IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Employee Attrition Prediction Using Machine Learning And Simpy Simulation

Student Name: Fatima Abdullah Mohammed Aldosari
Student ID: 443300852

CS 516 – System Modeling and Simulation

Doctor Name: Ahmed M.Alkhadhr Almasabi

Abstract

This project focuses on employee attrition prediction using machine learning models and SimPy simulation.

Employee attrition prediction is important because it helps organizations identify employees who may leave the company and supports better Human Resources decision-making.

Three machine learning classification models were implemented and compared:

1. Logistic Regression
2. Decision Tree
3. Random Forest

The models were evaluated using Accuracy, Precision, Recall, F1 Score, and Confusion Matrix. Logistic Regression achieved the best overall performance based on F1 Score and Accuracy.

The second section of the project involved building a SimPy simulation that represents a real-world HR prediction system. The simulation demonstrated how the trained machine learning model can be integrated into an employee management environment.

Introduction

Employee attrition means employees leaving a company. Predicting employee attrition is important because it helps the Human Resources department identify employees who may leave and take early actions to reduce employee turnover.

In this project, machine learning models are used to predict whether an employee will stay or leave the company. The project also includes a simulation using SimPy to represent a real-world HR prediction system.

The project has two main sections:

1. Building machine learning models.
2. Performing simulation using SimPy.

Dataset Description

The dataset used in this project is the IBM HR Analytics Employee Attrition Dataset.

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

The dataset contains employee information such as age, monthly income, job role, overtime, years at company, job satisfaction, and work-life balance.

Item	Description
Dataset name	IBM HR Analytics Employee Attrition Dataset
Number of records	1,470
Number of features	35
Target variable	Attrition
Target meaning	0 = Employee stays, 1 = Employee leaves

Key Features

Feature	Description
Age	Employee age in years
JobRole	Employee job position
MonthlyIncome	Employee monthly salary
YearsAtCompany	Number of years at the company
JobSatisfaction	Job satisfaction rating
WorkLifeBalance	Work-life balance rating
OverTime	Whether the employee works overtime
TotalWorkingYears	Total years of work experience
Department	Employee department

