HR ANALYTICS

PROJECT ANALYSIS

JUNE 25, 2024

ANNAS KHALID & ZAIN SHEIKH
CSC-21F-064 & CSC-21F-134

Abstract

We utilized the ‘HR Analytics’ dataset from Kaggle. This dataset is specifically designed to analyze employee attrition and other related factors within an organization. It contains 38 attributes spanning demographic information, job roles, education, performance, and satisfaction metrics for employees. Key variables include **EmpID, Age, Attrition, BusinessTravel, Department, Education, JobRole, MonthlyIncome, OverTime, PerformanceRating, WorkLifeBalance,** among others. The dataset is structured to facilitate comprehensive analysis of the factors influencing employee turnover, including job satisfaction, career progression, and personal demographics. Such analysis aims to identify patterns and predictors of attrition, aiding in the development of strategies to enhance employee retention and satisfaction. This study compares the performance of three machine learning algorithms - Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, on the task of HR Analytics. We evaluate these models based on key metrics such as accuracy, log loss, precision, recall, and F1 score. Through comprehensive analysis and visualization, we aim to identify the strengths and weaknesses of each model in handling this classification task.

1. Introduction

1.1 Background

The human resources (HR) function is pivotal in managing an organization's workforce, ensuring the right talent is recruited, retained, and developed. Modern HR practices heavily rely on data analytics to make informed decisions that can improve employee satisfaction, productivity, and retention. The dataset at hand provides a comprehensive view of employee attributes and their workplace dynamics, which can be crucial for HR analytics.

1.2 Topic

The primary focus of this dataset is to explore various factors that influence employee attrition within an organization. By analyzing these factors, organizations can devise strategies to enhance employee retention and satisfaction. This dataset includes variables related to personal demographics, job role, work environment, compensation, and professional development, which are essential in understanding the holistic employee experience.

1.3 Problem Statement

Employee attrition is a significant challenge for many organizations, leading to increased recruitment costs, loss of institutional knowledge, and disruption in team dynamics. The problem is to identify the key factors that contribute to employee attrition and understand the patterns and correlations that can help predict and prevent future attrition. This analysis aims to uncover insights that can aid in developing targeted interventions to retain valuable talent.

1.4 Scope

The dataset encompasses various attributes of employees such as age, department, education, job role, performance ratings, salary components, and years at the company, among others. The analysis will focus on:

1. Descriptive statistics to summarize the dataset.
2. Identification of key factors influencing attrition.
3. Predictive modeling to determine the likelihood of attrition based on these factors.
4. Recommendations for HR policies and practices to mitigate attrition.

1.5 Conclude

This dataset serves as a critical resource for HR professionals and data analysts aiming to enhance workforce management strategies. By delving into the underlying causes of employee attrition, organizations can foster a more stable and motivated workforce, ultimately driving better organizational performance and employee well-being. The insights derived from this analysis will be instrumental in guiding data-driven HR decisions and policies.

2. LITERATURE REVIEW

2.1 Introduction

Human Resource (HR) analytics has emerged as a pivotal tool in modern organizational management, offering significant strategic value. The integration of advanced statistical methods and data analysis techniques into HR functions aims to enhance decision-making processes, optimize workforce management, and contribute to the competitive advantage of organizations.

2.2 Integration and Implementation of HR Analytics

The integration of HR analytics within an organization involves several critical steps, including data collection, analysis, and interpretation. Effective implementation requires a robust IT infrastructure and technological interventions to handle data storage, mining, and processing. Organizations must also develop the necessary skills in data analysis and interpretation among HR professionals to fully leverage the benefits of HR analytics.

