Odisha University of Technology and Research
Bhubaneswar

# DATA ANALYSIS PROJECT

## Topic: Employee Attrition Prediction using Machine Learning

**PRESENTED BY:**
Adya Bramha Samantroy
24111061 , ECE(A3)

**PRESENTED TO:**
SIr Sidharth Dash

# Contents

# Introduction

Employee attrition has become a critical concern for modern organizations as it directly impacts productivity, project continuity, and recruitment costs. Understanding why employees leave and predicting who is likely to resign enables companies to adopt proactive retention strategies. This project explores the use of machine learning—specifically logistic regression—to analyze employee data and estimate the probability of attrition. By examining demographic, financial, and job-related variables, the system identifies key factors that influence employee turnover. Through statistical analysis, visualization, and predictive modeling, the project aims to convert raw employee information into actionable insights that can support HR decision-making and help reduce workforce loss.

# Objectives and Goals

The project's main overall objectives and goals.

### Goal # 1

To analyze employee data and identify major factors affecting attrition.

### Goal # 2

To apply statistical tests and logistic regression to predict whether an employee is likely to leave.

### Goal # 3

To provide a simple, interpretable prediction system that helps HR make better retention decisions.

# Motivation

The first step toward building a stronger, more stable workforce.

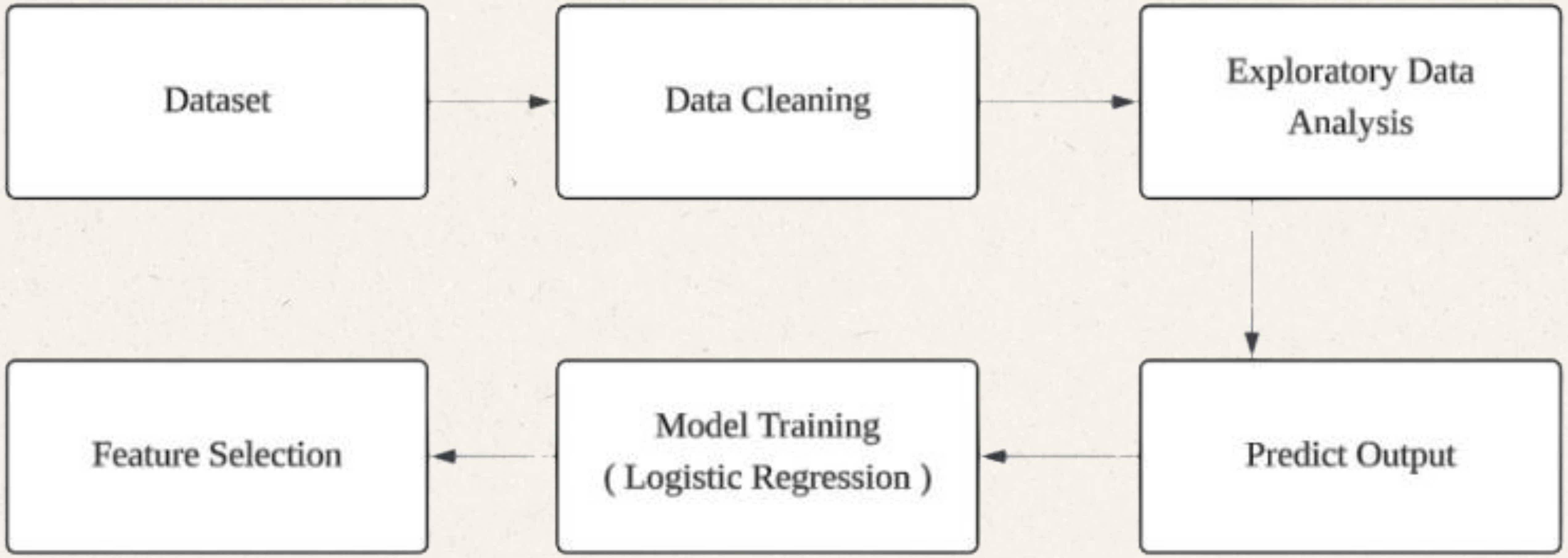
01 Employee attrition creates financial and productivity losses for organizations, making it important to understand the factors that drive employees to leave. Traditional methods often fail to capture deeper patterns behind turnover.

