Introduction of the Project

Employee attrition refers to the rate at which employees leave an organization. HR Analytics helps identify patterns, predict resignations, and support better retention strategies.

Objective

To analyze HR data, predict employee attrition using machine learning, and provide actionable insights to help HR managers reduce turnover.

Logistic Regression Model Results

Step 1: Import Libraries

👉 Imported Python libraries (Pandas, Seaborn, Sklearn, SHAP) for data analysis, visualization, and modeling.

```
[*]: # Step 1: Import Libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import shap
import warnings
warnings.filterwarnings("ignore")
```

Step 2: Load Dataset

👉 Loaded the HR attrition dataset (employee_attrition_data.csv) and displayed its shape & first few rows.

```
[3]: # Step 2: Load Dataset
file_path = "employee_attrition_data.csv"   # just the filename
df = pd.read_csv(file_path)

print("Dataset Shape:", df.shape)
print("\nFirst 5 Rows:\n", df.head())
```

Step 3: Exploratory Data Analysis (EDA)
👉 Checked dataset info, missing values, and visualized employee attrition distribution.

```
[4]: # Step 3: EDA
print("\nDataset Info:")
print(df.info())

print("\nMissing Values:\n", df.isnull().sum())
```

Step 4: Preprocessing
👉 Encoded categorical variables using Label Encoder and split data into training (80%) and testing (20%).

```
[6]: # Step 4: Preprocessing
# Encode categorical columns
le = LabelEncoder()
for col in df.select_dtypes(include=["object"]).columns:
    df[col] = le.fit_transform(df[col])
```

Step 5: Logistic Regression Model
👉 Trained Logistic Regression model; accuracy was **49%**, showing poor performance.

```
[8]: # Step 5: Logistic Regression Model
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred_log = log_model.predict(X_test)

print("\nLogistic Regression Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_log))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_log))
print("Classification Report:\n", classification_report(y_test, y_pred_log))
```

Step 6: Decision Tree Model
👉 Trained Decision Tree model; achieved better accuracy and classification results than Logistic Regression.

```
# Step 6: Decision Tree Model
tree_model = DecisionTreeClassifier(max_depth=5, random_state=42)
tree_model.fit(X_train, y_train)
y_pred_tree = tree_model.predict(X_test)

print("\nDecision Tree Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_tree))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_tree))
print("Classification Report:\n", classification_report(y_test, y_pred_tree))
```

