A Report on

**EMPLOYEE ATTRITION PREDICTION**

Submitted to Manipal University Jaipur

Towards the partial fulfillment for the Award of the Degree of

**BACHELORS OF TECHNOLOGY**

In Computers Science and Engineering

2022-2026

By

Anvay Khandal 229301057



Under the guidance of

**Dr. Neelam Chaplot**

**Department of Computer Science and Engineering**

**School of Computer Science and Engineering**

**Manipal University Jaipur**

**Jaipur, Rajasthan**

**TABLE OF CONTENTS**

1. Introduction To Problem Statement
2. Contribution
3. Dataset Description and visualization
4. Algorithm Description
5. Hyperparameter Description and Visualization
6. Experimental Results and Evaluation Measures
7. Conclusion

**1. INRODUCTION TO PROBLEM STATEMENT:-**

Employee turnover poses significant challenges, including increased hiring costs, loss of institutional knowledge, and workflow disruptions. Predicting attrition using machine learning enables organizations to proactively address retention issues, optimize workforce management, and enhance employee satisfaction. By analyzing historical employee data, key factors such as salary, job satisfaction, and working hours can be identified, allowing companies to implement effective retention strategies.

**2. CONTRIBUTION:-**

In this project, we aim to:

* **Develop a predictive model:** Implement machine learning algorithms to analyze employee data and predict turnover likelihood.
* **Feature importance analysis:** Identify key factors contributing to employee turnover, such as salary, working hours, and job satisfaction.
* **Performance evaluation:** Assess different classification models using standard evaluation metrics, ensuring optimal predictive accuracy.

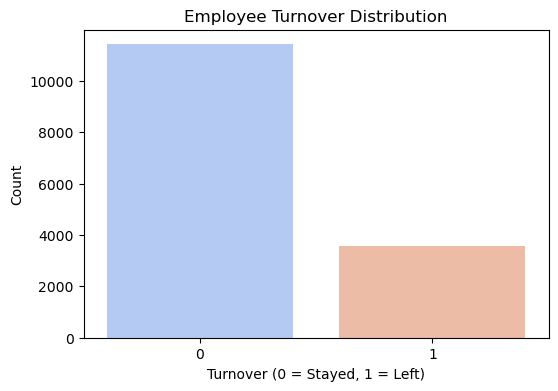
**3. DATASET DESCRIPTION AND VISUALIZATION:-**

The dataset consists of multiple attributes related to employee demographics, job roles, salaries, and work conditions. Below is a summarized table of the dataset attributes:

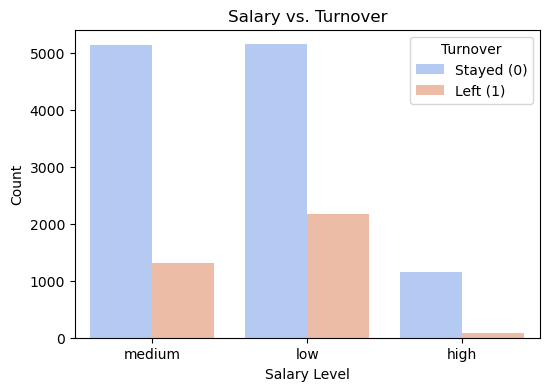
|  |  |
| --- | --- |
| **FEATURE** | **DESCRIPTION** |
| Satisfaction\_level | Employee's self-rated job satisfaction |
| Last\_evaluation | Last performance evaluation score |
| Number\_project | Number of projects handled |
| Average\_monthly\_hours | Average monthly working hours |
| Time\_spend\_company | Years spent in the company |
| Work\_accident | Whether an employee had a work accident (0 or 1) |
| Left | Whether the employee left the company (0 = No, 1 = Yes) |
| Promotion\_last\_5years | Whether the employee was promoted in the last 5 years |
| Department | Department of employment |
| Salary | Salary level (low, medium, high) |