02 With the rise of data analytics, machine learning provides an efficient way to analyse employee behaviour and identify risk factors early. This motivates the need for a predictive, data-driven approach.

03 By building an attrition prediction model, organizations can take timely actions–such as improving satisfaction or adjusting workload–to retain valuable employees and strengthen their workforce.

# Description of The System

Workflow of the proposed system


Dataset
Data Cleaning
Exploratory Data Analysis
Predict Output
Model Training ( Logistic Regression )
Feature Selection

# Techniques Used

To Build the Attrition Model

**01** Data Preprocessing

- Cleaned missing values and duplicates.
- Removed irrelevant fields.
- Encoded categorical variables for model compatibility.

**02** Exploratory Data Analysis (EDA)

- Visualized data patterns and distributions.
- Computed correlations and key statistical tests.
- Identified major factors influencing attrition.

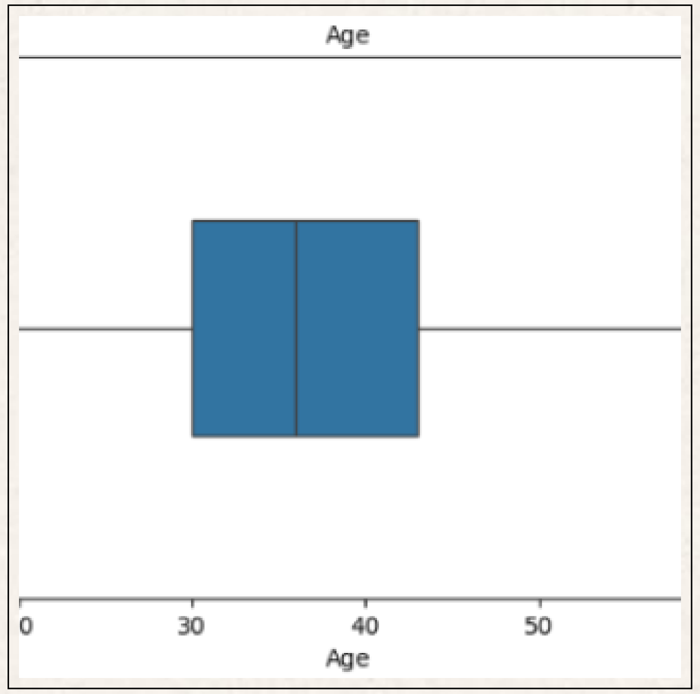
**03** Feature Selection

- Selected high-impact predictors based on correlation and significance.
- Final features included age, income, job satisfaction, overtime, and commute distance.

**04** Machine Learning Model

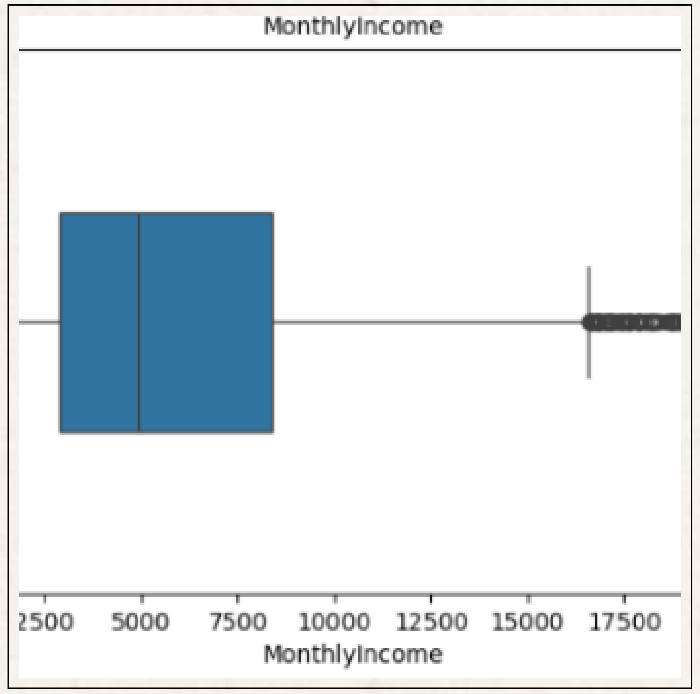
- Used Logistic Regression for prediction.
- Split dataset into train/test for performance validation.
- Evaluated using accuracy, precision, recall, and F1-score.

