EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING

Badugu Blessy
Vasavi College of Engineering
Hyderabad, Telangana – India

M. Akshitha
Vasavi College of Engineering
Hyderabad, Telangana – India

Dr. K. Srinivas
Vasavi College of Engineering
Hyderabad, Telangana – India

Dr. T. Adilakshmi
Vasavi College of Engineering
Hyderabad, Telangana – India

ABSTRACT

Employee attrition, whether voluntary or involuntary, marks the departure of an individual from a company's workforce. Businesses allocate significant resources to recruit and train talented personnel, underscoring the importance of minimizing turnover. When employees depart, it incurs costs for hiring and training replacements, disrupting productivity as new hires adjust to their roles. Every employee contributes uniquely to a company's success, rendering turnover a critical concern for HR departments. The financial repercussions of turnover are significant, with studies consistently showing an uptick in turnover rates in recent years. The challenge of finding qualified replacements for experienced staff further complicates the issue. Hence, our goal is to construct a predictive framework for employee churn by examining various behaviours and attributes. Through the application of classification techniques, our project aims to furnish organizations with valuable insights into the factors influencing attrition. Ultimately, this will empower companies to refine their retention strategies and alleviate the adverse impacts of turnover on their operations.

Keywords: Dataset, Machine Learning, Ensemble Learning Methods, Decision Trees, K-Fold Cross Validation.

1. INTRODUCTION

Employee attrition, the departure of employees from a company, presents significant hurdles for businesses. It disrupts operations, reduces productivity and adds financial strain due to frequent recruitment and training. Losing experienced staff also means losing valuable knowledge hindering organizational growth. Therefore, accurately predicting and understanding the causes of employee attrition is crucial.

Research indicates that analyzing various personal and professional variables stored within organizational databases can forecast attrition. These variables include Daily Rate, Distance from Home, Total Working Years, Training Times Last Year, Work Life Balance, Years in Current Role, and Years Since Last Promotion. By utilizing these features, organizations can develop predictive models to identify employees likely to leave, enabling proactive retention strategies.

To achieve this, organizations often employ machine learning techniques, particularly ensemble learning, to build robust predictive models. Ensemble learning combines multiple models' predictions to enhance accuracy and generalization. Commonly used algorithms include KNN, Random Forest, Logistic Regression, and Decision Tree, often combined with ensemble methods like Stacking, Bagging, and Boosting. The process involves data collection, preprocessing, splitting into training and testing sets, and training various models. Performance metrics like accuracy, precision, recall, and F1 score gauge model effectiveness in predicting attrition. By analyzing and comparing these models, organizations can select the most suitable one.

2. LITERATURE SURVEY

The literature review provides a comprehensive overview of previous research on employee attrition, guiding our current investigation. We incorporate insights from various studies to build and evaluate predictive models using machine learning techniques.

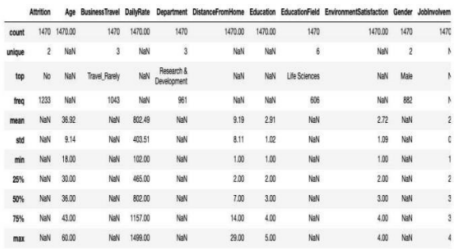
Our project aims to reduce employee attrition rates and identify contributing factors through machine learning and ensemble learning methods. We utilize a real dataset from IBM analytics, preprocess the data by converting categorical variables and removing irrelevant features, then train base models using algorithms like KNN, Decision Trees, Random Forest, and Logistic Regression. Ensemble techniques such as stacking, bagging, and boosting are applied to improve model performance.

Hyper parameter tuning and k-fold cross-validation are employed to enhance model accuracy, with evaluation metrics including Accuracy, F1 Score, Precision, and Recall used to assess model performance. Finally, we integrate ensemble learning techniques with Gradio for visualization of model output.

3.DATASET

We've obtained the IBM HR Analytics dataset, which contains 35 attributes, with "Attrition" as the dependent variable. Our aim is to utilize this dataset to address the issue at hand effectively. The "Attrition" variable serves as the target feature, where "No" indicates employees who remain with the company, and "Yes" denotes those who

have left. Leveraging this dataset allows the machine learning system to learn from real-world data, rather than relying solely on programmed rules. Through iterative training on relevant samples, the system can improve its predictive accuracy over time, resulting in more dependable outcomes. This dataset offers valuable insights into employee behaviour and the factors influencing attrition, empowering organizations to implement targeted retention strategies and effectively mitigate attrition risks.



	Attrition	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	JobInvolvem
count	1470	1470.00	1470	1470.00	1470	1470.00	1470.00	1470	1470.00	1470	1470
unique	2	NaN	3	NaN	3	NaN	NaN	6	NaN	2	N
top	No	NaN	Travel_Rarely	NaN	Research & Development	NaN	NaN	Life Sciences	NaN	Male	N
freq	1233	NaN	1043	NaN	961	NaN	NaN	606	NaN	882	N
mean	NaN	36.92	NaN	802.49	NaN	9.19	2.91	NaN	2.72	NaN	2
std	NaN	9.14	NaN	403.51	NaN	8.11	1.02	NaN	1.09	NaN	0
min	NaN	18.00	NaN	102.00	NaN	1.00	1.00	NaN	1.00	NaN	1
25%	NaN	30.00	NaN	465.00	NaN	2.00	2.00	NaN	2.00	NaN	2
50%	NaN	36.00	NaN	802.00	NaN	7.00	3.00	NaN	3.00	NaN	3
75%	NaN	43.00	NaN	1157.00	NaN	14.00	4.00	NaN	4.00	NaN	3
max	NaN	60.00	NaN	1499.00	NaN	29.00	5.00	NaN	4.00	NaN	4

Figure 1

strategically combining their results to enhance overall predictive performance. Our objective was to determine the most dependable approach for predicting attrition, empowering organizations to take proactive measures against employee turnover and deploy tailored retention strategies.

- Machine Learning models
- Ensemble learning techniques

Dataset split process

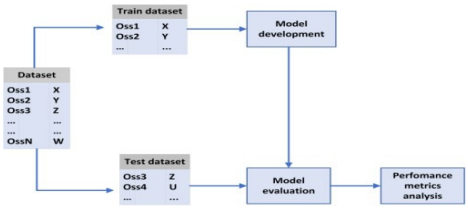


Figure 2

4. MODELS

To conducted a comprehensive analysis to compare the predictive effectiveness of K Nearest Neighbors, Logistic Regression, Decision Trees, and Random Forest algorithms in forecasting attrition. By meticulously adjusting hyperparameters and employing ensemble learning techniques like stacking, we pinpointed the algorithm with the highest predictive accuracy. This involved refining the parameters of each algorithm and

5. MACHINE LEARNING MODEL

5.1 K-Nearest Neighbours (KNN):

KNN, a non-parametric algorithm, excels in disease prediction by leveraging local similarities in the feature space. Our hyperparameter tuning focused on optimizing the number of neighbours (k) and distance metrics, enhancing its ability to capture localized disease patterns. This refinement enhances KNN's utility in healthcare settings, aiding clinicians in early diagnosis and intervention.

5.2 Logistic Regression:

Logistic Regression is fundamental in binary classification tasks like disease prediction. Hyperparameter tuning improved its performance by optimizing regularization strength and selecting appropriate solvers, reducing overfitting and enhancing model interpretability. This approach provides valuable insights into disease occurrence and progression.

5.3 Decision Trees:

Decision Trees offer an intuitive approach to disease prediction by segmenting data based on feature values. Hyper parameter tuning optimized tree depth and impurity measures, improving prediction accuracy and interpretability. This refinement aids clinicians in understanding disease pathways and making informed treatment decisions.

5.4 Random Forest:

Random Forest, an ensemble learning technique, excels in handling complex datasets for disease prediction. Hyperparameter tuning optimized parameters like the number of trees and feature selection, improving prediction accuracy. This enhanced capability aids healthcare professionals in identifying key disease-related features for improved patient care.

6. ENSEMBLE LEARNING METHODS:

6.1 Bagging:

Bagging fosters diversity among ensemble members by training each model on different subsets of the training data, enhancing model robustness and generalization.

6.2 Boosting:

Boosting adapts training data to emphasize instances previously misclassified by weak learners, constructing a "strong learner" with improved predictive performance.

6.3 Stacking:

Stacking combines predictions from multiple models to improve overall predictive performance, leveraging the strengths of different model types**.**

7. K-FOLD CROSS VALIDATION

K-fold cross-validation partitions the dataset into k subsets and iteratively trains and evaluates the model k times. By averaging performance metrics from each fold, it provides a more accurate estimate of the model's generalization performance. This technique aids in assessing model effectiveness, guiding model selection and hyperparameter tuning, and gauging its robustness across various data subsets. Ultimately, K-fold cross-validation enhances confidence in the model's predictive ability in real-world scenarios.

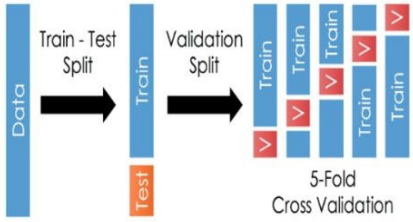


Figure 3

8. PERFORMANCE METRICS

Performance metrics quantify a model's effectiveness in predicting outcomes, offering valuable insights into its predictive capabilities. Key metrics include accuracy, precision, recall, and the F1 score. Accuracy measures the ratio of correctly predicted instances to the total evaluated. Precision assesses the accuracy of positive predictions by calculating the proportion of true positives among all predicted positives. Recall evaluates the model's ability to identify positive instances among all actual positives. The F1 score provides a balanced assessment of precision and recall, particularly useful when both are crucial for model evaluation or when class distribution is uneven.

9. ARCHITECTURE

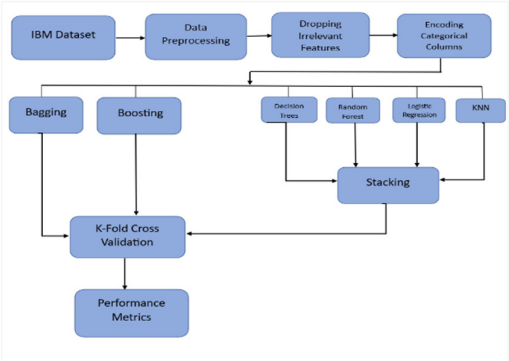


Figure 4

10. OVERVIEW TECHNOLOGY

Our methodology for predicting employee attrition integrates key steps seamlessly, ensuring robust and relevant results.

Beginning with thorough data preprocessing and strategic partitioning, and culminating in rigorous hyperparameter tuning and ensemble learning, our approach navigates the complexities of attrition prediction with precision. This comprehensive method delivers actionable,

Model	Accuracy
1. Random Forest	0.829
2. Logistic Regression	0.857
3. K Nearest Neighbors	0.806
4. Decision Trees	0.809
5. Stacking(Before K Fold Cross Validation)	0.870
6. Stacking(After K Fold Cross Validation)	0.892
7. Bagging	0.860
8. Boosting	0.872

comprehensive method delivers actionable , driving informed decision-making. insights, driving informed decision-making within organizations. Data pre-processing establishes the foundation for data quality by performing tasks like cleaning and normalization, enhancing the reliability of our predictive models. Splitting our dataset into training and testing sets enables thorough model evaluation, ensuring accurate predictions on new data. We employ a variety of machine learning algorithms, such as K Nearest Neighbours and Decision Trees, each trained individually to capture distinct patterns. Ensemble techniques like Bagging and Boosting enhance predictive accuracy by aggregating model outputs.

Hyperparameter tuning optimizes model performance by fine-tuning specific parameters, maximizing predictive power.

11.2 Precision

Model	Precision
1. Random Forest	0.785
2. Logistic Regression	0.785
3. K Nearest Neighbors	0.6
4. Decision Trees	0.571
5. Stacking(Before K Fold Cross Validation)	0.794
6. Stacking(After K Fold Cross Validation)	0.709
7. Bagging	0.454
8. Boosting	0.677

Table 2

11.3 F1 Score

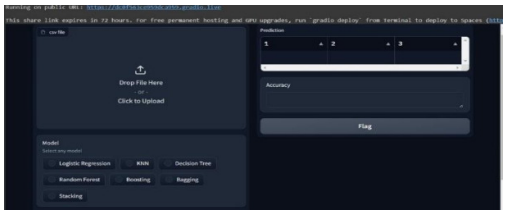
Table 3

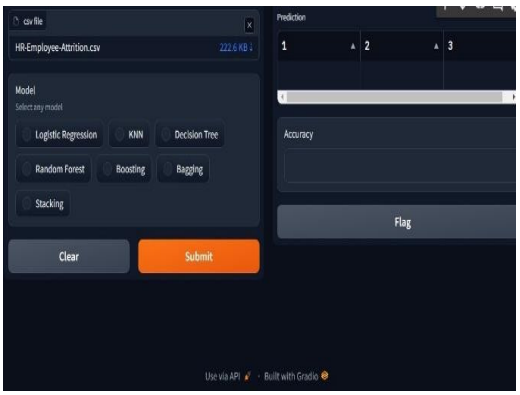
Model	F1 Score
1. Random Forest	0.305
2. Logistic Regression	0.511
3. K Nearest Neighbors	0.095
4. Decision Trees	0.222
5. Stacking(Before K Fold Cross Validation)	0.586
6. Stacking(After K Fold Cross Validation)	0.538
7. Bagging	0.250
8. Boosting	0.471

Table 4

Model	Recall
1. Random Forest	0.189
2. Logistic Regression	0.379
3. K Nearest Neighbors	0.051
4. Decision Trees	0.137
5. Stacking(Before K Fold Cross Validation)	0.465
6. Stacking(After K Fold Cross Validation)	0.440
7. Bagging	0.172
8. Bagging	0.362

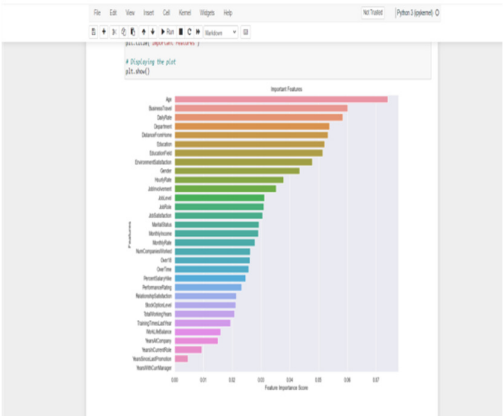
12. SUCCESSPLOT





13. CONCLUSION

The integration of ensemble techniques in



predicting employee attrition represents a significant advancement, promising improved prediction accuracy. Methods like Bagging, Boosting, and Stacking

combine multiple models' outputs, resulting in a more robust predictive framework that generalizes well to new data. This approach not only enhances prediction reliability but also provides deeper insights into the factors driving turnover.

Machine learning methodologies offer the advantage of uncovering crucial features associated with attrition by analysing large datasets, empowering HR managers to address risks proactively. Ensemble techniques have gained widespread adoption due to their superior performance in prediction accuracy, surpassing traditional models.

Ensemble methods are versatile and adaptable, making them well-suited for addressing the complexities of attrition in various organizational settings. By using predictive analytics, companies can identify at-risk individuals and implement measures to retain them, preserving talent and enhancing organizational stability and performance.

14. REFERENCES

1. "Machine Learning Approaches for Predicting Employee Attrition" - Written by A. Raza, K. Munir, M. Almutairi, F. Younas - Published in Applied Sciences, 2022.

2. "Deep Learning-Based Forecasting of Employee Attrition and Performance" - Authored by S.M. Arqawi, M. Rumman, E. Zitawi - Appeared in the Journal of Theoretical , 2022.

3. "Forecasting Employee Attrition and Identifying High-Risk Employees Using Big Data and Machine Learning" - Authored by A. Mhatre, A. Mahalingam, M. Narayanan - Presented at a conference on advances in ..., 2020.

4. S. Rogers and M. Girolami. (2016). "Introduction to Machine Learning." CRC Press.

5. M. Maisuradze. (2017). "Predictive Analysis Applied to Employee Turnover" (Master’s thesis). Tallinn: Tallinn University of Technology.

6. S. Kaur and R. Vijay. (2016). "Job Satisfaction – A Key Factor in Attrition or Retention in the Retail Industry." Imperial Journal of Interdisciplinary Research, 2(8).