2.3 Predictive Modelling and Decision-Making

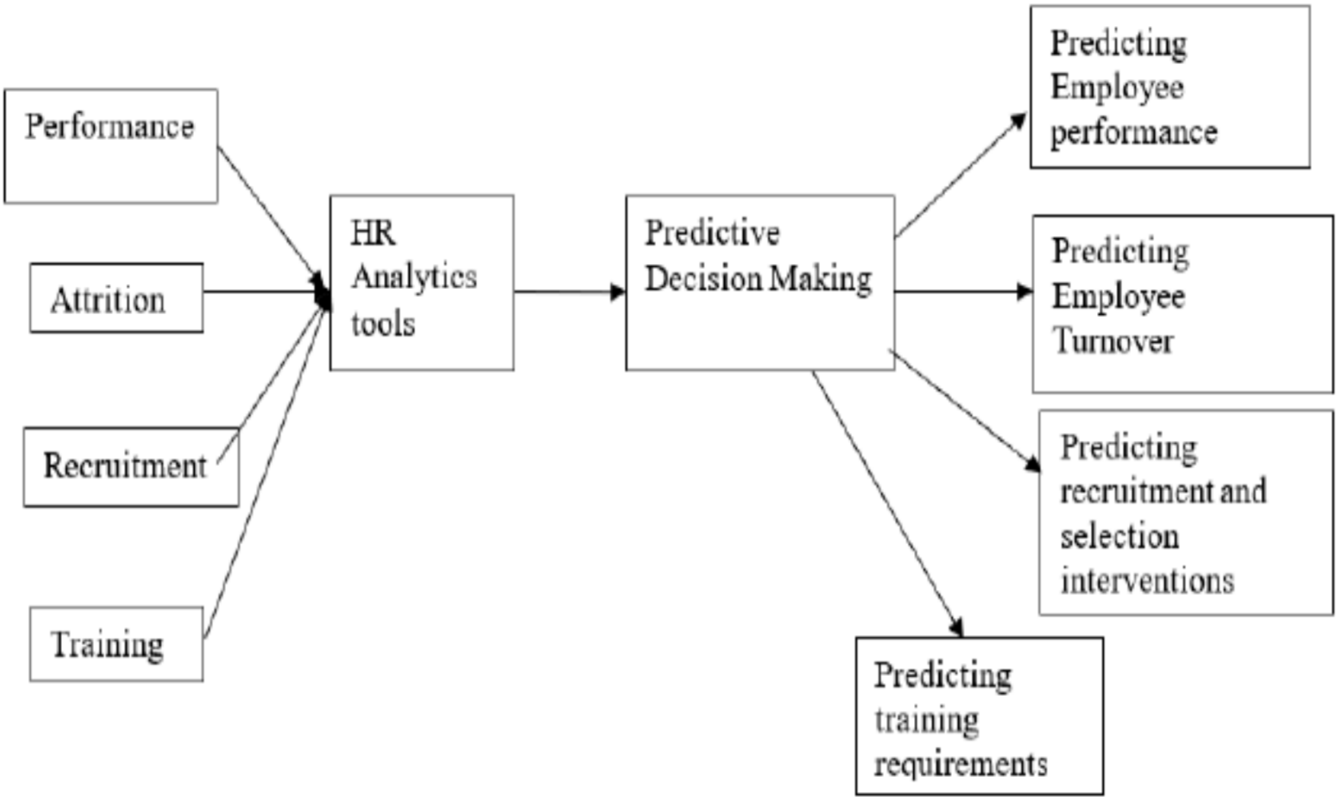
One of the significant contributions of HR analytics is in predictive modelling, which helps in forecasting employee behavior, attrition rates, training needs, and overall workforce dynamics. Predictive models use historical data to predict future outcomes, enabling organizations to make proactive decisions. However, the effectiveness of these models varies across different industries and organizational contexts, necessitating tailored approaches for different settings.

