Rapport Machine Learning

Made by:

-Oussema Khemiri

-Houda Chambi

-Nourhene Ouhichi

-Wassef Ammar

-Fourat Chawech

Table of Contents

1.Introduction

In today's competitive business landscape, human resources (HR) departments face growing pressure to enhance recruitment efficiency, retain top talent, and foster employee development. Traditional HR practices often rely on manual processes and subjective decision-making, which can lead to inefficiencies and missed opportunities. Machine learning (ML) offers a transformative approach by leveraging data-driven insights to optimize critical HR functions.

This project focuses on harnessing ML techniques to address three core challenges: predicting salary evolution to inform compensation strategies, identifying employees requiring targeted training programs, and proactively detecting attrition risks to mitigate turnover. By analyzing the IBM HR Analytics dataset—which includes variables such as job roles, performance metrics, and employee demographics—we aim to build models that empower HR teams with actionable intelligence. A regression model will forecast future salaries based on experience and performance, clustering will segment employees for personalized development plans, and a classification algorithm will predict attrition likelihood.

Aligned with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, this initiative systematically navigates data exploration, model development, and deployment strategies. The outcomes not only demonstrate the value of ML in HR but also provide scalable solutions to enhance organizational agility and employee satisfaction.

2. Business Understanding :

The Human Resources (HR) department aims to leverage machine learning to address three strategic challenges:

estimating future salaries to inform compensation fairness (BO1), identifying employees requiring targeted training to bridge skill gaps (BO2), and detecting employees at risk of attrition to mitigate turnover (BO3). By analyzing historical employee data, the initiative aims to deliver actionable insights that align with organizational goals.

2.1.Business Objectives:

1. **BO1: Predict Future Salary Evolution**
    - **Purpose**: Enable fair and competitive compensation planning by forecasting salary adjustments based on employee experience, education, and performance.
    - **Business Impact**: Reduce salary discrepancies, retain high performers, and align budgets with talent expectations.
2. **BO2: Identify Employees Needing Additional Training**
    - **Purpose**: Group employees requiring supplemental training (e.g., younger workers with limited experience) and provide tailored recommendations.
    - **Business Impact**: Improve workforce competency, reduce skill gaps, and enhance productivity.
3. **BO3: Detect At-Risk Employees Likely to Leave**
    - **Purpose**: Proactively address attrition by identifying employees at high risk of departure.
    - **Business Impact**: Mitigate turnover costs (e.g., recruitment, onboarding) and retain institutional knowledge.

2.2.Data Science Objectives:

Each business objective maps to a machine learning task using the IBM HR Analytics dataset:

1. **DSO1: Regression Model for Salary Prediction**
    - **Target Variable**: MonthlyIncome (current salary).
    - **Features**: TotalWorkingYears, YearsAtCompany, JobLevel, PerformanceRating, Education, TrainingTimesLastYear.
    - **Technique**: Regression algorithms (e.g., Linear Regression, Random Forest Regressor).
2. **DSO2: Clustering for Employee Segmentation**
    - **Features**: Education, TotalWorkingYears, MonthlyIncome, JobLevel.
    - **Technique**: Unsupervised clustering (e.g., K-Means) to group employees with similar development needs.
3. **DSO3: Classification Model for Attrition Prediction**
    - **Target Variable**: Attrition (binary: Yes/No).
    - **Features**: Age, MonthlyIncome, JobSatisfaction, YearsAtCompany, WorkLifeBalance, OverTime, JobRole.
    - **Technique**: Classification algorithms (e.g., Logistic Regression, XGBoost).

2.3.Success Criteria

To ensure alignment with business objectives, success will be measured as follows:

- **BO1 (Salary Prediction)**:

- *Accuracy*: Model predictions within a reasonable margin of actual salaries
- *Business Impact*: Reduced discrepancies in compensation planning.

