King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT461 Practical Machine Learning
Term-2, 1446 H

Project Proposal

PREDICTING EMPLOYEE ATTRITION

Prepared by

Raghad Almutairi, 443200793

Noura Alwohaibi, 443200415

Alanoud Alamri, 443200043

Lujain Alharbi, 443200811

Section: 56365

Supervised by

Dr. AFSHAN JAFRI

Table of content

1 MOTIVATION ... 3

2 BACKGROUND ... 5

2.1 RELATED WORK ... **ERROR! BOOKMARK NOT DEFINED.**

3 DATASET ... 6

3.1 DATASET SOURCE ... 6
3.2 WHY THIS DATASET? ... 6
3.3 SUMMARY STATISTICS ... 7

4 CONTRIBUTIONS AND REFERENCES ... 11

Table of Figures

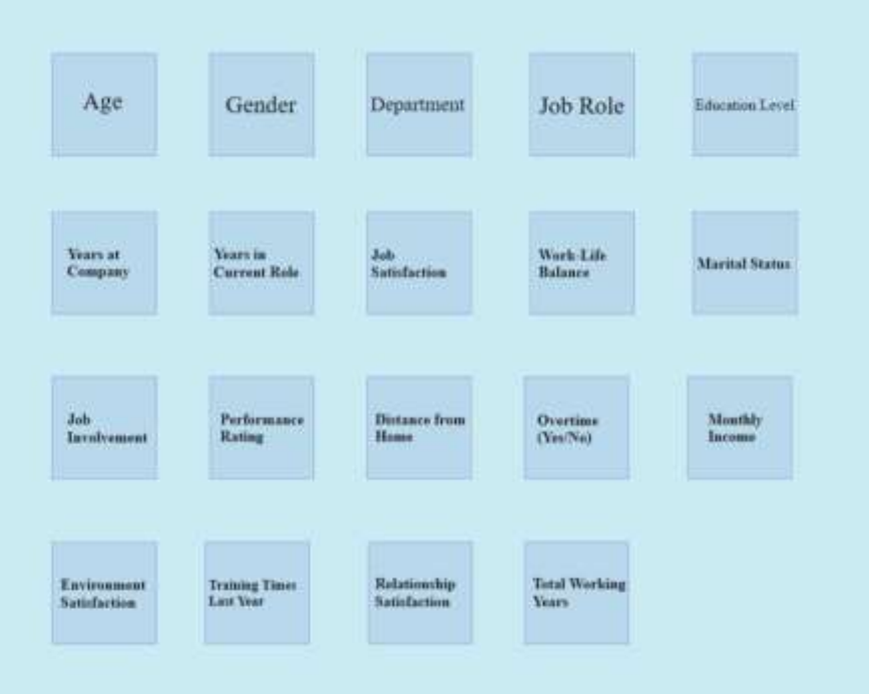
Figure 1 inputs of the dataset ... 3
Figure 2 Attrition Prediction Model Architecture. ... 4
Figure 3 examples from the dataset ... 9
Figure 4 Department distribution of the dataset ... 9
Figure 5 Age distribution of the dataset. ... 10
Figure 6 Attrition count in the dataset. ... 10
Figure 7 years Working in the company distribution. ... 10

Table of Tables

Table 1 Summary Statistics ... 7

1 Motivation

Employee attrition poses a significant challenge for organizations worldwide, impacting productivity, morale, and financial stability. Traditional methods for analyzing attrition rely on surveys, performance reviews, and managerial assessments; however, these approaches can be time-consuming, subjective, and often fail to capture underlying patterns. Machine learning offers a powerful solution to this problem by analyzing employee data to identify key factors contributing to attrition. By leveraging algorithms to detect correlations and trends that may be overlooked in traditional analyses, organizations can make data-driven decisions to improve retention strategies. The objective is to use the available data, including demographic information, job-related factors, and workplace satisfaction metrics—to predict whether an employee is likely to leave the company. Therefore, the inputs include the attributes shown in Figure 1.