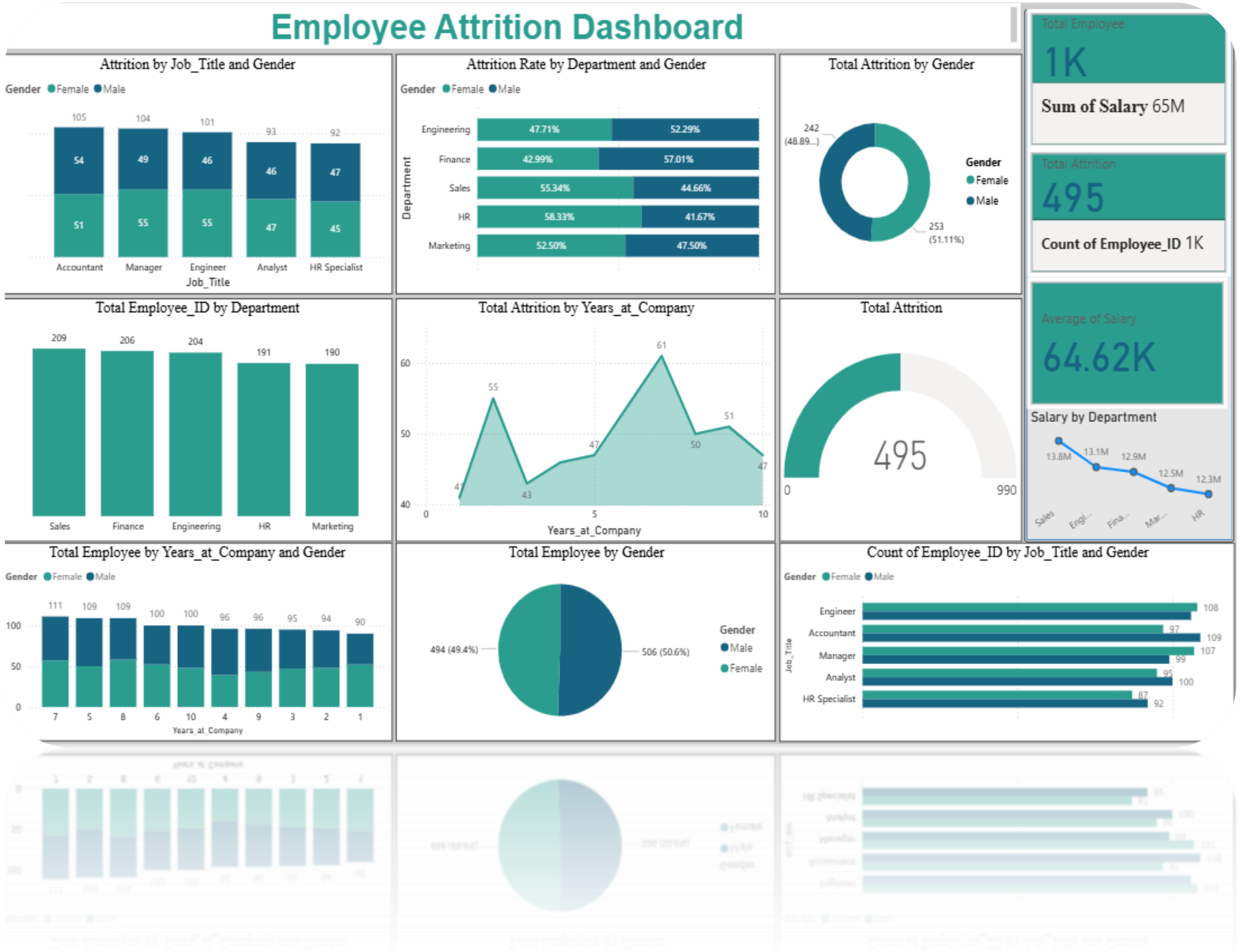
Step 7: SHAP Analysis (Explainability)
👉 Used SHAP values to explain Decision Tree predictions; identified key factors (Overtime, Age, Salary).

```
[10]:
# Step 7: SHAP Value Analysis for Explainability (using Decision Tree)
explainer = shap.TreeExplainer(tree_model)
shap_values = explainer.shap_values(X_test)

print("\nGenerating SHAP summary plot...")
shap.summary_plot(shap_values, X_test, plot_type="bar")
```

Insights

•**Attrition by Job Title & Gender** – Attrition is evenly split across job roles, with accountants and managers slightly higher.
•**Attrition Rate by Department & Gender** – HR and Sales show the highest attrition, while Finance has the lowest.
•**Total Attrition by Gender** – Male and female attrition is nearly equal, showing no major gender gap.
•**Total Employee_ID by Department** – Sales, Finance, and Engineering have the largest employee counts.
•**Attrition by Years at Company** – Attrition peaks around 5 years of service.
•**Total Employees by Years & Gender** – Employee tenure distribution is balanced between males and females.
•**Total Employees by Gender** – Workforce is almost equally split between males and females.
•**Total Attrition Gauge** – Nearly half of the total employees (495 out of 1000) have left.
•**Salary by Department** – Sales and Engineering have the highest salary expenses, HR the lowest.



Project Summary

This project focuses on predicting employee attrition using HR data analytics. Exploratory Data Analysis (EDA) revealed high attrition in HR and Sales departments, with a major spike around 5 years of service. Gender showed almost equal attrition rates, while salary variations across departments influenced turnover. A classification model (Logistic Regression/Decision Tree) was built to predict attrition, and SHAP analysis explained the key drivers such as salary, promotions, and job role. Insights were visualized through a Power BI dashboard, and recommendations were provided to help HR managers reduce attrition and improve employee retention.