2.4 Challenges and Limitations

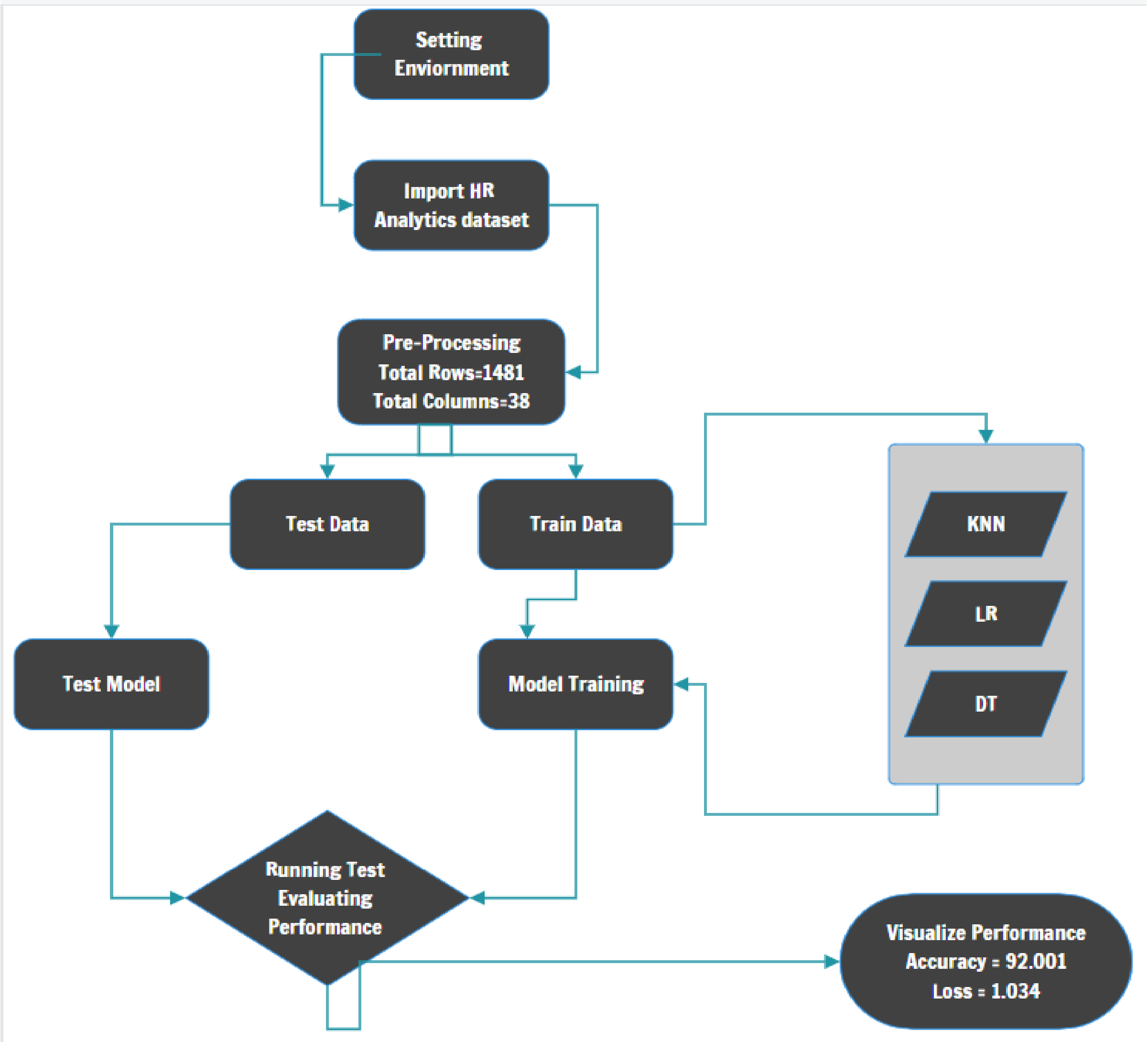
Despite its potential, HR analytics faces several challenges. There is a notable gap in the literature regarding the real-world applicability of theoretical models and the success of these models when implemented in organizational contexts. Additionally, the availability of comprehensive case studies and empirical data is limited, which hinders the development of universally applicable models. Future research should focus on addressing these gaps through detailed case studies and quantitative analysis.

2.5 Conclusion

HR analytics is a transformative tool that offers significant prospects for improving HR and managerial decision-making processes. The integration of analytics into HR functions supports evidence-based strategies, enhances the strategic value of HR, and contributes to organizational sustainability. However, to fully realize these benefits, organizations must address existing challenges and invest in developing the necessary skills and infrastructure.