- **BO2 (Employee Training Needs)**:
  - *Actionability*: Clear identification of employee groups (e.g., low-experience, high-potential) for targeted training.
  - *Business Impact*: Increased participation in development programs.

- **BO3 (Attrition Risk)**:
  - *Reliability*: Proactive identification of at-risk employees for retention interventions.
  - *Business Impact*: Reduced turnover and associated costs.

3. Data Understanding

This section outlines the dataset's structure, variables, and initial insights to guide modeling decisions.

3.1. Data Sources

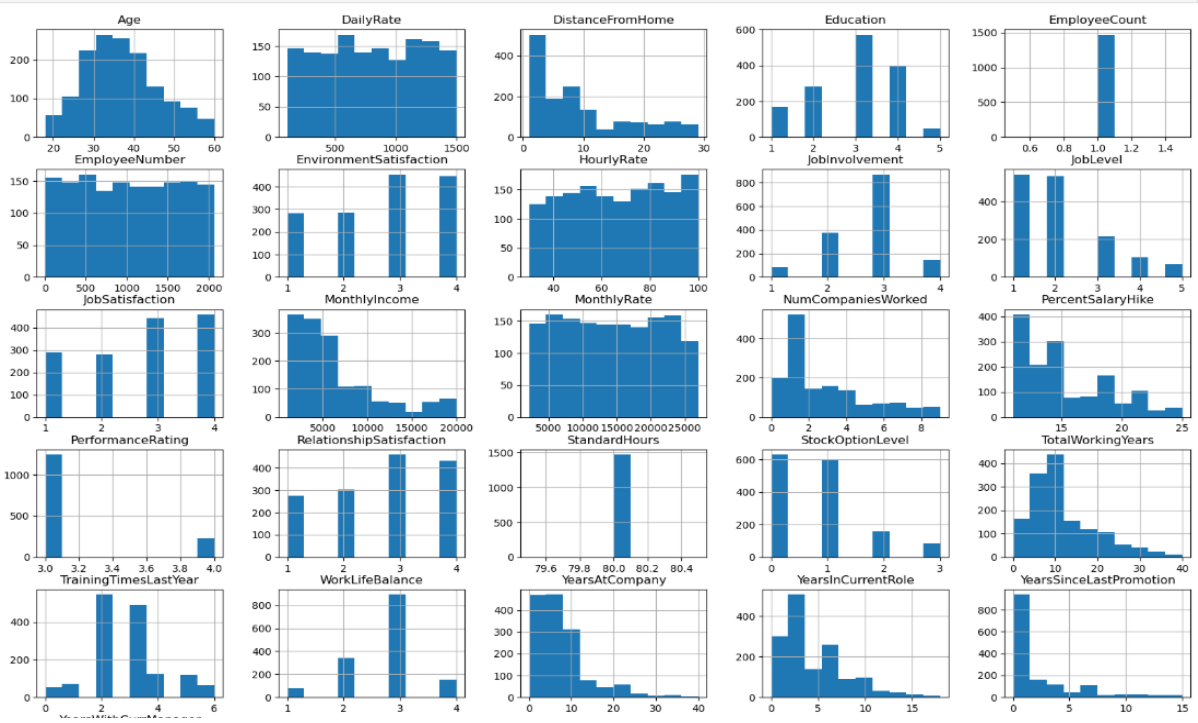
- **Dataset**: IBM HR Analytics Employee Attrition & Performance (Kaggle).
- **Link**: IBM HR Analytics Dataset
- **Scope**:
  - **Entries**: 1,470 employee records.
  - **Variables**: 35 columns, including demographic, job-related, and behavioral attributes.
  - **Purpose**: Supports analysis of attrition, salary trends, and training needs.