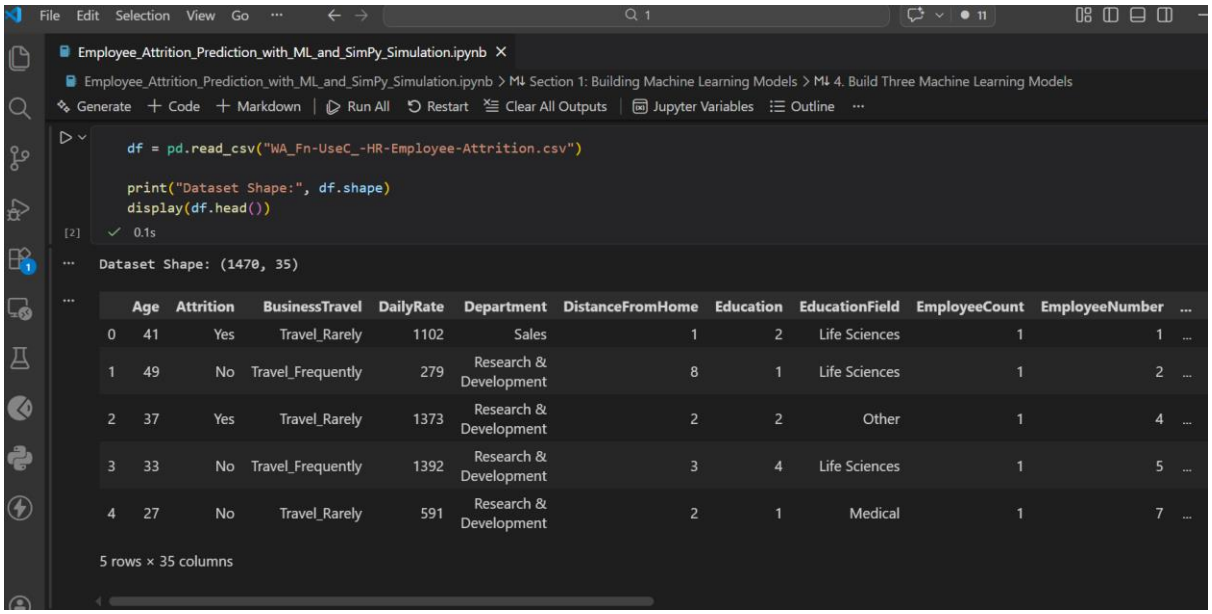
Data Overview

The dataset shape was:

1470 rows × 35 columns

The first rows of the dataset showed different employee information such as Age, Attrition, BusinessTravel, Department, MonthlyIncome, OverTime, JobRole, and YearsAtCompany.

First Five Rows of Dataset



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...

This figure shows the first five rows of the dataset. It helps to understand the structure of the data and the type of features used for employee attrition prediction.

Data Preprocessing

Data preprocessing was applied before building the machine learning models.

The preprocessing steps included:

1. Checking missing values.
2. Removing irrelevant features.
3. Encoding categorical variables.
4. Defining features and target.
5. Scaling the features.
6. Splitting the dataset into training and testing sets.

The removed columns were:

- EmployeeCount
- EmployeeNumber
- Over18
- StandardHours

These columns were removed because they do not provide useful information for prediction.

Missing Values

The dataset was checked for missing values.

The result showed that there were no missing values in the dataset.

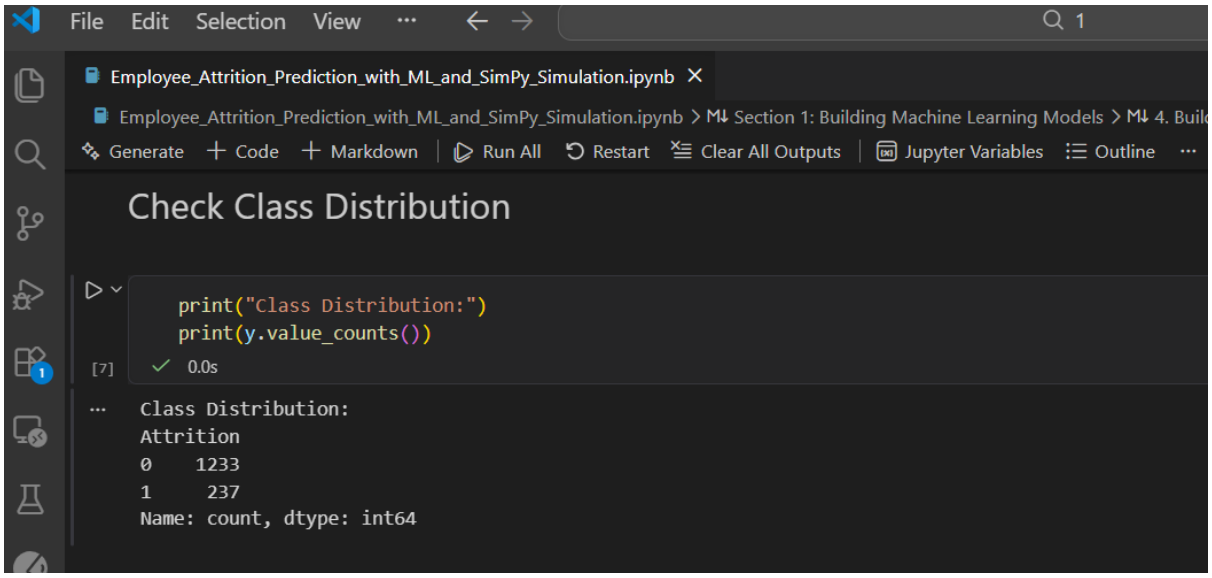
Class Distribution

The target variable is Attrition.