Performance
Attrition
Recruitment
Training
HR Analytics tools
Predictive Decision Making
Predicting Employee performance
Predicting Employee Turnover
Predicting recruitment and selection interventions
Predicting training requirements

3. METHODOLOGY


Setting Enviornment
Import HR Analytics dataset
Pre-Processing Total Rows=1481 Total Columns=38
Test Data
Train Data
KNN
LR
DT
Test Model
Model Training
Running Test Evaluating Performance
Visualize Performance Accuracy = 92.001 Loss = 1.034

1. **Data Collection & Pre-Processing**

   * The dataset for this study was obtained from KAGGLE. This HR Analytics dataset consists of 1481 Rows (No. of Employees) & 38 Columns (Complete Employee Details). Descriptive statistics were calculated for numerical variables like Age, MonthlyRate, JobLevel, Jobsatisfaction, Salary & many more.

   * First the dataset was cleaned by removing unwanted columns with missing values. Then, we checked the info of data that how many columns have null values, then replaced those missing values with random values according to column's datatype. Then we changed the datatype of columns to make data more reliable.

2. Model Selection and Implementation

2.1 Logistic Regression (LR)

Logistic Regression is a fundamental classification technique that models the probability of a class based on input features. For multi-class tasks like digit recognition, it is extended using methods like SoftMax regression. It is simple, efficient, and serves as a strong baseline.

Parameters:

max_iter: The maximum number of iterations taken for the solvers to converge.
solver: The solver algorithm used for optimization. In this study, the **'lbfgs'** solver was chosen.
multi_class: The method used to handle multinomial logistic regression. We used the **'multinomial'** option for multiclass classification.

2.2 K-Nearest Neighbors (KNN)

KNN classifies data points based on the majority class of their nearest neighbors. It is non- parametric and straightforward.

Parameters:

n_neighbors: The number of neighbors to consider for classification. We used a value of **3** in this study.

2.3 Decision Tree

Constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness.

Parameters:

n_estimators: The number of trees in the forest. We utilized **100** decision trees in this study.

3. Model Training & Evaluation

3.1 Training

Each model is trained on the training set using appropriate training techniques (e.g., gradient descent for LR)

3.2 Evaluation Metrics

The following evaluation metrics are calculated for each model on the test set:

- **Accuracy:** The proportion of correctly classified instances.
- **Log Loss:** The negative log-likelihood of the predicted probabilities.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.
- **F1 Score:** The harmonic mean of precision and recall.

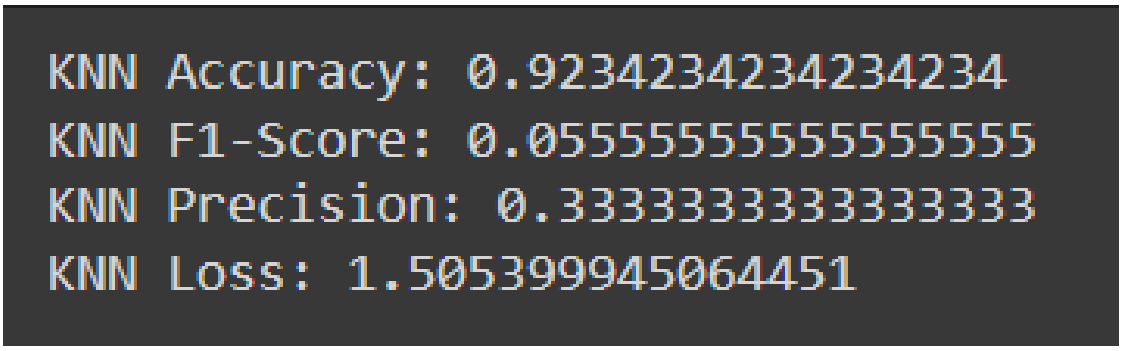
4. Performance Comparison

	LR	KNN	Decision Tree
Accuracy	0.9256	0.923	0.887
F1-Score	0.0	0.055	0.264
Precision	0.0	0.333	0.257
Loss	0.203	1.505	4.058