3.2.Data Description and Schema

Key Variables by Category:

- **Numerical Features** (26 columns):
    - *Demographic*: Age, DistanceFromHome.
    - *Compensation*: MonthlyIncome, DailyRate, HourlyRate, PercentSalaryHike.
    - *Tenure/Experience*: TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion.
    - *Satisfaction Metrics*: JobSatisfaction, WorkLifeBalance (ordinal scales).
- **Categorical Features** (9 columns):
    - *Job Details*: Department, JobRole, BusinessTravel.
    - *Attrition*: Attrition (target for classification: Yes/No).
    - *Personal Details*: Gender, MaritalStatus, OverTime.

3.3. Exploratory Data Analysis



4.Data Preparation

This phase focuses on refining the dataset for modeling, including cleaning, encoding, and transforming features.

4.1. Data Cleaning

Removed Columns (Non-informative or redundant features):

- **Rationale**:
    - EmployeeNumber: Unique identifier (no predictive value).
    - EmployeeCount, StandardHours, Over18: Constant values across all records.

4.2. Handling Missing Values

- **Completeness Check**:
    - No missing values detected in any column (all 1,470 entries complete).
- **Action**: No imputation required.

4.3. Feature Engineering

Rare Category Grouping:

- Categorical variables (e.g., Department, BusinessTravel) with categories representing <5% of the data were grouped into an "Other" bucket to reduce noise.

```
# Regroupement des catégories rares (<5%)
def group_rare_categories(serie, threshold=0.05):
    counts = serie.value_counts(normalize=True)
    return serie.where(counts > threshold, 'Other')

df['EducationField'] = group_rare_categories(df['EducationField'])
df['JobRole'] = group_rare_categories(df['JobRole'])
```

Binary Encoding:

- Converted categorical targets and features to binary (0/1) for model compatibility.

```
binary_mappings = {
    'Attrition': {'Yes': 1, 'No': 0},
    'Gender': {'Female': 1, 'Male': 0},
    'OverTime': {'Yes': 1, 'No': 0}
}
```

Outlier Treatment:

- Removed extreme values in numerical features (e.g., MonthlyIncome, Age) using the IQR method to reduce skewness.