# Exploratory Data Analysis



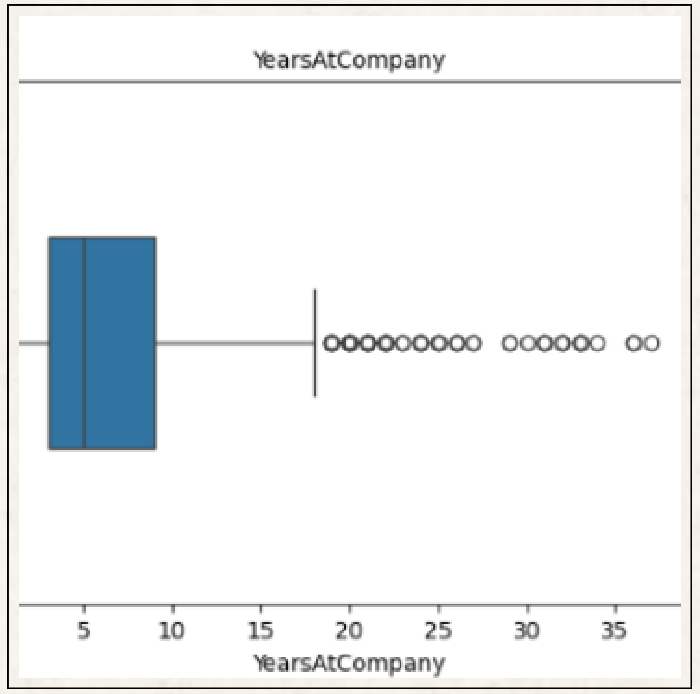
**BoxPlot for Age**

No Outliers



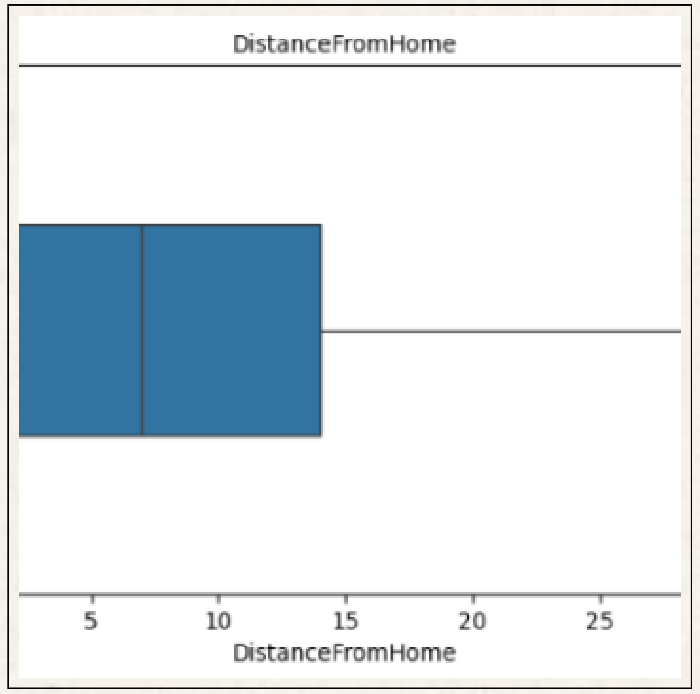
**BoxPlot for Monthly income**

Many Outliers



**BoxPlot for YearsAtCompany**

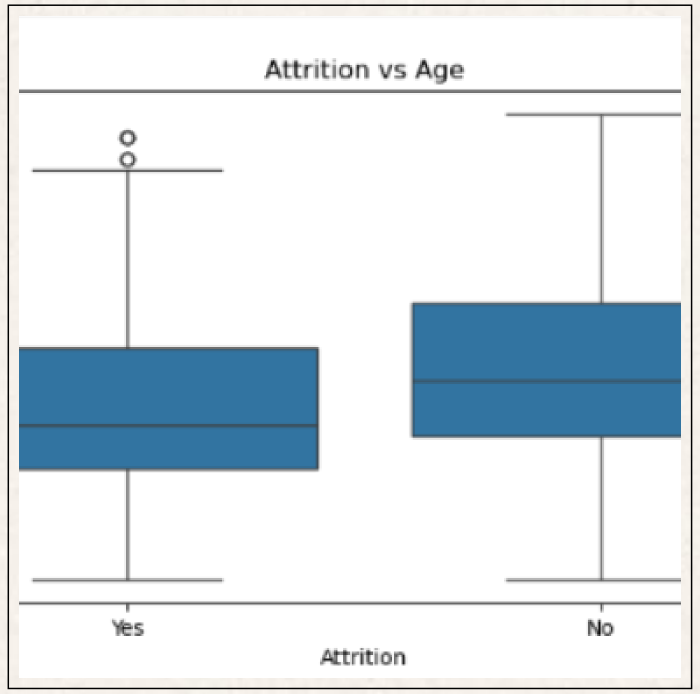
Some Outliers



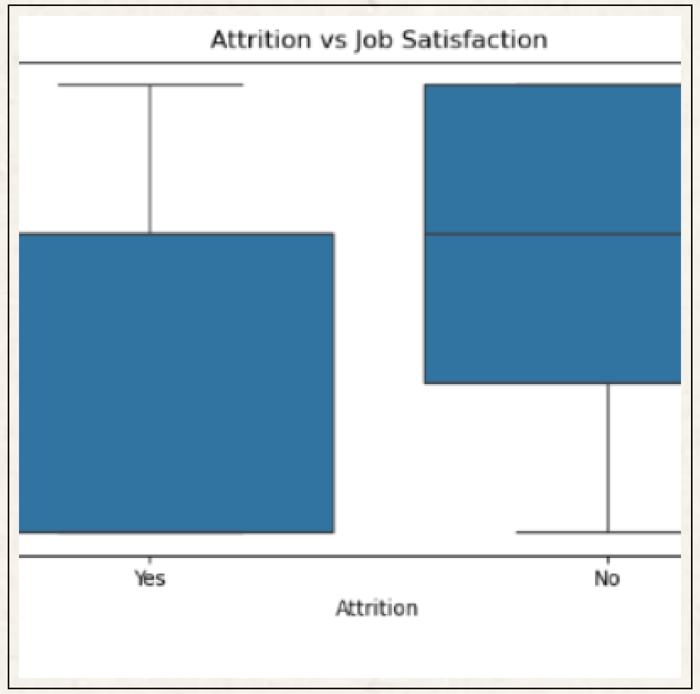
**BoxPlot for Distance**

No Outliers

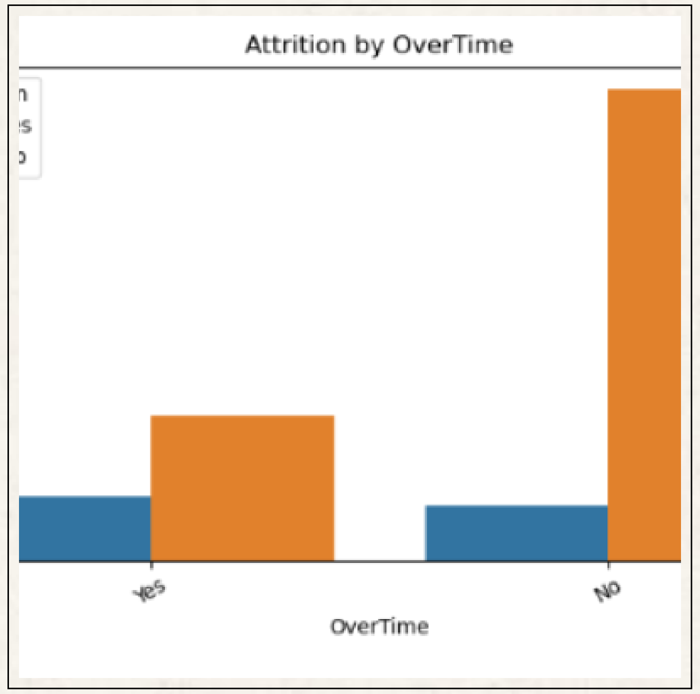
# Exploratory Data Analysis



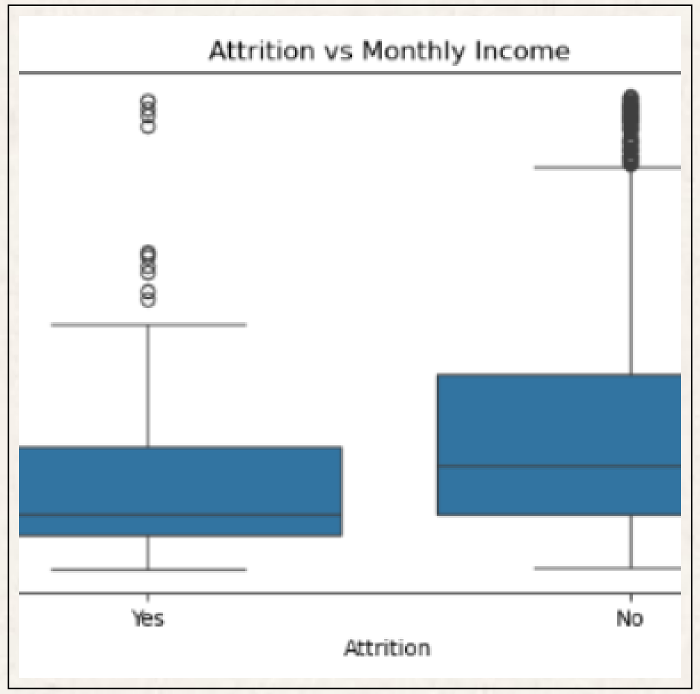
Attrition vs Age



Attrition vs Job Satisfaction



Attrition vs Overtime



Attrition vs Monthly Income

# Feature Selection

It’s the most important part for creating a perfect prediction model.

### Overtime

Employees who work overtime regularly show very high attrition rates due to stress and burnout.

### Job Satisfaction

Employees unhappy with their role, manager, or workload tend to leave earlier.

### Monthly Income

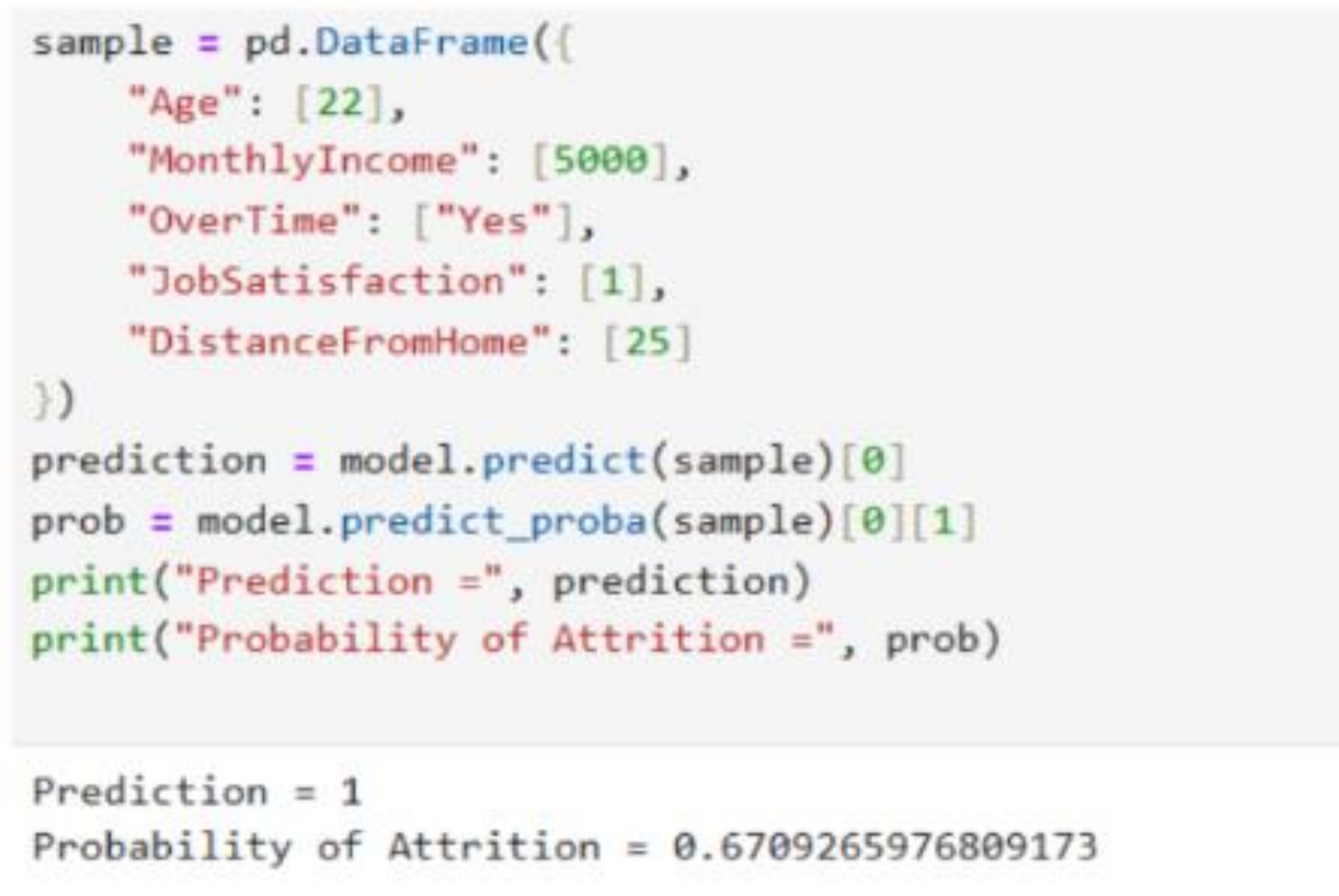
Employees with lower salaries relative to peers have a higher risk of attrition.

### Mental Stress

Employees experiencing sustained stress tend to show reduced engagement.

# Model Training & Prediction Output

The Logistic Regression model was trained using Age, Monthly Income, Job Satisfaction, OverTime, and Distance From Home. The dataset was split 80/20 for training and testing, with categorical features encoded via One-Hot Encoding. The model predicts both the attrition class (Yes/No) and the probability, helping HR identify high-risk employees for proactive retention measures.

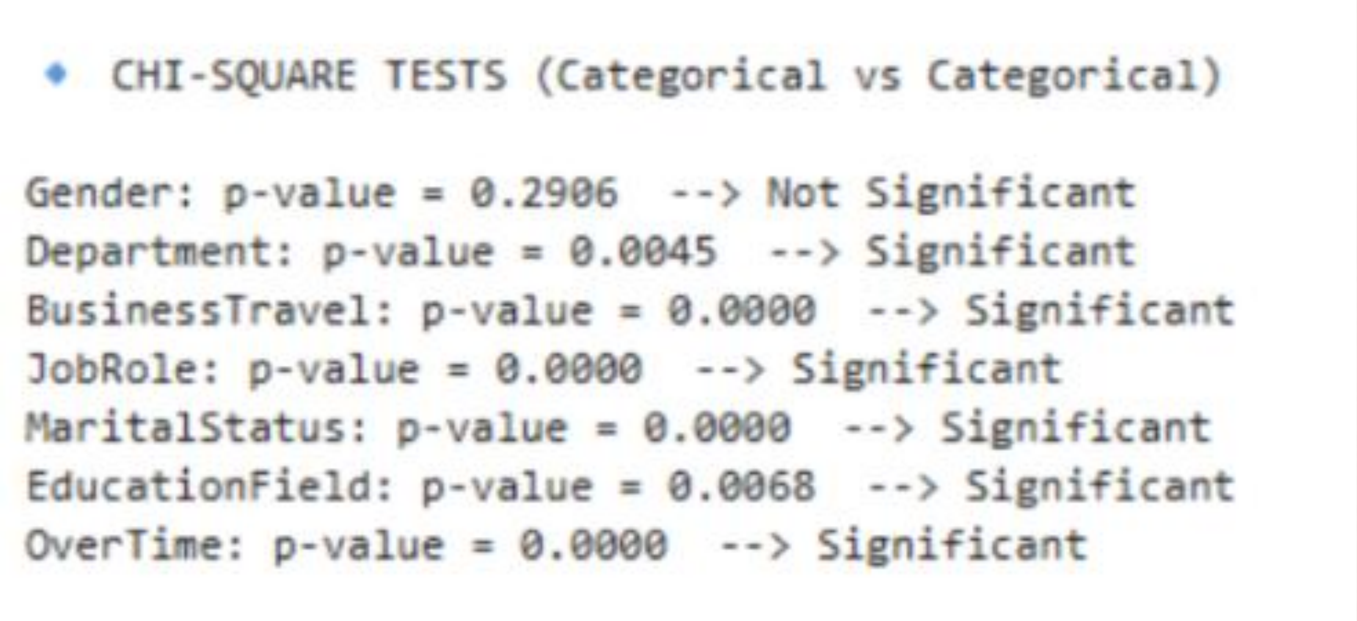


```
sample = pd.DataFrame({
    "Age": [22],
    "MonthlyIncome": [5000],
    "OverTime": ["Yes"],
    "JobSatisfaction": [1],
    "DistanceFromHome": [25]
})
prediction = model.predict(sample)[0]
prob = model.predict_proba(sample)[0][1]
print("Prediction =", prediction)
print("Probability of Attrition =", prob)
```

```
Prediction = 1
Probability of Attrition = 0.6709265976809173
```

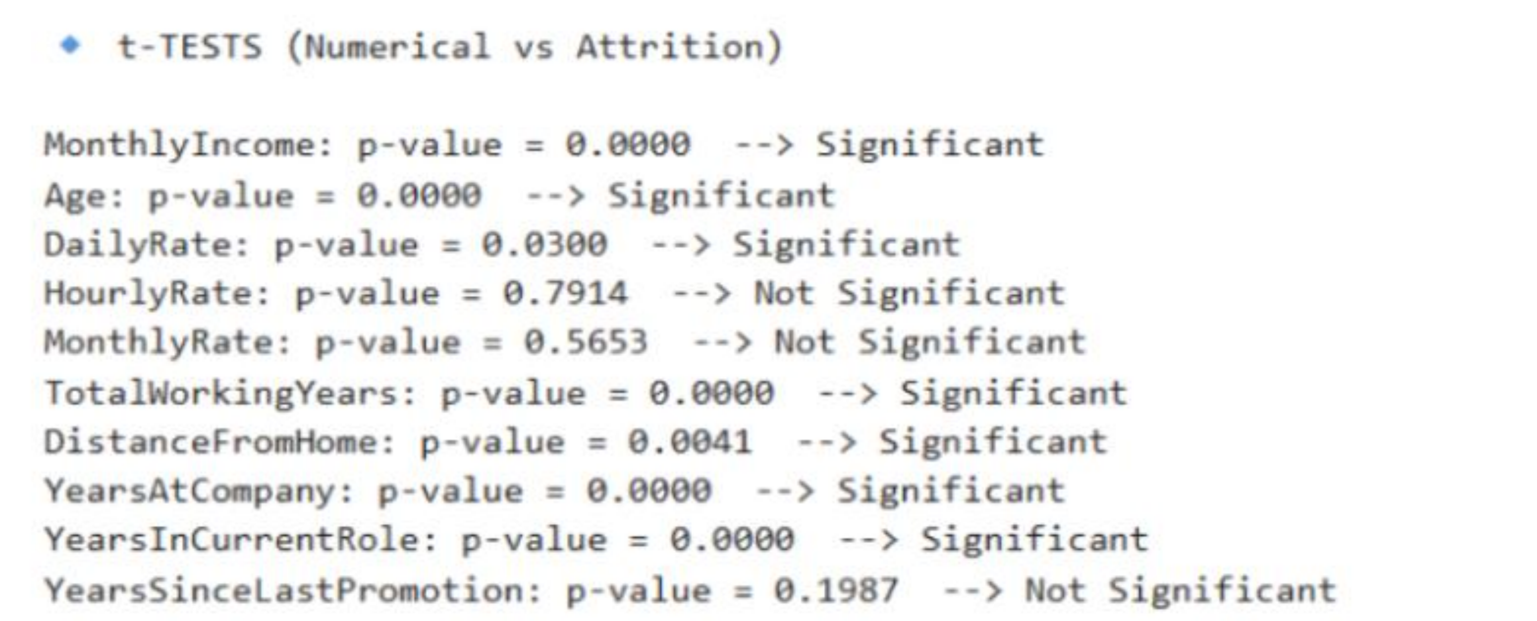
# Result & Interpretation

The model achieved an overall accuracy of 87%, performing well in identifying employees who are likely to stay, but less effective for predicting those who will leave due to class imbalance. Statistical analysis shows that factors like Age, Monthly Income, Job Satisfaction, OverTime, and Distance From Home significantly influence attrition. Predicted probabilities provide actionable insights for HR to plan retention strategies.



```
◆ CHI-SQUARE TESTS (Categorical vs Categorical)

Gender: p-value = 0.2906  --> Not Significant
Department: p-value = 0.0045  --> Significant
BusinessTravel: p-value = 0.0000  --> Significant
JobRole: p-value = 0.0000  --> Significant
MaritalStatus: p-value = 0.0000  --> Significant
EducationField: p-value = 0.0068  --> Significant
OverTime: p-value = 0.0000  --> Significant
```



```
◆ t-TESTS (Numerical vs Attrition)

MonthlyIncome: p-value = 0.0000  --> Significant
Age: p-value = 0.0000  --> Significant
DailyRate: p-value = 0.0300  --> Significant
HourlyRate: p-value = 0.7914  --> Not Significant
MonthlyRate: p-value = 0.5653  --> Not Significant
TotalWorkingYears: p-value = 0.0000  --> Significant
DistanceFromHome: p-value = 0.0041  --> Significant
YearsAtCompany: p-value = 0.0000  --> Significant
YearsInCurrentRole: p-value = 0.0000  --> Significant
YearsSinceLastPromotion: p-value = 0.1987  --> Not Significant
```

# Advantages

It can help a lot...

01 **Quantitative Attrition Probabilities**: The model outputs a probability score for each employee, indicating the likelihood of attrition. This allows HR to prioritize interventions based on risk levels rather than relying solely on intuition.

02 I**nterpretable Features**: The prediction uses clear and understandable features like Age, Monthly Income, Job Satisfaction, OverTime, and Distance From Home. HR personnel can easily comprehend how each factor contributes to attrition, making the insights actionable.

03 **Early Identification of High-Risk Employees**: By highlighting employees with a higher probability of leaving, the system enables proactive measures such as counseling, salary revision, or workload adjustments, reducing unexpected turnover.

04 **Support for HR Dashboards and Decision-Making**: The model's outputs can be integrated into HR analytics dashboards, providing visualizations of attrition risks and key influencing factors. This helps managers make informed decisions and track the effectiveness of retention strategies over time.

# Future Scope

### Addressing Class Imbalance

Currently, the model struggles to accurately predict employees who are likely to leave due to the imbalance between “Yes” and “No” attrition classes. Applying techniques such as SMOTE (Synthetic Minority Oversampling Technique) or using class-weighted models can help balance the dataset, improving the model’s recall and precision for attrition cases.

### Advanced Machine Learning Models

While logistic regression provides a simple and interpretable solution, more sophisticated algorithms like Random Forest, Gradient Boosting, or XGBoost can capture non-linear relationships and complex interactions between features. These models have the potential to improve overall predictive performance and better identify high-risk employees.

### Incorporating Additional HR Data

Including richer HR-related features such as past performance ratings, training history, promotion records, and manager feedback can enhance the predictive power of the model. This additional context allows the system to make more informed predictions, helping HR implement targeted retention strategies.

# Conclusion

The Employee Attrition Prediction project demonstrates the effectiveness of logistic regression combined with statistical analysis to model employee turnover. It identifies key factors influencing attrition and provides HR teams with actionable insights. While overall accuracy is high, further improvements can be achieved by addressing class imbalance and integrating additional data features.