The class distribution was:

Class	Meaning	Count
0	Employee stays	1233
1	Employee leaves	237

Class Distribution



The figure shows that most employees stayed in the company, while fewer employees left. This means the dataset is imbalanced because the number of employees who stayed is much higher than the number of employees who left.

Machine Learning Models

Three machine learning models were built and compared:

1. Logistic Regression
2. Decision Tree
3. Random Forest

These models were selected because the project requirement asks for at least three machine learning models covered in the lab.

Logistic Regression Results

Logistic Regression achieved the following results:

Metric	Score
Accuracy	0.894558
Precision	0.700000
Recall	0.358974
F1 Score	0.474576

Confusion Matrix:

Actual / Predicted	Stay	Leave
Stay	249	6
Leave	25	14

Logistic Regression achieved the highest accuracy and the highest F1 Score among the three models. It correctly predicted many employees who stayed and also detected some employees who may leave.

Decision Tree Results

Decision Tree achieved the following results:

Metric	Score
Accuracy	0.799320
Precision	0.236842
Recall	0.230769
F1 Score	0.233766

Confusion Matrix:

Actual / Predicted	Stay	Leave
Stay	226	29
Leave	30	9

Decision Tree had the lowest performance compared with the other models. It made more incorrect predictions, especially when predicting employees who may leave.

Random Forest Results

Random Forest achieved the following results:

Metric	Score
Accuracy	0.880952
Precision	0.833333
Recall	0.128205
F1 Score	0.222222

Confusion Matrix:

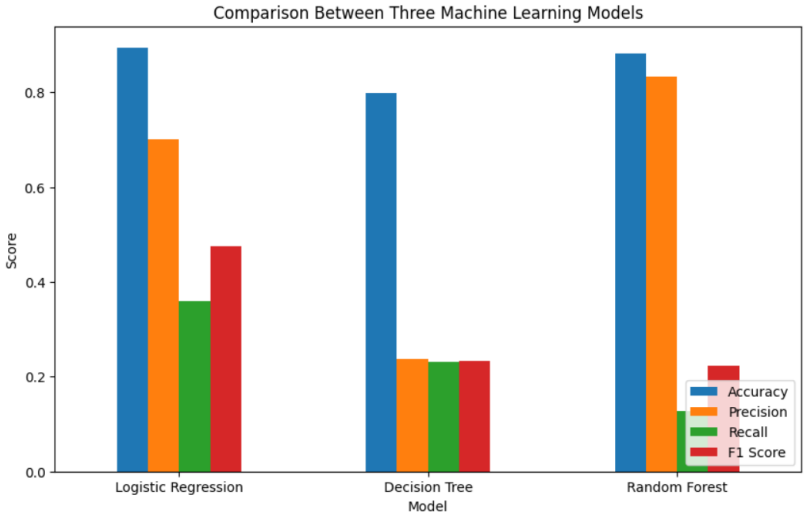
Actual / Predicted	Stay	Leave
Stay	254	1
Leave	34	5

Random Forest achieved high accuracy and high precision. However, its recall was low, meaning it did not detect many employees who actually left the company.

Model Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.894558	0.700000	0.358974	0.474576
Decision Tree	0.799320	0.236842	0.230769	0.233766
Random Forest	0.880952	0.833333	0.128205	0.222222

Model Comparison Plot



This figure compares the three models using Accuracy, Precision, Recall, and F1 Score. Logistic Regression performed best overall because it achieved the highest accuracy and the highest F1 Score.