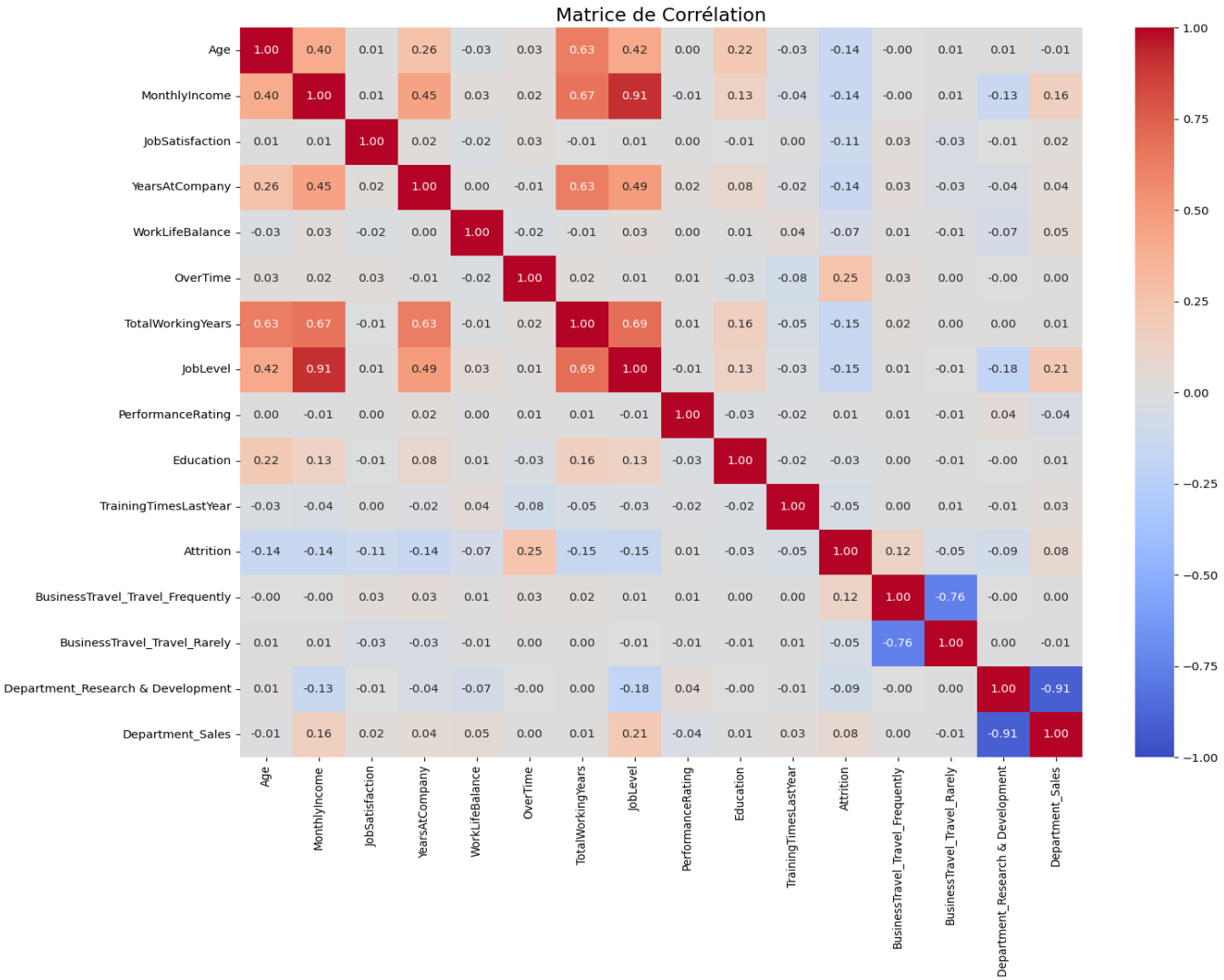
```
# Traitement des outliers
#Éliminer les valeurs extrêmes qui faussent les analyses.
Q1 = df['MonthlyIncome'].quantile(0.25)
Q3 = df['MonthlyIncome'].quantile(0.75)
IQR = Q3 - Q1
df = df[(df['MonthlyIncome'] >= Q1 - 1.5 * IQR) & (df['MonthlyIncome'] <= Q3 + 1.5 * IQR)]
```

Normalization:

- Scaled numerical features (e.g., Age, MonthlyIncome, YearsAtCompany) using StandardScaler to ensure equal weighting in models.

```
# Normalisation des variables numériques
numeric_cols = ['MonthlyIncome', 'TotalWorkingYears', 'Age', 'YearsAtCompany']
scaler = StandardScaler()
df[numeric_cols] = scaler.fit_transform(df[numeric_cols])
```

Correlation Matrix:

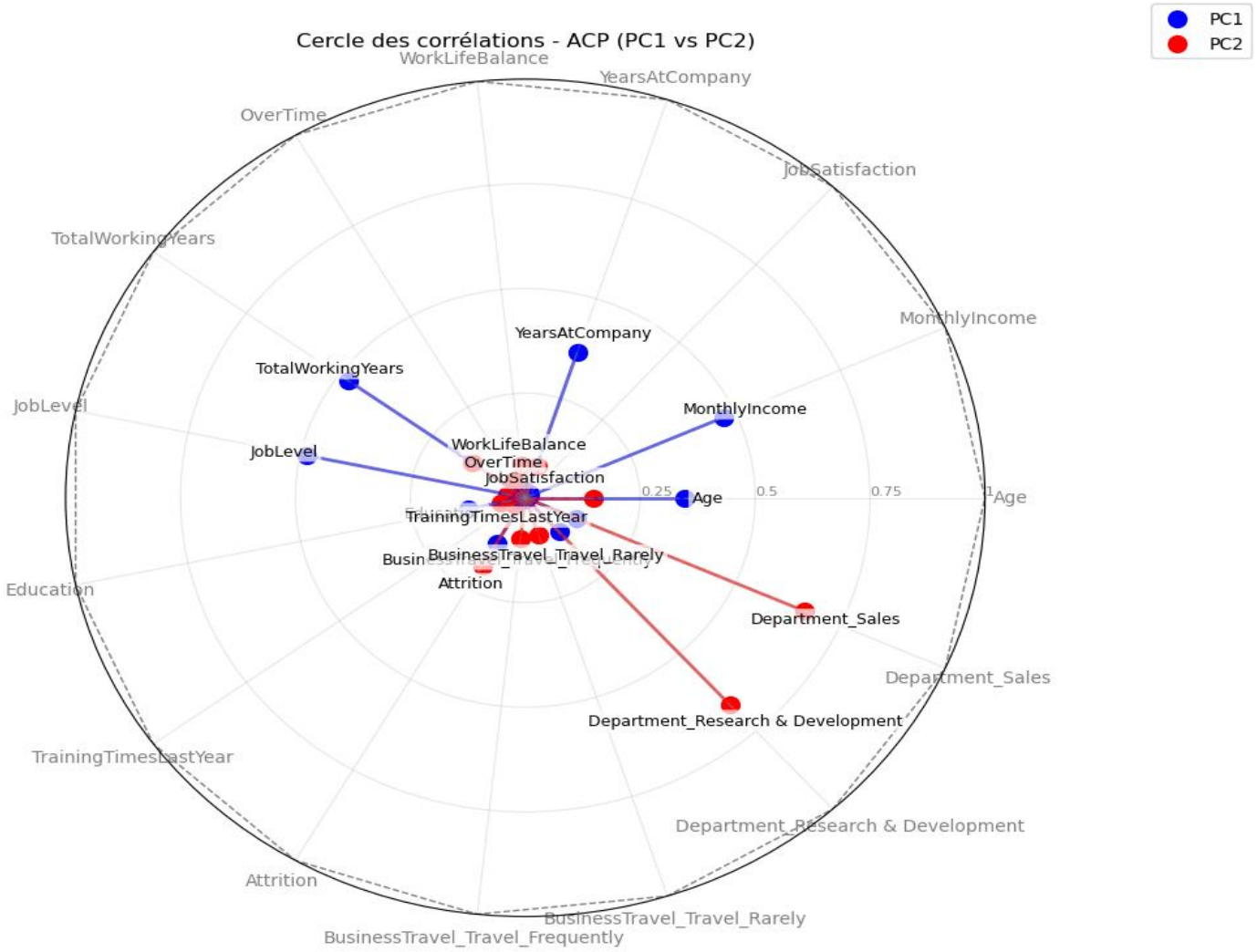


. Dimensionality Reduction (PCA)

Number of Components

	PC1	PC2	PC3
JobLevel	0.486401	-0.041026	-0.004152
TotalWorkingYears	0.475242	0.142788	0.035447
MonthlyIncome	0.473076	-0.009389	-0.008825
YearsAtCompany	0.365461	0.080065	0.050657
Age	0.346345	0.149198	0.013614
Education	0.128047	0.055458	0.012781
Department_Sales	0.121601	-0.663990	-0.107924
WorkLifeBalance	0.015467	-0.077709	-0.012763
JobSatisfaction	0.013507	0.000088	0.038895
BusinessTravel_Travel_Frequently	0.004862	-0.097471	0.691866
OverTime	0.000311	-0.047776	0.067040
BusinessTravel_Travel_Rarely	-0.003063	0.093711	-0.684223
TrainingTimesLastYear	-0.026740	-0.028274	-0.029998
Department_Research & Development	-0.109686	0.665628	0.111206
Attrition	-0.125932	-0.186408	0.133453

Correlation Circle:



PC1 reflects a dimension of professional experience (age, tenure, income, job level).
PC2 captures differences between departments, particularly between Sales and R&D.

Well-represented variables (long arrows close to the circle):

- **MonthlyIncome, Age, TotalWorkingYears, JobLevel** → strongly correlated with PC1 (horizontal axis).
- **Department_Sales, Department_Research & Development** → strongly correlated with PC2 (vertical axis).

5. Modeling

This phase focuses on selecting and training algorithms to address the three data science objectives (DSO1, DSO2, DSO3).

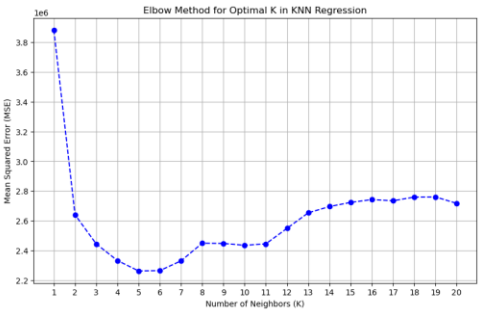
5.1. Algorithms Selected

Objective	Algorithms
DSO1 (Regression)	- K-Nearest Neighbors (KNN)
	- Linear Regression
	- Random Forest Regressor
DSO2 (Clustering)	- K-means
	- spectral
	-ACH
	-GMM
DSO3 (Classification)	-KNN
	- Random Forest
	-SVM
	-Logistic Regression

5.2.Hyperparameter Tuning

- **DSO1 (Regression)**:
    - **KNN**:

      Selecting K value :



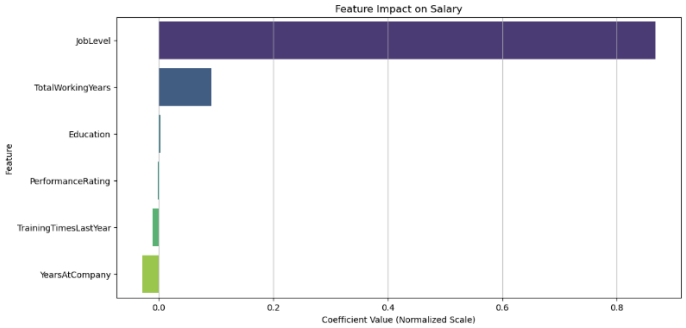
```
KNeighborsRegressor(n_neighbors=5)  # Use 5 nearest neighbors
```

- **Random Forest**:

```
RandomForestRegressor(random_state=42)
```

- **Linear Regression**:

  Feature Analysis:



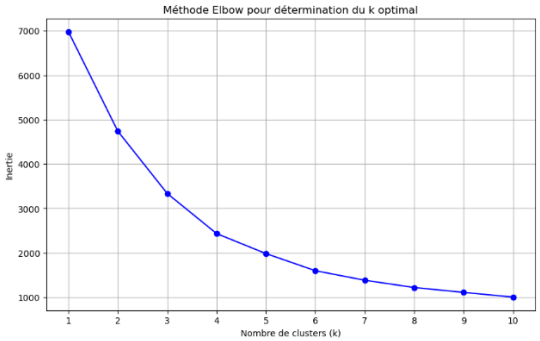
	Features	MSE	R2
0	(JobLevel,)	0.179126	0.833303
1	(TotalWorkingYears, JobLevel)	0.179074	0.833268
2	(YearsAtCompany, JobLevel)	0.179253	0.833185
3	(YearsAtCompany, JobLevel, PerformanceRating)	0.179281	0.833159
4	(JobLevel, TrainingTimesLastYear)	0.179374	0.833072
5	(JobLevel, PerformanceRating, TrainingTimesLas...	0.179396	0.833052
6	(JobLevel, PerformanceRating)	0.179147	0.833283
7	(TotalWorkingYears, JobLevel, PerformanceRating)	0.179436	0.833015
8	(TotalWorkingYears, JobLevel, TrainingTimesLas...	0.179451	0.833001
9	(TotalWorkingYears, JobLevel, PerformanceRatin...	0.179481	0.832973

```
LinearRegression()
```