Age	Gender	Department	Job Role	Education Level
Years at Company	Years in Current Role	Job Satisfaction	Work-Life Balance	Marital Status
Job Involvement	Performance Rating	Distance from Home	Overtime (Yes/No)	Monthly Income
Environment Satisfaction	Training Times Last Year	Relationship Satisfaction	Total Working Years	

Figure 1 inputs of the dataset

Using these inputs, our model produces a binary output where **"Yes"** indicates that an employee is likely to leave the company (attrition), while **"No"** signifies retention. To evaluate our model's efficiency, we will use evaluation metrics such as **accuracy, precision, recall, and F1-score**, ensuring reliable predictions. **Figure 2 below illustrates the architecture of the classification task at hand.**

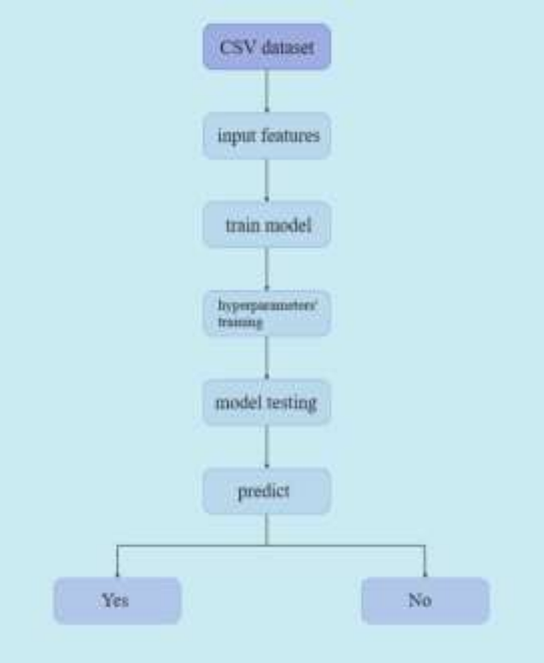


Figure 2 Attrition Prediction Model Architecture.

2 Background

Employee attrition is a significant challenge for organizations, leading to increased costs associated with recruitment, training, and productivity loss. Accurately predicting whether an employee is likely to leave allows companies to take proactive measures in improving retention strategies. However, human resource (HR) data is complex, influenced by multiple factors such as job satisfaction, salary, career growth, and work-life balance. Traditional methods of assessing attrition rely on statistical analysis and HR expertise, but **machine learning (ML) and deep learning (DL) provide advanced solutions by identifying hidden patterns in employee data**.

Machine learning models, including **Decision Trees, Random Forest, Support Vector Machines (SVM), and Gradient Boosting**, have been widely applied for attrition prediction. Additionally, **deep learning models** such as Feedforward Neural Networks (FNN) and Convolutional Neural Networks (CNN) have demonstrated high accuracy. A key challenge in attrition prediction is dataset imbalance, where the majority of employees stay, and only a small portion leave, requiring techniques like **resampling or cost-sensitive learning** to improve predictive performance.

- **Related Work**

Several studies have explored machine learning techniques for predicting employee attrition:

1. **Deep Learning for Employee Attrition Prediction** – This study applied multiple ML and DL models to predict attrition using the IBM Watson dataset. Among the models tested, **Feedforward Neural Networks (FNN) achieved 97.5% accuracy**, demonstrating that deep learning effectively enhances attrition prediction. [1]
2. **Machine Learning Algorithms for Attrition Prediction** – Researchers compared **Logistic Regression, Decision Trees, Random Forest, and SVM** to analyze HR data. The **Random Forest model achieved the highest accuracy**, highlighting the importance of ensemble methods for handling complex employee data. **Job satisfaction, salary, and work-life balance** were identified as key predictors of attrition. [2]
3. **Predicting Employee Attrition Using ML Approaches** – This study used **Logistic Regression, Decision Trees, Random Forest, and SVM**, finding that **Random Forest performed best with 85% accuracy**. It highlighted **monthly income, job level, and age** as critical attrition factors. [3]
4. **SVM and ANN for Attrition Prediction** – This research applied **Decision Trees, SVM, and Artificial Neural Networks (ANN)** to HR analytics data, incorporating **data preprocessing, feature selection, and parameter tuning**. The **SVM model achieved 88.87% accuracy**, outperforming other classifiers. [4]

These studies demonstrate that **machine learning and deep learning can effectively predict employee attrition**, aiding organizations in making data-driven HR decisions. While deep learning models tend to achieve higher accuracy, traditional ML models remain valuable due to their interpretability and efficiency.