Best Model Selection

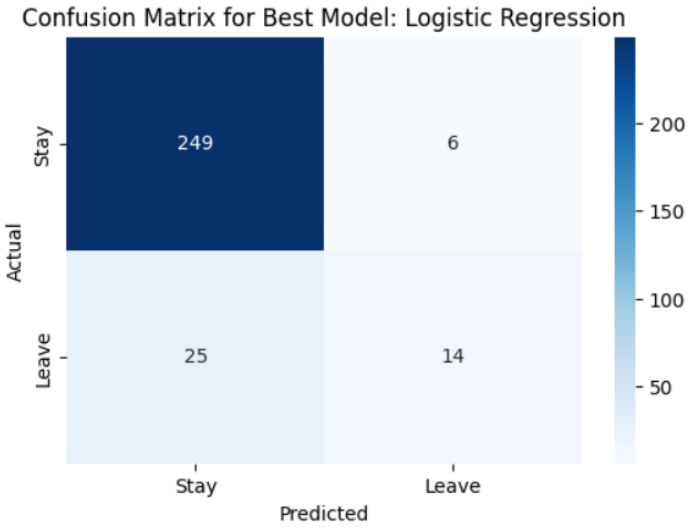
The best model selected was:

Logistic Regression

The model was selected based on F1 Score.

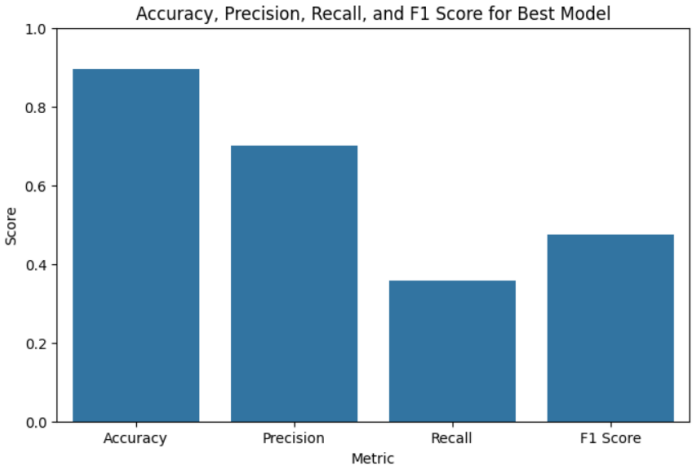
F1 Score is important because it balances Precision and Recall. Since the dataset is imbalanced, using F1 Score is better than depending only on Accuracy.

Confusion Matrix for Best Model



The confusion matrix shows that the Logistic Regression model correctly predicted 249 employees who stayed and 14 employees who left. It incorrectly predicted 6 employees as leaving when they stayed, and 25 employees as staying when they actually left.

Best Model Metrics Plot
Accuracy, Precision, Recall, and F1 Score for Best Model



This figure shows the performance metrics of the best model. Logistic Regression achieved high accuracy, good precision, and the best F1 Score compared with the other models. However, recall is still lower because the dataset is imbalanced.

SimPy Simulation

The second section of the project is the simulation part.