- **DSO2 (Clustering)**:

  - **K-Means**:

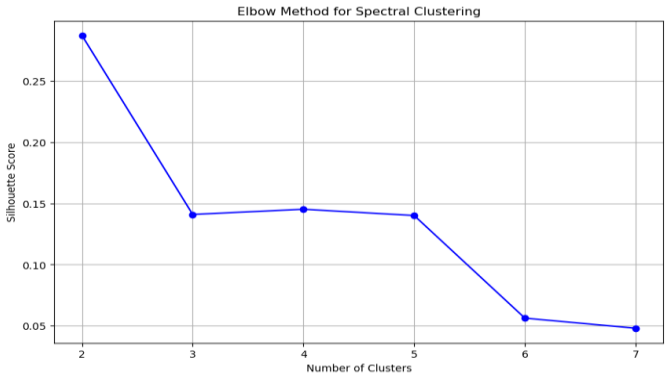
    Elbow for K-means




```
KMeans(n_clusters=3, random_state=42, n_init=20)
```

- **Spectral**:

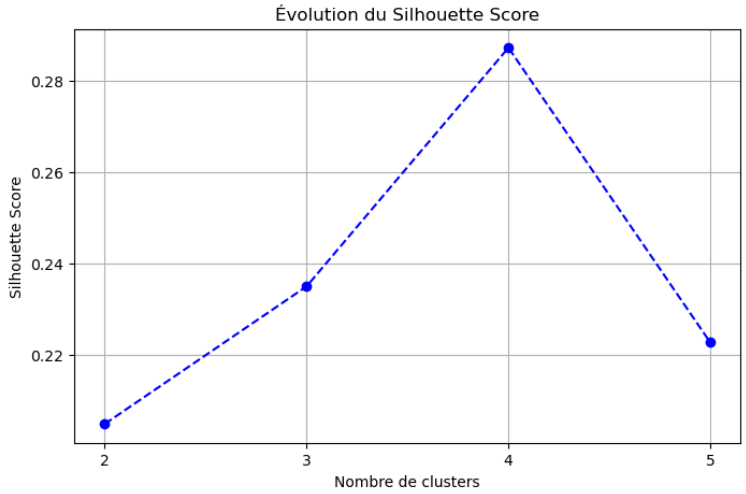
  Elbow for Spectral




```
SpectralClustering(n_clusters=3, random_state=42)
```

- **ACH:**

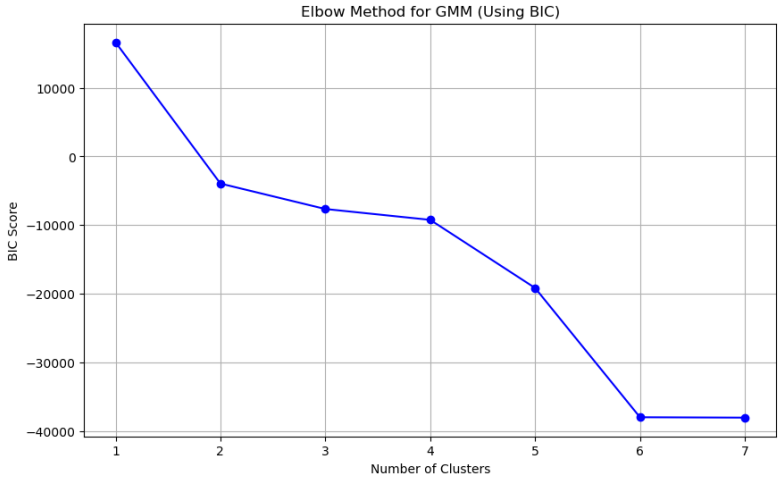
  Clusters:




```
AgglomerativeClustering(n_clusters=3)
```