4. RESULTS

4.1 Applying KNN ALGORITHM:

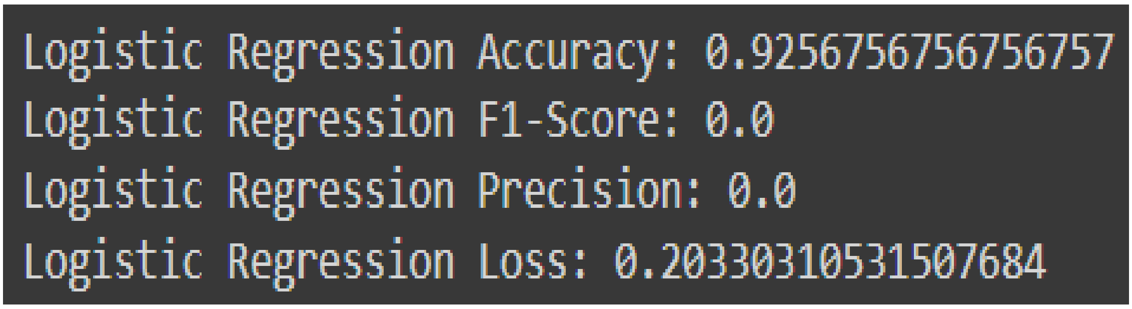
It’s a machine learning algorithm used for both classification and regression tasks. In KNN, the prediction for a new data point is based on the labels or values of its k nearest neighbors in the training data.



```
KNN Accuracy: 0.9234234234234234
KNN F1-Score: 0.05555555555555555
KNN Precision: 0.3333333333333333
KNN Loss: 1.505399945064451
```

4.2 Applying LOGISTIC REGRESSION:

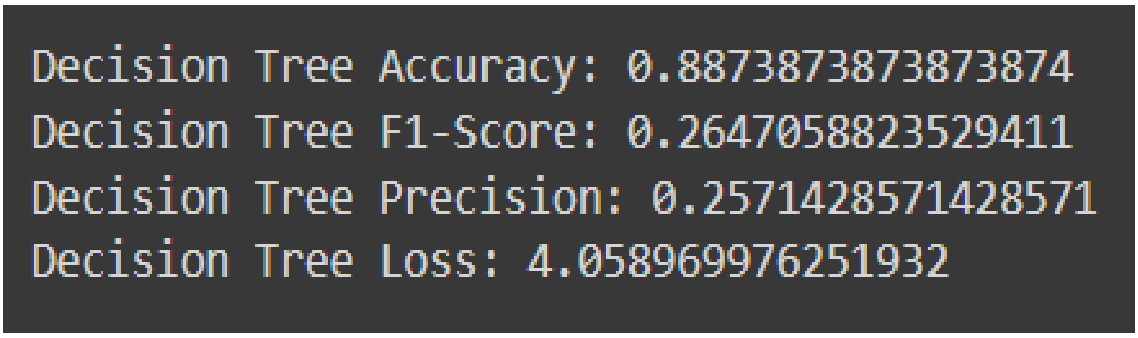
Logistic regression is a method used to predict a binary outcome (one with two possible results). For example, it can predict whether an email is spam (yes or no) or if a customer will buy a product (yes or no).



```
Logistic Regression Accuracy: 0.9256756756756757
Logistic Regression F1-Score: 0.0
Logistic Regression Precision: 0.0
Logistic Regression Loss: 0.20330310531507684
```

