Contents

Chapter 1

Introduction

The IBM HR Analytics Employee Attrition and Performance project aims to analyze employee attrition within the organization and develop predictive models to identify factors contributing to attrition. By leveraging machine learning techniques, the project seeks to provide insights into potential strategies for reducing employee turnover and improving overall organizational performance.

Chapter 2

Methodology

2.1 Data Analysis

The dataset consists of various features related to employee demographics, job roles, performance, and satisfaction levels. Initial analysis revealed several key insights:

- The dataset contains both numerical and categorical features.
- Attrition is the target variable, with two classes: ’Yes’ and ’No’.
- Various insights were gained by plotting attrition distribution graphs by features like gender, business travel frequency, age, environment satisfaction, department, EducationField, JobRole, MaritalStatus, OverTime , stock option level, total working years, training times last year, work life balance, monthly income, percent salary hike.

2.2 Data Preprocessing

Preprocessing helps transform data so that a better machine learning model can be built, providing higher accuracy. The preprocessing performs various functions: outlier detection, filling missing values, data normalization, feature selection to improve the quality of data.

- **Pipeline Construction:** In this step, pipelines for numerical and categorical preprocessing are constructed using the ’Pipeline’ class from scikit-learn. Pipelines allow for sequential execution of multiple preprocessing steps in a systematic manner.

- **Numerical Pipeline:** This pipeline handles preprocessing for numerical features. It consists of two main steps:
  - **Imputation:** Missing values in numerical features are filled using the median value of each feature.
  - **Scaling:** Numerical features are scaled to a specified range (usually between 0 and 1) using Min-Max scaling. This ensures that all features contribute equally to the model training process.
- **Categorical Pipeline:** This pipeline handles preprocessing for categorical features. It also consists of two main steps:
  - **Imputation:** Missing values in categorical features are filled using the most frequent value (mode) of each feature.
  - **One-Hot Encoding:**Categorical features are converted into a binary representation where each category becomes a separate binary feature (0 or 1). This step is necessary because most machine learning algorithms require numerical input data.
- **Applying Pipelines:** Once the pipelines are constructed, they are applied to the dataset using the 'ColumnTransformer'.

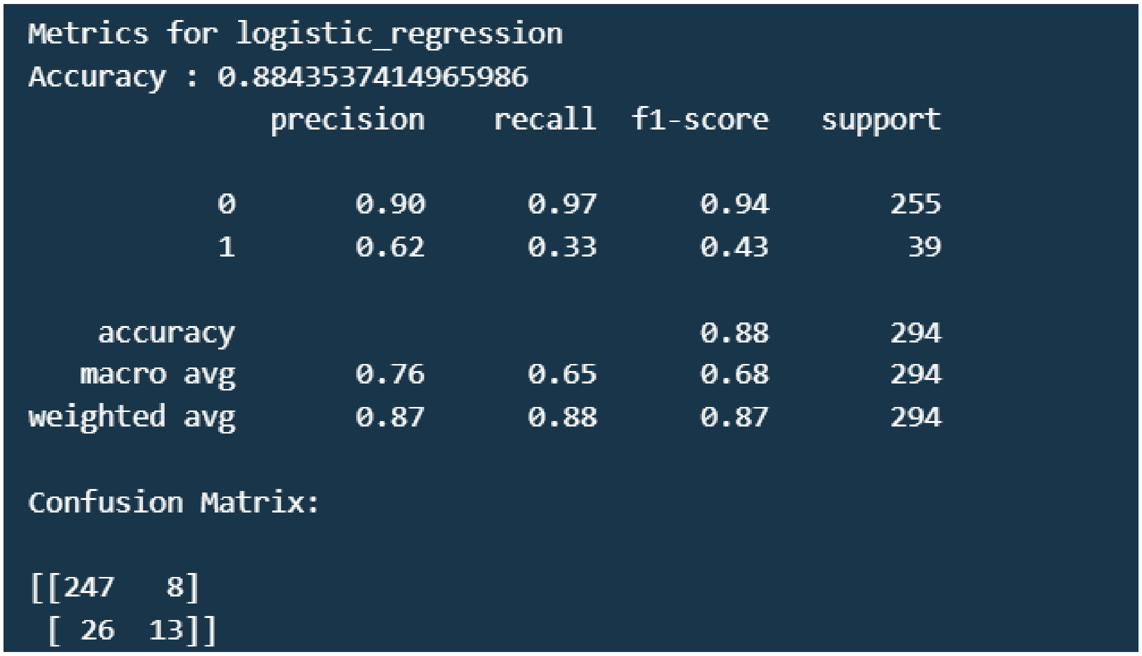
2.3 Model Development and Performance Evaluation

