رواد مصر الرقمية

Employee Attrition Prediction & Analysis

Using Machine Learning to Improve Retention

Meet Our Team

- Hussam Ahmed
- Amr Mohamed
- Reham Hassan
- Mohamed Tamer

Under supervision: Eng. Noor El-Deen Magdy

View The Full Code Notebooks From Here:

View The Project Web Application From Here:

1

In this project, we aim to predict employee attrition that is, whether an employee is likely to leave the company using machine learning techniques. By analyzing features such as job roles, monthly income, tenure, and others, organizations can identify at-risk employees and design effective retention strategies.

Introduction

▫ Libraries and Tools Used

Pandas and NumPy: For efficient data handling and numerical computations.

Matplotlib, Seaborn, Plotly: For advanced and interactive data visualization.

Scikit-learn: For preprocessing, modeling, and evaluation.

SMOTE: For handling class imbalance by generating synthetic examples of the minority class.

XGBoost: For building an accurate, robust machine learning model.

Executive Summary

Employee attrition poses a significant challenge to organizations, impacting operational efficiency and increasing recruitment costs. This project leverages advanced machine learning techniques to predict employee attrition and uncover key drivers influencing employee decisions to leave.

Using the IBM HR Analytics dataset, we conducted extensive exploratory data analysis (EDA), applied data preprocessing techniques, and implemented a powerful XGBoost classifier to build an effective prediction model. Key findings highlight the impact of factors such as overtime work, monthly income, job role, and tenure on attrition likelihood.

The insights derived from this analysis enable businesses to design targeted retention strategies, improving employee satisfaction and organizational stability.

Dataset Overview

Source: IBM HR Analytics Employee Attrition & Performance dataset

Records: 1,470 entries

Features: 35 columns (demographics, job roles, performance, etc.)

Sample Data

	T	U	V	W	X	Y	Z	AA	AB
	Distance From Home	Education	Employee Count	Environment Satisfaction	Hourly Rate	Job Involvement	Job Level	Job Satisfaction	Monthly Income
102	1	Associates Deg	1	2	94	3	2	4	5
279	8	High School	1	3	61	2	2	2	5
373	2	Associates Deg	1	4	92	2	1	3	2
392	3	Master's Degre	1	4	56	3	1	3	2
591	2	High School	1	1	40	3	1	2	3
005	2	Associates Deg	1	4	79	3	1	4	3
324	3	Bachelor's Deg	1	3	81	4	1	1	2
358	24	High School	1	4	67	3	1	3	2
216	23	Bachelor's Deg	1	4	44	2	3	3	9
299	27	Bachelor's Deg	1	3	94	3	2	3	5
809	16	Bachelor's Deg	1	1	84	4	1	2	2
153	15	Associates Deg	1	4	49	2	2	3	4
670	26	High School	1	1	31	3	1	3	2
346	19	Associates Deg	1	2	93	3	1	4	2
103	24	Bachelor's Deg	1	3	50	2	1	3	2
389	21	Master's Degre	1	2	51	4	3	1	9
334	5	Associates Deg	1	1	80	4	1	2	3
123	16	Associates Deg	1	4	96	4	1	4	2
219	2	Master's Degre	1	1	78	2	4	4	15
371	2	Bachelor's Deg	1	4	45	3	1	4	3

Metadata:

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   Age                       1470 non-null   int64 
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64 
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64 
 6   Education                 1470 non-null   int64 
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64 
 9   EmployeeNumber            1470 non-null   int64 
 10  EnvironmentSatisfaction   1470 non-null   int64 
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64 
 13  JobInvolvement            1470 non-null   int64 
 14  JobLevel                  1470 non-null   int64 
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64 
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64 
 19  MonthlyRate               1470 non-null   int64 
 20  NumCompaniesWorked        1470 non-null   int64 
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64 
 24  PerformanceRating         1470 non-null   int64 
 25  RelationshipSatisfaction  1470 non-null   int64 
 26  StandardHours             1470 non-null   int64 
 27  StockOptionLevel          1470 non-null   int64 
 28  TotalWorkingYears         1470 non-null   int64 
 29  TrainingTimesLastYear     1470 non-null   int64 
 30  WorkLifeBalance           1470 non-null   int64 
 31  YearsAtCompany            1470 non-null   int64 
 32  YearsInCurrentRole        1470 non-null   int64 
 33  YearsSinceLastPromotion   1470 non-null   int64 
 34  YearsWithCurrManager      1470 non-null   int64 
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
None
```

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.00	3.0	4.00	5.0
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	1555.75	2068.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.00	3.0	4.00	4.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.00	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.00	3.0	4.00	4.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	3.00	15.0

Describe Data

2

Dashboard

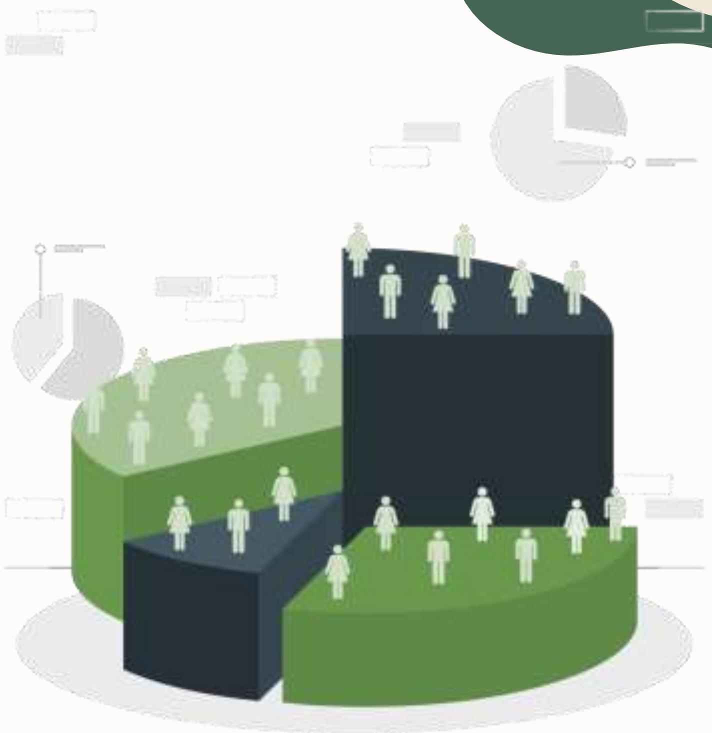
Dashboard using
Power bi
:

Dashboard


Overall Employees
1470
Attrition
237
Attriton Rate
16.12%
Active Employee
1233
Average Age
37
Department wise Attrition
92 (38.82%)
12 (5.06%)
133 (56.12%)
Department
R&D
Sales
HR
No of Employee by Age group
Gender Female Male
60
337
217
309
196
132
113
Under 25
25 - 34
35 - 44
45 - 54
Over 55
Age Band
Job Satisfaction Rating
Job Role 1 2 3 4 Total
Sales Representative 12 21 27 23 83
Sales Executive 69 54 91 112 326
Research Scientist 54 53 90 95 292
Research Director 15 16 27 22 80
Manufacturing Director 26 32 49 38 145
Manager 21 21 27 33 102
Laboratory Technician 56 48 75 80 259