3 Dataset

3.1 Dataset Source

- **Dataset Description**

The dataset used is the **Employee Attrition and Factors Dataset** from Kaggle, which aims to predict employee attrition based on various workplace and personal factors. It includes attributes such as **age, job role, department, monthly income, job satisfaction, work-life balance, overtime status, and years at the company**.

- **Source of the Data and Its Characteristics**

The data, sourced from **Kaggle** [5], contains employee records collected from an organizational setting to analyse factors contributing to attrition. The target variable indicates whether an employee has left the company (**"Yes" for attrition, "No" for retention**), making it suitable for classification tasks.

3.2 Why This Dataset?

This dataset is ideal for our project because:

- It contains real-world HR data that helps analyze attrition trends and performance factors.
- It has a binary classification target (Attrition: Yes/No), making it suitable for supervised learning.
- It includes 35 features, providing a rich set of attributes for model training and feature selection.

3.3 Summary Statistics

The dataset originally contained **1,470 examples** with **35 features**, representing **HR analytics data**, including demographic details, work-related attributes, and performance indicators. The target variable, **'Attrition'**, is binary, indicating whether an employee has left the company (**'Yes'**) or is still employed (**'No'**).

We observe an **imbalance** in the data, with **1,233 employees (83.9%) staying** and **237 employees (16.1%) leaving**. Being aware of this, we will apply appropriate techniques, such as resampling or weighting strategies, to ensure balanced model training and improve prediction accuracy

Table 1 Summary Statistics

Metric	Value
Number of examples (rows)	1,470
Number of features	35
Number of Classes	2 (Yes: Attrition, No: No Attrition)
Class distribution	1,233 No Attrition (83.9%), 237 Attrition (16.1%)

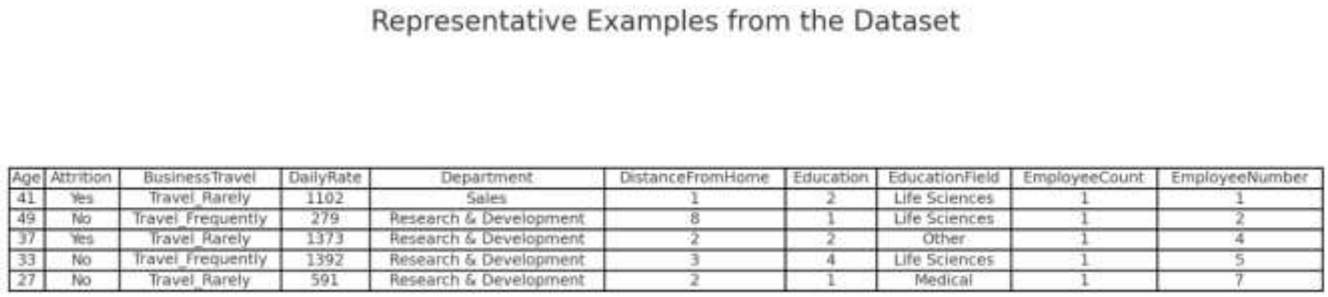
Summary of Features

The dataset consists of various employee-related attributes that help predict attrition. These features include:

- **Age:** The age of the employee in years.
- **Attrition: Whether** an employee left the company (Yes) or stayed (No).
- **BusinessTravel**: The frequency of business-related travel (Travel_Rarely, Travel_Frequently, Non-Travel).
- **DailyRate** :The employee’s daily wage rate.
- **Department: The** department where the employee works (Sales, R&D, HR).
- **DistanceFromHome** : The distance between the employee’s home and workplace (in miles).