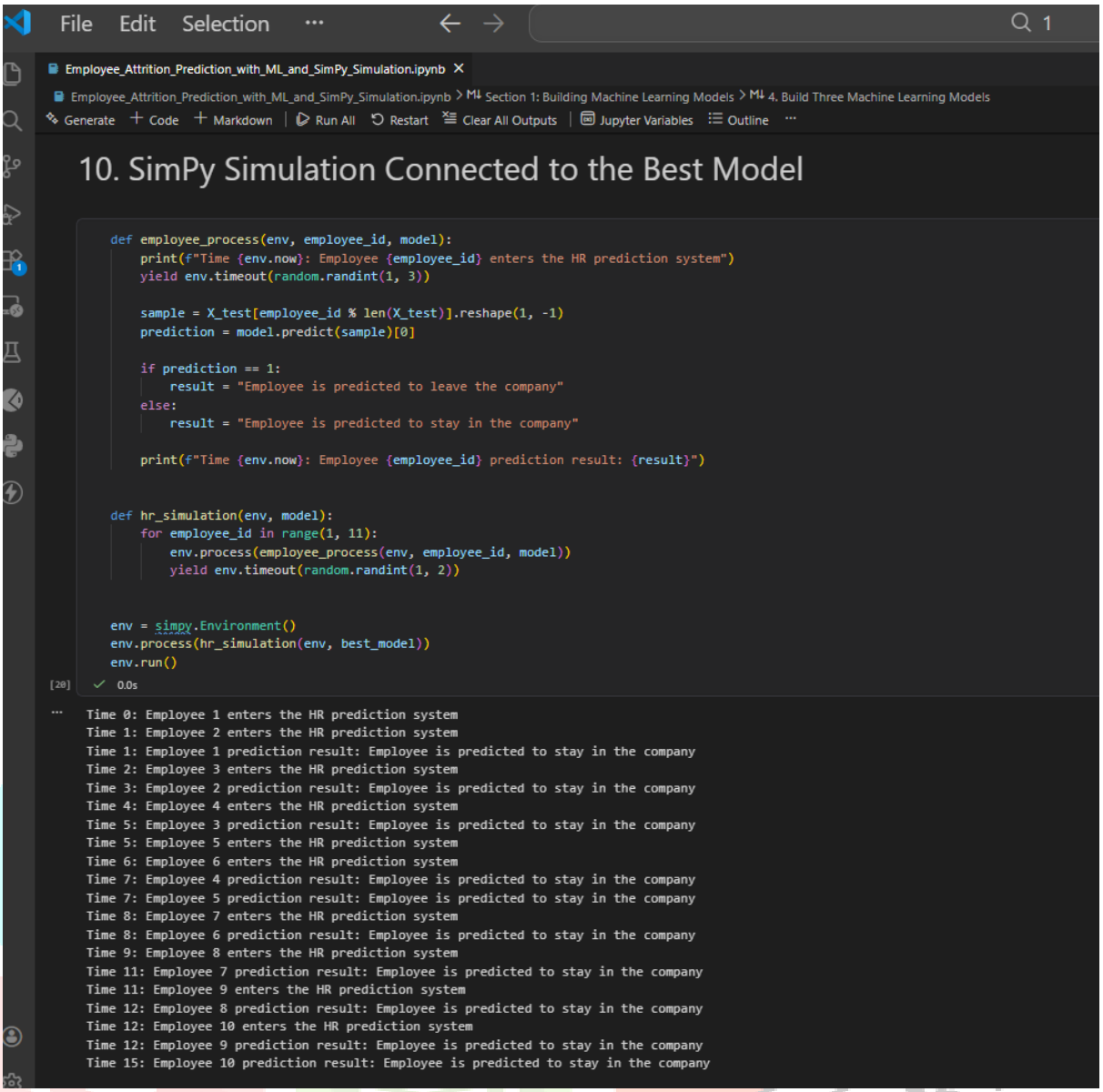
SimPy was used to simulate a real-world HR prediction system.

In the simulation:

1. Employees enter the HR prediction system.
2. Each employee waits for a short processing time.
3. The best machine learning model predicts whether the employee will stay or leave.
4. The prediction result is displayed.

The best model used in the simulation was: **Logistic Regression**

Simulation Output



The simulation output shows employees entering the HR prediction system at different times. The trained Logistic Regression model predicts whether each employee is expected to stay or leave the company. This connects the machine learning model to a real-world HR scenario.

Discussion

The project successfully built and compared three machine learning models for employee attrition prediction.

The models were evaluated using:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

Logistic Regression was selected as the best model because it achieved the highest F1 Score and the highest accuracy.

The SimPy simulation connected the best model to a real-world HR prediction system. This shows how the model can be used in practice to support decision-making in Human Resources.

Conclusion

This project successfully developed a machine learning system for employee attrition prediction using classification models and SimPy simulation.

Three machine learning models were trained and compared:

1. Logistic Regression
2. Decision Tree
3. Random Forest

The models were evaluated using Accuracy, Precision, Recall, F1 Score, and Confusion Matrix.

Logistic Regression achieved the best overall performance because it obtained the highest F1 Score and Accuracy. The model was able to predict employee attrition more effectively than the other models.

The project also included a SimPy simulation to represent a real-world HR prediction environment. The simulation demonstrated how machine learning models can support Human Resources departments in predicting employee attrition and improving decision-making.

Overall, the project successfully satisfied the project requirements by combining machine learning and simulation in a real-world scenario.