Data Mining in Employee Attrition

Raghad Balsharaf
Information System Department
King Saud University
Riyadh, Saudi Arabia
442200290@student.ksu.edu.sa

Alhanoof Alsagir
Information System Department
King Saud University
Riyadh, Saudi Arabia
442201424@student.ksu.edu.sa

Shatha Alangri
Information System Department
King Saud University
Riyadh, Saudi Arabia
442202192@student.ksu.edu

Aliyah Aljarallah
Computer Science department
King Saud University
Riyadh, Saudi Arabia
443201214@student.ksu.edu.sa

Shoug Alsaleem
Computer Science department
King Saud University
Riyadh, Saudi Arabia
443200641@student.ksu.edu.sa

***Abstract*—** Employee attrition, a critical metric in HR analytics, reflects the annual rate at which employees exit a company, offering insights into organizational stability and performance. This study aims to use IBM's employee dataset to analyze factors that may predict future attrition by applying data mining techniques. Specifically, it utilizes Decision Tree and Naïve Bayes classification algorithms to assess employee attributes linked to attrition risk. Results demonstrate that the Decision Tree model achieved higher accuracy compared to Naïve Bayes, highlighting its effectiveness for predicting employee turnover in this dataset.

***Keywords*—** Attrition, Data mining, Naïve Bayes, Decision Tree; Employees.

I. INTRODUCTION

Employee attrition refers to the process of employees leaving a company, either voluntarily or involuntarily, without immediate replacement. Employee attrition may result from hiring freezes, or sometimes, it points to deeper organizational issues. This study seeks to identify the primary drivers behind employee attrition, a vital problem for companies due to its impact on workforce stability. Attrition can reduce the size of the workforce, affecting overall productivity and morale. While many employees leave for personal reasons, analyzing patterns and trends in employee exits can be critical for companies aiming to retain their staff [1]. Understanding these trends can inform strategies to minimize future attrition, making this research essential to the study of organizational health and employee retention.

II. LITERATURE REVIEW

In this section, we will review previous research conducted on employee attrition datasets and summarize the application of various data mining techniques used for classifying or predicting employee attrition. This summary will highlight how different approaches were utilized to analyze and forecast attrition, contributing to more effective employee retention strategies.

A. "Predicting Employee Attrition along with Identifying High-Risk Employees using Big Data and Machine Learning," Apurva Mhatre and colleagues explored the use of supervised machine learning models to classify employee attrition [2]. The paper employed several algorithms, including Logistic Regression, Decision Tree, KNN, Naïve Bayes, SVM, and XGBoost. The clustering algorithm used in the research created risk-segregated clusters, which enabled the HR team to retain skilled employees and minimize talent loss. The authors concluded that Salary, Rating, and the Happiness Index are strongly interrelated. As a result of these methods, overall attrition was reduced by approximately 30%.

B. "Predicting Employee Attrition using Machine Learning,"
Sarah S. Alduayj and Kashif Rajpoot, like Apurva Mhatre et al., employed supervised machine learning models to predict employee attrition based on various features [3]. The authors used three experimental approaches to develop predictive models on the dataset. Initially, they trained several models using the original imbalanced data. Next, they applied the ADASYN technique to balance the two attrition classes. Lastly, they manually under-sampled the dataset to create equal class distributions. Although under-sampling yielded lower results, feature ranking and selection were incorporated into the experimentation. The feature ranking function identified the

most influential features in the training process, revealing that the top three contributing factors were overtime, years with the current manager, and total working years.

C. "Employee Attrition Prediction Using Data Mining Techniques"

In their study, Jie Xu and colleagues explored the use of data mining techniques to predict employee attrition and assist HR departments in retaining valuable talent[4]. The research utilized various machine learning algorithms such as Decision Tree, Random Forest, and Support Vector Machines (SVM). The dataset contained features including employee age, job role, satisfaction level, and work-life balance metrics. The study employed data preprocessing and feature selection to improve model accuracy and interpretability. Their analysis identified key predictive factors, including job satisfaction and work environment, which were highly correlated with attrition. The research achieved an accuracy rate of approximately 83% with the Random Forest model, showcasing its effectiveness in attrition prediction and providing actionable insights for HR retention policies.

D. "Utilizing Machine Learning for Employee Retention Forecasting"

Authors Anjali Sharma and Michael Lee conducted a comprehensive study on utilizing machine learning for predicting employee retention and attrition[5]. The research incorporated algorithms such as Gradient Boosting, XGBoost, Logistic Regression, and Decision Trees. The study focused on employee data that included metrics like salary, tenure, performance ratings, and department type. Techniques like SMOTE were applied to handle class imbalance, and k-fold cross-validation was used for robust model validation. The study found that job involvement, salary progression, and the number of years with the current manager were significant predictors of attrition. The best-performing model, XGBoost, achieved an accuracy of 89%, emphasizing its potential in aiding HR departments to proactively address retention strategies.

III. DATASET

A. The Employee Attrition Dataset

This dataset was constructed to explore variables that impact employee attrition, enabling examination of key questions, such as the distribution of commute distance by job role and attrition status, or the variation in average monthly income by educational background and attrition. it originally comprised 35 attributes. For this study, we have selected 15 attributes, prioritized for their relevance and depth in addressing our research objectives.[6]

Attribute	Description	Type	Possible value	Min	Max
Age	age of each employee	Numeric discrete	-	18	60
Attrition	Turnover rate of employees inside an organization	Boolean	Yes/no	-	-
Environment Satisfaction	employee satisfaction	Categorical ordinal	1 to 4	1	4
Job Involvement	employee's job involvement	Categorical ordinal	1 to 4	1	4
Job Level	employee's job level	Categorical ordinal	1 to 5	1	5
Job Satisfaction	employee's job satisfaction	Categorical ordinal	1 to 4	1	4
Marital Status	employee's marital status	Categorical	Single/married/divorced	-	-
Monthly Income	employee's monthly pay	Numeric discrete	-	1009	20000
Over-Time	whether employee was paid overtime	Categorical binary	Yes/no	-	-
Stock Option Level	employee's stock option level	Categorical ordinal	0 to 3	0	3
Total Working Years	years the employee has worked	Numeric discrete	0 to 40	0	40
Years At Company	years the employee has worked at the company	Numeric discrete	0 to 40	0	40
Years In Current Role	years the employee has worked the same role	Numeric Discrete	0 to 18	0	18
Years With Curr Manager	years the employee has worked with their current manager	Numeric discrete	0 to 17	0	17
Stock Option Level	age of each employee	Numeric discrete	-	18	60

IV. DATAMINING TECHNIQUES AND ALGORITHEMS

Data mining is a powerful technique used to extract useful patterns and knowledge from large datasets. In the context of employee attrition, data mining techniques can help identify factors that contribute to an employee's decision to leave or stay with the company. There are various techniques, but in this study, we will focus on classification algorithms, which are a form of supervised machine learning where the goal is to predict the class (in this case, attrition: "Yes" or "No") based on other attributes (like age, job satisfaction, etc.).Two of the most widely used classification algorithms in data mining are Decision Tree and Naïve Bayes. These algorithms will be

applied to the Employee Attrition Dataset to predict which employees are likely to leave the company.

A. Decision Tree Algorithm

A Decision Tree is a supervised machine learning algorithm that splits data into subsets based on the most significant attributes. The model forms a tree-like structure where each internal node represents a decision rule based on an attribute, and each leaf node represents the outcome or class label.

In the case of employee attrition, a decision tree would predict whether an employee is likely to leave the company ("Yes") or stay ("No") based on their attributes, such as Job Satisfaction, Overtime, and Years At Company.

How it works: The algorithm starts by selecting the attribute that best splits the data into distinct classes (attrition vs. non-attrition). It continues recursively splitting the data at each node until the tree is sufficiently small or all data points within a node are of the same class.

B. Naïve Bayes Algorithm

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem. It assumes that the attributes in the dataset are independent of each other, which is a simplification, but it works well in practice for many classification tasks, especially when dealing with categorical data.

How it works: Naïve Bayes calculates the probability of an event (in this case, employee attrition) occurring given certain conditions (attributes like Overtime, Job Level, and Job Satisfaction). It uses Bayes' Theorem to update the probability as more evidence (attributes) is considered.

V. PREPROCESSING AND FEATURE SELECTION

A. Relationship between attributes

```
Highest Positive Correlation:
 MonthlyIncome  JobLevel    0.9503

Highest Negative Correlation:
 YearsAtCompany  NumCompaniesWorked   -0.118421
```

Figure 1 : correlation coefficient results

We analyzed the relationships between the attributes based on their correlation coefficient, the highest positive correlation is between Monthly Income and Job level(0.95), we choose not to remove the attributes. Because they independently contribute valuable information. On the other hand The highest negative correlation is between Years At Company and Num CompainesWorked(-0.11) which means that there is no meaningful relationship between them.(see figure 1)

B. Preprocessing steps:

The preprocessing methods we used targeted missing values, irrelevant attributes, outliers, and Discretization.