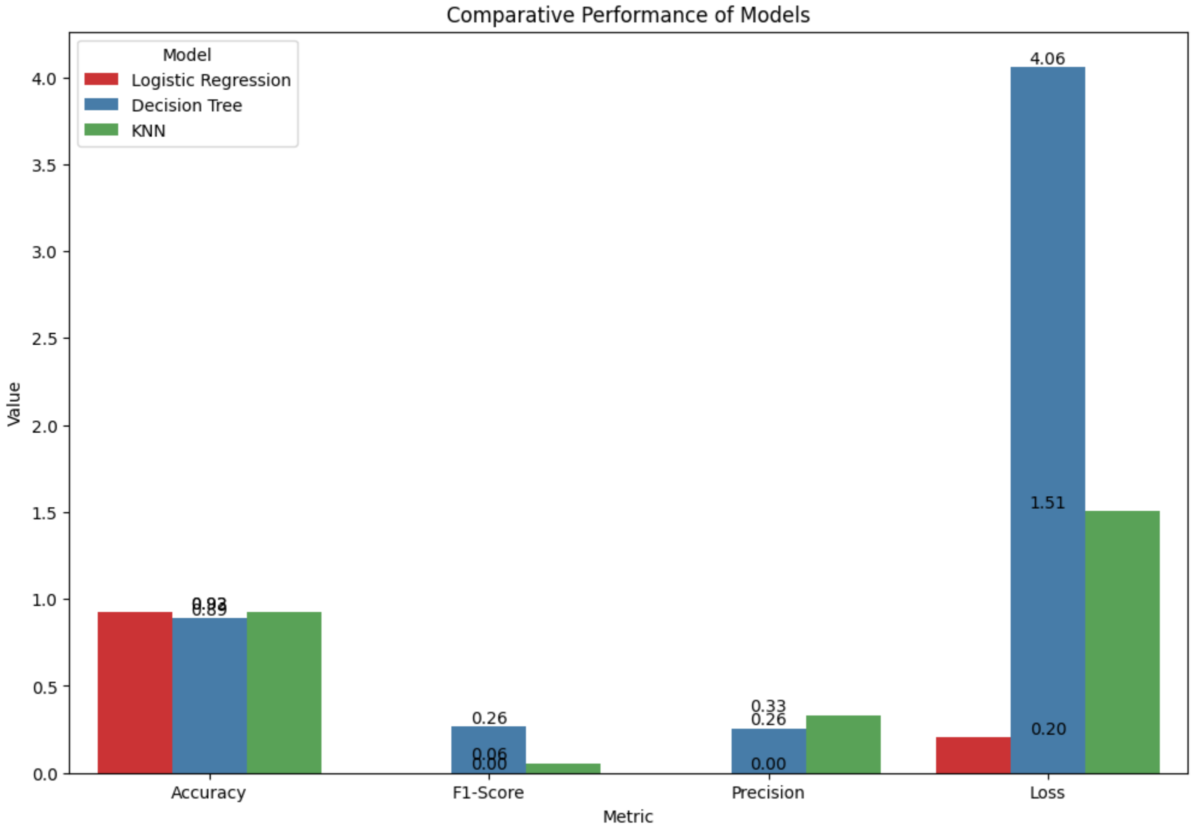
4.3 Applying DECISION TREE:

A decision tree is a way to make decisions by breaking a problem down into smaller, more manageable parts. It looks like a tree with branches, where each branch represents a choice leading to a decision.