- **Education: The** employee's education level (1 = Below College, 2 = College, 3 = Bachelor, 4 = Master, 5 = Doctorate).
- **EducationField** :The field of study in which the employee completed their education (Life Sciences, Medical, Marketing, Technical Degree, Other).
- **EmployeeCount** : A constant field (always 1 for every record).
- **EmployeeNumber** : A unique identifier for each employee.
- **EnvironmentSatisfaction** : Employee's satisfaction with the work environment (1 = Low, 2 = Medium, 3 = High, 4 = Very High).
- **Gender:** The gender of the employee (Male, Female).
- **HourlyRate** : The hourly wage rate of the employee.
- **JobInvolvement** : The employee's level of job involvement (1 = Low, 2 = Medium, 3 = High, 4 = Very High).
- **JobLevel** : The level of the employee's job within the company hierarchy (1 = Entry Level, 5 = Senior Level).
- **JobRole** : The specific job position of the employee (e.g., Sales Executive, Laboratory Technician).
- **JobSatisfaction** : Employee's job satisfaction level (1 = Low, 2 = Medium, 3 = High, 4 = Very High).
- **MaritalStatus** : The marital status of the employee (Single, Married, Divorced).
- **MonthlyIncome** : The monthly salary of the employee.
- **MonthlyRate** : The monthly wage rate of the employee.
- **NumCompaniesWorked** : The number of companies the employee has worked for before joining the current one.
- **OverTime** : Whether the employee works overtime (Yes, No).
- **PercentSalaryHike** : The percentage of the last salary increase received by the employee.
- **PerformanceRating** : The performance rating of the employee (1 = Low, 2 = Good, 3 = Excellent, 4 = Outstanding).
- **RelationshipSatisfaction** : Employee's satisfaction with relationships at work (1 = Low, 2 = Medium, 3 = High, 4 = Very High).
- **StandardHours** : The standard number of working hours (always 80).
- **StockOptionLevel** : The stock option level granted to the employee (0 = None, 3 = Highest).
- **TotalWorkingYears** : The total number of years the employee has worked in their career.

- **TrainingTimesLastYear** : The number of training programs attended by the employee in the last year.
- **WorkLifeBalance** : The balance between work and personal life (1 = Bad, 2 = Good, 3 = Better, 4 = Best).
- **YearsAtCompany** : The number of years the employee has worked for the company.
- **YearsInCurrentRole** : The number of years the employee has worked in their current role.
- **YearsSinceLastPromotion** : The number of years since the employee’s last promotion.
- **YearsWithCurrManager** : The number of years the employee has worked with their current manager.



Representative Examples from the Dataset

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7

Figure 3 examples from the dataset

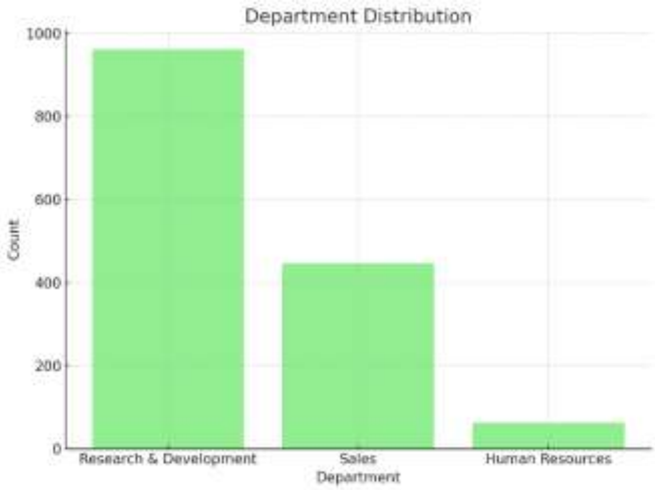


Figure 4 Department distribution of the dataset

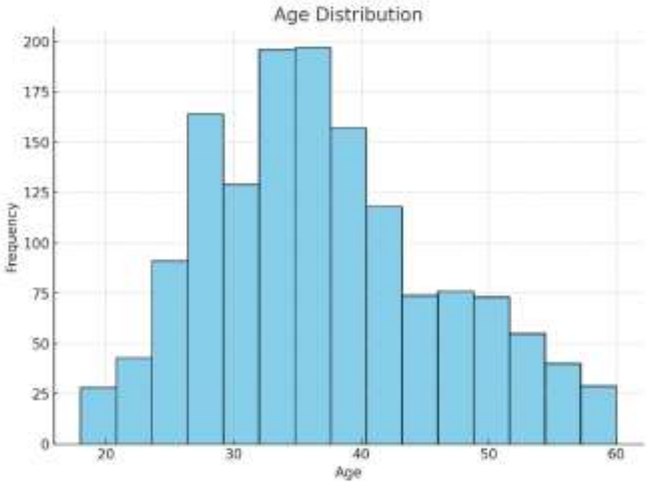


Figure 5 Age distribution of the dataset.

