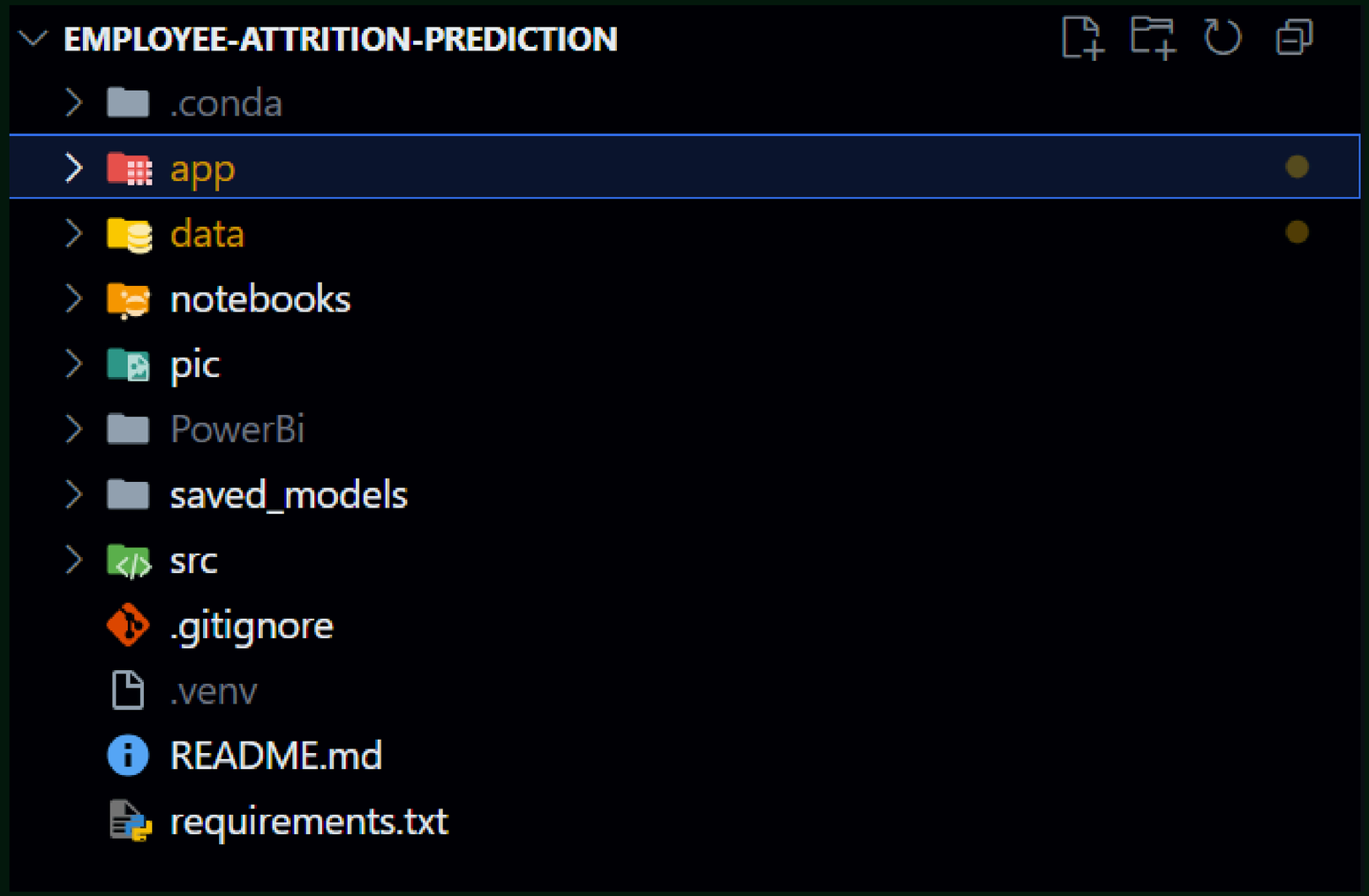
Project Report – Employee Attrition Prediction

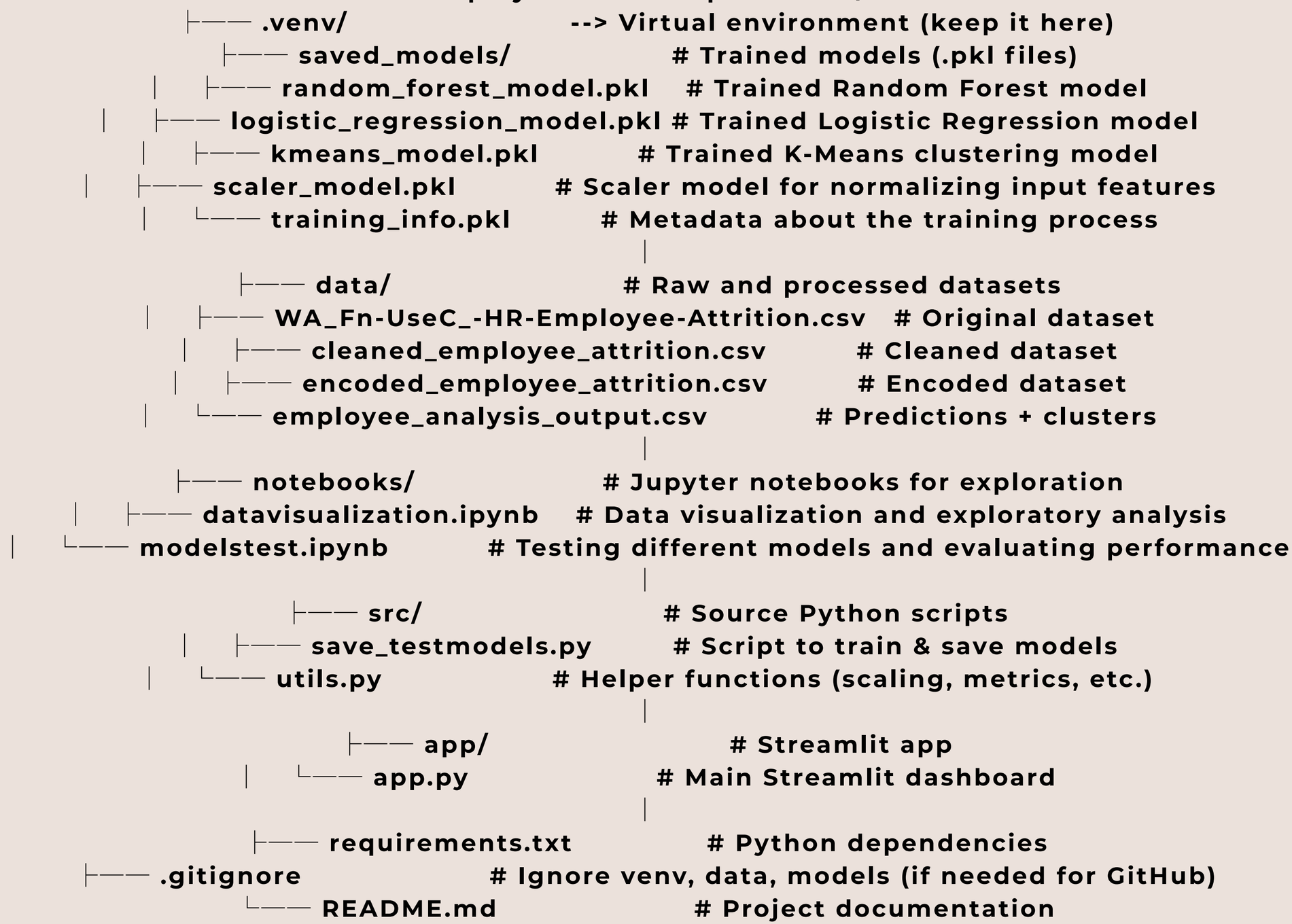
Presented by: SAID Salma
Supervised by: Mr. KHAMJANE Aziz

1. Introduction

Employee attrition is when employees leave a company, either voluntarily or involuntarily. High attrition can affect productivity, increase recruitment costs, and lower team morale.
The goal of this project is to build a predictive machine learning model that identifies employees who are at risk of leaving. This helps HR take proactive retention actions.