Human Resources 10 16 13 13 52
Healthcare Representative 26 19 43 43 131
Education Field wise Attrition
Education Field
Life Scien... 89
Medical 63
Marketing 35
Technical ... 32
Other 11
Human Re... 7
Attrition Rate by Gender for different Age Group
Under 25
25 - 34
35 -44
45 - 54
Over 55
38
20 (52.63%)
18 (47.37%)
112
69 (61.61%)
43 (38.39%)
51
37 (72.55%)
14 (27.45%)
25
16 (64%)
9 (36%)
11
8 (72.7...)
3 (27.27%)

3

Data Visualization



Data Visualization

Data visualization is the graphical representation of data to facilitate understanding and analysis, It helps in identifying patterns, trends, and outliers that might go unnoticed in raw data, enabling more informed decision-making.

- ## Components of Data Visualization :

Data: The raw information that is visualized.

Visual Elements: Charts, graphs, maps, colors, and symbols used to represent data.

Axes: Provide a frame of reference for measurements.

Legends: Explain the meaning of colors, symbols, or patterns used.

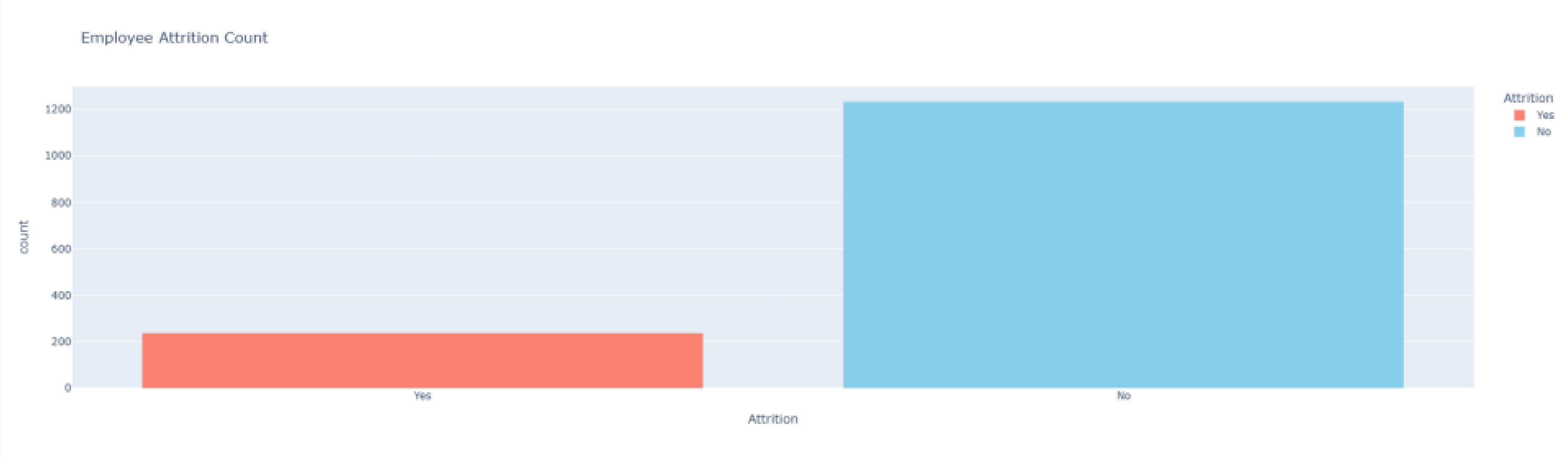
Titles and Labels: Clarify what the visualization represents.

Context: Additional information that helps in understanding the data's relevance.

Benefits of Data Visualization:

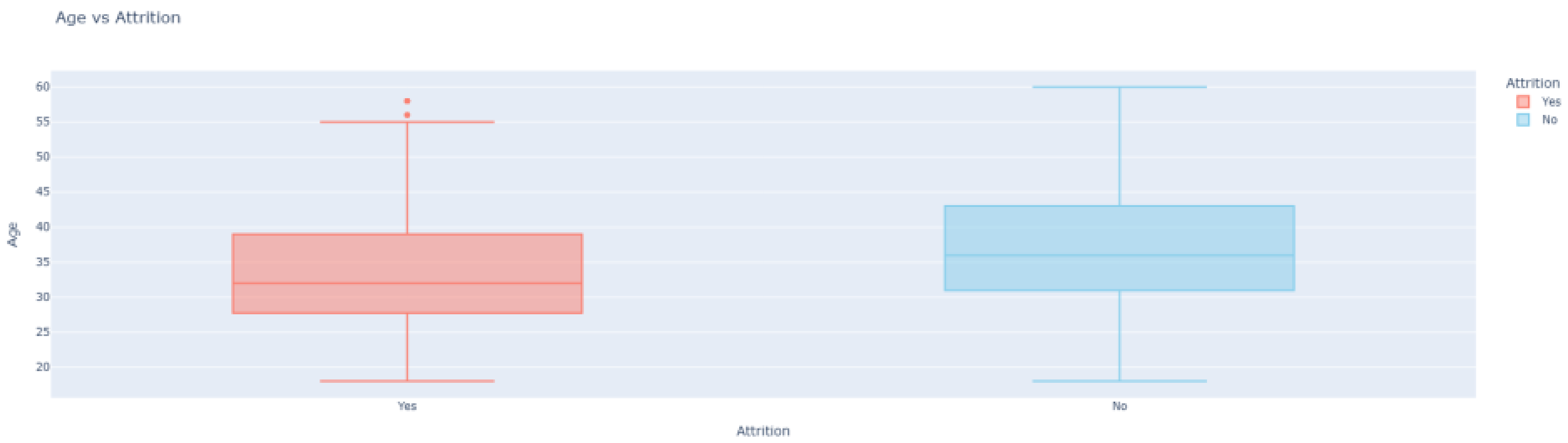
