] Tigga, N.P.; Garg, S. Predicting Type 2 Diabetes Using Logistic Regression. Lect. Notes Electr. Eng. 2021, 673, 491–500. [Google Scholar] [CrossRef]

[32] Maswadi, K.; Ghani, N.A.; Hamid, S.; Rasheed, M.B. Human activity classification using Decision Tree and Naïve Bayes classifiers. Multimed. Tools Appl. 2021, 80, 21709–21726. [Google Scholar] [CrossRef]

[33] Azad, C.; Bhushan, B.; Sharma, R.; Shankar, A.; Singh, K.K.; Khamparia, A. Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. Multimed. Syst. 2021, 1–19. [Google Scholar] [CrossRef]

[34] Ossai, C.I.; Wickramasinghe, N. GLCM and statistical features extraction technique with Extra-Tree Classifier in Macular Oedema risk diagnosis. Biomed. Signal Process. Control 2022, 73, 103471. [Google Scholar] [CrossRef]