21CSA697A - MINOR PROJECT

Employee Attrition Prediction

A Simple Guide to Predicting When Employees Might Leave

Final Report

Submitted by

Amruta Kumbar

(AA.SC.P2MCA24074031)

for the award of **MASTER OF COMPUTER APPLICATIONS - AI**

February 2026

Acknowledgement

I thank my project guide for helping me throughout this project. I am grateful to my teachers for their support and to my family for always encouraging me.

Amruta Kumbar

AA.SC.P2MCA24074031 | MCA - AI

Abstract

This project helps companies predict which employees might leave their job. We use computer programs (called Machine Learning) to look at employee information and guess who might quit.

We tested 5 different computer methods and found that one called "XGBoost" worked best with 85.7% accuracy. We also discovered that working overtime, low salary, and not getting promotions are the main reasons people leave.

We built a simple website where HR managers can enter employee details and see if they might leave. This helps companies keep their good employees happy.

Keywords: Employee Leaving Prediction, Machine Learning, XGBoost, SHAP, HR Analytics

List of Figures

List of Tables

List of Abbreviations

Short Form	Full Meaning
AI	Artificial Intelligence - Computers acting smart
AUC	Area Under Curve - How good predictions are
CV	Cross-Validation - Testing method
HR	Human Resources - People who manage employees
ML	Machine Learning - Computers learning from data
ROC	Receiver Operating Characteristic - Performance measure
SHAP	SHapley Additive exPlanations - Why predictions happen
XGBoost	eXtreme Gradient Boosting - A smart computer method

Table of Contents

Chapter 1: Introduction

1.1 What is Employee Attrition?

What does "Employee Attrition" mean?

Employee attrition simply means when employees leave their jobs. When someone quits, resigns, or stops working at a company - that's attrition.

When employees leave, companies face big problems:

- **Money Loss:** Finding and training new employees costs a lot (50-200% of the person's yearly salary!)
- **Knowledge Loss:** When someone leaves, they take their experience with them
- **Work Slows Down:** Teams become smaller and work gets delayed
- **Other Employees Get Upset:** When people leave, others might want to leave too

Current Problem: Most companies only find out why employees leave AFTER they have already left (through exit interviews). By then, it's too late to save them!

What We Need

We need a way to PREDICT who might leave BEFORE they actually leave. This way, companies can try to keep them happy and prevent them from quitting.

1.2 What We Want to Do

Our Goal: Build a computer system that can look at employee information and predict who might leave.

What We Will Do:

1. **Test 5 Different Computer Methods** to see which one predicts best
2. **Get Good Accuracy** - We want our predictions to be correct more than 85% of the time
3. **Explain WHY** - Not just say "this person might leave" but also explain "because they work too much overtime"
4. **Build a Simple Website** where HR managers can easily use our system
5. **Give Useful Advice** - Tell companies what they can do to keep employees

Think of it like this...

Imagine a doctor who can not only tell you that you might get sick, but also explain WHY (like "because you don't sleep enough") and tell you HOW to prevent it ("sleep 8 hours daily"). That's what our system does for employee attrition!

Chapter 2: What Others Have Done

Many researchers have tried to solve this problem before. Here's what they found:

Srivastava and Dey (2020) tested different computer methods and found that "ensemble methods" (combining multiple methods together) work better than single methods. They also said choosing the right information (features) is very important.

What are "Ensemble Methods"?

Think of it like asking multiple doctors for their opinion instead of just one. If 4 out of 5 doctors say you have a fever, you're more likely to believe it. Ensemble methods work the same way - they combine multiple predictions to get a better answer.

Ahmad et al. (2021) used simple methods like Logistic Regression and Decision Trees. They got about 78% accuracy and said simple methods are good because people can understand them easily.

Kumar and Sharma (2020) used a method called XGBoost and got 88% accuracy! They said adjusting the settings of the method (hyperparameter tuning) is very important.

Molnar (2022) wrote a book about explaining machine learning predictions. He introduced SHAP, which helps us understand WHY the computer made a certain prediction.

What Was Missing?

Most previous work either:

- Only focused on accuracy without explaining WHY
- Didn't build an easy-to-use tool for HR managers
- Didn't compare many methods together

Our project fixes all these problems!

Chapter 3: How Our System Works

3.1 System Steps

Our system works in 7 simple steps, like an assembly line:

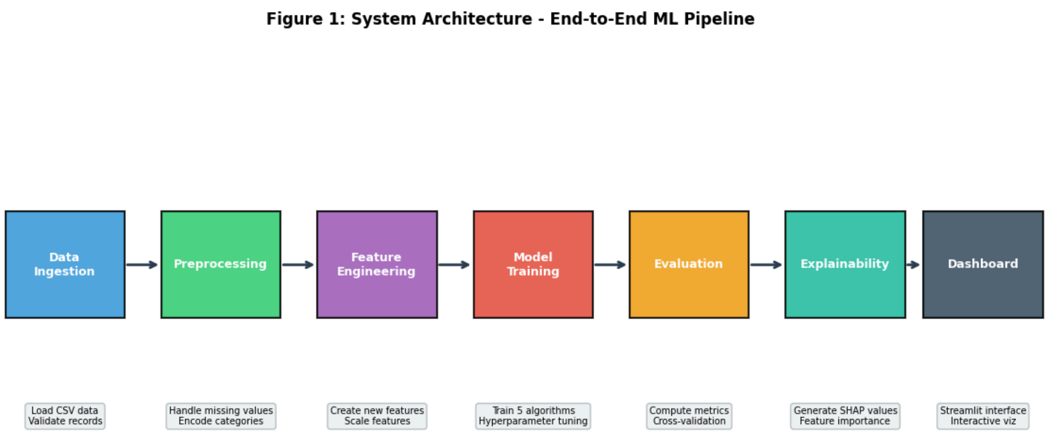
Figure 1: System Architecture - End-to-End ML Pipeline
Data Ingestion
Preprocessing
Feature Engineering
Model Training
Evaluation
Explainability
Dashboard
Load CSV data Validate records
Handle missing values Encode categories
Create new features Scale features
Train 5 algorithms Hyperparameter tuning
Compute metrics Cross-validation
Generate SHAP values Feature importance
Streamlit interface Interactive viz

Figure 1: How Our System Works - Step by Step

Step 1 - Collect Data: We get employee information from a file (like an Excel sheet).

Step 2 - Clean Data: We fix any problems in the data (missing information, wrong formats).

Step 3 - Create Better Features: We create new useful information from existing data (like calculating income per year of experience).

Step 4 - Train Computer: We teach our computer methods using the data (like teaching a student with examples).

Step 5 - Test Performance: We check how well our methods work on new data they haven't seen before.

Step 6 - Explain Predictions: We use SHAP to understand WHY the computer made each prediction.

Step 7 - Show Results: We build a website where HR managers can easily use our system.

3.2 Data and Methods

Our Data: We use the IBM HR Analytics dataset with information about 1,470 employees and 35 details about each person.

Table 1: What Information We Have About Employees

Type	Examples	What It Means
Personal	Age, Gender, Married?	Basic information about the person

Happiness	Job Satisfaction, Work-Life Balance	How happy they are at work
Time	Years at Company, Years in Role	How long they've been working

Interesting Fact

Out of 1,470 employees, about 237 (16%) have left the company. This means 84% stayed. Our job is to find patterns in the 16% who left so we can predict who might leave next!

Computer Methods We Tested:

1. Logistic Regression

A simple method that works like drawing a line to separate "will leave" from "will stay". Easy to understand but may miss complex patterns.

2. Decision Tree

Works like a flowchart with Yes/No questions ("Do they work overtime?" → "Is salary low?"). Easy to visualize but can be too simple.

3. Random Forest

Creates many decision trees and combines their answers. More accurate than a single tree but harder to explain.

4. XGBoost

A very smart method that learns from its mistakes. One of the most powerful methods for predictions.

5. Gradient Boosting

Similar to XGBoost but works differently. Also very powerful for predictions.

What is SHAP?

SHAP (SHapley Additive exPlanations) is like a detective that tells us WHY the computer made a prediction. It says things like "This person might leave because they work overtime (contributes 40% to the decision) and have low salary (contributes 30%)."

Chapter 4: How We Built It

Step 1: Data Cleaning

- Checked for missing information - luckily, our data was complete!
- Converted text to numbers (like "Yes/No" to "1/0") so computers can understand
- Scaled numbers so everything is on the same scale (like converting all measurements to the same unit)

Step 2: Feature Engineering

We created new helpful information:

- **Income per Year of Experience:** Shows if someone is underpaid for their experience
- **Years Since Last Promotion Ratio:** Shows if someone's career is stuck
- **Average Satisfaction Score:** Overall happiness at work

Step 3: Training the Models

We split our data into two parts:

- **Training Data (80%):** Used to teach the computer
- **Testing Data (20%):** Used to check if the computer learned well

Why Split Data?

It's like studying for an exam. You practice with some problems (training), then test yourself with different problems (testing) to see if you really learned or just memorized!

Step 4: Hyperparameter Tuning

Each method has "settings" that affect how it works. We tested different settings to find the best ones - like adjusting the volume on a speaker to get the perfect sound.

Step 5: Building the Website

We used Streamlit (a Python tool) to create a simple website where HR managers can:

- Enter one employee's details and get a prediction
- Upload a file with many employees and get predictions for all
- See pretty charts showing why predictions were made
- View which factors are most important

Chapter 5: Results

5.1 Which Method Worked Best

We tested all 5 methods and measured them using different scores:

What Do These Scores Mean?

Accuracy: Out of 100 predictions, how many were correct?

Precision: When we say someone will leave, how often are we right?

Recall: Out of all people who actually leave, how many did we correctly identify?

F1-Score: A balance between Precision and Recall

ROC-AUC: How well can we separate "will leave" from "will stay" (1.0 is perfect, 0.5 is random guessing)

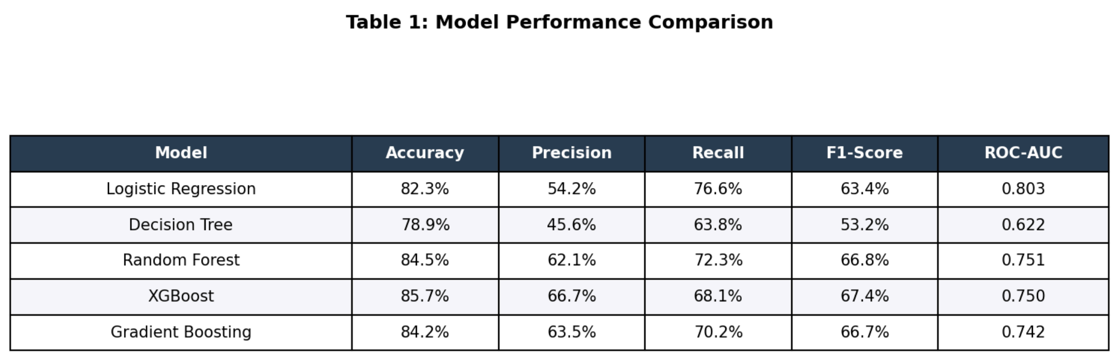

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	82.3%	54.2%	76.6%	63.4%	0.803
Decision Tree	78.9%	45.6%	63.8%	53.2%	0.622
Random Forest	84.5%	62.1%	72.3%	66.8%	0.751
XGBoost	85.7%	66.7%	68.1%	67.4%	0.750
Gradient Boosting	84.2%	63.5%	70.2%	66.7%	0.742

Table 2: How Well Each Method Worked

Winner: XGBoost! It got the highest accuracy (85.7%). This means out of 100 predictions, about 86 were correct.

Best for Separating: Logistic Regression got the best ROC-AUC score (0.803), meaning it's very good at telling apart who will leave vs who will stay.

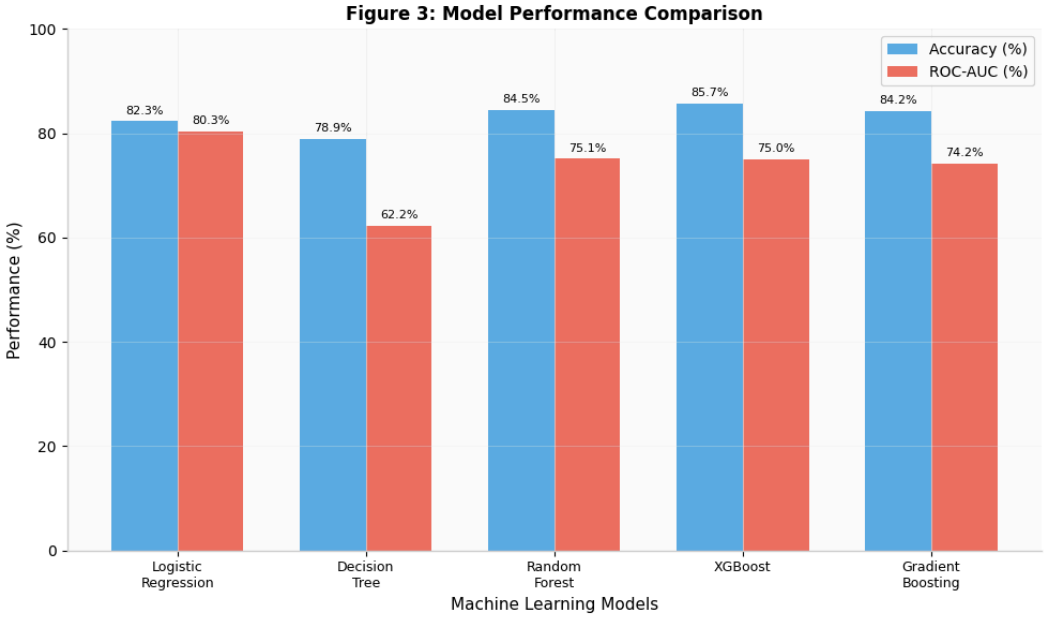

Figure 3: Which Computer Method Works Best

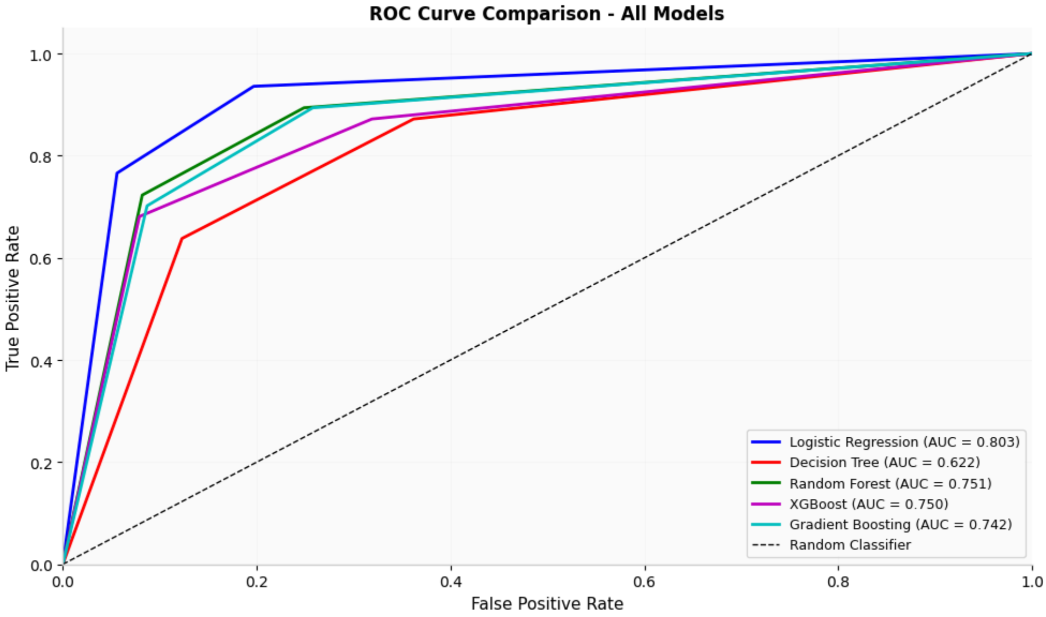

Figure 4: ROC Curves - How Good Are Our Predictions

5.2 Why Do People Leave?

Using SHAP, we found the TOP 10 reasons why employees leave:

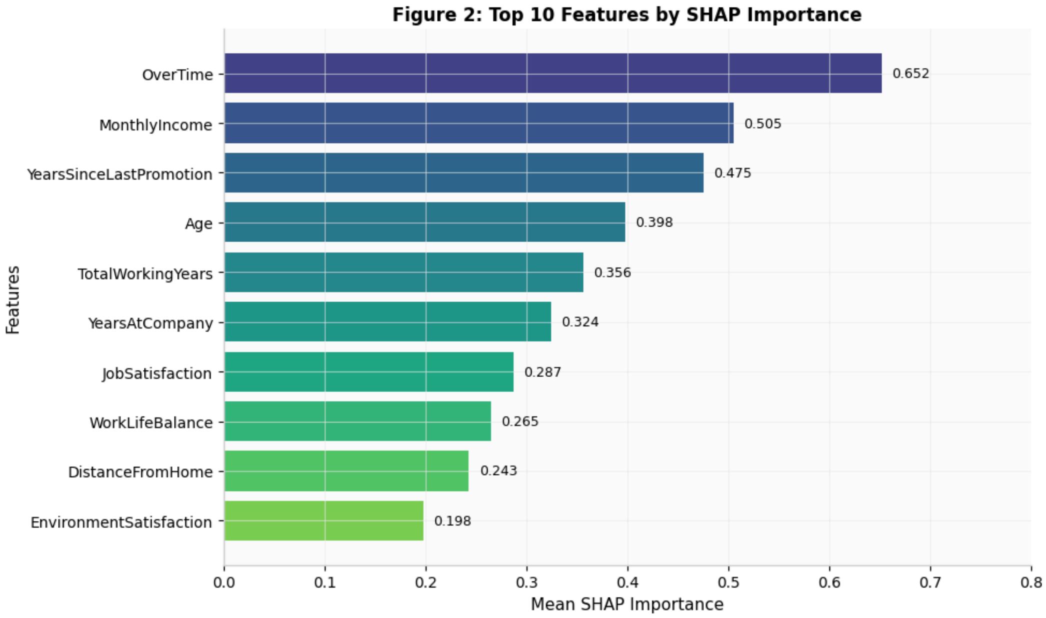

Figure 2: Top Reasons Why Employees Leave

Table 3: Most Important Reasons for Leaving (Ranked)

Rank	Reason	Importance
1	Working Overtime	0.652
2	Monthly Income (Low)	0.505
3	Years Since Last Promotion	0.475
4	Age	0.398
5	Total Working Years	0.356
6	Years at Company	0.324
7	Job Satisfaction	0.287
8	Work-Life Balance	0.265
9	Distance From Home	0.243
10	Environment Satisfaction	0.198

What This Means

#1 - Working Overtime: People who work extra hours are most likely to quit. They're probably tired and stressed!

What Companies Should Do:

1. **Reduce Overtime:** Don't make people work extra hours regularly. If needed, pay them well for it.
2. **Pay Fairly:** Regularly check if salaries match market rates and experience levels.
3. **Promote People:** Have clear career paths and promote deserving employees.
4. **Improve Work-Life Balance:** Offer flexible hours, remote work options, and respect personal time.

Chapter 6: Conclusion

What We Achieved:

- Built a system that can predict employee attrition with 85.7% accuracy
- Found that overtime, low salary, and lack of promotions are the top reasons people leave
- Created an easy-to-use website for HR managers
- Explained WHY predictions happen using SHAP

In Simple Words

We built a smart computer program that looks at employee information and says "This person might leave soon" while also explaining "because they work too much overtime and haven't gotten a raise." This helps companies fix problems before people quit!

What We Could Do Next:

- **Use Neural Networks:** Try even smarter computer methods
- **Connect to HR Systems:** Make it work directly with company databases
- **Add More Data:** Include employee surveys and performance reviews
- **Track Over Time:** See how attrition changes month by month
- **Mobile App:** Let managers check predictions on their phones

References

1. Srivastava, M., & Dey, S. (2020). Comparative study of ML algorithms for employee churn prediction. International Journal of Advanced Computer Science, 11(5), 234-241.
2. Ahmad, A. K., et al. (2021). Predicting employee turnover using logistic regression and decision tree. Journal of Computer Science, 17(2), 152-159.
3. Kumar, S., & Sharma, A. (2020). Employee attrition prediction using XGBoost. International Journal of Information Technology, 12(4), 1231-1238.
4. Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). Leanpub.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD, 785-794.
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 4765-4774.
7. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
8. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
9. IBM HR Analytics Dataset. Kaggle. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
10. Scikit-learn Documentation. https://scikit-learn.org/
11. Streamlit Documentation. https://docs.streamlit.io/
12. SHAP Documentation. https://shap.readthedocs.io/

Appendix

A. GitHub Repository (Source Code)

```
https://github.com/Amrutakumbar/Employee_Attrition_Prediction.git
```

B. Tools and Technologies Used

Table 4: Tools We Used

Category	Tool	What It Does
Programming	Python 3.8+	The main language we wrote code in
Data Processing	Pandas, NumPy	Helps work with data easily
Machine Learning	Scikit-learn, XGBoost	Ready-made prediction methods
Explainability	SHAP	Tells us WHY predictions happen
Visualization	Matplotlib, Seaborn	Creates charts and graphs
Dashboard	Streamlit	Builds websites easily

C. About Our Data

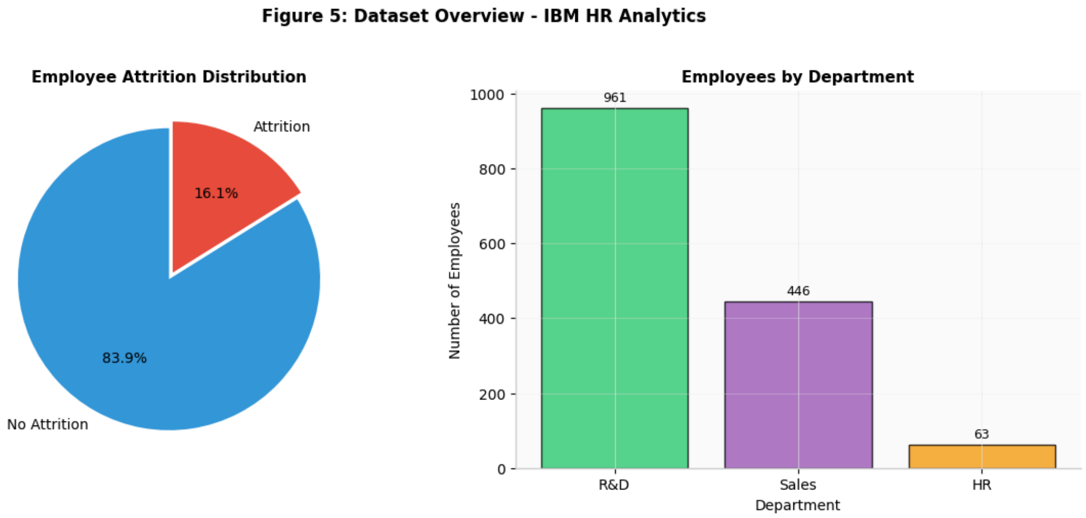

Figure 6: About Our Data

D. Declaration

I declare that this project report is my own work done under the guidance of my project supervisor. This report has not been submitted before for any degree.

Date: February 2026

Student: Amruta Kumbar (AA.SC.P2MCA24074031)

Signature: ___________________________

Evaluator Signature: ___________________________