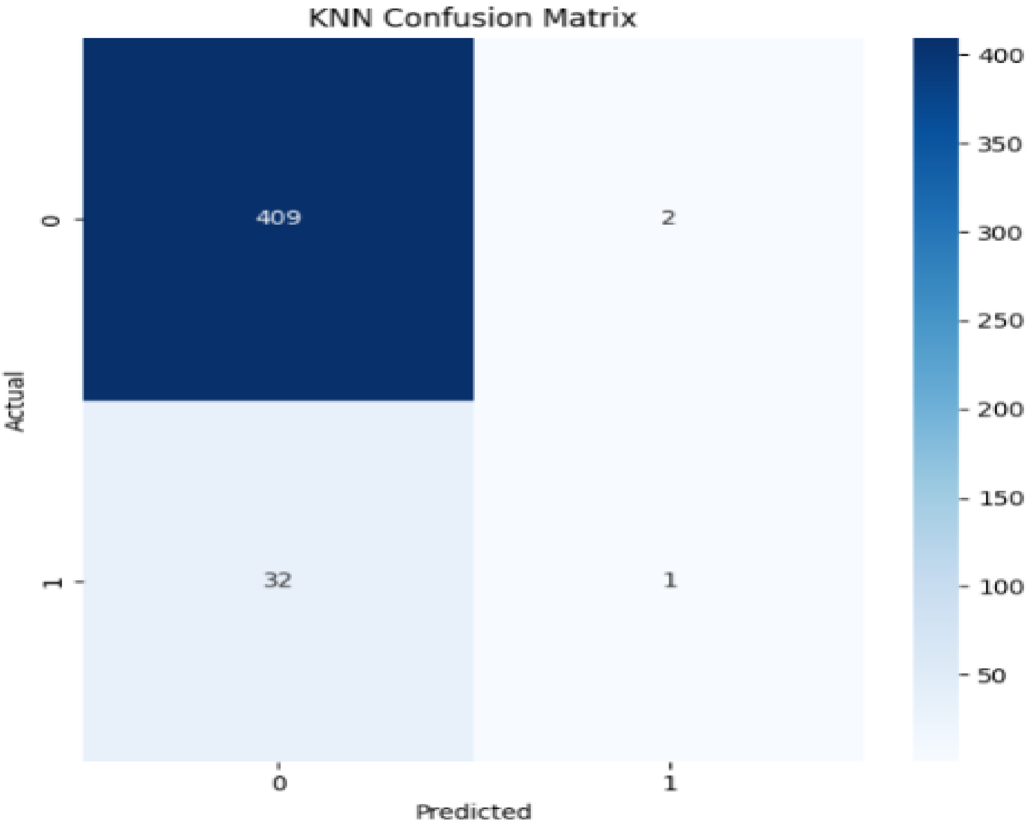


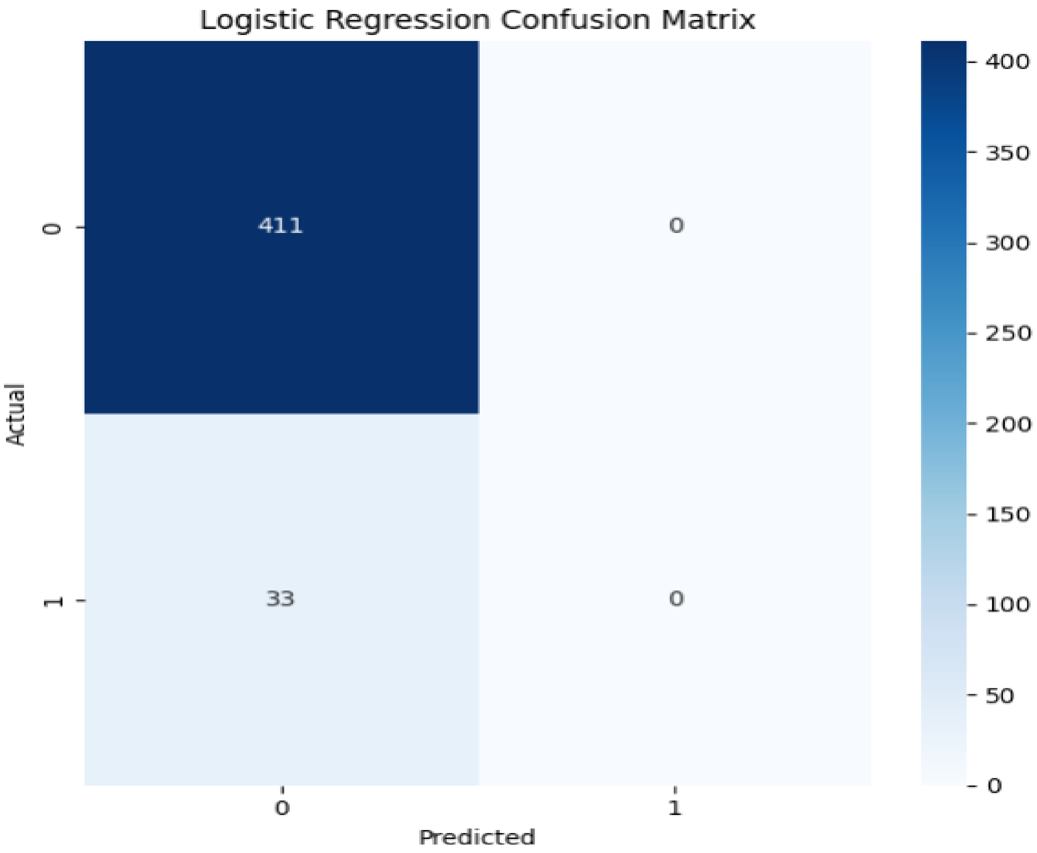
```
Decision Tree Accuracy: 0.8873873873873874
Decision Tree F1-Score: 0.2647058823529411
Decision Tree Precision: 0.2571428571428571
Decision Tree Loss: 4.058969976251932
```