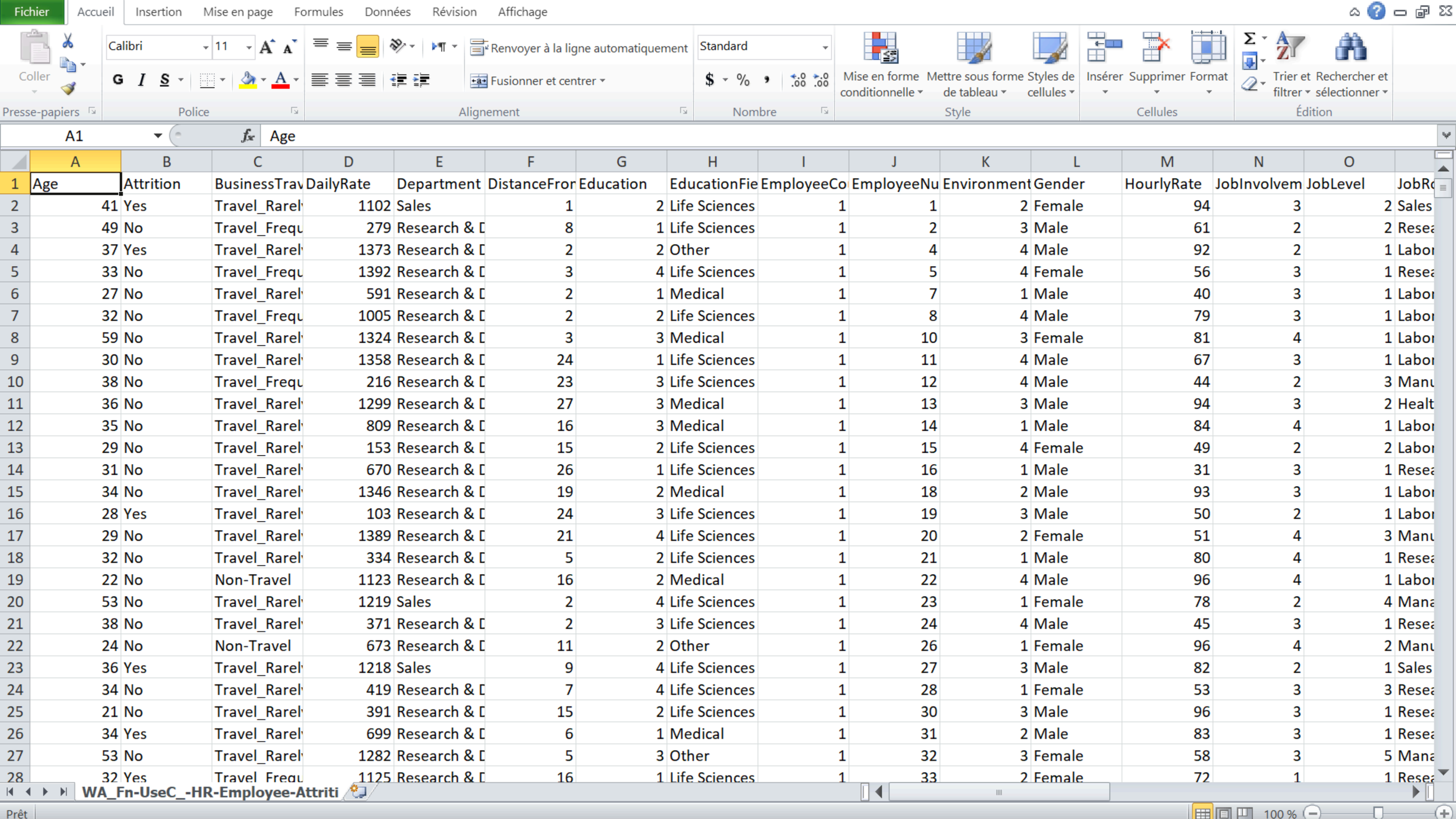
Project Structure




employee-attrition-prediction/
├── .venv/ --> Virtual environment (keep it here)
├── saved_models/ # Trained models (.pkl files)
│ ├── random_forest_model.pkl # Trained Random Forest model
│ ├── logistic_regression_model.pkl # Trained Logistic Regression model
│ ├── kmeans_model.pkl # Trained K-Means clustering model
│ ├── scaler_model.pkl # Scaler model for normalizing input features
│ └── training_info.pkl # Metadata about the training process
│
├── data/ # Raw and processed datasets
│ ├── WA_Fn-UseC_-HR-Employee-Attrition.csv # Original dataset
│ ├── cleaned_employee_attrition.csv # Cleaned dataset
│ ├── encoded_employee_attrition.csv # Encoded dataset
│ └── employee_analysis_output.csv # Predictions + clusters
│
├── notebooks/ # Jupyter notebooks for exploration
│ ├── datavisualization.ipynb # Data visualization and exploratory analysis
│ └── modelstest.ipynb # Testing different models and evaluating performance
│
├── src/ # Source Python scripts
│ ├── save_testmodels.py # Script to train & save models
│ └── utils.py # Helper functions (scaling, metrics, etc.)
│
├── app/ # Streamlit app
│ └── app.py # Main Streamlit dashboard
│
├── requirements.txt # Python dependencies
├── .gitignore # Ignore venv, data, models (if needed for GitHub)
└── README.md # Project documentation

2. Dataset Description

The dataset used is the IBM HR Analytics Employee Attrition dataset. It contains around 1,470 employee records.
The target variable is Attrition (Yes = 1, No = 0). The dataset includes demographic, job-related, compensation, and worklife balance features. Around 16% of employees left the company, showing class imbalance.



Fichier Accueil Insertion Mise en page Formules Données Révision Affichage

Coller Calibri 11 Renvoyer à la ligne automatiquement Standard Mise en forme conditionnelle Mettre sous forme de tableau Styles de cellules Insérer Supprimer Format Trier et filtrer Rechercher et sélectionner

Presse-papiers Police Alignement Nombre Style Cellules Édition

A1 Age

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Age	Attrition	BusinessTrav	DailyRate	Department	DistanceFror	Education	EducationFie	EmployeeCo	EmployeeNu	Environment	Gender	HourlyRate	JobInvolvem	JobLevel	JobRo
2	41	Yes	Travel_Rarel	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales
3	49	No	Travel_Frequ	279	Research & D	8	1	Life Sciences	1	2	3	Male	61	2	2	Resea
4	37	Yes	Travel_Rarel	1373	Research & D	2	2	Other	1	4	4	Male	92	2	1	Labor
5	33	No	Travel_Frequ	1392	Research & D	3	4	Life Sciences	1	5	4	Female	56	3	1	Resea
6	27	No	Travel_Rarel	591	Research & D	2	1	Medical	1	7	1	Male	40	3	1	Labor
7	32	No	Travel_Frequ	1005	Research & D	2	2	Life Sciences	1	8	4	Male	79	3	1	Labor
8	59	No	Travel_Rarel	1324	Research & D	3	3	Medical	1	10	3	Female	81	4	1	Labor
9	30	No	Travel_Rarel	1358	Research & D	24	1	Life Sciences	1	11	4	Male	67	3	1	Labor
10	38	No	Travel_Frequ	216	Research & D	23	3	Life Sciences	1	12	4	Male	44	2	3	Manu
11	36	No	Travel_Rarel	1299	Research & D	27	3	Medical	1	13	3	Male	94	3	2	Healt
12	35	No	Travel_Rarel	809	Research & D	16	3	Medical	1	14	1	Male	84	4	1	Labor
13	29	No	Travel_Rarel	153	Research & D	15	2	Life Sciences	1	15	4	Female	49	2	2	Labor
14	31	No	Travel_Rarel	670	Research & D	26	1	Life Sciences	1	16	1	Male	31	3	1	Resea
15	34	No	Travel_Rarel	1346	Research & D	19	2	Medical	1	18	2	Male	93	3	1	Labor
16	28	Yes	Travel_Rarel	103	Research & D	24	3	Life Sciences	1	19	3	Male	50	2	1	Labor
17	29	No	Travel_Rarel	1389	Research & D	21	4	Life Sciences	1	20	2	Female	51	4	3	Manu
18	32	No	Travel_Rarel	334	Research & D	5	2	Life Sciences	1	21	1	Male	80	4	1	Resea
19	22	No	Non-Travel	1123	Research & D	16	2	Medical	1	22	4	Male	96	4	1	Labor
20	53	No	Travel_Rarel	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4	Mana
21	38	No	Travel_Rarel	371	Research & D	2	3	Life Sciences	1	24	4	Male	45	3	1	Resea
22	24	No	Non-Travel	673	Research & D	11	2	Other	1	26	1	Female	96	4	2	Manu
23	36	Yes	Travel_Rarel	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1	Sales
24	34	No	Travel_Rarel	419	Research & D	7	4	Life Sciences	1	28	1	Female	53	3	3	Resea
25	21	No	Travel_Rarel	391	Research & D	15	2	Life Sciences	1	30	3	Male	96	3	1	Resea
26	34	Yes	Travel_Rarel	699	Research & D	6	1	Medical	1	31	2	Male	83	3	1	Resea
27	53	No	Travel_Rarel	1282	Research & D	5	3	Other	1	32	3	Female	58	3	5	Mana
28	32	Yes	Travel_Frequ	1125	Research & D	16	1	Life Sciences	1	33	2	Female	72	1	1	Resea

WA_Fn-UseC_-HR-Employee-Attriti

Prêt 100 %

3. Data Preprocessing

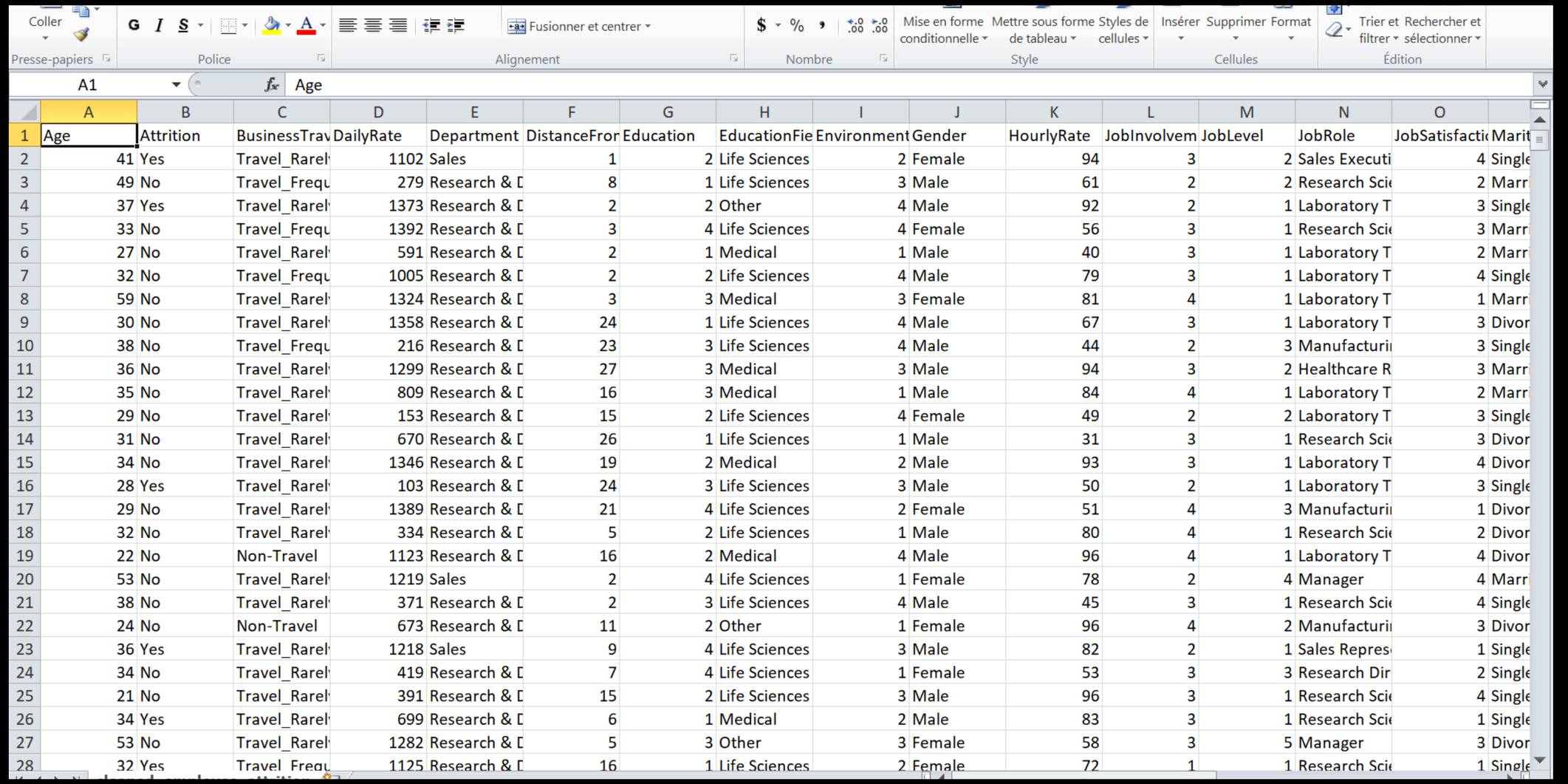
The data was cleaned by removing irrelevant columns such as EmployeeNumber and StandardHours. No missing values were found.

Categorical variables were encoded (e.g., OverTime: Yes = 1, No = 0). One-hot encoding was applied to multi-category features like JobRole.

A correlation matrix was computed to study the relationship between features and the target variable (Attrition). Features with very weak correlation (close to zero) were considered less useful and removed to reduce noise. This step helped keep only the most relevant predictors for the models.

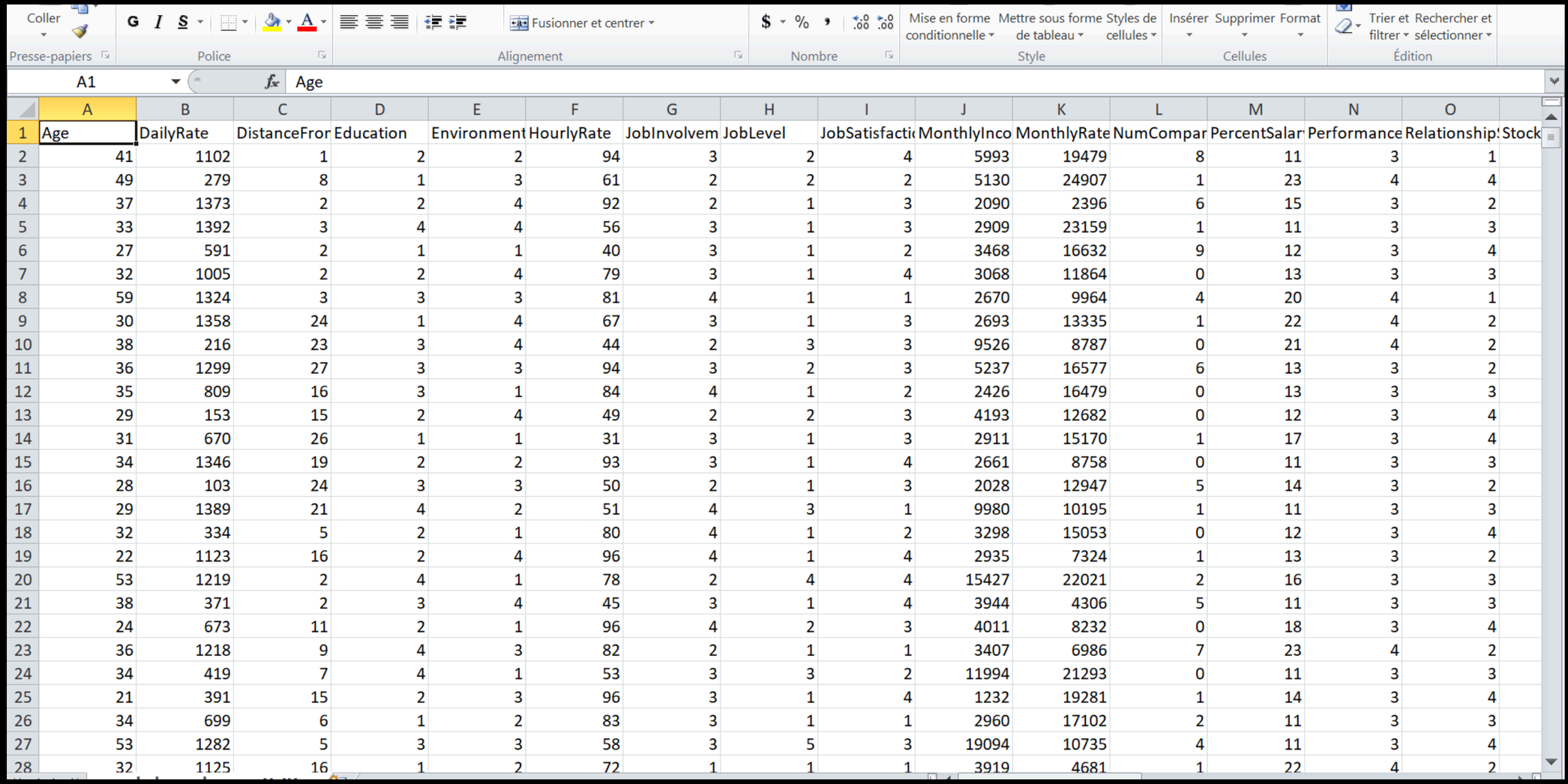
Continuous variables were then scaled using StandardScaler to prepare for clustering and improve model performance.

Cleaned-data



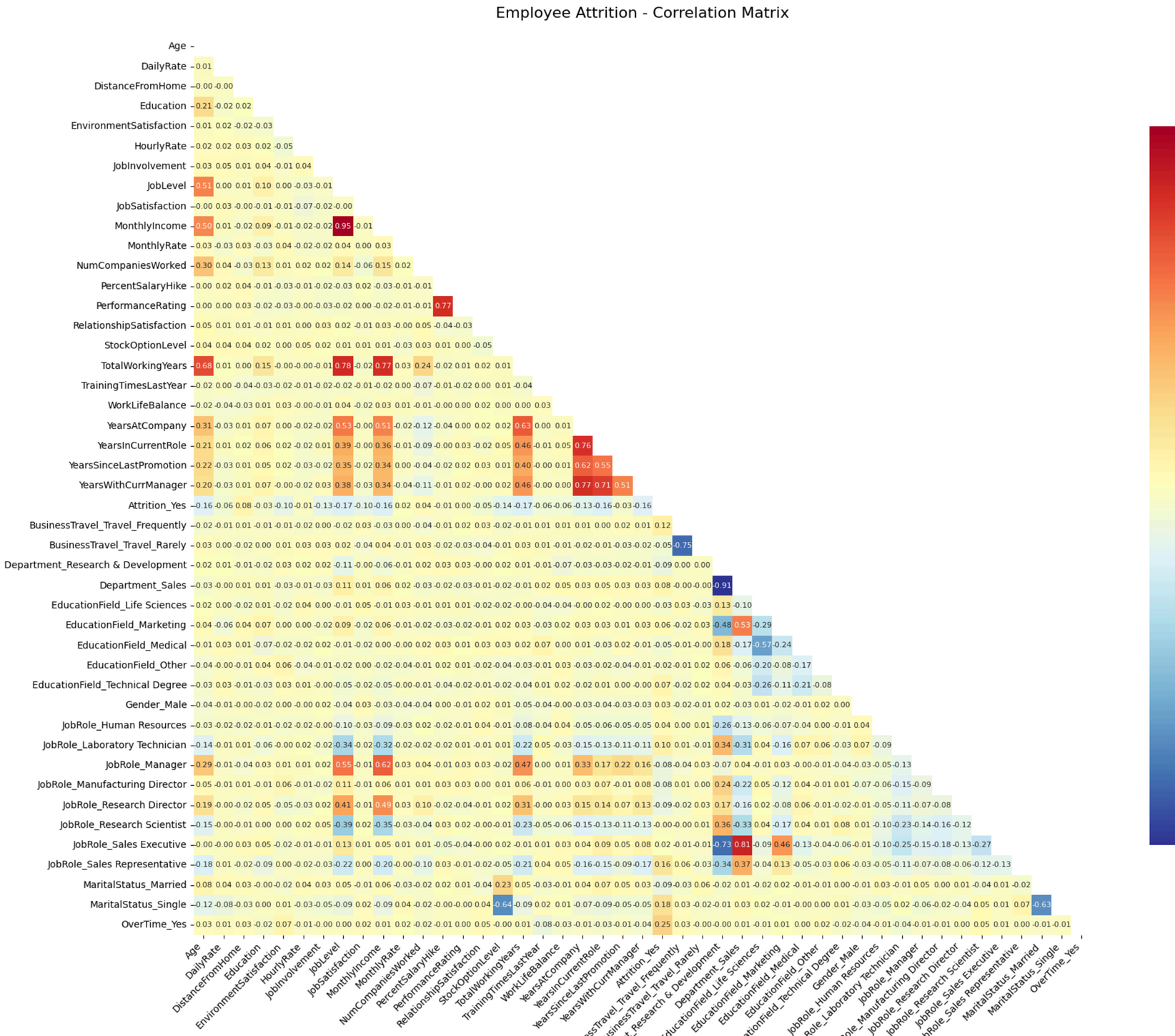
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Age	Attrition	BusinessTrav	DailyRate	Department	DistanceFror	Education	EducationFie	Environment	Gender	HourlyRate	JobInvolvem	JobLevel	JobRole	JobSatisfacti	Marit
2	41	Yes	Travel_Rarel	1102	Sales	1	2	Life Sciences	2	Female	94	3	2	Sales Executi	4	Single
3	49	No	Travel_Frequ	279	Research & D	8	1	Life Sciences	3	Male	61	2	2	Research Scie	2	Marri
4	37	Yes	Travel_Rarel	1373	Research & D	2	2	Other	4	Male	92	2	1	Laboratory T	3	Single
5	33	No	Travel_Frequ	1392	Research & D	3	4	Life Sciences	4	Female	56	3	1	Research Scie	3	Marri
6	27	No	Travel_Rarel	591	Research & D	2	1	Medical	1	Male	40	3	1	Laboratory T	2	Marri
7	32	No	Travel_Frequ	1005	Research & D	2	2	Life Sciences	4	Male	79	3	1	Laboratory T	4	Single
8	59	No	Travel_Rarel	1324	Research & D	3	3	Medical	3	Female	81	4	1	Laboratory T	1	Marri
9	30	No	Travel_Rarel	1358	Research & D	24	1	Life Sciences	4	Male	67	3	1	Laboratory T	3	Divor
10	38	No	Travel_Frequ	216	Research & D	23	3	Life Sciences	4	Male	44	2	3	Manufacturi	3	Single
11	36	No	Travel_Rarel	1299	Research & D	27	3	Medical	3	Male	94	3	2	Healthcare R	3	Marri
12	35	No	Travel_Rarel	809	Research & D	16	3	Medical	1	Male	84	4	1	Laboratory T	2	Marri
13	29	No	Travel_Rarel	153	Research & D	15	2	Life Sciences	4	Female	49	2	2	Laboratory T	3	Single
14	31	No	Travel_Rarel	670	Research & D	26	1	Life Sciences	1	Male	31	3	1	Research Scie	3	Divor
15	34	No	Travel_Rarel	1346	Research & D	19	2	Medical	2	Male	93	3	1	Laboratory T	4	Divor
16	28	Yes	Travel_Rarel	103	Research & D	24	3	Life Sciences	3	Male	50	2	1	Laboratory T	3	Single
17	29	No	Travel_Rarel	1389	Research & D	21	4	Life Sciences	2	Female	51	4	3	Manufacturi	1	Divor
18	32	No	Travel_Rarel	334	Research & D	5	2	Life Sciences	1	Male	80	4	1	Research Scie	2	Divor
19	22	No	Non-Travel	1123	Research & D	16	2	Medical	4	Male	96	4	1	Laboratory T	4	Divor
20	53	No	Travel_Rarel	1219	Sales	2	4	Life Sciences	1	Female	78	2	4	Manager	4	Marri
21	38	No	Travel_Rarel	371	Research & D	2	3	Life Sciences	4	Male	45	3	1	Research Scie	4	Single
22	24	No	Non-Travel	673	Research & D	11	2	Other	1	Female	96	4	2	Manufacturi	3	Divor
23	36	Yes	Travel_Rarel	1218	Sales	9	4	Life Sciences	3	Male	82	2	1	Sales Repres	1	Single
24	34	No	Travel_Rarel	419	Research & D	7	4	Life Sciences	1	Female	53	3	3	Research Dir	2	Single
25	21	No	Travel_Rarel	391	Research & D	15	2	Life Sciences	3	Male	96	3	1	Research Scie	4	Single
26	34	Yes	Travel_Rarel	699	Research & D	6	1	Medical	2	Male	83	3	1	Research Scie	1	Single
27	53	No	Travel_Rarel	1282	Research & D	5	3	Other	3	Female	58	3	5	Manager	3	Divor
28	32	Yes	Travel_Frequ	1125	Research & D	16	1	Life Sciences	2	Female	72	1	1	Research Scie	1	Single

Encoded-data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	Age	DailyRate	DistanceFror	Education	Environment	HourlyRate	JobInvolvem	JobLevel	JobSatisfacti	MonthlyInco	MonthlyRate	NumCompar	PercentSalar	Performance	Relationship	Stock
2	41	1102	1	2	2	94	3	2	4	5993	19479	8	11	3	1	
3	49	279	8	1	3	61	2	2	2	5130	24907	1	23	4	4	
4	37	1373	2	2	4	92	2	1	3	2090	2396	6	15	3	2	
5	33	1392	3	4	4	56	3	1	3	2909	23159	1	11	3	3	
6	27	591	2	1	1	40	3	1	2	3468	16632	9	12	3	4	
7	32	1005	2	2	4	79	3	1	4	3068	11864	0	13	3	3	
8	59	1324	3	3	3	81	4	1	1	2670	9964	4	20	4	1	
9	30	1358	24	1	4	67	3	1	3	2693	13335	1	22	4	2	
10	38	216	23	3	4	44	2	3	3	9526	8787	0	21	4	2	
11	36	1299	27	3	3	94	3	2	3	5237	16577	6	13	3	2	
12	35	809	16	3	1	84	4	1	2	2426	16479	0	13	3	3	
13	29	153	15	2	4	49	2	2	3	4193	12682	0	12	3	4	
14	31	670	26	1	1	31	3	1	3	2911	15170	1	17	3	4	
15	34	1346	19	2	2	93	3	1	4	2661	8758	0	11	3	3	
16	28	103	24	3	3	50	2	1	3	2028	12947	5	14	3	2	
17	29	1389	21	4	2	51	4	3	1	9980	10195	1	11	3	3	
18	32	334	5	2	1	80	4	1	2	3298	15053	0	12	3	4	
19	22	1123	16	2	4	96	4	1	4	2935	7324	1	13	3	2	
20	53	1219	2	4	1	78	2	4	4	15427	22021	2	16	3	3	
21	38	371	2	3	4	45	3	1	4	3944	4306	5	11	3	3	
22	24	673	11	2	1	96	4	2	3	4011	8232	0	18	3	4	
23	36	1218	9	4	3	82	2	1	1	3407	6986	7	23	4	2	
24	34	419	7	4	1	53	3	3	2	11994	21293	0	11	3	3	
25	21	391	15	2	3	96	3	1	4	1232	19281	1	14	3	4	
26	34	699	6	1	2	83	3	1	1	2960	17102	2	11	3	3	
27	53	1282	5	3	3	58	3	5	3	19094	10735	4	11	3	4	
28	32	1125	16	1	2	72	1	1	1	3919	4681	1	22	4	2	

Correlation-matrix :



EXPLORER

OPEN EDITORS

datavisualizatio...

EMPLOYEE-ATTRITION-PR...

.conda
app
app.py M
data
notebooks
datavisualization...
modelstest.ipynb
pic
PowerBi
saved_models
src
save_testmodels...
utils.py
.gitignore
.venv
README.md
requirements.txt

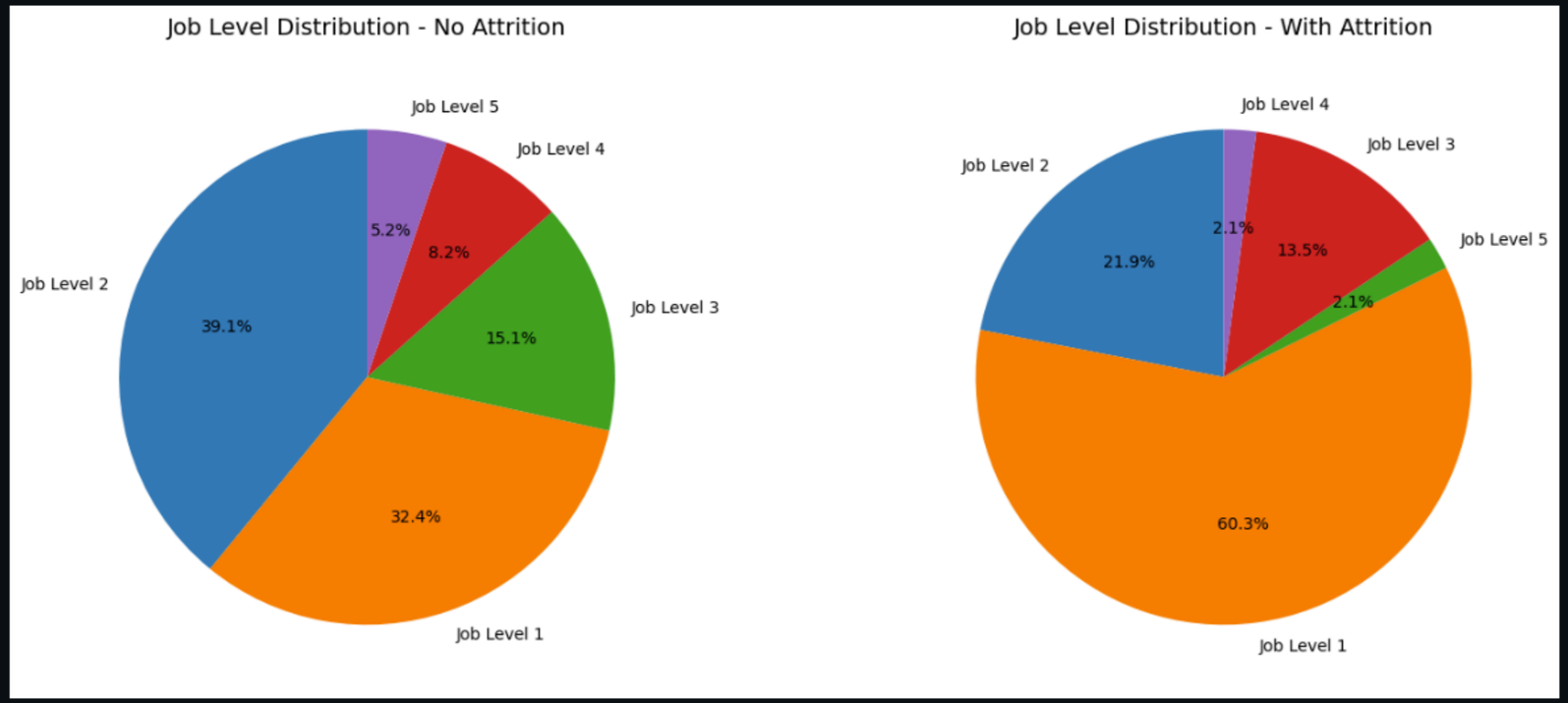
OUTLINE
TIMELINE
DEEPSEEK

datavisualization.ipynb

notebooks > datavisualization.ipynb > correlation_matrix = df_encoded.corr()

Generate + Code + Markdown | Run All Clear All Outputs | Outline ... Select Kernel

```
[2, 1, 1, 1, 1, 1, 3, 2, 1, 2, 1, 1, 3, 1, 1, 4, 1, 2, 3, 1, 5, 2, 3, 5, 1, 2, 1, 1, 1, 1, 2, 1, 1, 3, 2, 2
[2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 5, 3, 1, 1, 3, 1, 1, 2, 2, 1, 2, 3, 1, 2, 3, 1, 1, 1, 1, 2, 2, 3, 3, 1, 2, 1
1470
Counter({2: 482, 1: 400, 3: 186, 4: 101, 5: 64})
Counter({1: 143, 2: 52, 3: 32, 5: 5, 4: 5})
```




```
Job Level Distribution - No Attrition:
  Level 1: 400 employees
  Level 2: 482 employees
  Level 3: 186 employees
```

Training models.

With the progress bar, they see:

```
Training Random
Training Logisti
Training K-Means
Saving models...
✅ Models train
```

The `progress_bar.progress(100)` line specifically signals **"Mission Accomplished!"** to the user interface.

now explain me the code from line 111

Add Context...

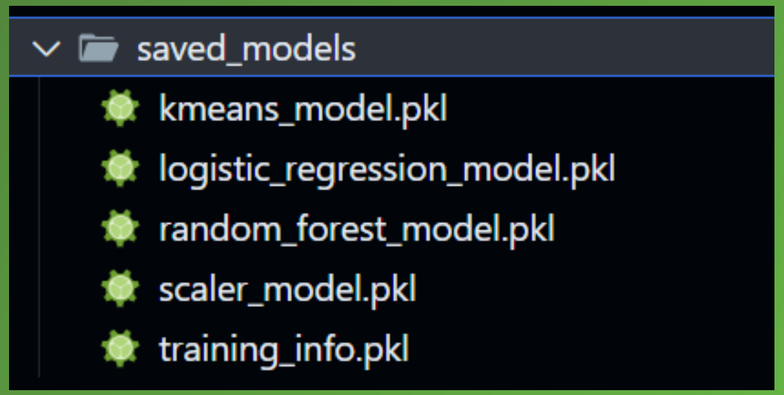
datavisualizatio +

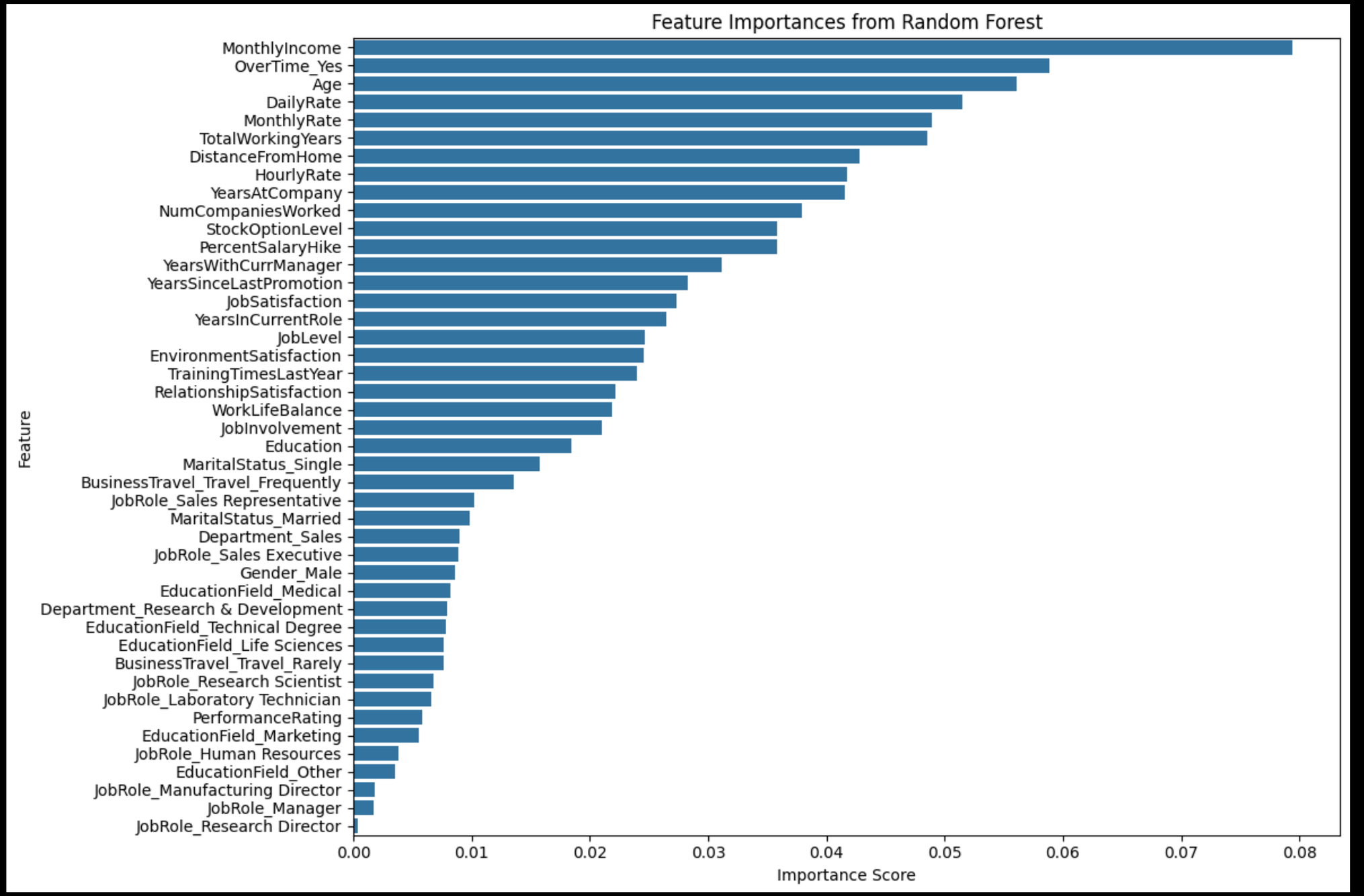
Add context (#), exter

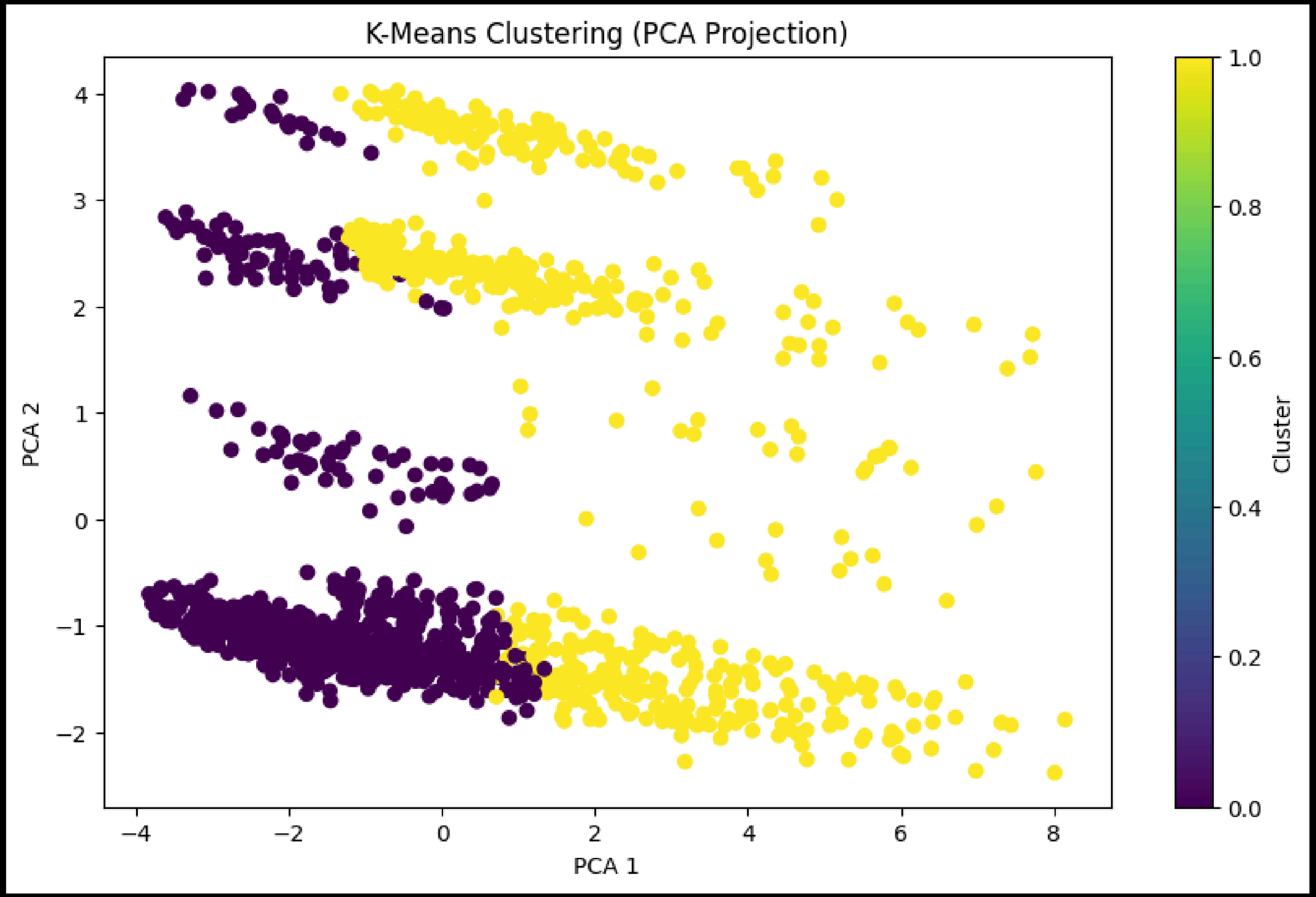
main* 0 0 Connect Spaces: 4 Cell 5 of 7 Go Live

4. Model Building

Two supervised learning models were used: Logistic Regression and Random Forest. Logistic Regression was chosen for its simplicity and interpretability, while Random Forest was used for its ability to handle complex patterns and feature importance. Unsupervised learning with K-Means clustering was also applied. Although it does not predict attrition directly, it groups employees into clusters, providing extra insights into hidden patterns.




Feature Importances from Random Forest
MonthlyIncome
OverTime_Yes
Age
DailyRate
MonthlyRate
TotalWorkingYears
DistanceFromHome
HourlyRate
YearsAtCompany
NumCompaniesWorked
StockOptionLevel
PercentSalaryHike
YearsWithCurrManager
YearsSinceLastPromotion
JobSatisfaction
YearsInCurrentRole
JobLevel
EnvironmentSatisfaction
TrainingTimesLastYear
RelationshipSatisfaction
WorkLifeBalance
JobInvolvement
Education
MaritalStatus_Single
BusinessTravel_Travel_Frequently
JobRole_Sales Representative
MaritalStatus_Married
Department_Sales
JobRole_Sales Executive
Gender_Male
EducationField_Medical
Department_Research & Development
EducationField_Technical Degree
EducationField_Life Sciences
BusinessTravel_Travel_Rarely
JobRole_Research Scientist
JobRole_Laboratory Technician
PerformanceRating
EducationField_Marketing
JobRole_Human Resources
EducationField_Other
JobRole_Manufacturing Director
JobRole_Manager
JobRole_Research Director
Feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Importance Score


K-Means Clustering (PCA Projection)
PCA 1
PCA 2
Cluster
−4
−2
0
2
4
6
8
4
3
2
1
0
−1
−2
1.0
0.8
0.6
0.4
0.2
0.0

5. Model Evaluation

The models were evaluated using accuracy, precision, recall, and F1-score. Cross-validation was performed to ensure robustness.
Random Forest achieved higher recall and F1-score compared to Logistic Regression, making it more reliable for detecting employees at risk of leaving. Confusion matrices were used to visualize prediction performance.

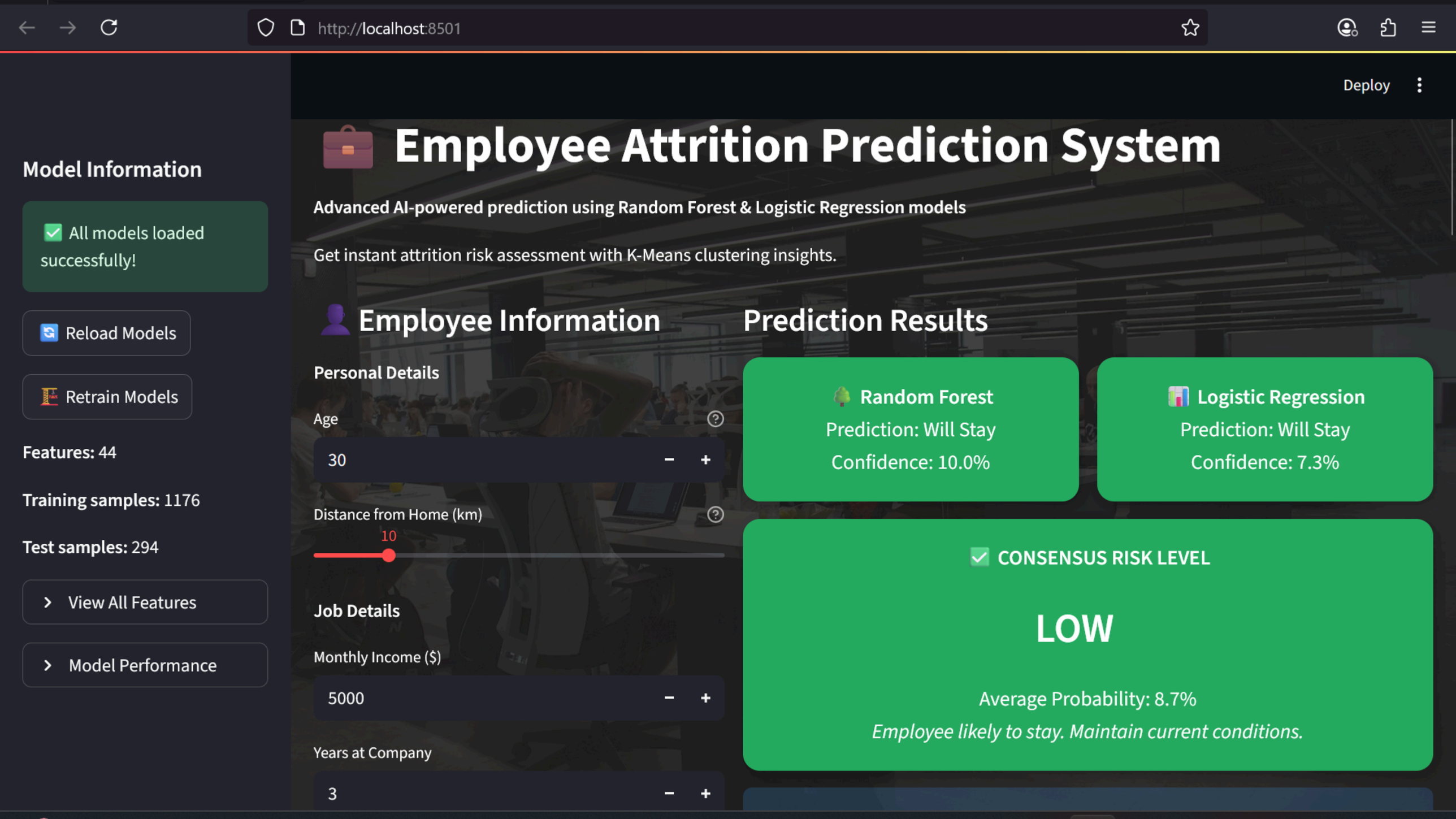
6. Results & Insights

- Random Forest outperformed Logistic Regression, especially in detecting attrition cases.
- Feature importance analysis showed that OverTime, MonthlyIncome, JobSatisfaction, and YearsAtCompany were the most influential factors.
- K-Means clustering revealed two main groups of employees, one with high attrition risk and another with low risk.

These insights confirm that workload, compensation, and job satisfaction play key roles in attrition.

7. Deployment

The model was deployed using Streamlit, an interactive web app framework. Users can enter employee details through a form and instantly get predictions with probabilities.
The app also visualizes model outputs with 3D-style charts, feature importance, and risk levels, making it accessible to HR professionals without technical background.


http://localhost:8501
Deploy
Model Information
✅ All models loaded successfully!
Reload Models
Retrain Models
Features: 44
Training samples: 1176
Test samples: 294
View All Features
Model Performance
Employee Attrition Prediction System
Advanced AI-powered prediction using Random Forest & Logistic Regression models
Get instant attrition risk assessment with K-Means clustering insights.
Employee Information
Personal Details
Age
30
Distance from Home (km)
10
Job Details
Monthly Income ($)
5000
Years at Company
3
Prediction Results
Random Forest
Prediction: Will Stay
Confidence: 10.0%
Logistic Regression
Prediction: Will Stay
Confidence: 7.3%
CONSENSUS RISK LEVEL
LOW
Average Probability: 8.7%
Employee likely to stay. Maintain current conditions.

http://localhost:8501

Deploy

3

Satisfaction & Work-Life

Job Satisfaction

3

Work-Life Balance

3

Works Overtime

No

Advanced Options

Predict Attrition Risk

Analysis Dashboard

Key Factors Analysis

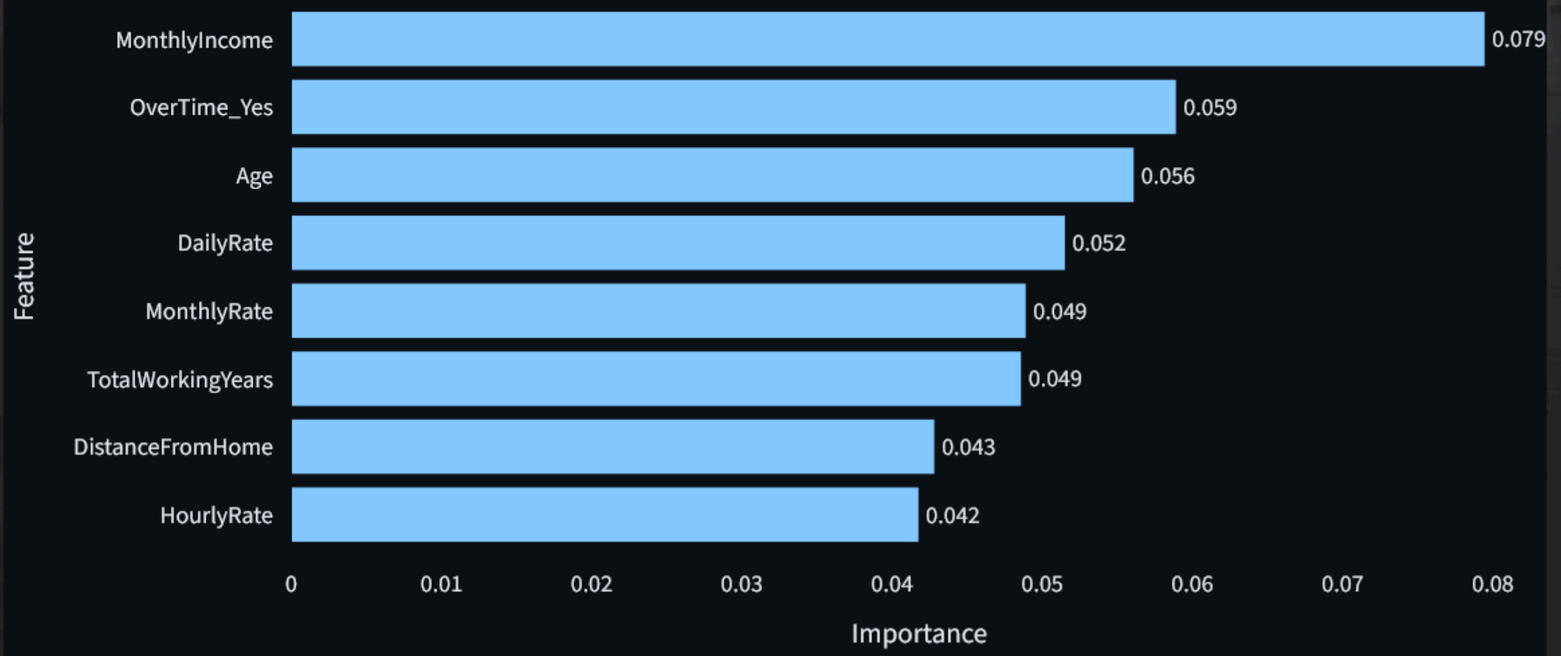
http://localhost:8501

Deploy

Predict Attrition Risk

🔍 Key Factors Analysis

🎯 Top Factors Influencing Prediction



Employee's Values for Key Factors:

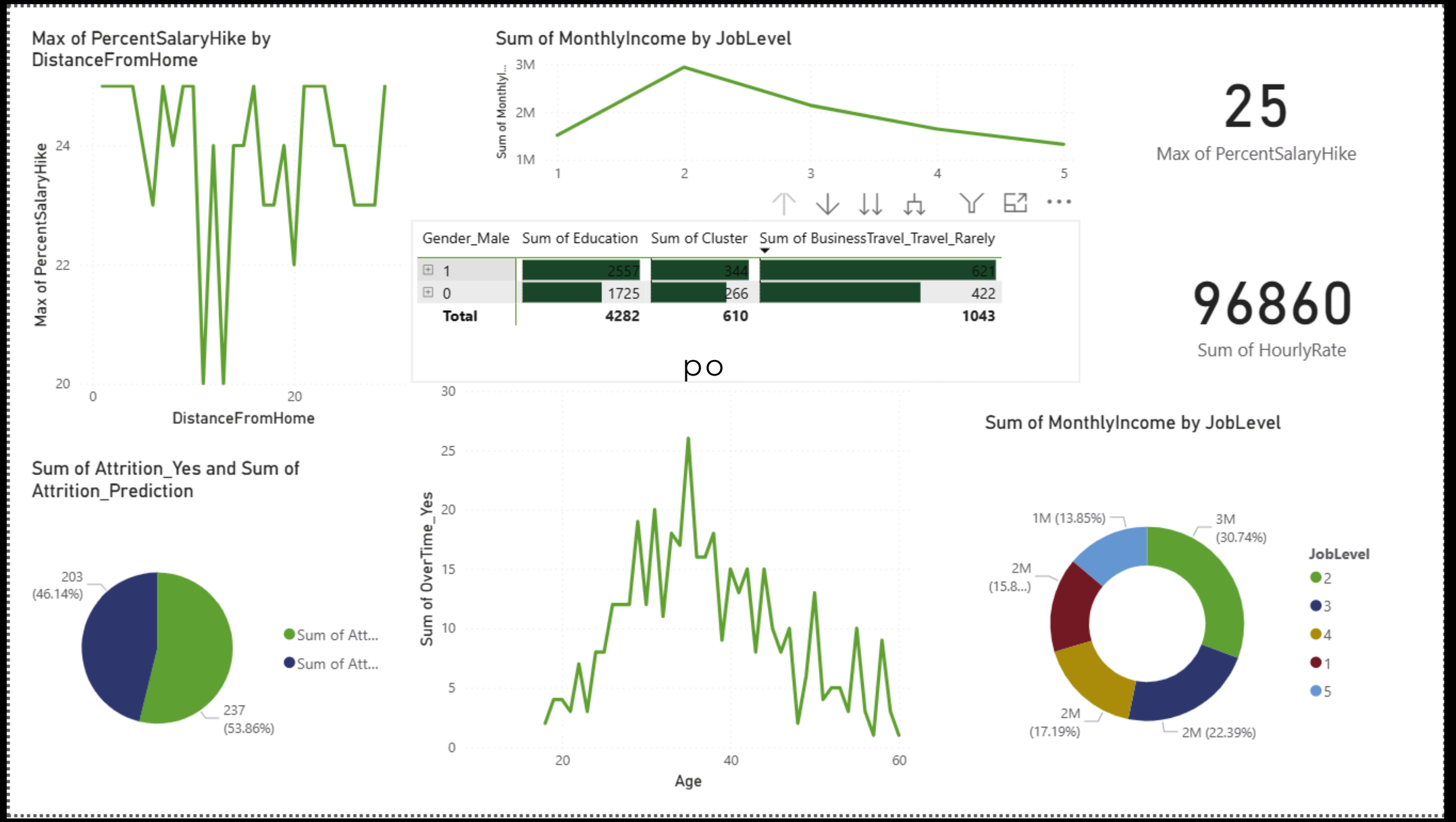
	Factor	Employee Value	Model Importance
9	MonthlyIncome	5000	0.079
43	OverTime_Yes	0	0.059
0	Age	30	0.056
1	DailyRate	0	0.052
10	MonthlyRate	0	0.049
16	TotalWorkingYears	8	0.049
2	DistanceFromHome	10	0.043
5	HourlyRate	0	0.042

Recommendations

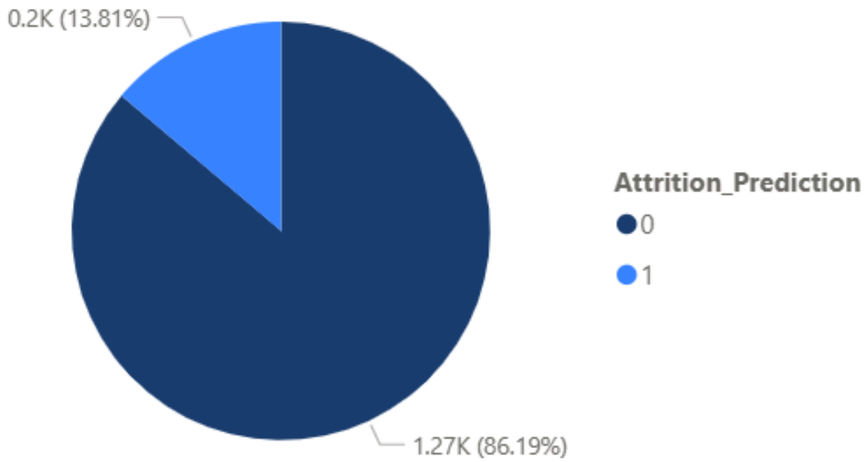
Continue Current Strategy:

- Employee is well-engaged
- Maintain current management approach

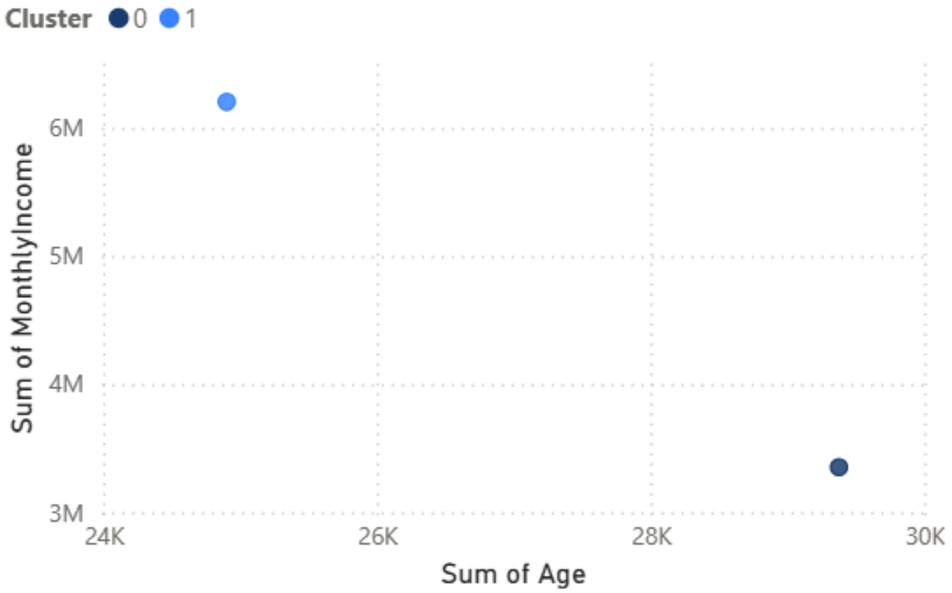
Power BI Visualization Results



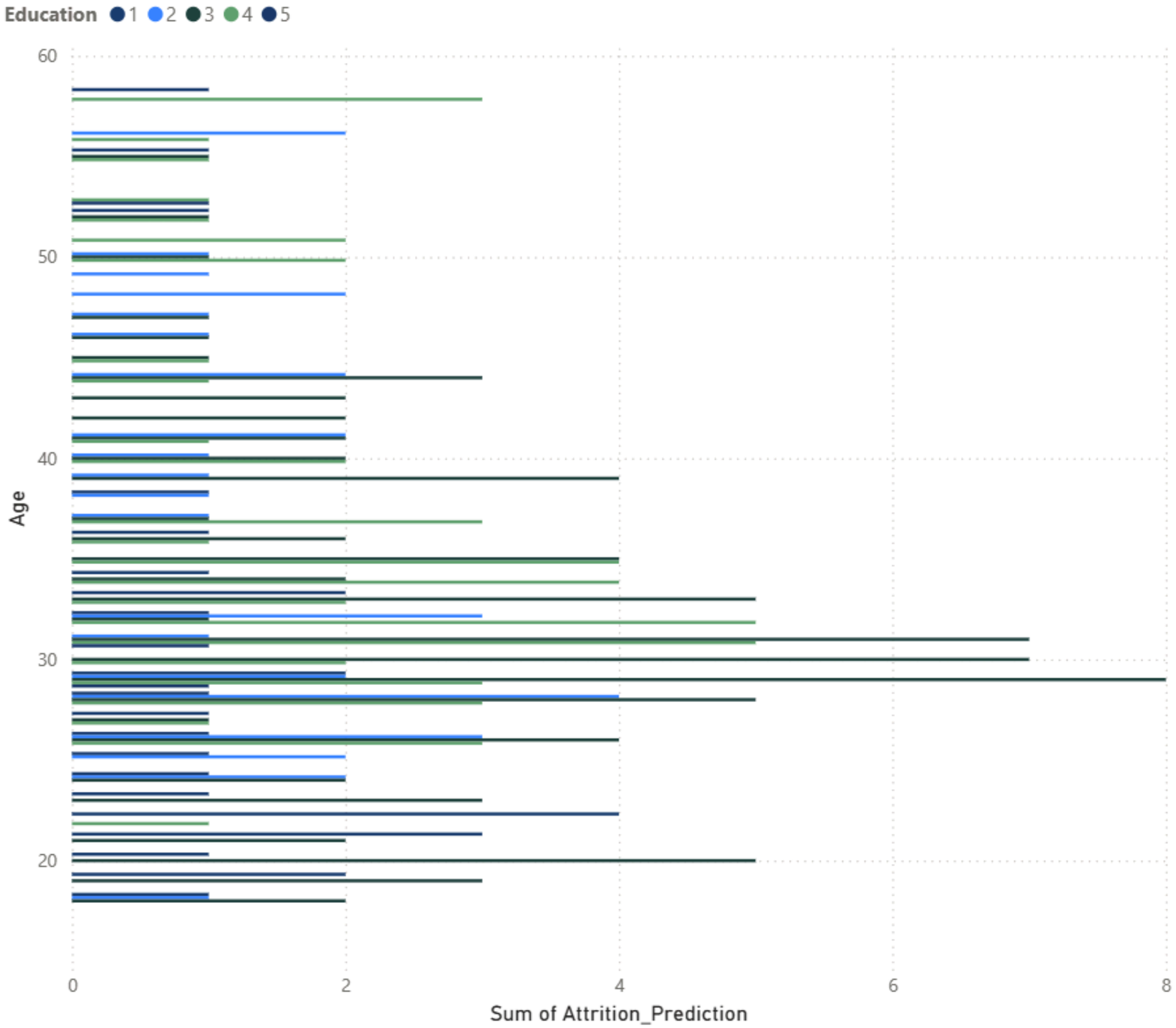
Count of Gender_Male by Attrition_Prediction



Sum of Age and Sum of MonthlyIncome by Cluster



Sum of Attrition_Prediction by Age and Education



8. Conclusion

This project successfully built a predictive system to estimate employee attrition. Random Forest provided the best performance, while K-Means added exploratory insights.
The solution can help HR departments monitor employees, identify high-risk cases, and take early actions to improve retention. Future work could include testing more advanced models and integrating real-time HR data.