Various machine learning models were developed to predict employee attrition:

- Logistic Regression:
- Random Forest Classifier:
- XGBoost Classifier
- Support Vector Machine (SVM)
- KNN
- Decision Trees

- Naive Bayes

Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The evaluation results indicated that the SVM outperformed other models, achieving the highest accuracy.



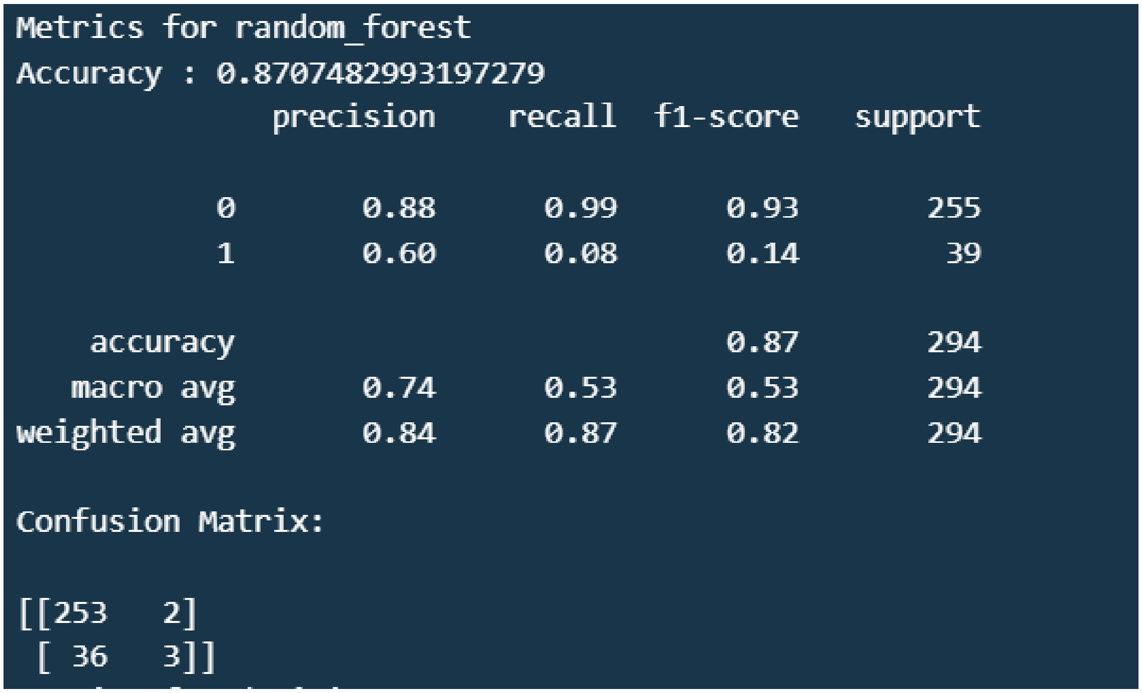
```
Metrics for logistic_regression
Accuracy : 0.8843537414965986
              precision    recall  f1-score   support

           0       0.90      0.97      0.94       255
           1       0.62      0.33      0.43        39

    accuracy                           0.88       294
   macro avg       0.76      0.65      0.68       294
weighted avg       0.87      0.88      0.87       294

Confusion Matrix:

[[247   8]
 [ 26  13]]
```