4.4 COMPARATIVE PERFORMANCE OF MODELS:

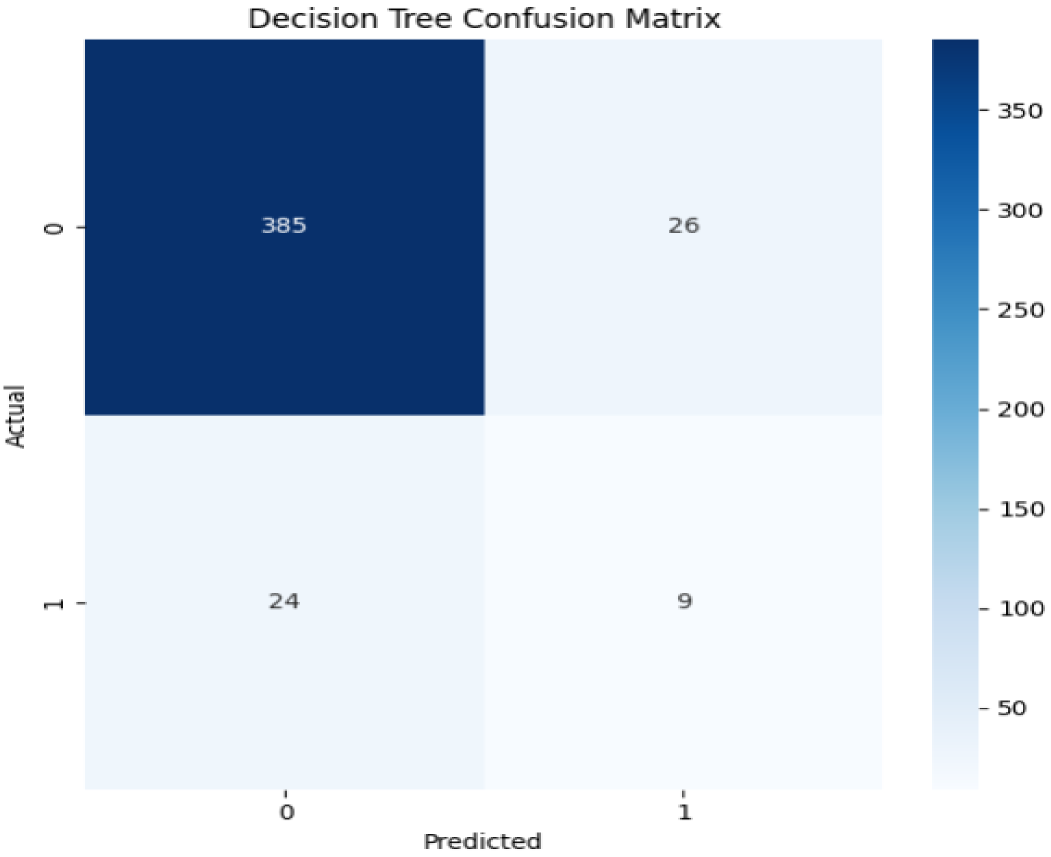


4.5 CONFUSION MATRIX:





Logistic Regression Confusion Matrix



Decision Tree Confusion Matrix

5. REFERENCES

1. Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: why HR are set to fail the big data challenge. Human Resource Management Journal, 26(1), 1-11.

2. Ben-Gal, H. C. (2018). An ROI-based review of HR analytics: practical implementation tools. Personnel Review.

3. Bharti, A. (2017). Human resource analytics. South Asian Journal of Marketing & Management Research, 7(5), 68-77.

4. Fitz-enz, J., & Mattox, J. R. (2014). Predictive analytics for human resources. Hoboken, NJ: Wiley.

5. Griffin, R. W., & Moorhead, G. (2011). Organizational behavior. Cengage Learning.

6. King, K. G. (2016). Data Analytics in Human Resources: A Case Study and Critical Review. Human Resource Development Review, 15(4).

7. Reddy, P., & Lakshmikeerthi, B. (2017). Big data in HR.

8. chat.openai.com