- **Increases Efficiency:** Reduces time spent on data analysis by highlighting key insights.
- **Promotes Engagement:** Captures attention and maintains interest with visual elements.
- **Enables Better Retention:** Visual information is often remembered more effectively than text alone.
- **Supports Collaboration:** Creates a common understanding among team members or stakeholders.

1-Distribution of employees who stayed vs. left :



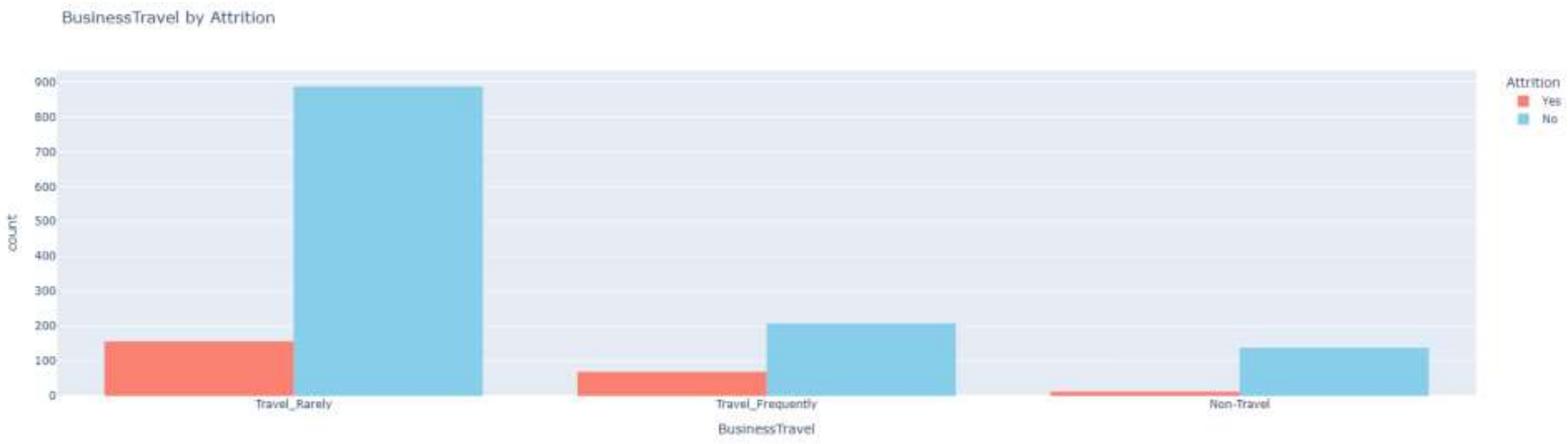
- Shows the number of employees who stayed ("No" in blue) versus those who left ("Yes" in red).

2-Age Vs Attrition:



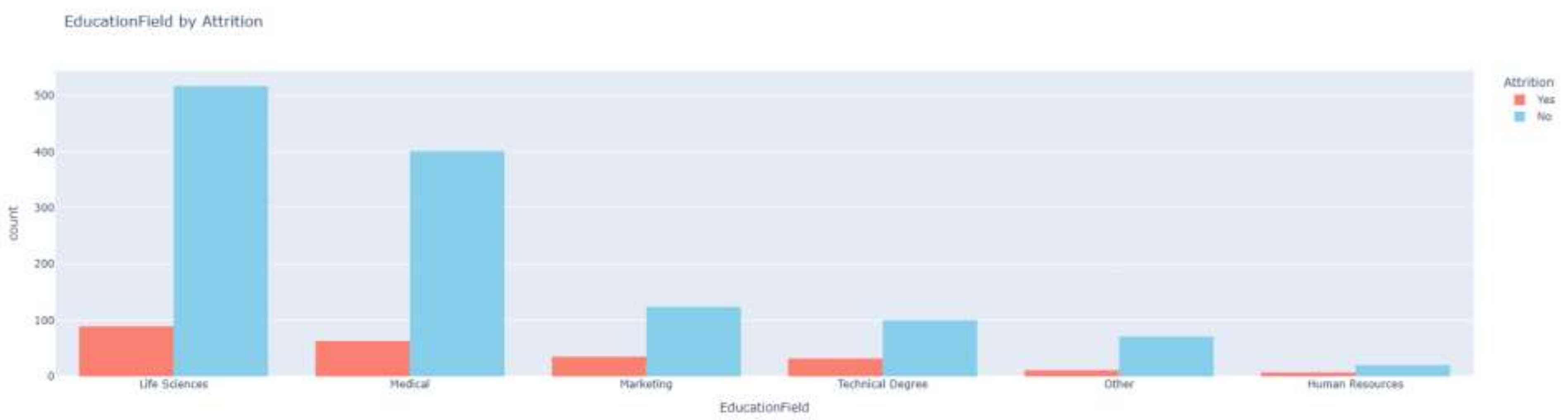
- Compares the age distribution of employees who stayed ("No" in blue) versus those who left ("Yes" in red).

3.1-BusinessTravel by Attrition:



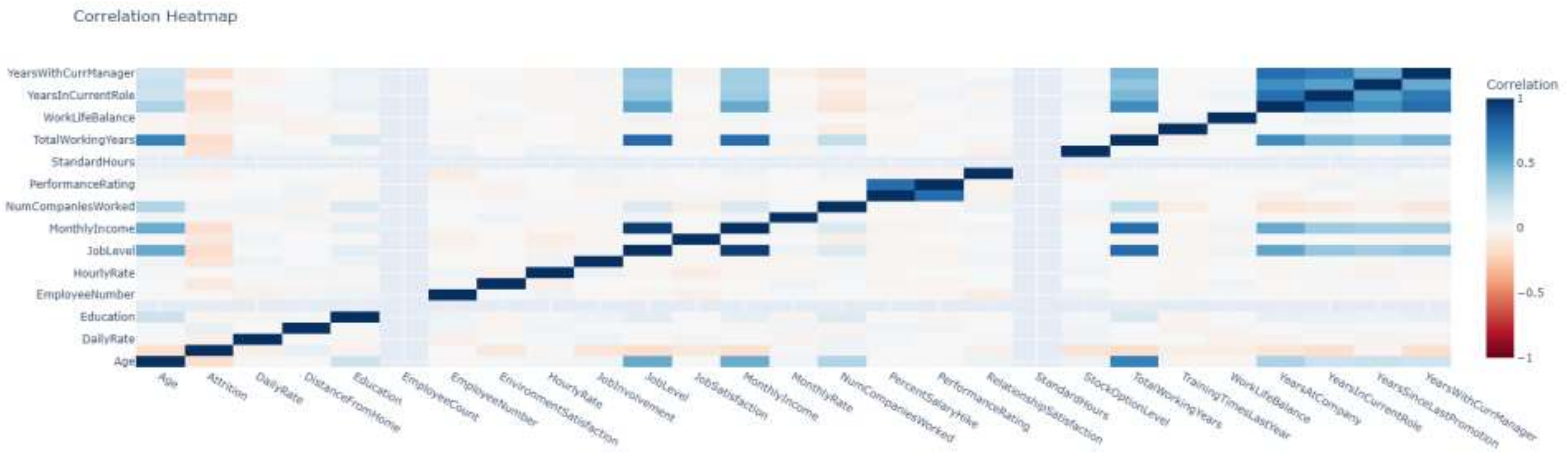
- Compares the count of employees who stayed ("No" in blue) and those who left ("Yes" in red) across different business travel frequencies. It shows that most employees who travel rarely stayed, while non-travelers had the fewest leavers.

3.2-EducationField by Attrition:



- Compares the count of employees who stayed ("No" in blue) and those who left ("Yes" in red) across different education fields. It highlights how attrition varies by educational background.

4-Employee Data Correlation Heatmap:

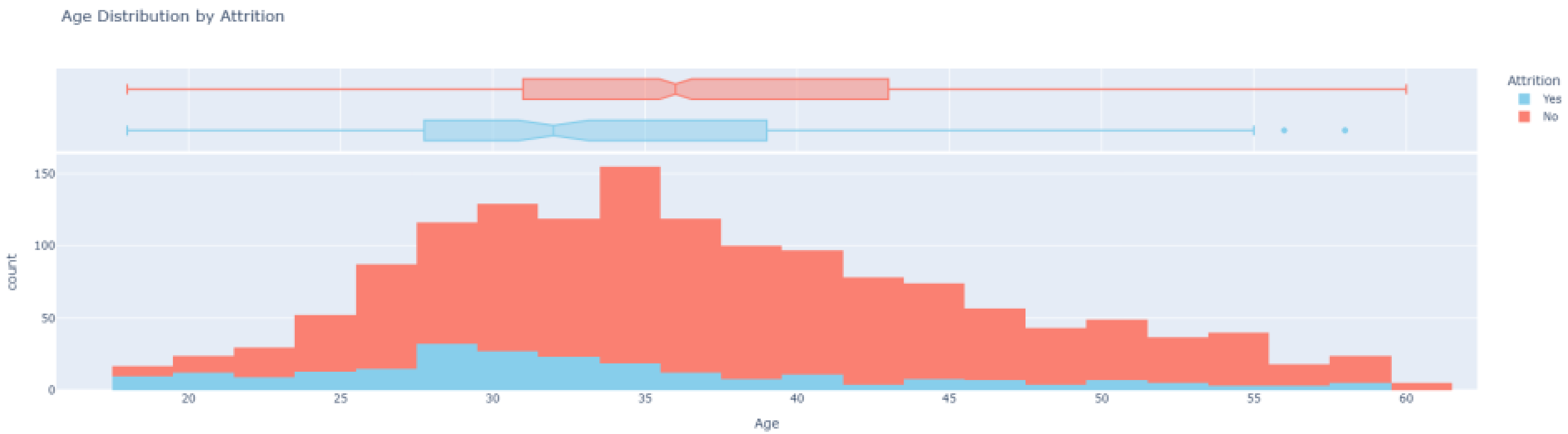


- Visualizes the relationships between various employee attributes. The color scale indicates the strength and direction of correlations, ranging from -1 to 1. It helps identify which factors are strongly related.

5-Monthly Income vs Total Working Years by Attrition:

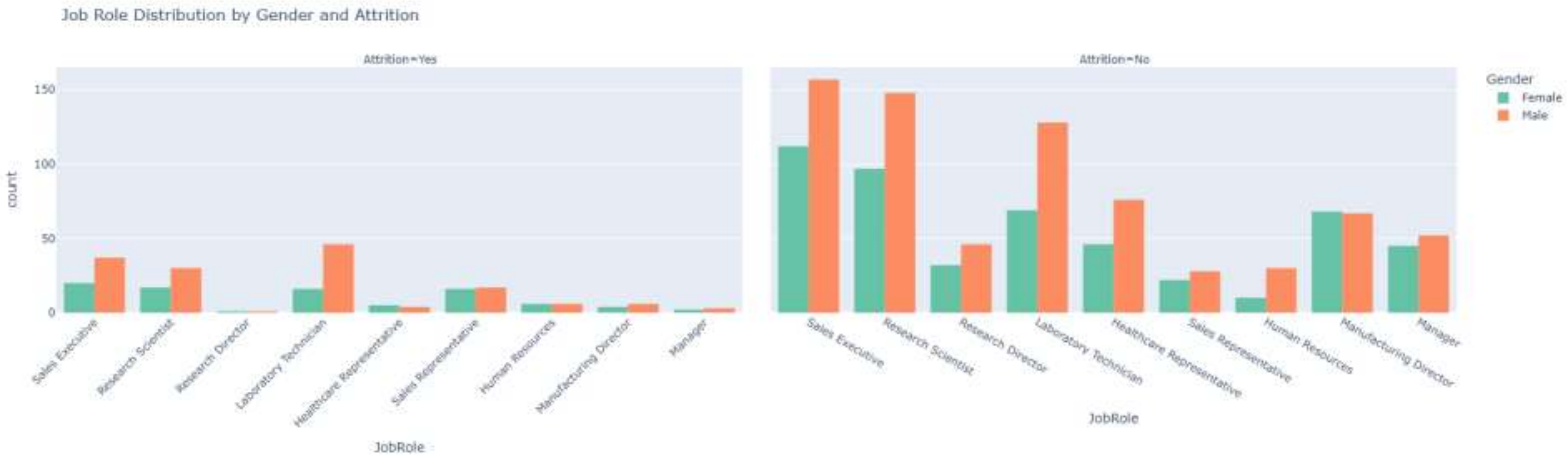
- Scatter plot shows the relationship between employees' total working years and their monthly income, colored by attrition status. It helps identify if income grows with experience and if this relates to employees leaving or staying.

6-Age Distribution by Attrition:



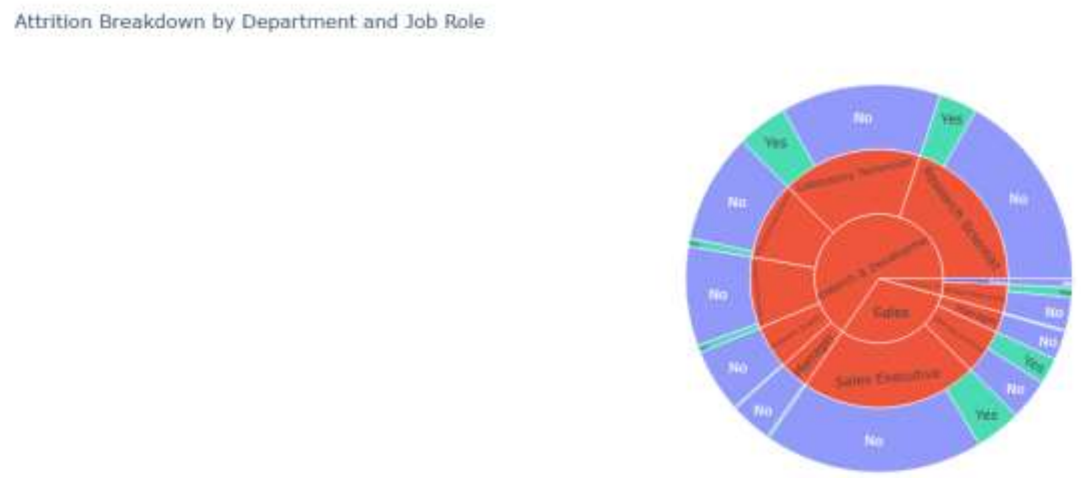
- It shows the comparing of those who left ("Yes" in blue) with those who stayed ("No" in red). It shows how age varies between the two groups.

7-Job Role Distribution by Gender and Attrition:



- Grouped bar chart displays the distribution of job roles by gender for employees who stayed ("Attrition=No") and those who left ("Attrition=Yes"). It compares the number of male and female employees in each job role across both attrition categories, highlighting any gender disparities and their relation to employee turnover.

8-Attrition Breakdown by Department and Job Role:



- Represents a hierarchy level, with departments on the outside, job roles in the middle, and attrition status ("Yes" or "No") at the center. It highlights which departments and roles have higher turnover.

4

Predictions

Project Objectives

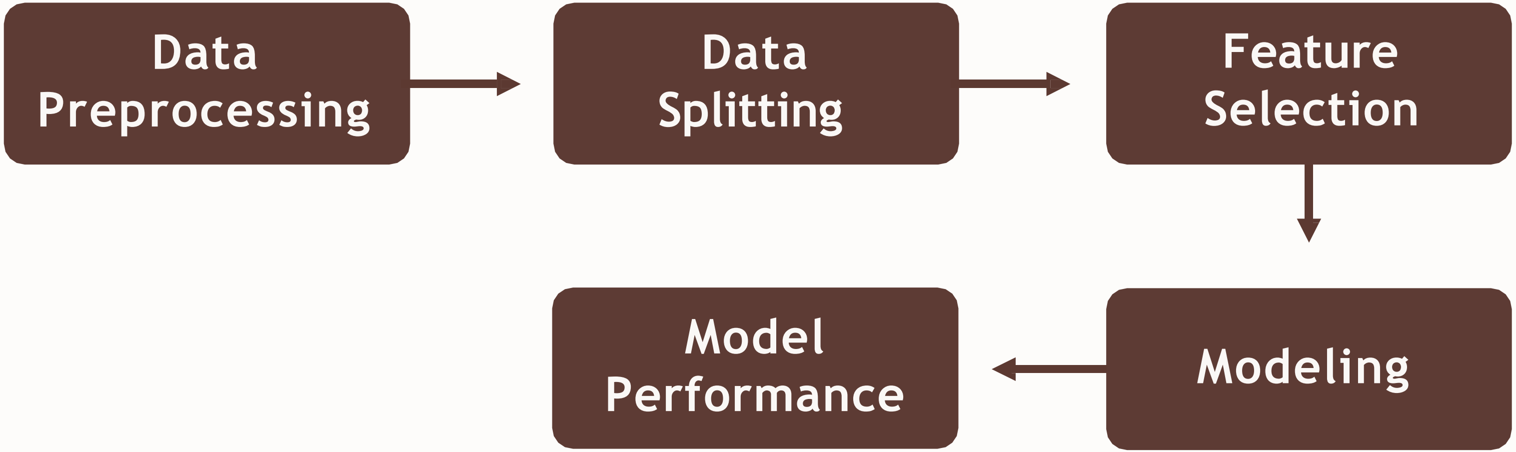
- **Predict Employee Attrition:**
  Develop a machine learning model to predict the likelihood of employee attrition, enabling proactive retention strategies.

- **Develop a User-Friendly Dashboard:**
  Create an interactive dashboard to visualize predictions and key insights, facilitating decision-making for HR stakeholders.

- **Enhance Feature Engineering Techniques:**
  Implement advanced feature engineering methods to create meaningful features (e.g., PromotionRate, JobRole_Stability) that improve model performance and interpretability.

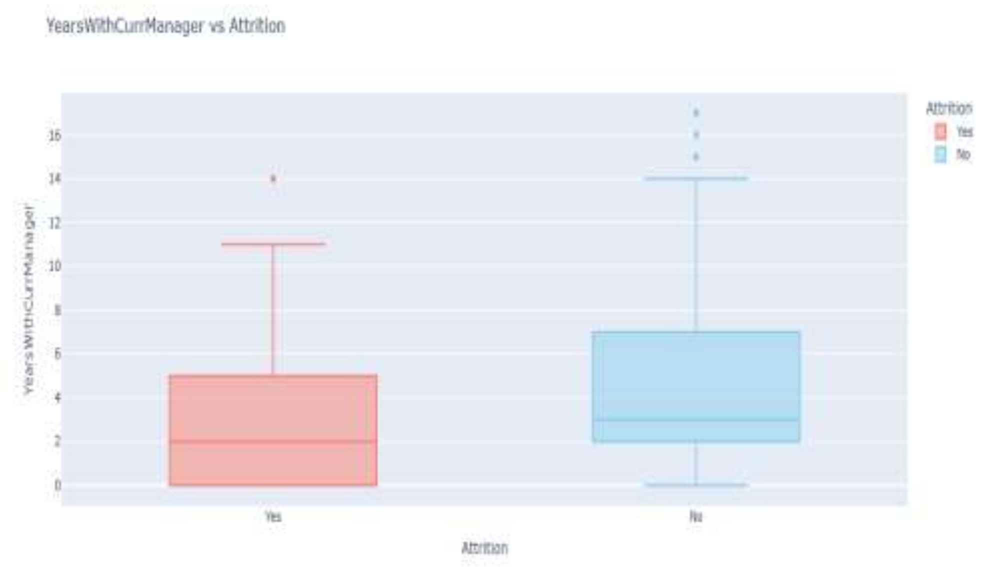
- **Provide Strategic Recommendations:**
  Based on model findings, offer actionable recommendations to reduce attrition rates and improve employee satisfaction and retention.

Process to Build the Model:



Data Preprocessing

- Convert Data Types: Ensure correct data types (e.g., convert string to category)
- Feature Selection & Dropping Irrelevant Columns
- Encoding Categorical Variables



Feature Selection:

This new feature helps model the relationship between tenure and promotion history.