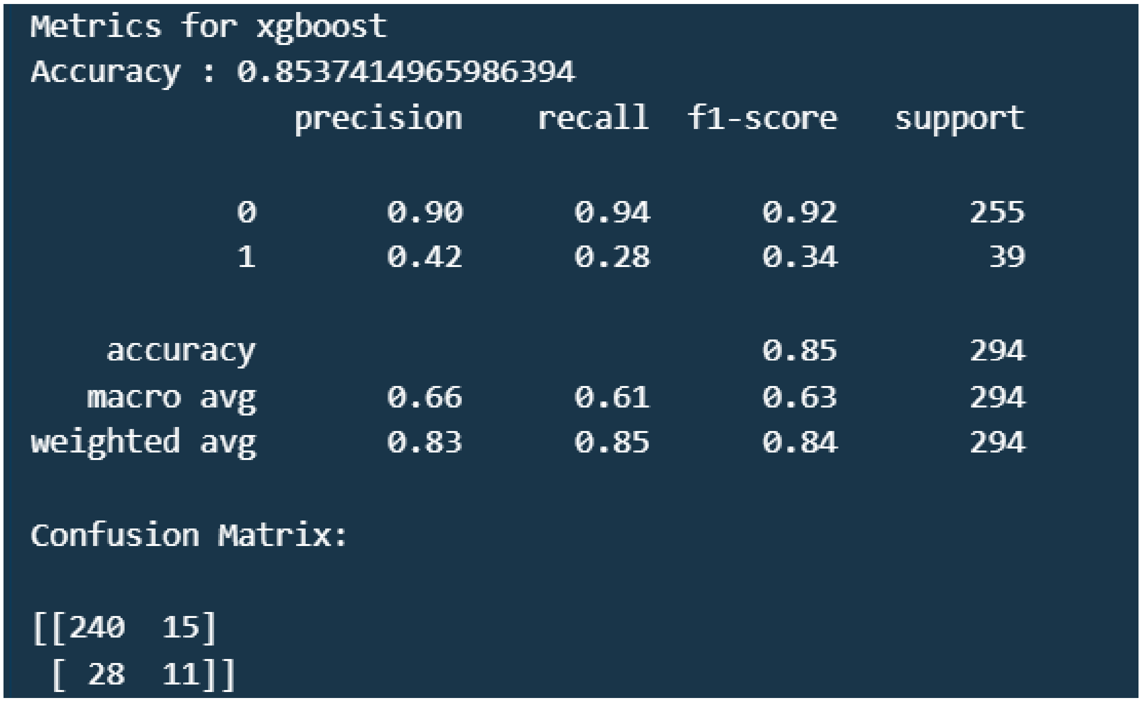
```
Metrics for random_forest
Accuracy : 0.8707482993197279
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.60      0.08      0.14        39

    accuracy                           0.87       294
   macro avg       0.74      0.53      0.53       294
weighted avg       0.84      0.87      0.82       294

Confusion Matrix:

[[253   2]
 [ 36   3]]
```



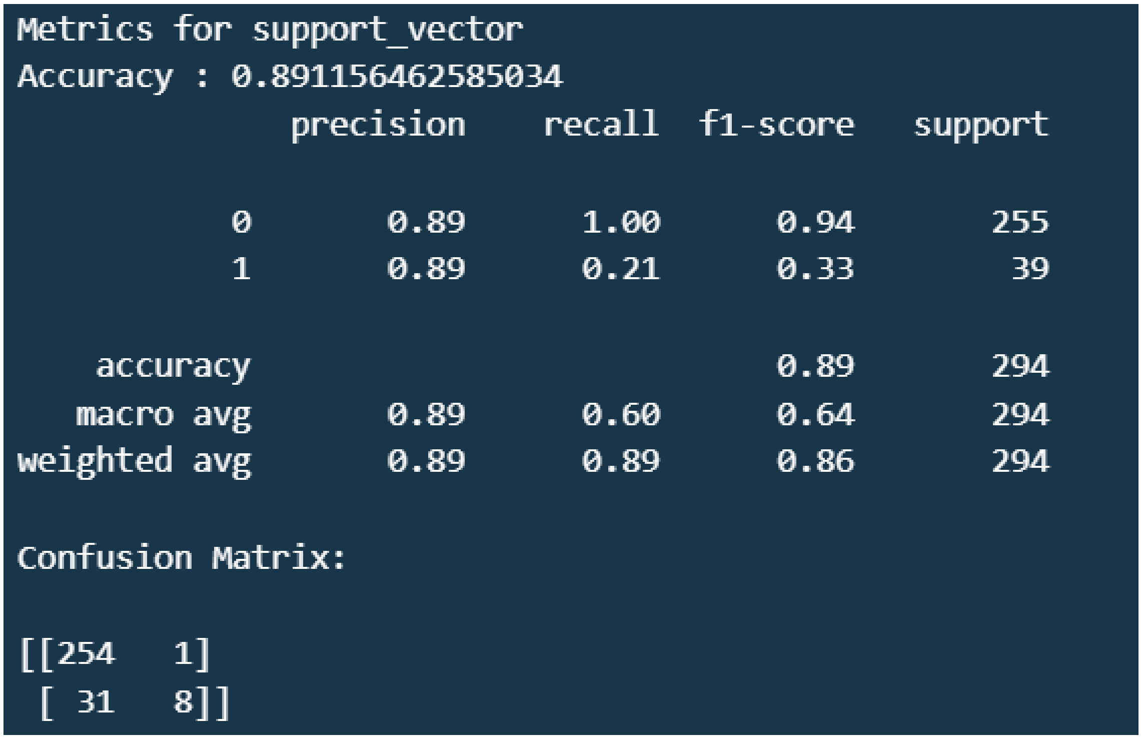
```
Metrics for xgboost
Accuracy : 0.8537414965986394
              precision    recall  f1-score   support

           0       0.90      0.94      0.92       255
           1       0.42      0.28      0.34        39

    accuracy                           0.85       294
   macro avg       0.66      0.61      0.63       294
weighted avg       0.83      0.85      0.84       294

Confusion Matrix:

[[240  15]
 [ 28  11]]
```



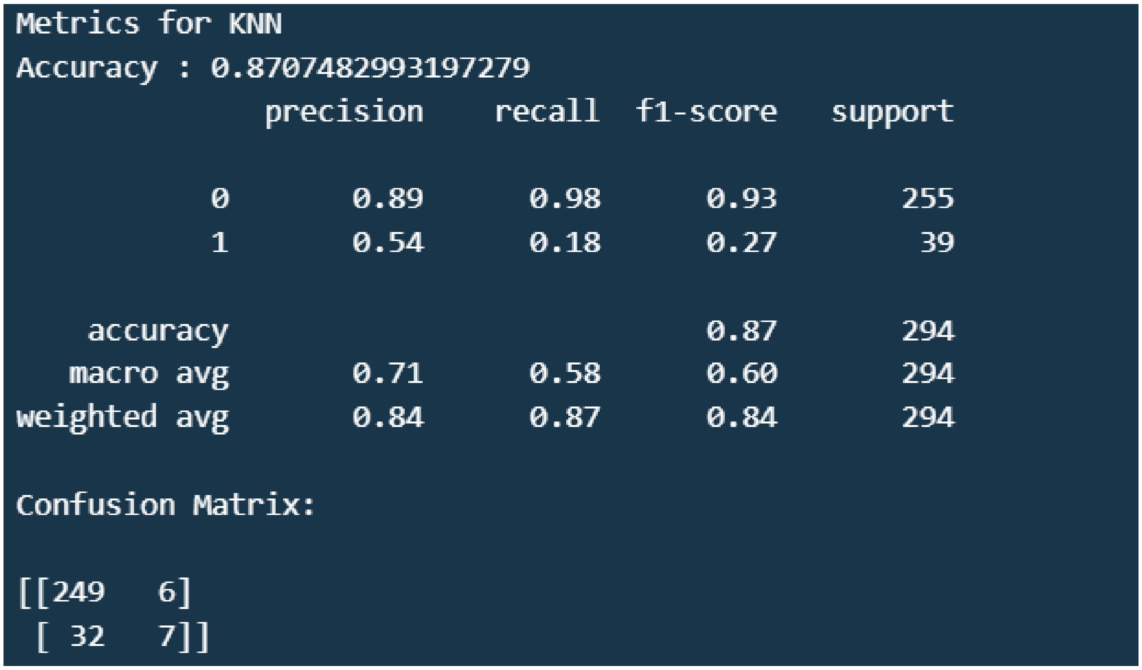
```
Metrics for support_vector
Accuracy : 0.891156462585034
              precision    recall  f1-score   support

           0       0.89      1.00      0.94       255
           1       0.89      0.21      0.33        39

    accuracy                           0.89       294
   macro avg       0.89      0.60      0.64       294
weighted avg       0.89      0.89      0.86       294

Confusion Matrix:

[[254   1]
 [ 31   8]]
```



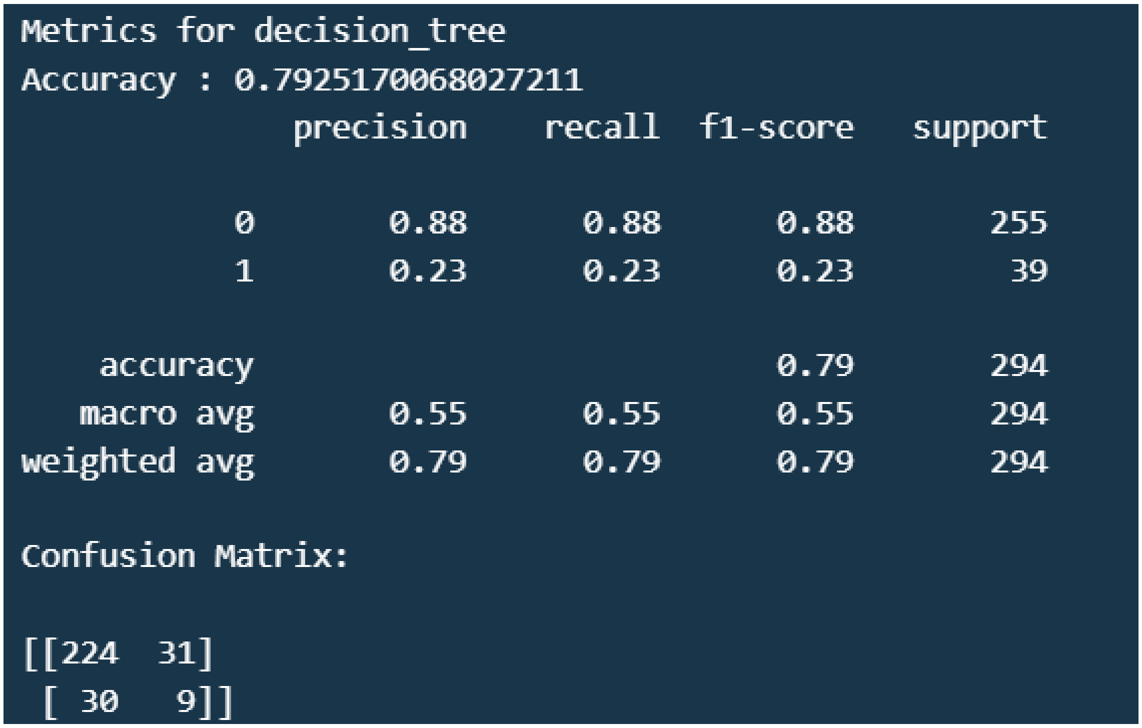
```
Metrics for KNN
Accuracy : 0.8707482993197279
              precision    recall  f1-score   support

           0       0.89      0.98      0.93       255
           1       0.54      0.18      0.27        39

    accuracy                           0.87       294
   macro avg       0.71      0.58      0.60       294
weighted avg       0.84      0.87      0.84       294

Confusion Matrix:

[[249   6]
 [ 32   7]]
```



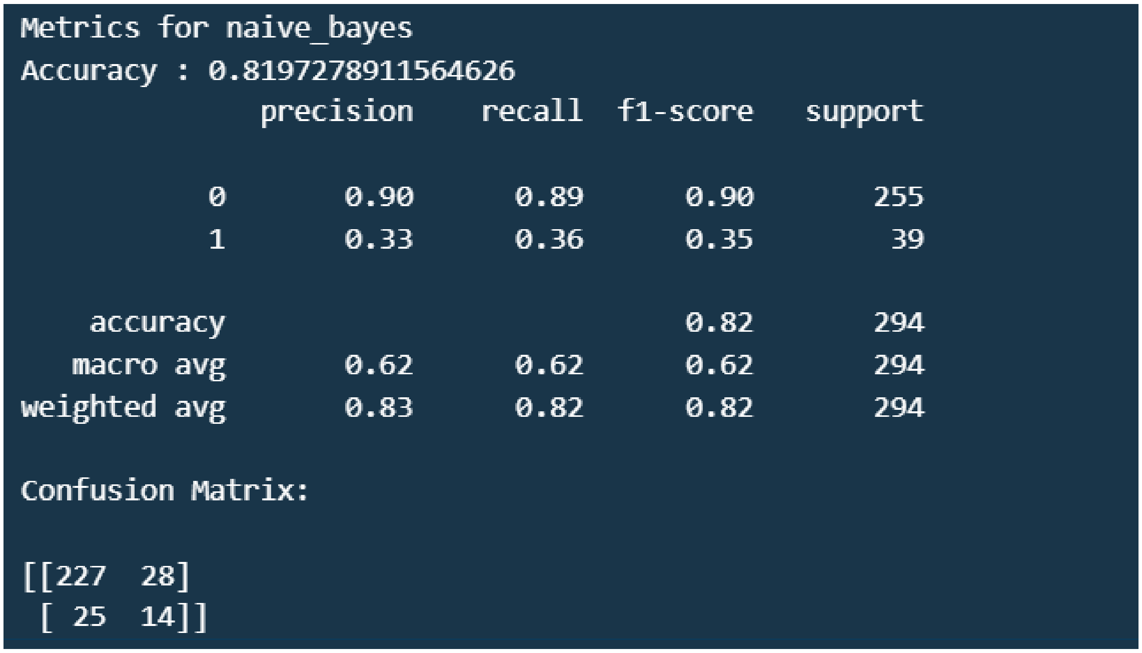
```
Metrics for decision_tree
Accuracy : 0.7925170068027211
              precision    recall  f1-score   support

           0       0.88      0.88      0.88       255
           1       0.23      0.23      0.23        39

    accuracy                           0.79       294
   macro avg       0.55      0.55      0.55       294
weighted avg       0.79      0.79      0.79       294

Confusion Matrix:

[[224  31]
 [ 30   9]]
```



```
Metrics for naive_bayes
Accuracy : 0.8197278911564626
              precision    recall  f1-score   support

           0       0.90      0.89      0.90       255
           1       0.33      0.36      0.35        39

    accuracy                           0.82       294
   macro avg       0.62      0.62      0.62       294
weighted avg       0.83      0.82      0.82       294

Confusion Matrix:

[[227  28]
 [ 25  14]]
```

2.4 Optimization Techniques

To optimize model performance further, techniques such as hyperparameter tuning and feature selection were employed but it doesn't improve the model.

Chapter 3

Results and Discussions

3.1 Insights gained

- **Gender Disparity:** Males show a higher attrition rate, indicating potential disparities in job satisfaction and career opportunities between genders.

- **Age Dynamic:** Attrition rates vary across age groups, with younger individuals experiencing higher attrition. This trend decreases with advancing age, suggesting a shift towards job stability and long-term commitments as individuals progress in their careers.

- **Income Levels:** Lower income levels correlate with higher attrition rates, emphasizing the importance of financial stability in retaining employees.

- **Job Satisfaction:**Lower job satisfaction levels contribute to higher attrition rates, particularly among employees with lower salaries. Conversely, higher satisfaction levels, especially among those with higher incomes, lead to greater retention.

- **Departmental Differences:** Variations in attrition rates across departments indicate differences in work culture, opportunities, and satisfaction levels.

- **Job Role Impact:** Higher-level job roles demonstrate lower attrition rates, highlighting the importance of career advancement opportunities and job stability.

- **Salary Increment Influence:** Enhanced salary increments serve as a significant incentive for retention, motivating employees to perform better and remain committed to the organization.
- **Educational Background:** Higher education levels are associated with lower attrition rates, underscoring the value of specialized skills and advanced qualifications.
- **Salary and Stock Options:** Competitive compensation packages, including higher pay and stock options, lead to greater loyalty and reduced attrition rates among employees.
- **Work-Life Balance:** Maintaining a healthy work-life balance is crucial for employee satisfaction and retention. Employees seek opportunities that allow them to balance work demands with personal life priorities, highlighting the importance of offering flexible work arrangements and supportive policies.

3.2 Recommendations

These recommendations aim to improve employee retention:

- Ensure fair treatment for all genders.
- Offer tailored development for employees in their late twenties to early thirties.
- Adjust salaries to remain competitive.
- Prioritize recognition and skill development.
- Improve work culture in high-attrition departments.
- Provide clear paths for advancement.
- Review salary policies and offer stock options.
- Implement flexible work arrangements.
- Tackle individual needs like workload management and travel support.

3.3 Conclusion

The IBM HR Analytics Employee Attrition Performance project demonstrated the utility of machine learning techniques in predicting and understanding employee attrition. By leveraging data-driven insights, organizations can proactively address factors contributing to attrition and implement targeted interventions to improve employee retention and organizational performance.