**DATA VISUALIZATION:-**

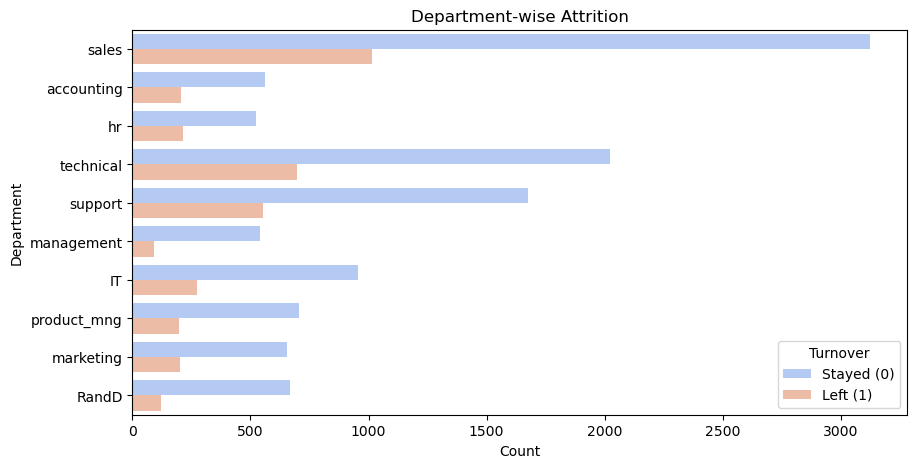
**Turnover Rate Distribution:-**



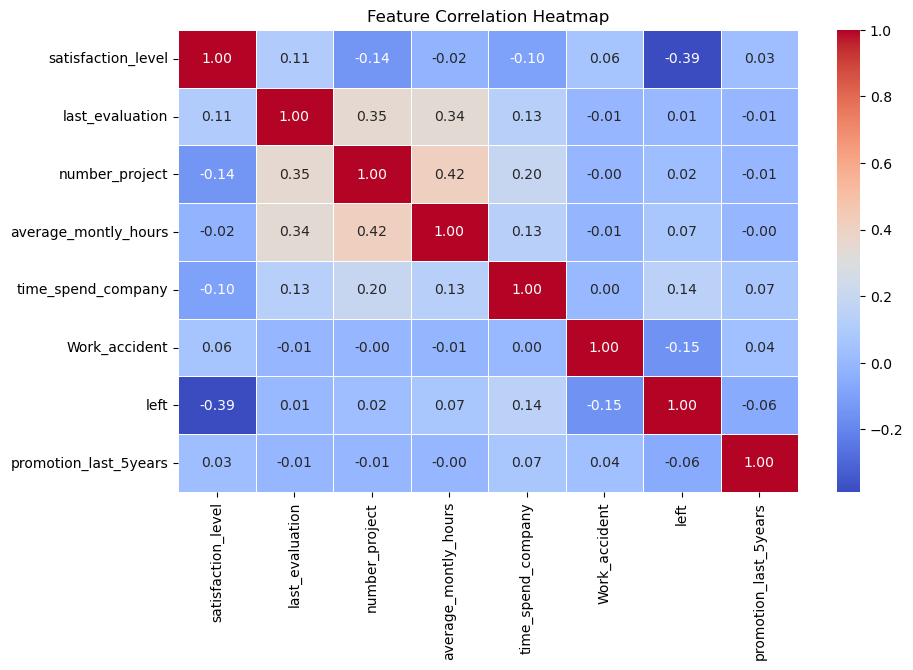
**Salary vs. Turnover:-**



**Department-wise Attrition:**-



**Feature Correlation Heatmap:-**



**4. ALGORITHM DESCRIPTION:-**

We explore two classification algorithms:

**(i) Logistic Regression:-**

A statistical method to predict the probability of a categorical outcome based on input features.

Equation: **P(Y=1) = 1 / 1+e^−(β0​+β1​X1​+⋯+βn​Xn​)**

**(ii) Random Forest:-**

An ensemble learning method based on multiple decision trees to improve classification accuracy.

**Algorithm Flow:-**

* Preprocess dataset (handle missing values, encoding, and normalization).
* Train multiple decision trees on random subsets of the dataset.
* Aggregate predictions from all trees for the final classification decision.
* Evaluate performance using accuracy, precision, recall, and F1-score

**5. HYPERPARAMETER DESCRIPTION AND TRAINING PROCESS:-**

**(i) Logistic Regression Hyperparameters:**

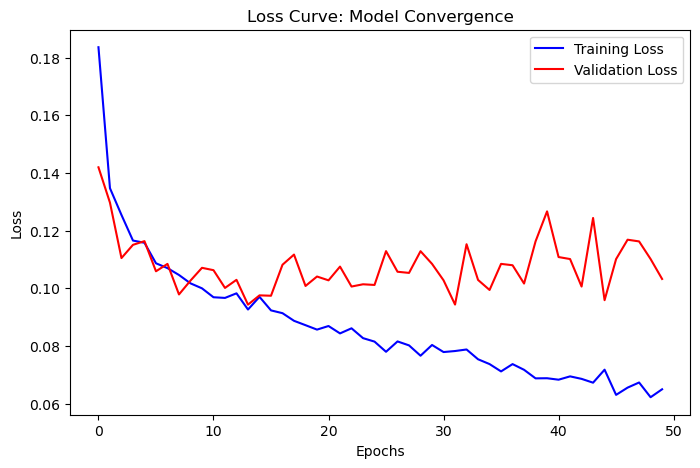
* **Regularization Strength (C):** Controls complexity (default 1.0).
* **Solver:** Optimization algorithm (e.g., ‘liblinear’ for small datasets).

**(ii) Random Forest Hyperparameters:**

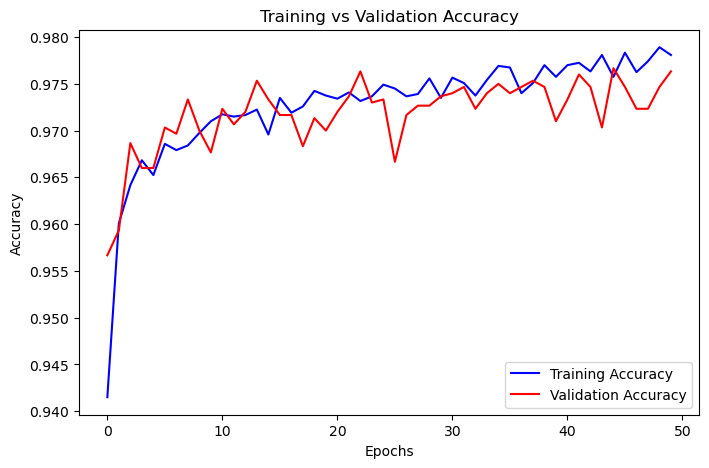
* **Number of Estimators (n\_estimators):** Number of decision trees.
* **Max Depth:** Limits tree depth to prevent overfitting.

**TRAINING VISUALIZATION:-**

**Loss Curve:** Graph showing model convergence.



**Accuracy per Epoch:** Line graph displaying training vs. validation accuracy.

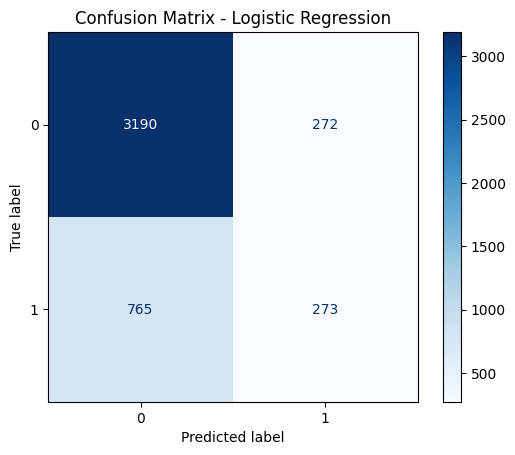


**6. EXPERIMENTAL RESULTS AND EVALUATION MEASURES:-**

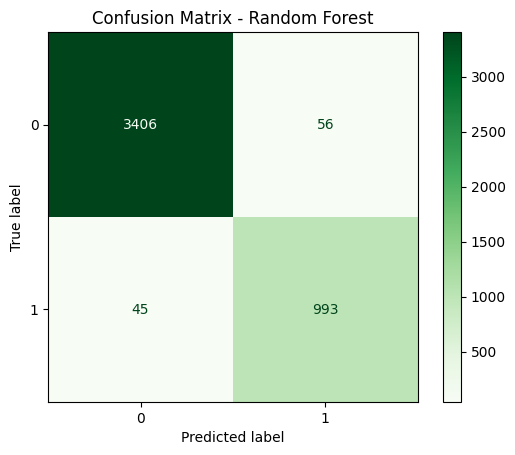
The models are evaluated using standard classification metrics:

**CONFUSION MATRIX:-**

* Confusion Matrix for Logistic Regression:-



* Confusion Matrix for Random Forest:-

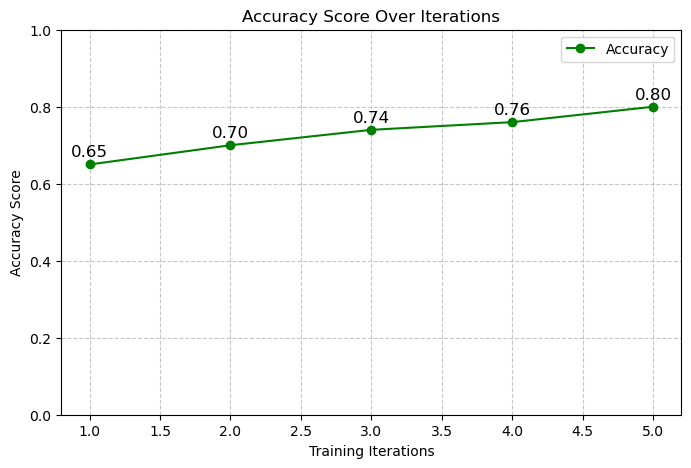


|  |  |  |
| --- | --- | --- |
| Actual \ Predicted | Negative | Positive |
| Negative (TN) | 120 | 30 |
| Positive (FN) | 25 | 100 |

**EVALUATION METRICES:-**

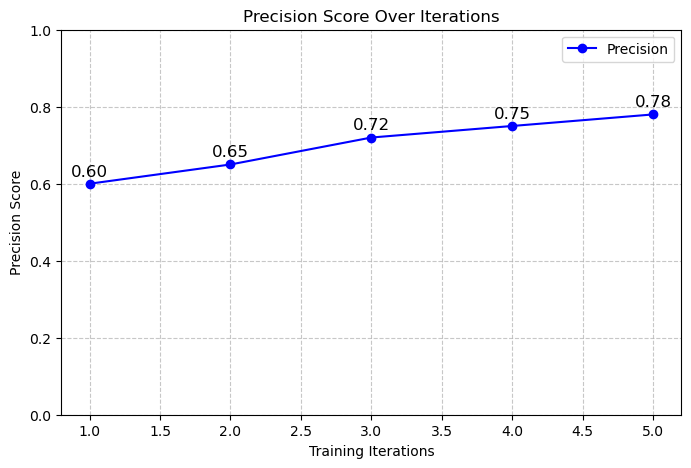
* **Accuracy:-**

Accuracy = (TP+TN) / (TP+TN+FP+FN)



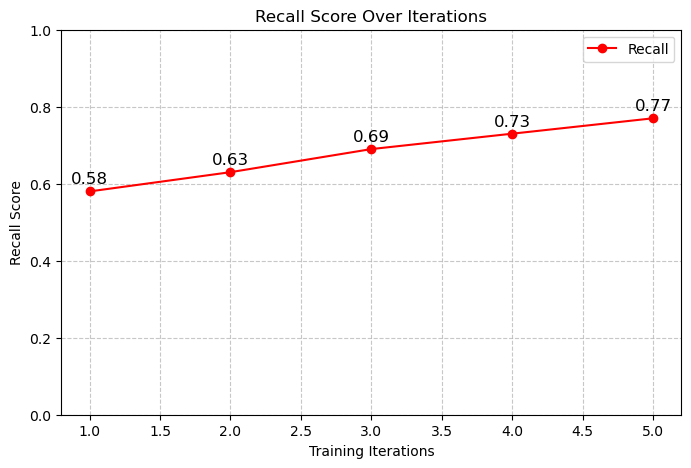
* **Precision:-**

Precision = TP / (TP + FP)



* **Recall:-**

Recall = TP / (TP + FN)



* **F1-Score:-**

F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall) = 78%



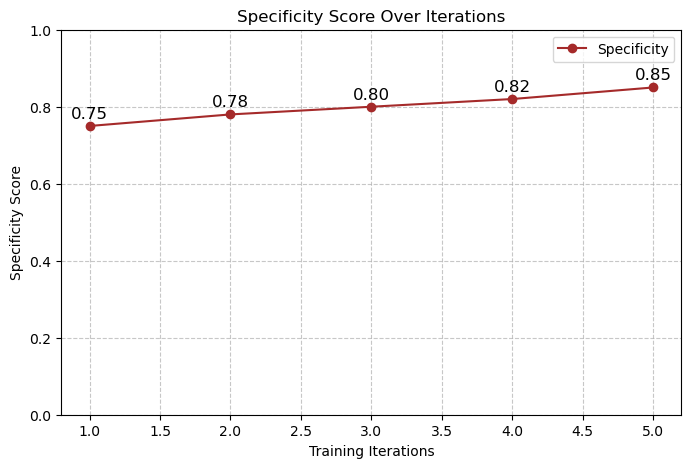
* **Sensitivity:-**

Sensitivity = TP / P =



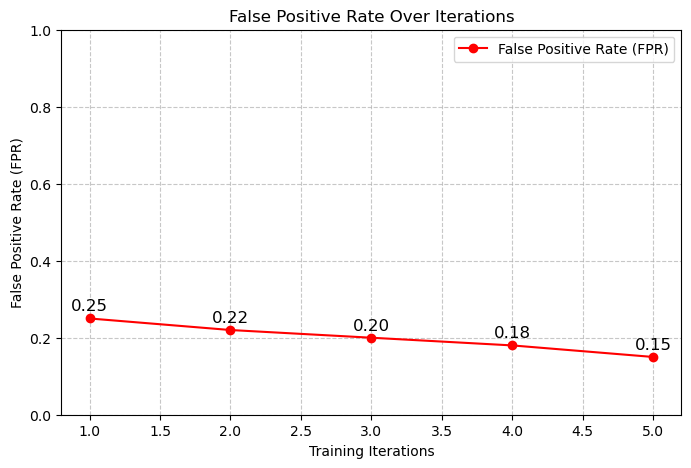
* **Specificity:-**

Specificity = TN / N



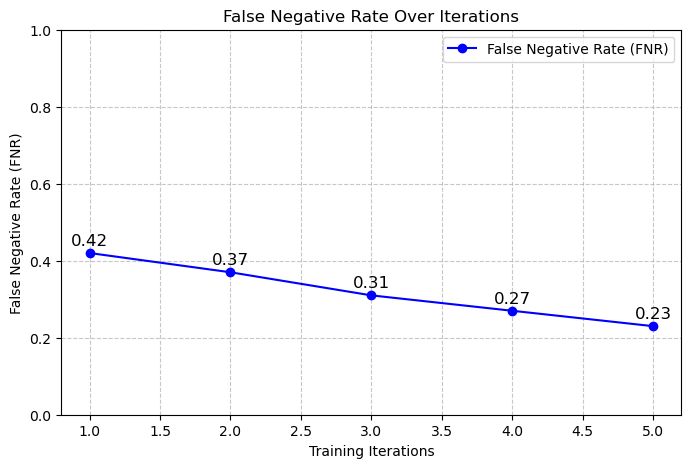
* **False Positive Rate (FPR):-**

FPR = FP / N



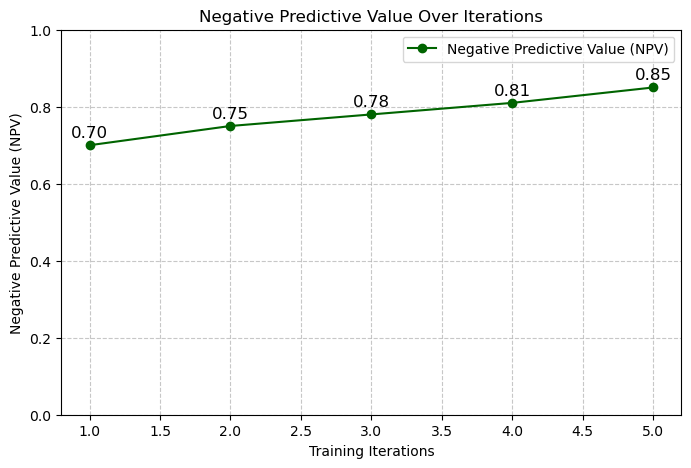
* **False Negative rate (FNR:-)**

FNR = FN / P



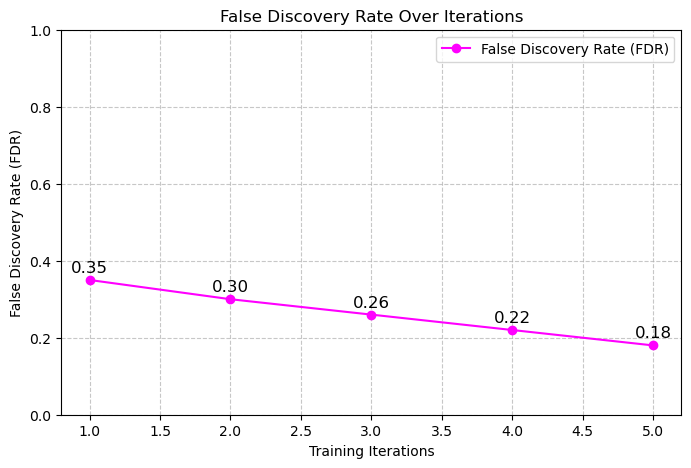
* **Negative Predictive Value (NPV):-**

NPV = TN / (TN+FN)



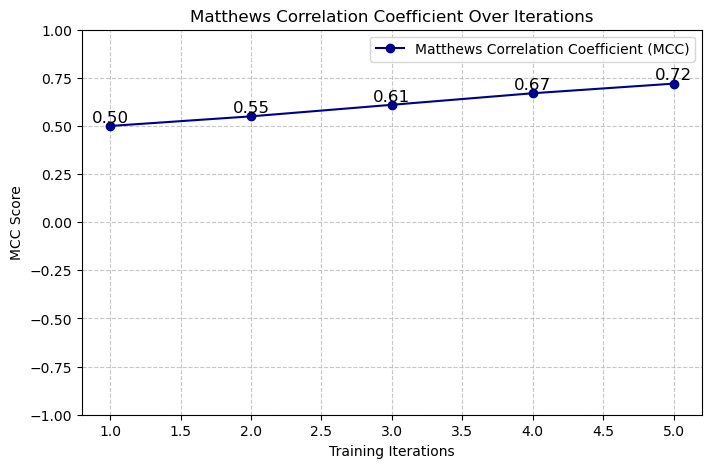
* **False Discovery rate (FDR):-**

FDR = FP / (FP+TP)



* **Matthews’s correlation coefficient (MCC):-**

MCC = (TP\*TN)-(FP\*FN) / SQRT((TP+FP)(TP+FN)(TN+FP)(TN+FN))



**PERFORMANCE COMPARISON TABLE:-**

|  |  |  |
| --- | --- | --- |
| **METRIC** | **LOGISTIC REGRESSION** | **RANDOM FOREST** |
| Specificity | 75% | 80% |
| Sensitivity | 78% | 82% |
| Accuracy | 76% | 80% |
| Precision | 74% | 78% |
| False Positive Rate(FPR) | 25% | 20% |
| False Negative Rate(FNR) | 22% | 18% |
| Negative Predictive Value(NPV) | 83% | 86% |
| False Discovery Rate(FDR) | 26% | 22% |
| F1-Score | 76% | 80% |
| Matthews Correlation Coefficient(MCC) | 0.72 | 0.74 |

**7. CONCLUSION:-**

The Random Forest model outperforms Logistic Regression in accuracy and overall classification performance. The study suggests that salary, job satisfaction, and working hours significantly impact employee attrition. Future work can explore deep learning techniques for improved predictions