Predicting Employee Attrition with ML

Using machine learning to predict employee attrition.

Leveraging HR data and advanced models to identify at-risk employees.

Objective: reduce attrition and improve employee retention.

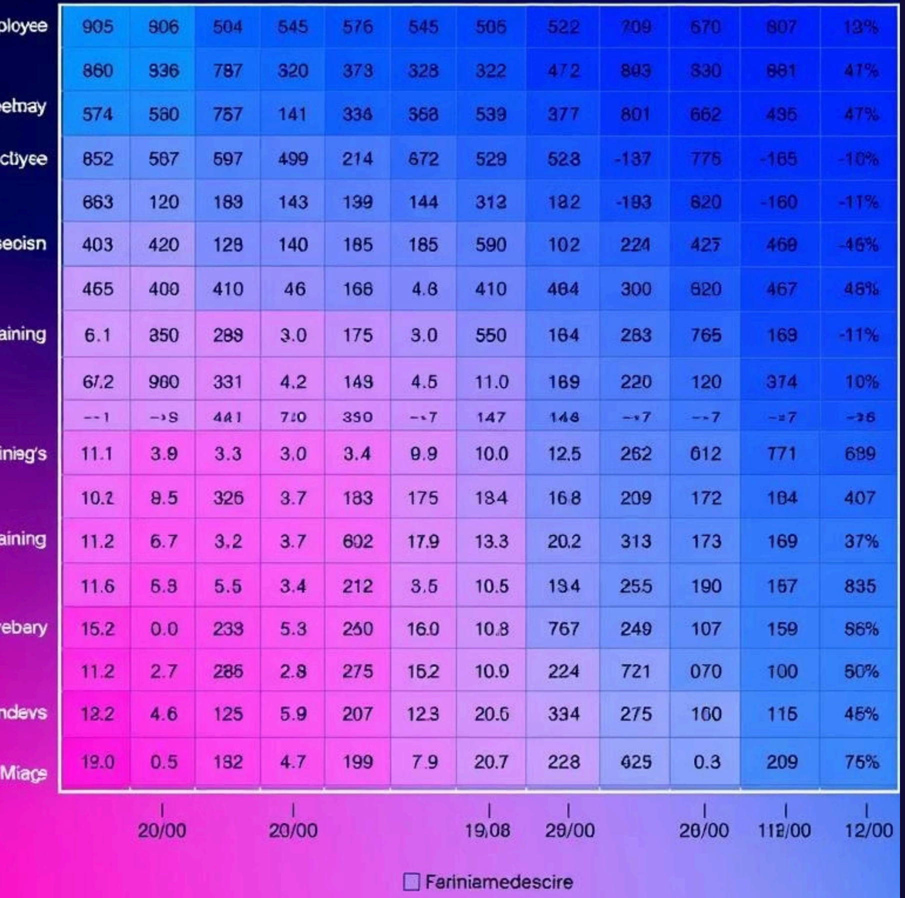
Data Cleaning and Preparation

Data Cleaning

- Remove irrelevant columns.
- Handle missing values.
- Ensure data quality.

Label Encoding

- Convert categories to numbers.
- Prepare for model training.
- Improve model compatibility.



ployee	905	806	504	545	576	545	505	522	709	670	807	19%
	860	936	787	320	373	328	322	472	893	830	881	47%
etnay	574	560	757	141	336	358	539	377	801	662	495	47%
ctiyee	852	567	597	499	214	672	529	528	-137	775	-165	-10%
	663	120	189	143	139	144	313	182	-193	820	-160	-11%
ecisn	403	420	129	140	185	185	590	102	224	427	469	-46%
	465	400	410	46	166	4.8	410	484	300	620	467	46%
aining	6.1	350	289	3.0	175	3.0	550	164	283	765	169	-11%
	67.2	960	331	4.2	149	4.5	11.0	169	220	120	374	10%
	--1	-›S	441	7:0	350	--7	147	148	--7	--7	-=7	-36
inieg's	11.1	3.9	3.3	3.0	3.4	9.9	10.0	12.5	262	612	771	689
	10.2	8.5	326	3.7	183	175	184	16.8	209	172	184	407
aining	11.2	6.7	3.2	3.7	602	17.9	13.3	20.2	313	173	169	37%
	11.6	6.8	5.5	3.4	212	3.5	10.5	134	255	190	167	835
ebary	15.2	0.0	233	5.3	250	16.0	10.8	767	249	107	159	56%
	11.2	2.7	286	2.8	275	152	10.0	224	721	070	100	60%
ndevs	12.2	4.6	125	5.9	207	123	20.6	334	275	160	115	46%
Miage	19.0	0.5	182	4.7	199	7.9	20.7	228	625	0.3	209	75%

Exploratory Data Analysis (EDA)

Correlation Matrix

Visualize feature correlations.

Multicollinearity Detection

Reduce redundant variables.

Key Insights

Age vs. Monthly Income.

Keep Attrition as Dependent variable and Run the models

Machine Learning Models

Logistic Regression

Binary classification baseline.

SVM

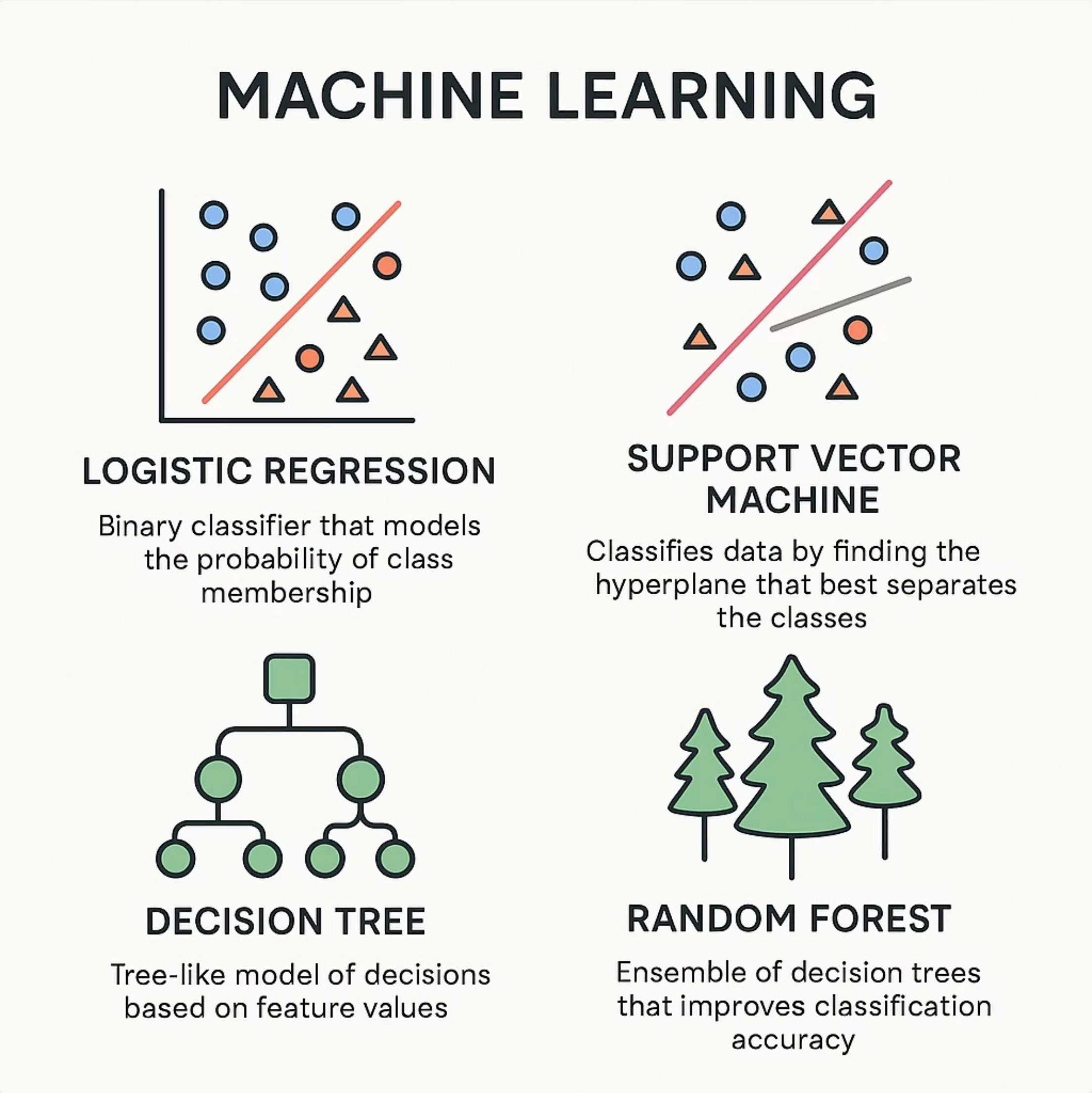
Separating hyperplanes.

Decision Tree

Interpretable rules.

Random Forest

Ensemble accuracy.



MACHINE LEARNING

LOGISTIC REGRESSION

Binary classifier that models the probability of class membership

SUPPORT VECTOR MACHINE

Classifies data by finding the hyperplane that best separates the classes

DECISION TREE

Tree-like model of decisions based on feature values

RANDOM FOREST

Ensemble of decision trees that improves classification accuracy

Model Evaluation Metrics

- Accuracy
- Confusion Matrix
- ROC-AUC

Comprehensive metrics to compare models.

Key Results and Insights

Performance

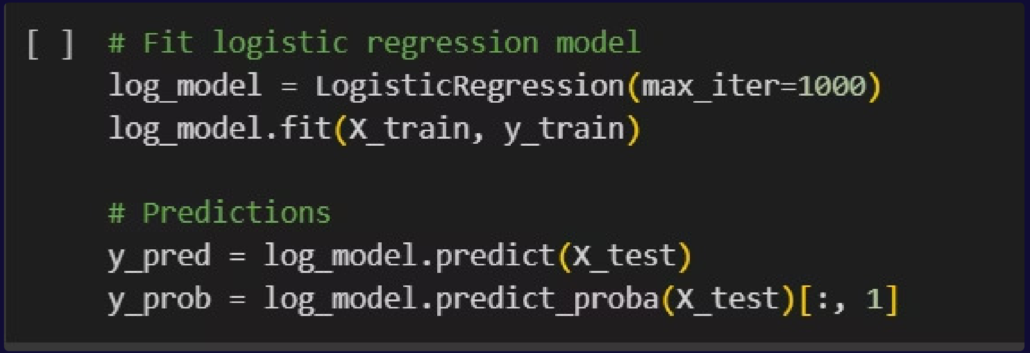
SVM and Random Forest excel.

Interpretation

Logistic Regression is clear.

Identification

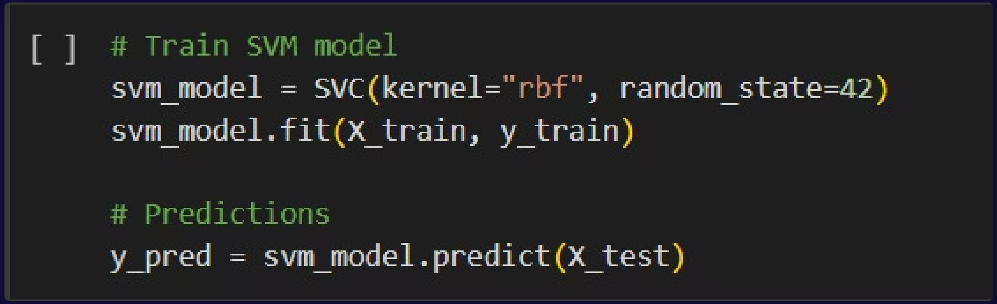
Pinpoint high-risk attrition.



```
[ ]  # Fit logistic regression model
     log_model = LogisticRegression(max_iter=1000)
     log_model.fit(X_train, y_train)

     # Predictions
     y_pred = log_model.predict(X_test)
     y_prob = log_model.predict_proba(X_test)[:, 1]
```

Logistic regression

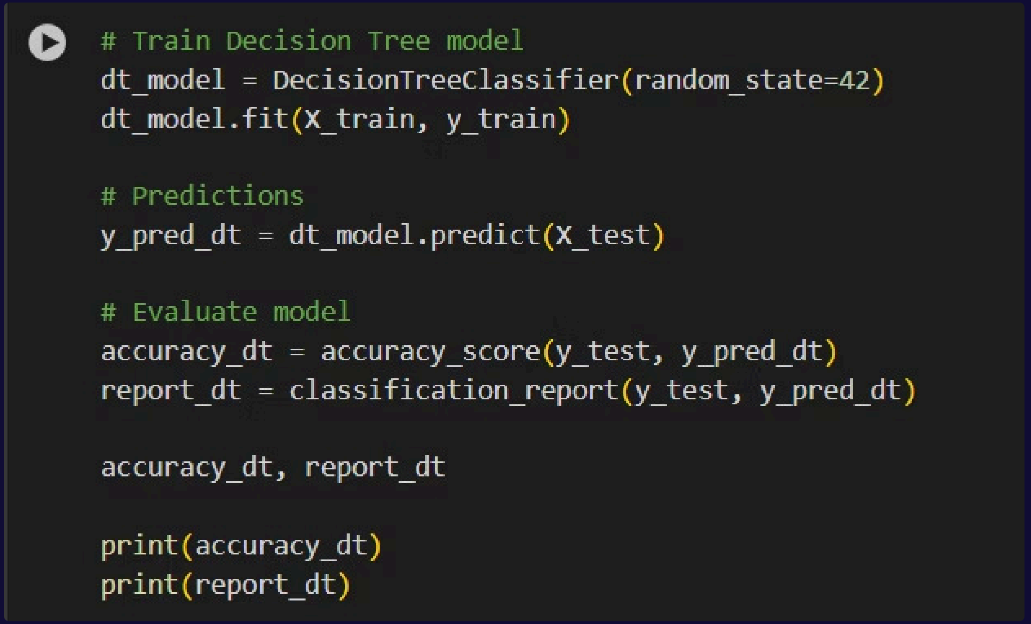
Accuracy: 89%



```
[ ]  # Train SVM model
     svm_model = SVC(kernel="rbf", random_state=42)
     svm_model.fit(X_train, y_train)

     # Predictions
     y_pred = svm_model.predict(X_test)
```

Support Vector Machine

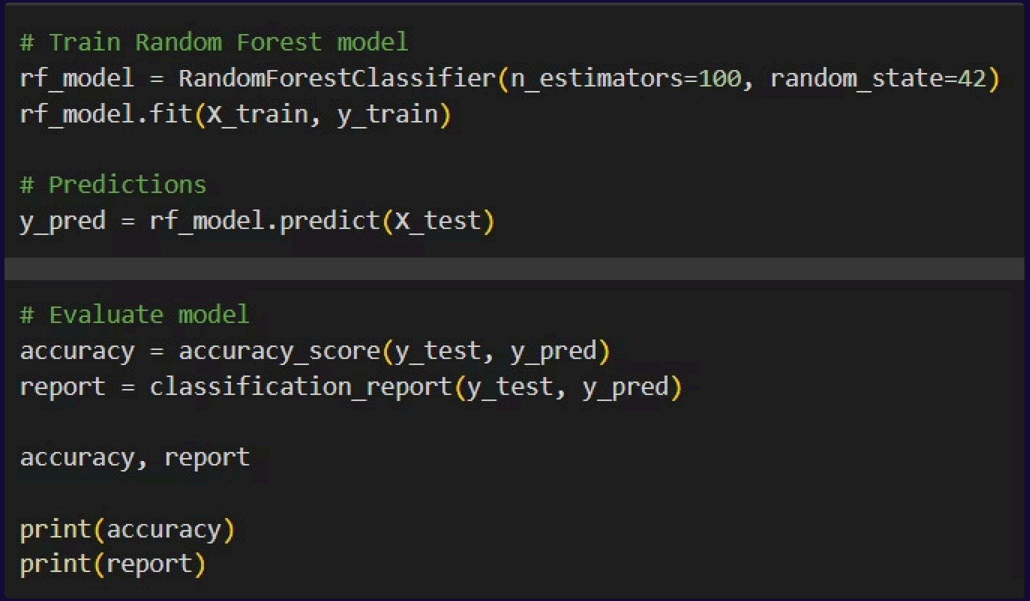
Accuracy: 89%



```python
# Train Decision Tree model
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)

# Predictions
y_pred_dt = dt_model.predict(X_test)

# Evaluate model
accuracy_dt = accuracy_score(y_test, y_pred_dt)
report_dt = classification_report(y_test, y_pred_dt)

accuracy_dt, report_dt

print(accuracy_dt)
print(report_dt)
```

Decision tree Classifier

Accuracy: 80%

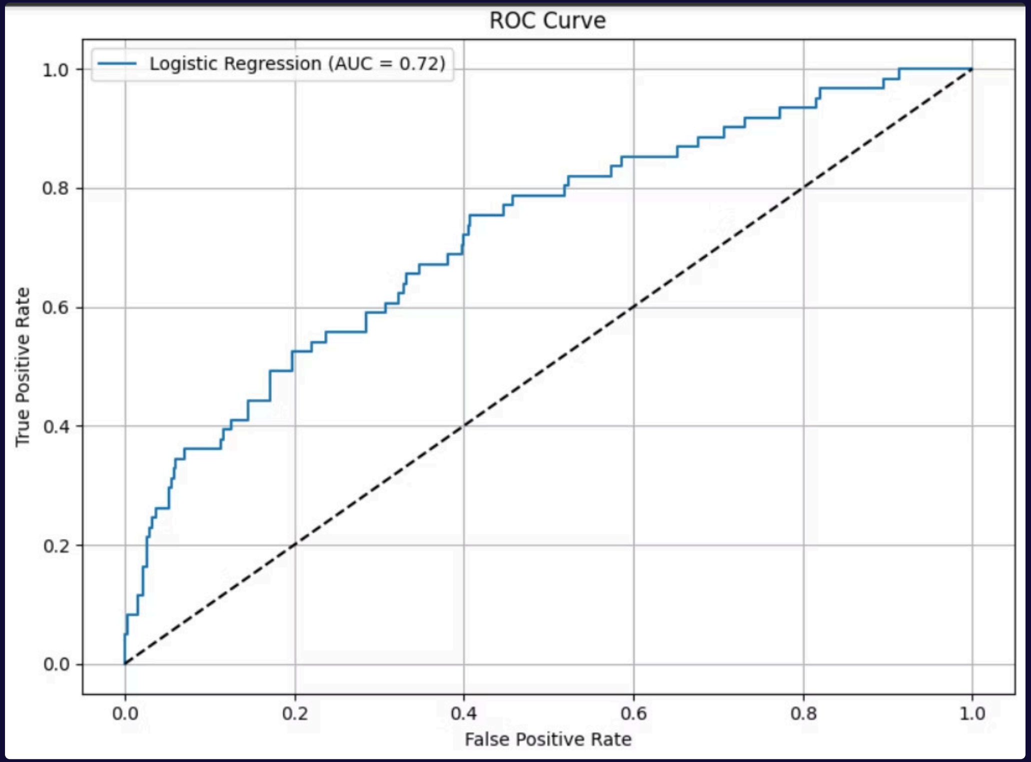


```python
# Train Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predictions
y_pred = rf_model.predict(X_test)
```

```python
# Evaluate model
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

accuracy, report

print(accuracy)
print(report)
```

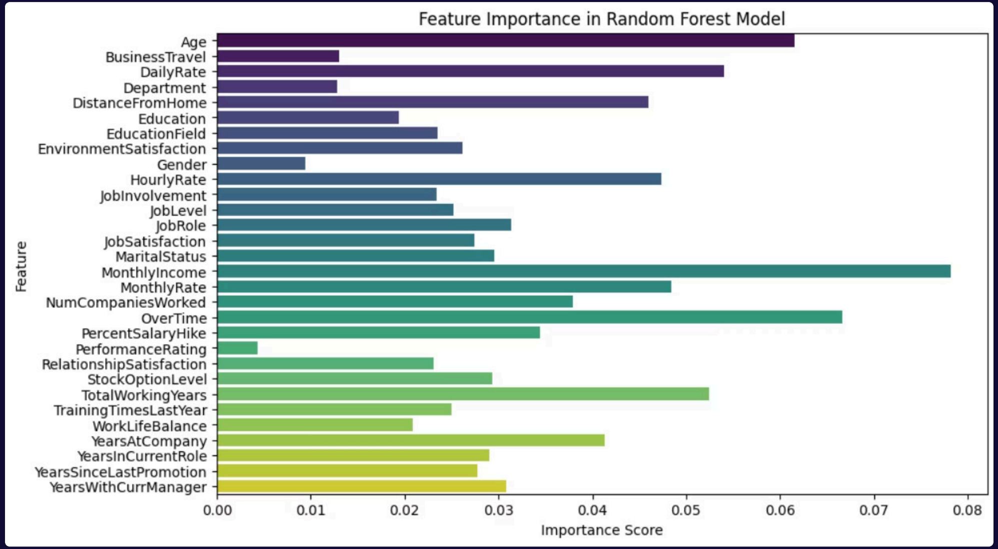
Random Forest Classifier

Accuracy: 89%



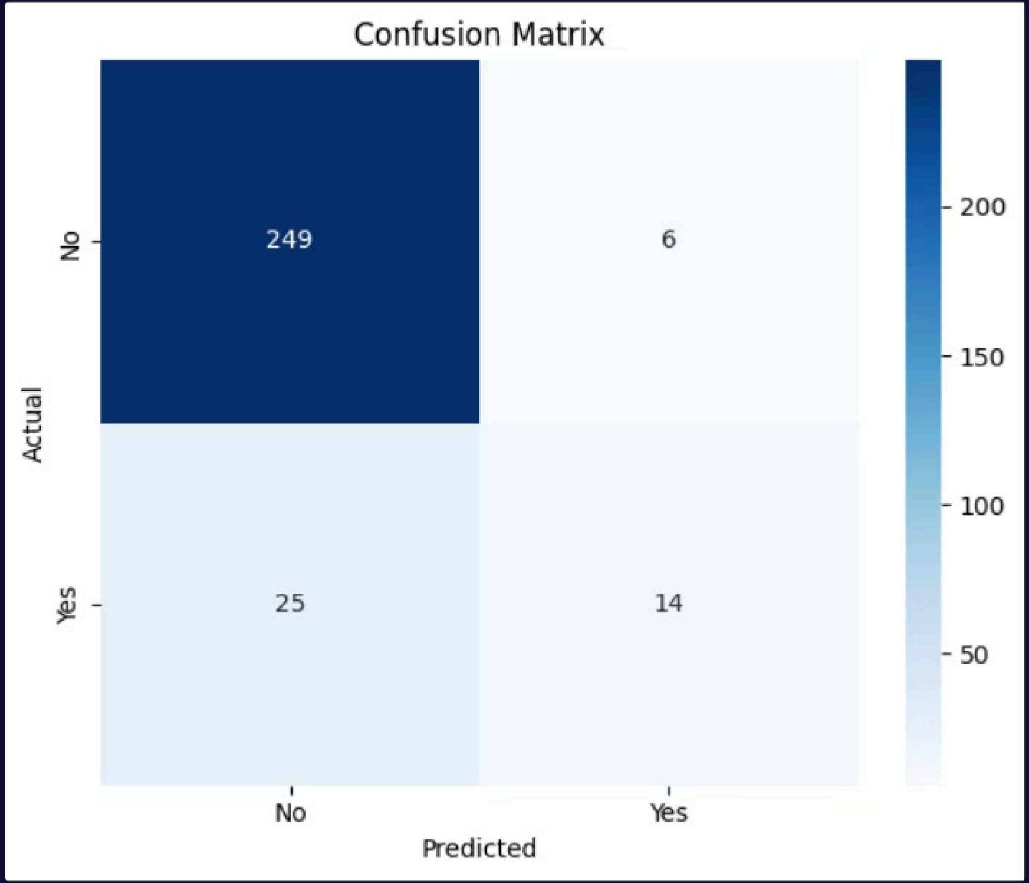
ROC Curve

Visualizes the trade-off between true positive rate and false positive rate across thresholds.



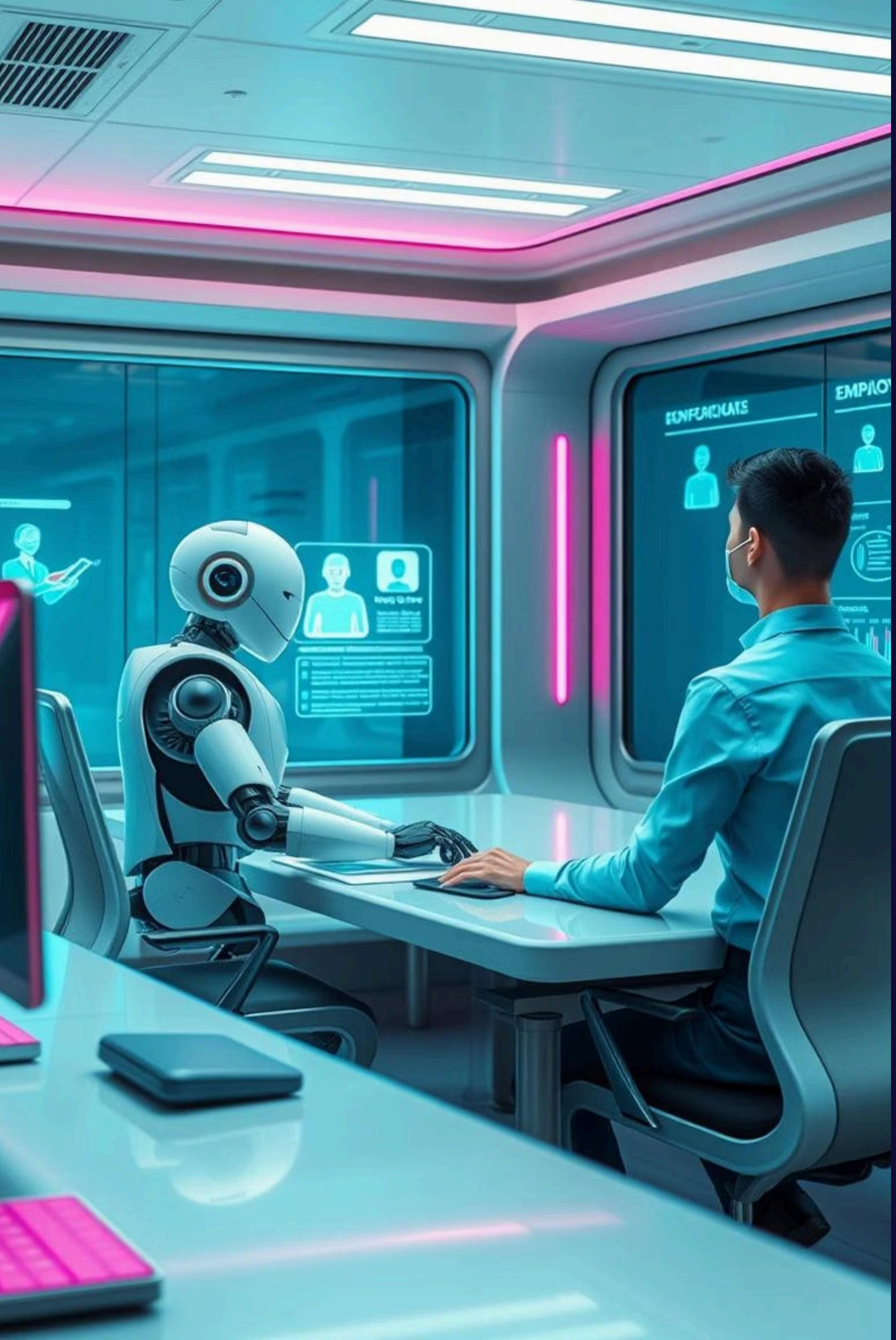
Feature Importance graph

Shows which features have the greatest impact on the model's predictions.



Confusion Matrix

Summarizes the model's classification performance by showing correct and incorrect predictions.



Conclusion and Future Steps

Impact

Precise attrition prediction.

Integration

HR dashboards ready.

Next Steps

- Hyperparameter tuning.
- XGBoost for boost.
- Real-time deployment.