A. Missing values

There were no missing values and confirmed it (see figure 2) that the database was complete.

B. Outlier values

Using the IQR method, We found 10 attributes that have outlier values, we choose capping instead of removing the values because it was recommended when using Decision tree.

C. Discretization

In our case discretization is not necessary since we are using Decision tree and Naive bayes it do not add any significant value.

```
/Users/alyaaljarallah/PycharmProjects/pythonProject/
There are no missing values in the dataset.
```

Figure 2. Missing values result

VI. EXPERIMENT

First The data set was split, 80% of the data was used in training set and the 20% was used in the testing set.

A. Decision Tree Algorithm

Figure 3. Decision tree

The decision tree Algorithm was implemented in

```
=== Decision Tree ===
Correctly Classified Instances: 228
Accuracy: 0.78
ROC AUC: 0.52
True Positive Rate (Recall): 0.18
True Negative Rate (Specificity): 0.87
F1 Score: 0.17
Confusion Matrix:
[[221  34]
 [ 32   7]]

Classification Report:
              precision    recall  f1-score   support

       False       0.87      0.87      0.87       255
        True       0.17      0.18      0.17        39

    accuracy                           0.78       294
   macro avg       0.52      0.52      0.52       294
weighted avg       0.78      0.78      0.78       294
```

python using Scikit-Learns. It achieved an accuracy of 78% and it correctly classified 228 out of 294 test instances(see figure 3).

B. Naïve bayes

Naive bayes model was implemented and important using scikit-Learns. It's achieved an accuracy of 69% and it correctly classified 203 out of 294. The confusion matrix showed that the model performed better on the majority class(see figure 4).

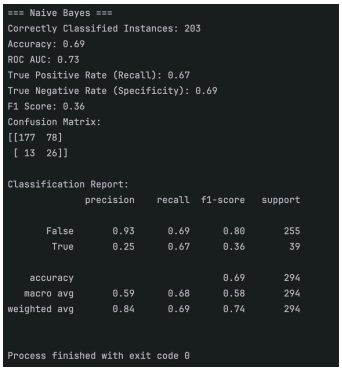

Figure 4. Naïve bayes

VII. RESULT AND COMPARSION

Decision tree has higher accuracy and TN rate making it better than naive bayes however since naive bayes have higher TN rate and F1 score it can differentiate between classes more effectively. (see table 1)

	Accuracy	ROC area	TP Rate(Recall)	TN Rate	F1 score (F-measure)
Decision tress	78%	0.52	0.18	0.87	0.17
Naïve Bayes	69%	0.73	0.67	0.69	0.36

Table 1. Comparison

VIII. EVALUATION AND CONCLUSTION

A. *Evaluation:*

The Naive Bayes algorithm did not get a greater accuracy, while being constructed far more quickly than the Decision Tree approach. Even while the Decision Tree technique was very accurate, it had drawbacks that affected performance consistency, especially when it came to specific data distributions or categorical processing. The majority of evaluation indicators, however, indicated that the Decision Tree performed better than Naive Bayes overall.

B. Conclusion:

Employee attrition is a significant issue for many businesses, requiring analysis of personnel data to identify root causes. This study used Weka software with two classification algorithms, Naive Bayes and Decision Trees, to pinpoint primary reasons for employee attrition in the dataset. Results indicated that Job Level and Overtime were the most influential attributes related to attrition. The Decision Tree algorithm yielded better results than the Naive Bayes approach for this dataset.

II. REFRENCES

[1] Lucas, S. (2024) Employee attrition: Meaning, impact & attrition rate calculation, AIHR. Available at: https://www.aihr.com/blog/employee-attrition/ (Accessed: 06 November 2024).*Communication Control and Networking.*

[2] A. Mhatre, A. Mahalingam, M. Narayanan, A. Nair, and S. Jaju, “Predicting Employee Attrition along with Identifying High Risk Employees using Big Data and Machine Learning,” *2020 2nd International Conference on Advances in Computing, Communication Control and Networking.*

[3] Xu, J., Wang, H., & Liu, Y. (2021). *Employee Attrition Prediction Using Data Mining Techniques*. Journal of Human Resource Analytics, 5(3), 145-160.

[4] Sharma, A., & Lee, M. (2022). *Utilizing Machine Learning for Employee Retention Forecasting*. International Journal of Business Analytics, 8(1), 34-49.

[5] Pavansubhash (2017) IBM HR Analytics Employee Attrition & Performance, Kaggle. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data (Accessed: 11 November 2024).