```
# This feature can be important in models predicting employee retention,
# especially when analyzing loyalty or age-related patterns in workforce retention.
data['YearsAtCompany_to_Age'] = data['YearsAtCompany'] / data['Age']
```

```
# This feature is useful to understand role stability.
data['JobRole_Stability'] = data['YearsInCurrentRole'] / (data['TotalWorkingYears'] + 1)
```

```
# This new feature helps model the relationship between tenure and promotion history.
data['PromotionRate'] = data['YearsSinceLastPromotion'] / (data['YearsAtCompany'] + 1)
```

```
data["Attrition"] = data["Attrition"].map({"Yes": 1, "No":0})
```

Preprocessing :

- Reasoning: Simplify dataset, prepare it for modeling by encoding categorical features and cleaning irrelevant columns.

```
data = pd.get_dummies(data, drop_first= True)
```

◆ Reasoning: Simplify dataset, prepare it for modeling by encoding categorical features and cleaning irrelevant columns.

✓ Model Chosen: XGBoost

```python
# Stratified K-Fold setup
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
accuracy_scores = []
f1_scores = []
for train_idx, val_idx in skf.split(X_train, y_train):
    # Use .iloc for DataFrames/Series
    X_tr, X_val = X_train.iloc[train_idx], X_train.iloc[val_idx]
    y_tr, y_val = y_train.iloc[train_idx], y_train.iloc[val_idx]

    # Apply SMOTE
    smote = SMOTE(random_state=42)
    X_tr_resampled, y_tr_resampled = smote.fit_resample(X_tr, y_tr)

    # Compute scale_pos_weight
    scale_pos_weight = len(y_tr_resampled[y_tr_resampled == 0]) / len(y_tr_resampled[y_tr_resampled == 1]) * 1.5

    # Define and train the model
    xgb_model = XGBClassifier(
        n_estimators=100,
        max_depth=5,
        learning_rate=0.1,
        scale_pos_weight=scale_pos_weight,
        use_label_encoder=False,
        eval_metric='logloss',
        random_state=42
    )
    xgb_model.fit(X_tr_resampled, y_tr_resampled)
    # Predict and evaluate
    y_pred = xgb_model.predict(X_val)
```

Model Performance

```
Accuracy: 0.8605442176870748
Confusion Matrix:
 [[231  16]
 [ 25  22]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.94      0.92       247
           1       0.58      0.47      0.52        47

    accuracy                           0.86       294
   macro avg       0.74      0.70      0.72       294
weighted avg       0.85      0.86      0.85       294
```

86%
Accuracy

Best Model:

XGBoost Classifier

Why XGBoost?

Exceptional performance on structured/tabular data

Effective handling of class imbalance

High computational efficiency

View the full code Notebooks from here:

86%
Accuracy

5

Model Deployment

Model Deployment:

🔍 **Purpose:** Predict whether an employee is likely to leave the company based on personal, job-related, and satisfaction data.

Built With:

- Streamlit for the interactive web app
- XGBoost model (pre-trained and loaded with pickle)
- Pandas for handling user input

How It Works:

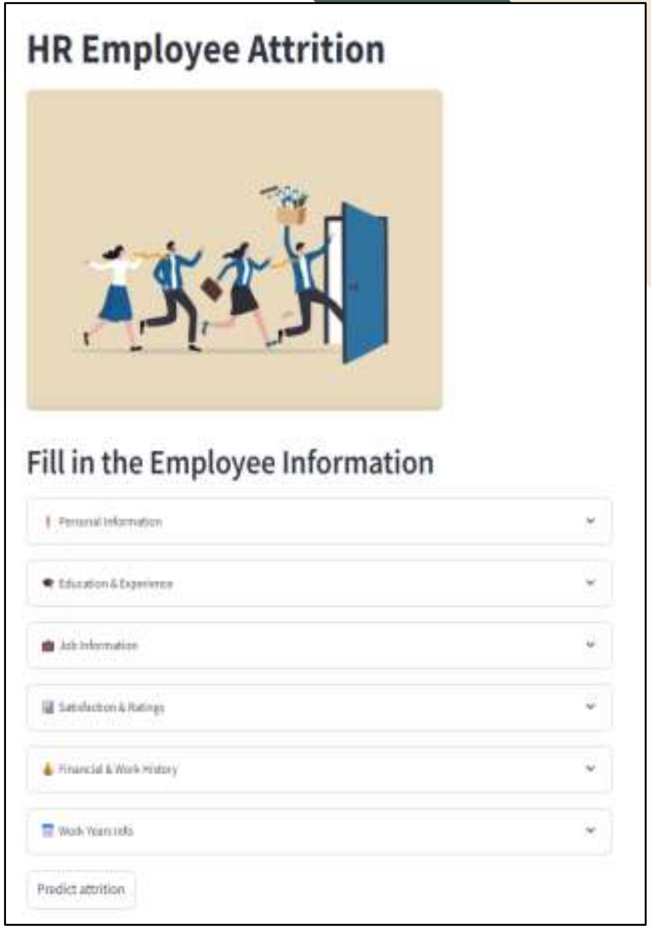
- HR inputs employee details (Age, Job Role, Satisfaction levels, etc.)
- Model Predicts Employee Attrition: Yes/No
- Shows suggestions if the employee is likely to leave the company.

Impact:

- Helps HR make data-driven decisions
- Supports employee retention strategies

View the web app from here:

View the code Notebook from here:



Thanks!

Do you have any questions?