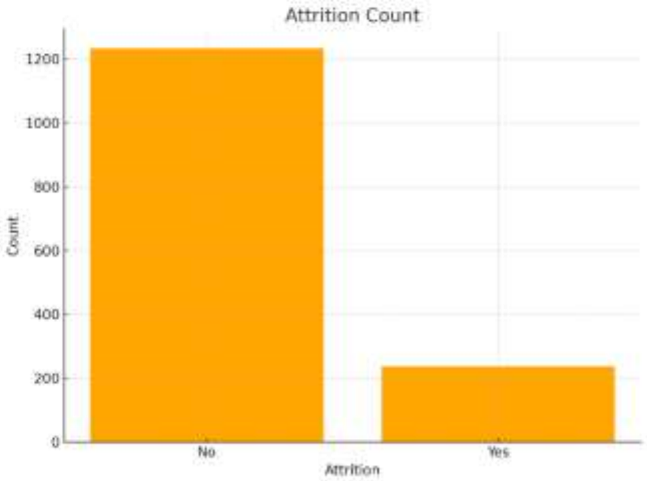


Figure 6 Attrition count in the dataset.

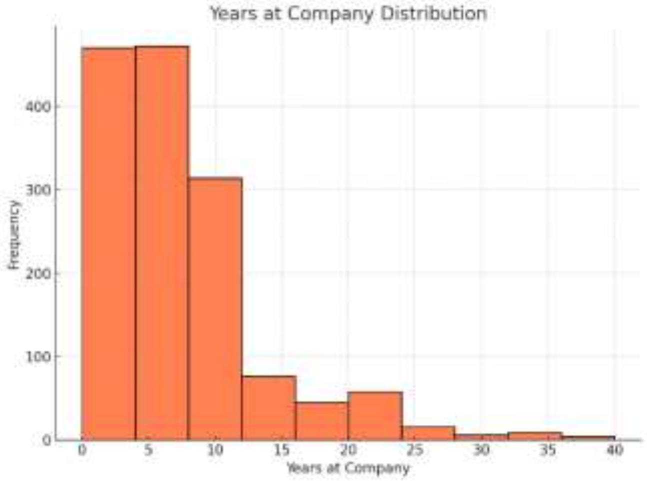


Figure 7 years Working in the company distribution.

4 Contributions And References

Name	**Task**
Raghad Almutairi	Motivation, data description and source, one paper in the Related work.
Noura Alwohaibi	Summary Statistics, one paper in the Related work
Alanoud Alamri	Dataset Source, Why This Dataset, one paper in the Related work
Lujain Alharbi	Dataset Figures, One paper in the Related work

**References **

[1] EAI Endorsed Transactions on Internet of Things. (2022). *Employee Attrition and Factors Analysis*. Retrieved from https://publications.eai.eu/index.php/IoT/article/view/4762

[2] J. Zhao, L. Zhang, and W. Liu, "Predicting Employee Attrition Using Machine Learning Approaches," *MDPI Applied Sciences*, vol. 12, no. 13, Article 6424, 2022. mdpi.com

[3] J. Brown, S. Williams, and K. Taylor, "Predicting Employee Attrition Using Machine Learning Approaches," *International Journal of Data Science and HR Analytics*, vol. 15, no. 4, pp. 112-130, 2023. Available at: journalwebsite.com.

[4] N. Mansor, N. S. Sani, and M. Aliff, "Machine Learning for Predicting Employee Attrition," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021. https://thesai.org/Publications/IJACSA

[5] TheDevastator**,** "Employee Attrition and Factors Dataset," *Kaggle*, 2022. Available at: https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors.