Employee Attrition Analysis and Prediction

1st Oghenero Monica Godwin
Western University
London, Ontario
ogodwin@uwo.ca

2nd Ramneek Kaur Arora
Western University
London, Ontario
rarora73@uwo.ca

Abstract
Organizations are constantly facing problems relating to employee control and managing replacement, which has an adverse effect on project result and team continuity. The objective of this study is to build models that can predict employee attrition using a variety of machine learning algorithms and give insights on factors causing this attrition rate. To build these models, we have taken critical steps such a data pre-processing, exploratory analysis, feature engineering, machine learning, and hyperparameter tuning to ensure high accuracy and reliability of these models. Having built different machine learning models using KNN, Naïve Bayes, Logistic regression, Random Forest, and decision Tree. This study aims to evaluate these models and provide a best fit for predicting employee attrition to help companies improve on areas that will reduce employee attrition rate.

Keywords- Attrition, Machine Learning, KNN, Naïve Bayes, Logistic regression, Random Forest, decision Tree.

I. INTRODUCTION

Employees are the bedrock of any business, playing a huge role in whether it thrives or just survives. When too many people leave a company, it affects the organization in different areas. Projects get delayed, and the whole team can be disrupted. This is not just about team disorientation, there is a serious financial effect to the organization too. In the last year alone, over 36% of blue-collar workers and more than one in five 21% of white-collar decided to pack up and move to new jobs, as pointed out by Randstad survey carried out in 2022. Every time an employee departs an organization, the organization expends resources to find and train a replacement for that employee. The team becomes disoriented, and customers could even start noticing the changes. It's like a domino effect that no business wants to deal with. Here are some factors causing employee attrition.

A. Salary

Compensation is important in an organisation. Employees are likely to look out for more opportunities when they are not well compensated for their work. When employees know they are not getting the right compensation that matches the industry standard for their work, they are more inclined to lookout for better opportunities that will match their efforts with the required compensation.

B. Growth Opportunities

Humans wants are insatiable, meaning an average employee will not be comfortable doing the same routine of work for a long period of time. Creating an avenue for career growth is important to reduce employee attrition rate. Employees are more likely to stay with an organization that provides an environment for advancement and professional development. When this environment is absent employees tend to feel undervalued and stagnant.

C. Recognition

Regular acknowledgment of employees' hard work and achievements creates some sense of value and an atmosphere of belonging. Recognition can be in various forms such as awards, verbal acknowledgment, and incentives.

D. Overall Job Satisfaction

This encompasses various aspects of the work environment, including work-life balance, company culture, relationships with colleagues and supervisors, and the nature of the work itself. High levels of job satisfaction can enhance employee loyalty and morale, while low job satisfaction can be a strong predictor of intent to leave.

In this study, we employed machine learning algorithms, specifically KNN, Naïve Bayes, Logistic Regression, Random Forest, and Decision Tree, using the IBM HR Analytics Employee Attrition Performance dataset sourced from Kaggle.org. This dataset encompasses 1470 employee records and offers a rich set of 35 attributes for a comprehensive analysis. The subsequent sections of the paper are organized as follows: Section II delves into related works, Section III provides details on the dataset, including exploratory data analysis and feature importance. Section IV outlines the methodology, Section V presents experimental results, Section VI concludes the study and outlines future work, and Section VII contains the references.

II. RELATED WORK

In this section, we researched different studies and literature relevant to employee attrition to help position our project amidst existing research. We'll focus on research about what drives employees to leave, the latest in predicting turnover using analytics, and how things like company culture and pay affect staff retention. We aim to gain insights from proven strategies and lay a strong groundwork for our own analysis and predictions regarding employee attrition.

The main goal of this study [1] was to analyze the different elements influencing employee attrition and pinpoint the main factors that lead to an employee's departure. The Gaussian Naïve Bayes classifier was employed for its effectiveness, displaying a high recall rate of 0.54. This rate reflects the classifier's ability to accurately identify true positive cases, along with a notably low false negative rate of only 4.5%. These results highlight the effectiveness of AI in predicting and comprehending the dynamics of employee turnover.

This work [2] explored various machine learning techniques within an ensemble model to identify causes of employee attrition. They made use of multiple performance metrics like cumulative lift, lift, accuracy, and F1 score, alongside fit statistic measures such as the Gini coefficient, misclassification rate, and average square error. These measures help in selecting the most effective model for implementation in real-world settings. Although no single model emerged as universally perfect for every business scenario.

The paper [3] explored various machine learning techniques for predicting employee turnover, aiming to identify a model that is both highly accurate and interpretable. Among the tested models, Logistic Regression emerged as the most effective, achieving an 88% accuracy and an 85% AUC-ROC, providing a promising approach for organizations to understand and address the factors leading to employee departures.

A comparative analysis of different machine learning approaches was done in this work [4]. Naïve Bayes, SVM, decision tree, random forest, and logistic regression techniques were implemented to understand employee behavior patterns indicative of potential attrition. The experimental findings highlight logistic regression as the most effective method, boasting an impressive 86% accuracy in predicting attrition over other techniques.

III. DATASET

In this comparative study we utilized the IBM HR Employee Attrition dataset (available at Kaggle.org.) for data analytics and the construction of a generalized machine learning model. This dataset consists of 35 attributes that provides a comprehensive insight and has 1470 records. Fig1 gives a more detailed information of the dataset.

#	Column	Non-Null Count	Dtype
0	Age	1470 non-null	int64
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64
10	EnvironmentSatisfaction	1470 non-null	int64
11	Gender	1470 non-null	object
12	HourlyRate	1470 non-null	int64
13	JobInvolvement	1470 non-null	int64
14	JobLevel	1470 non-null	int64
15	JobRole	1470 non-null	object
16	JobSatisfaction	1470 non-null	int64
17	MaritalStatus	1470 non-null	object
18	MonthlyIncome	1470 non-null	int64
19	MonthlyRate	1470 non-null	int64
20	NumCompaniesWorked	1470 non-null	int64
21	Over18	1470 non-null	object
22	OverTime	1470 non-null	object
23	PercentSalaryHike	1470 non-null	int64
24	PerformanceRating	1470 non-null	int64
25	RelationshipSatisfaction	1470 non-null	int64
26	StandardHours	1470 non-null	int64
27	StockOptionLevel	1470 non-null	int64
28	TotalWorkingYears	1470 non-null	int64
29	TrainingTimesLastYear	1470 non-null	int64
30	WorkLifeBalance	1470 non-null	int64
31	YearsAtCompany	1470 non-null	int64
32	YearsInCurrentRole	1470 non-null	int64
33	YearsSinceLastPromotion	1470 non-null	int64
34	YearsWithCurrManager	1470 non-null	int64

Fig 1

A. Exploratory Data Analysis

Our Exploratory Data analysis has illuminated a range of critical factors that organizations should prioritize to effectively manage employee attrition. These include understanding and addressing gender dynamics, recognizing the varying impacts of marital status, and evaluating the influence of different departmental roles. Additionally, educational background emerges as a significant factor, with its varied influence on employee retention. Age also plays a crucial role, with younger employees showing different attrition patterns compared to their older counterparts. Furthermore, compensation is a key element, where competitive and fair remuneration can significantly influence an employee's decision to stay. Work experience, both in terms of years and the nature of the job, has shown to be pivotal in determining employee loyalty and satisfaction.

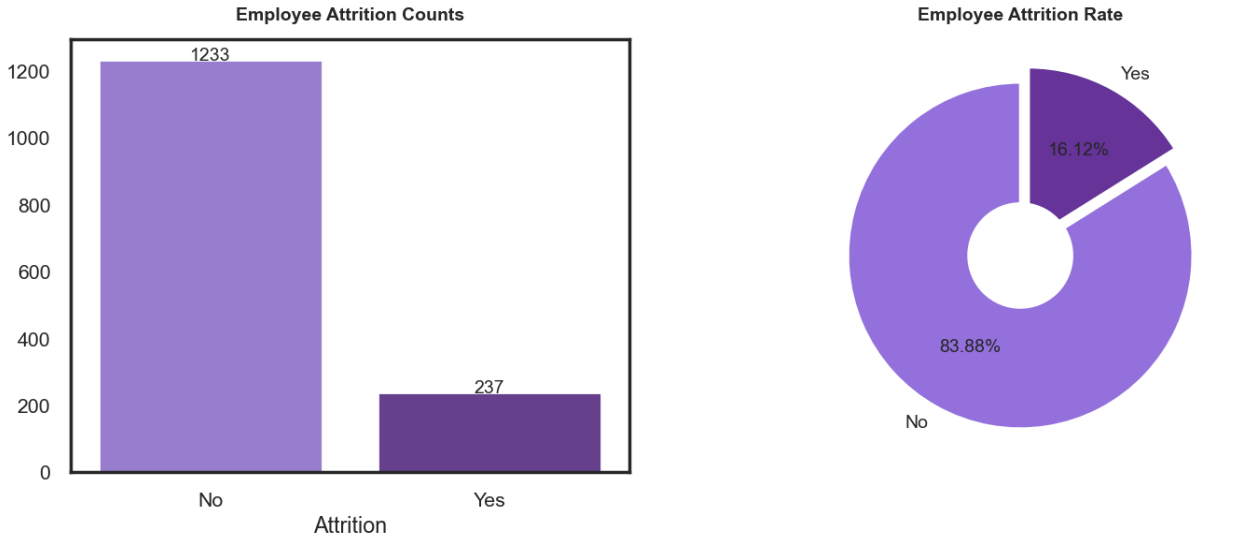


Fig 2

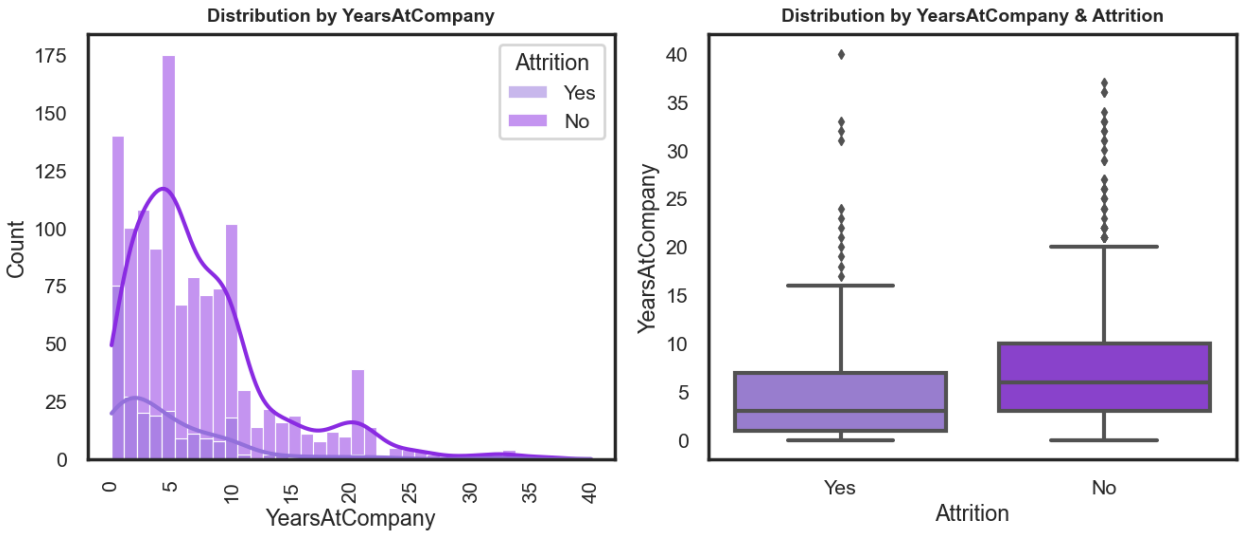


Fig 3

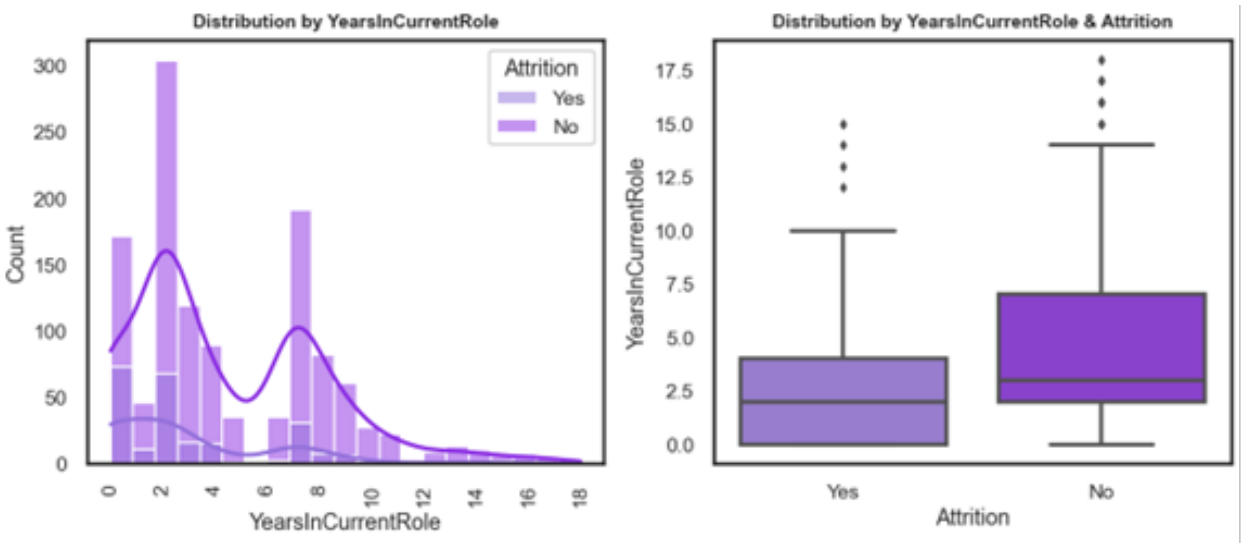

Distribution by YearsInCurrentRole
Attrition
Yes
No
Count
YearsInCurrentRole
Distribution by YearsInCurrentRole & Attrition
YearsInCurrentRole
Yes
No
Attrition

Fig 4

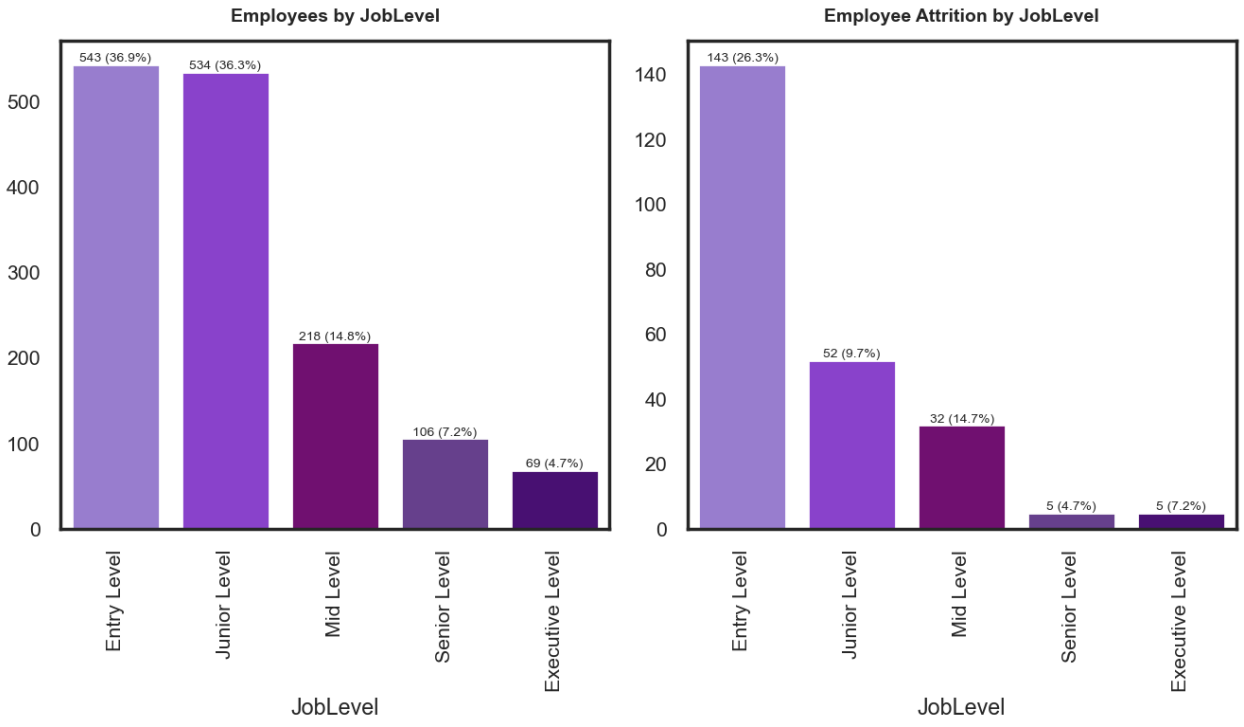

Employees by JobLevel
543 (36.9%)
534 (36.3%)
218 (14.8%)
106 (7.2%)
69 (4.7%)
Entry Level
Junior Level
Mid Level
Senior Level
Executive Level
JobLevel
Employee Attrition by JobLevel
143 (26.3%)
52 (9.7%)
32 (14.7%)
5 (4.7%)
5 (7.2%)
Entry Level
Junior Level
Mid Level
Senior Level
Executive Level
JobLevel

Fig 5

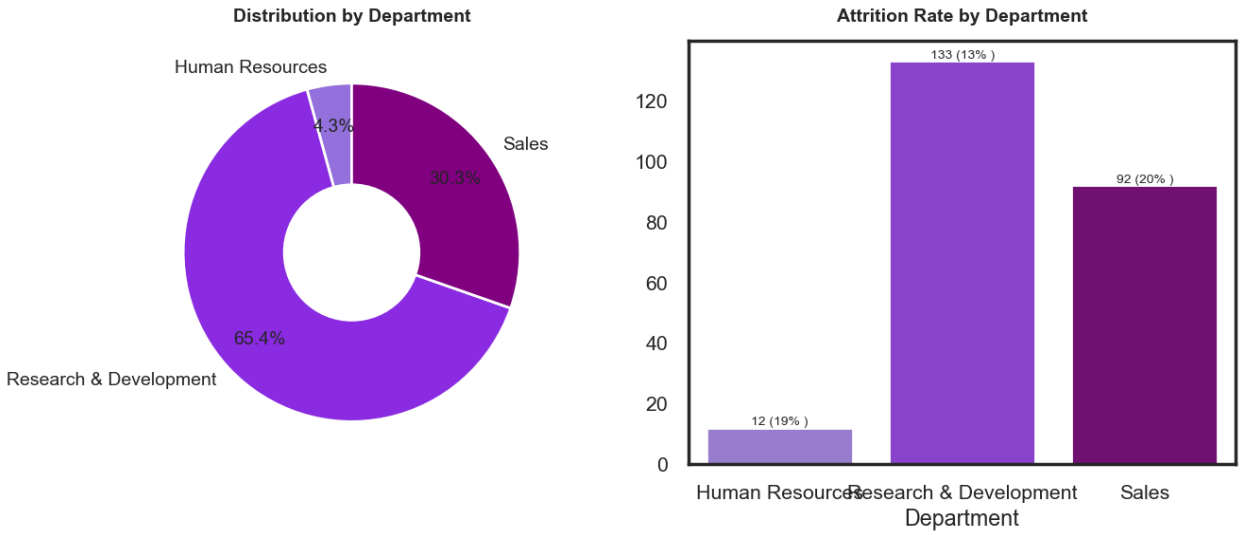


Fig 6

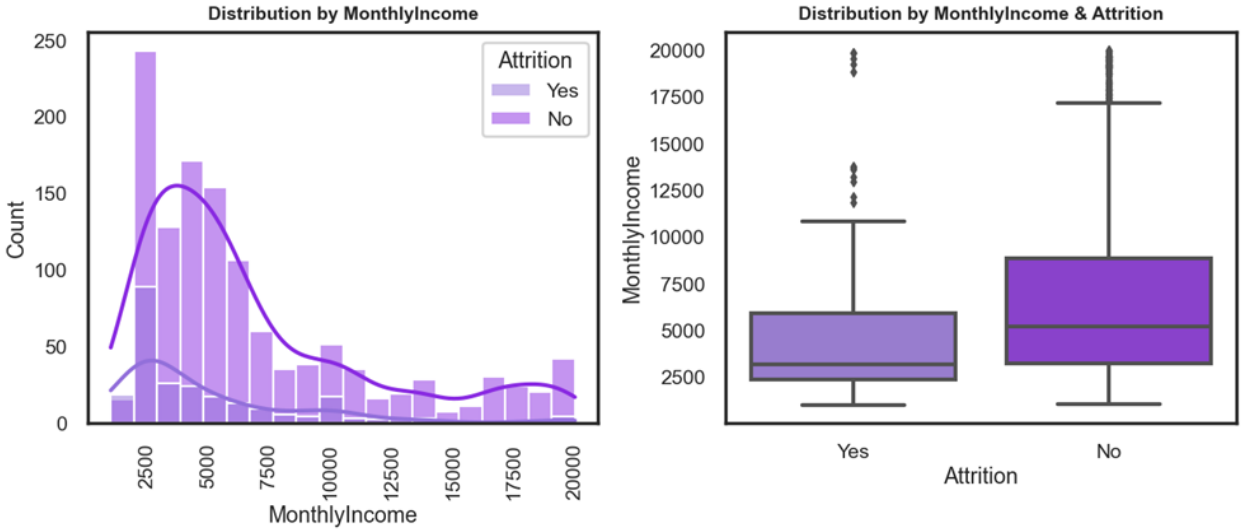


Fig 7

B. Feature Importance

In our statistical analysis of the Employee Attrition dataset, we employed the ANOVA test for numerical features to determine their significance and impact. This test specifically assessed whether there were notable differences in the means of these features between employees who left the company and those who stayed. For the categorical features, we used the Chi-Square test to evaluate their relevance in the context of employee attrition. This test was crucial in examining the relationships and associations between various categorical variables and the incidence of employee attrition, providing a comprehensive understanding of the factors influencing workforce turnover.

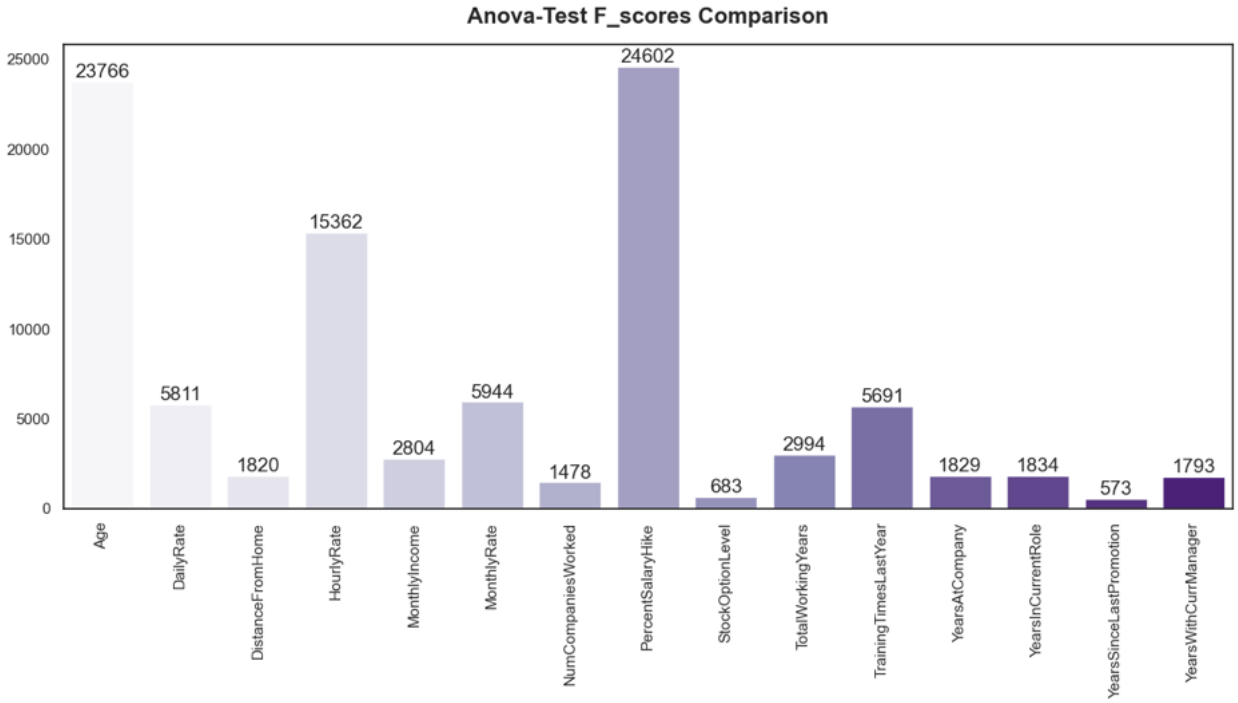

Anova-Test F_scores Comparison
25000
20000
15000
10000
5000
0
23766
5811
1820
15362
2804
5944
1478
24602
683
2994
5691
1829
1834
573
1793
Age
DailyRate
DistanceFromHome
HourlyRate
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Fig 8

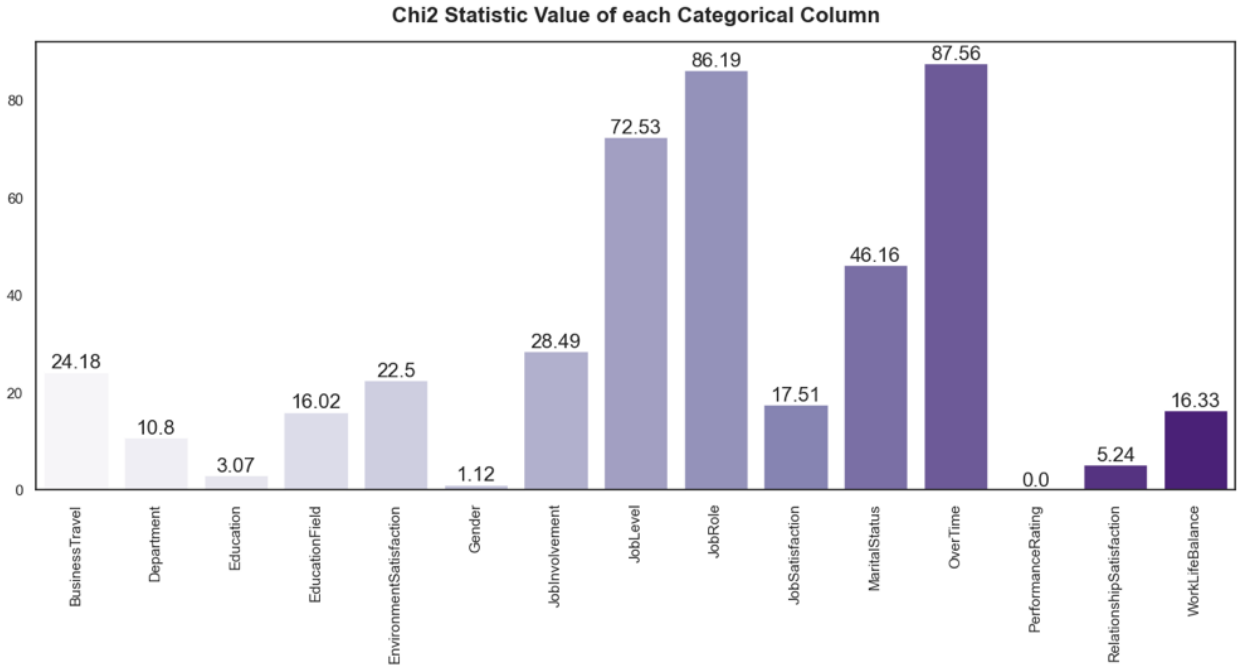

Chi2 Statistic Value of each Categorical Column
80
60
40
20
0
24.18
10.8
3.07
16.02
22.5
1.12
28.49
72.53
86.19
17.51
46.16
87.56
0.0
5.24
16.33
BusinessTravel
Department
Education
EducationField
EnvironmentSatisfaction
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
OverTime
PerformanceRating
RelationshipSatisfaction
WorkLifeBalance

Fig 9

IV. METHODOLGY

A. Data Preprocessing

In this preprocessing phase we first computed the dimensions of the dataset and provided statistical summaries, offering a detailed overview of the numerical features and their descriptive statistics for a thorough understanding of the data's characteristics. We then focused on data integrity by examining label categories within numerical features to pinpoint unique classifications. This was followed by an in-depth analysis that included addressing any duplicate records and missing values, as well as conducting descriptive analyses on both numerical and categorical attributes. Additionally, we explored the unique values in categorical attributes to gain deeper insights into the data's distinct aspects.

B. Model Building

This study encompasses the training and assessment of five machine learning models: logistic regression, Naive Bayes, k-Nearest Neighbors (KNN), Decision Tree, and Random Forest. Furthermore, each model undergoes hyperparameter tuning. Concise overviews of the foundational theories for these models are provided in the subsequent subsections.

- **Logistic Regression**

Logistic Regressions, a statistical technique that predicts the likelihood that an instance falls into a specific category, is employed in binary classification. The link between independent factors and the likelihood of a certain result is modelled by logistic regression, which is used for classification problems despite its name. An integer between 0 and 1, which represents the likelihood that an instance would belong to the positive class, is obtained by the logistic function from a linear combination of input data.[5]

- **Naive Bayes**

The Naive Bayes algorithm is a basic learning method that applies the Bayes rule and makes the strong assumption that, given the class, the attributes are conditionally independent. Even though the independence assumption is frequently broken in real-world scenarios, naïve Bayes frequently achieves competitive classification accuracy. When combined with several other appealing characteristics and its computing efficiency, naive Bayes is used extensively in real-world scenarios.

- **KNN**

k-Nearest Neighbors (kNN) is a versatile supervised learning algorithm that classifies a data point based on the majority class of its k nearest neighbors in the feature space.[6] It is non-parametric, instance-based, and commonly used for both classification and regression tasks. The algorithm's simplicity contributes to its effectiveness, particularly in scenarios with complex decision boundaries.

- **Decision Tree**

Decision trees (DT) stand as powerful algorithms with the ability to effectively model intricate datasets, finding applications in diverse tasks like medical diagnosis and assessing the credit risk of loan applications. The process of decision tree learning involves approximating a target function, depicted as a tree of "if-then" rules, aimed at enhancing human interpretability [7]. This approach involves iteratively breaking down a dataset into progressively smaller subsets, commencing from the topmost node known as the "root." The resulting decision tree comprises decision nodes and leaf nodes, proficiently handling both categorical and numerical data.

- **Random Forest**

Random Forest, as an ensemble learning algorithm, constructs numerous decision trees during training and amalgamates their outputs to enhance predictive accuracy. By introducing randomness using a random subset of features for each tree and bootstrapping samples, this methodology not only boosts performance but also mitigates overfitting, resulting in resilient outcomes across varied datasets. [8]

C. Hyperparameter Tuning

The practice of improving a machine learning model's hyperparameters for increased performance and generalisation is known as hyperparameter tuning. Hyperparameters are external settings that are defined before training begins and are not learnt from the training set. A neural network's number of hidden layers, regularisation strength, and learning rates are a few examples.[5]

Finding the collection of hyperparameter values that optimises the model's performance on a validation set or minimises a selected performance measure, such accuracy or mean squared error, is the goal of hyperparameter tuning. Until an ideal configuration is achieved, this iterative procedure entails modifying hyperparameter values, training the model, and assessing its performance.

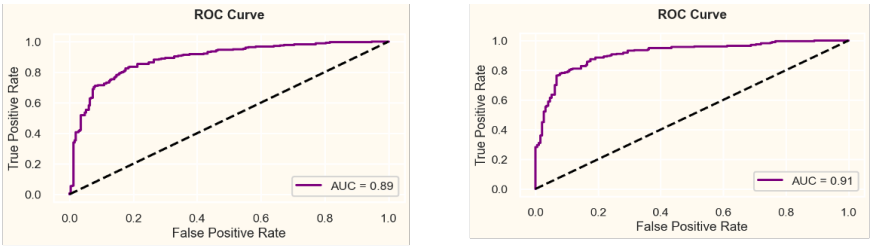
The introduced configurations for tuning include:

Model	**Parameter Grid**
Logistic Regression	•C: [0.001, 0.01, 0.1, 1, 10, 100] •Penalty: ['l1', 'l2']
Naive Bayes	•n_neighbors: [3, 5, 7, 9] •Weights: ['uniform', 'distance'] •p: [1, 2]
KNN	•var_smoothing: np.logspace(0, -9, num=10)
Decision Tree	•max_depth: [None, 10, 20, 30] •min_samples_split: [2, 5, 10] •min_samples_leaf: [1, 2, 4]
Random Forest	•n_estimators: [50, 100, 200] •max_depth: [None, 10, 20, 30] •min_samples_split: [2, 5, 10] •min_samples_leaf: [1, 2, 4]

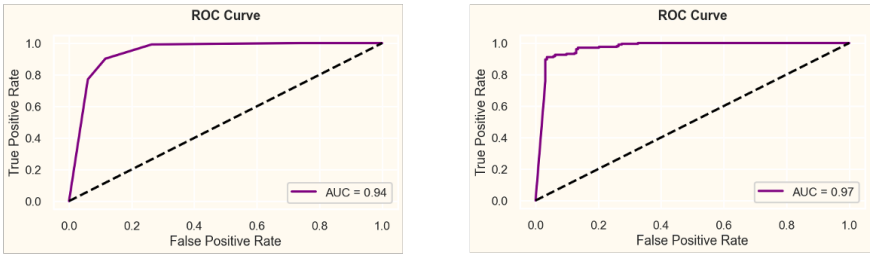
V. RESULTS

Algorithms	Accuracy	Precision	Recall
Logistic Regression	81.983806	0.818930	0.815574
Logistic Regression Tuned	84.810127	0.828704	0.886139
KNN	86.234818	0.785714	0.991803
KNN Tuned	86.329114	0.796000	0.985149
Naive Bayes	75.101215	0.731801	0.782787
Naive Bayes Tuned	77.215190	0.735294	0.866337
Decision Tree	79.352227	0.773077	0.823770
Decision Tree tuned	78.734177	0.758772	0.856436
Random Forest	90.485830	0.919149	0.885246
Random forest tuned	92.405063	0.921569	0.930693

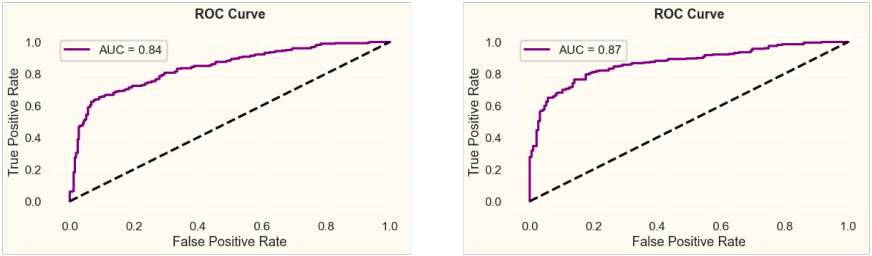
Significant data on predicting employee attrition has been gathered from the model evaluations. After hyperparameter adjustment, Logistic Regression showed a considerable increase in accuracy from 81.98% to 84.81%. The K-Nearest Neighbours (KNN) model demonstrated a little enhancement from 86.23% to 86.33%, highlighting the effectiveness of parameter modifications. After adjusting, Naive Bayes had a noteworthy increase from 75.10% to 77.22%. The Decision Tree model emphasises the significance of interpretability, although exhibiting a minor decline from 79.35% to 78.73%. After tweaking, the Random Forest ensemble showed outstanding accuracy, increasing from 90.49% to 92.41%. These findings offer complex information to help with strategic decision-making about staff retention and forecasts of loss of staff landscape.



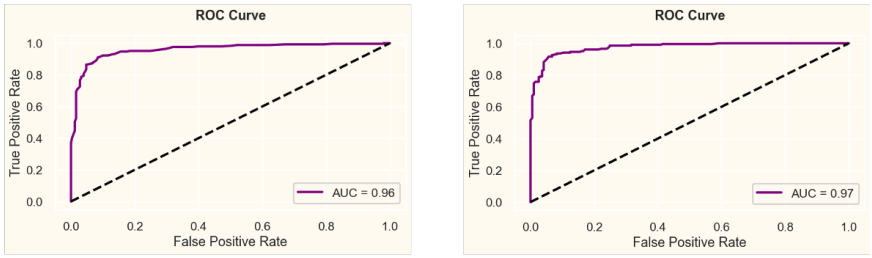
Logistic Regression Base vs Tuned



KNN Base vs Tuned



Naive Bayes Base vs Tuned



Random Forest Base vs Tuned

VI. CONCLUSION AND FUTURE WORK

In conclusion up, the evaluation of machine learning models for forecasting employee attrition produced complex results. The accuracy of the Random Forest, K-Nearest Neighbours, Naive Bayes, and Logistic Regression models was much enhanced by hyperparameter tweaking, highlighting the algorithms' capacity for attrition predictions. Even if accuracy is a little bit lower, the Decision Tree model highlights how important interpretability is. Strategic decision-making for staff retention is informed by these insights; two especially promising models are Logistic Regression and Random Forest. Subsequent investigations will concentrate on temporal analysis, model hyperparameter optimisation, and the integration of many data sources. Opportunities for development include investigating feature engineering, resolving class imbalance, and using ensemble approaches. Prediction algorithms are intended to gain more depth by including ethical issues and employee engagement indicators. Ongoing development will be facilitated by cross-validation studies, targeted treatments for high-risk personnel, and continual monitoring methods. Models' applicability in a variety of situations is ensured by accounting for global variables, which advances the developing subject of employee attrition prediction.

VII. REFERENCES

[1] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," Computers, vol. 9, p. 86, 2020. doi: 10.3390/computers9040086.

[2] F. Kamal Alsheref, I. Eldesoky, and W. Ead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms," Computational Intelligence and Neuroscience, 2022. doi: 10.1155/2022/7728668.

[3] F. Guerranti and G. Dimitri, "A Comparison of Machine Learning Approaches for Predicting Employee Attrition," Applied Sciences, vol. 13, p. 267, 2022. doi: 10.3390/app13010267.

[4] K. K. Mohbey, "Employee's Attrition Prediction Using Machine Learning Approaches," 2020. doi: 10.4018/978-1-7998-3095-5.ch005.

[5] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967. doi: 10.1109/TIT.1967.1053964.

[7] J. R. Quinlan, "Induction of decision trees," Mach. Learn., Kluwer Academic Publishers, Boston-Manufactured in The Netherland, pp. 81–106, 1978.

[8] A. Liaw and M. Wiener, "Classification and regression by Random Forest," R News, vol. 2, no. 3, pp. 18-22, 2002.