- **GMM**:

  Elbow For GMM




```
GaussianMixture(n_components=3, random_state=42)
```

- **DSO3 (Classification)**:
  - **Logistic Regression**:

```
LogisticRegression(class_weight='balanced', solver='lbfgs', C=10)
```

  - **SVM:**

```
SVC(class_weight='balanced', C=1, gamma='scale', kernel='rbf')
```

  - **randomForest:**

```
RandomForestClassifier(class_weight='balanced', max_depth=20, min_samples_leaf=4, n_estimators=200)
```

  - **KNN:**

```
KNeighborsClassifier(n_neighbors=9, metric='euclidean')
```

6.Evaluation

6.1. Performance Metrics

DSO1 (Regression):

KNN:

```
MSE: 2,263,013.578 (squared $)
RMSE: $1,504.33
R²: 0.809
```

Random Forest:

```
MSE: 1,668,431.848 (squared $)
RMSE: $1,291.68
R²: 0.859
```

Linear Regression:

```
Model Performance:
MSE: 0.179
R²: 0.833 (83.3% variance explained)
```

DSO2 (Clustering):

K-means:

```
Score de silhouette final pour K-Means: 0.488
```

Spectral:

```
Score de silhouette final pour Spectral: 0.141
```

ACH:

```
Nombre de clusters = 2 → Silhouette Score = 0.20
Nombre de clusters = 3 → Silhouette Score = 0.24
Nombre de clusters = 4 → Silhouette Score = 0.29
Nombre de clusters = 5 → Silhouette Score = 0.22
```

GMM:

```
Score de silhouette final pour GMM: 0.377
```

DSO3 (Classification):

Logistic Regression:

```
Model: Logistic Regression
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.73      0.82       224
           1       0.36      0.71      0.48        48

    accuracy                           0.73       272
   macro avg       0.64      0.72      0.65       272
weighted avg       0.82      0.73      0.76       272

Confusion Matrix:
[[164  60]
 [ 14  34]]
ROC-AUC: 0.7202380952380953
```

SVM:

```
Model: SVM
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.70      0.79       224
           1       0.33      0.69      0.45        48

    accuracy                           0.70       272
   macro avg       0.62      0.69      0.62       272
weighted avg       0.81      0.70      0.73       272

Confusion Matrix:
[[157  67]
 [ 15  33]]
ROC-AUC: 0.6941964285714286
```

RandomForest:

```
Model: Random Forest
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.92      0.90       224
           1       0.54      0.46      0.49        48

    accuracy                           0.83       272
   macro avg       0.71      0.69      0.70       272
weighted avg       0.83      0.83      0.83       272

Confusion Matrix:
[[205  19]
 [ 26  22]]
ROC-AUC: 0.6867559523809523
```

KNN:

```
Model: KNN
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.99      0.91       224
           1       0.78      0.15      0.25        48

    accuracy                           0.84       272
   macro avg       0.81      0.57      0.58       272
weighted avg       0.83      0.84      0.79       272

Confusion Matrix:
[[222   2]
 [ 41   7]]
ROC-AUC: 0.5684523809523809
```

6.2. Best-Model Results

DSO1 (Regression):

-KNN: $R^2 = 0.80$

-Linear Regression : $R^2 = 0.83$

-Random Forest : $R^2 = 0.86$

Random Forest is the best model. Highest R^2 (0.86) → Captures complex relationships

DSO2 (Clustering):

Model	Silhouette	Davies-Bouldin
KMeans	**0.703890**	**0.341208**
Agglomerative	**0.700392**	**0.350209**
GMM	**0.166460**	**1.131811**
Spectral	**-0.390419**	**2.444150**

KMeans is the best-performing model for the clustering task.

-Clear Superiority Over GMM/Spectral

DSO3 (Classification):

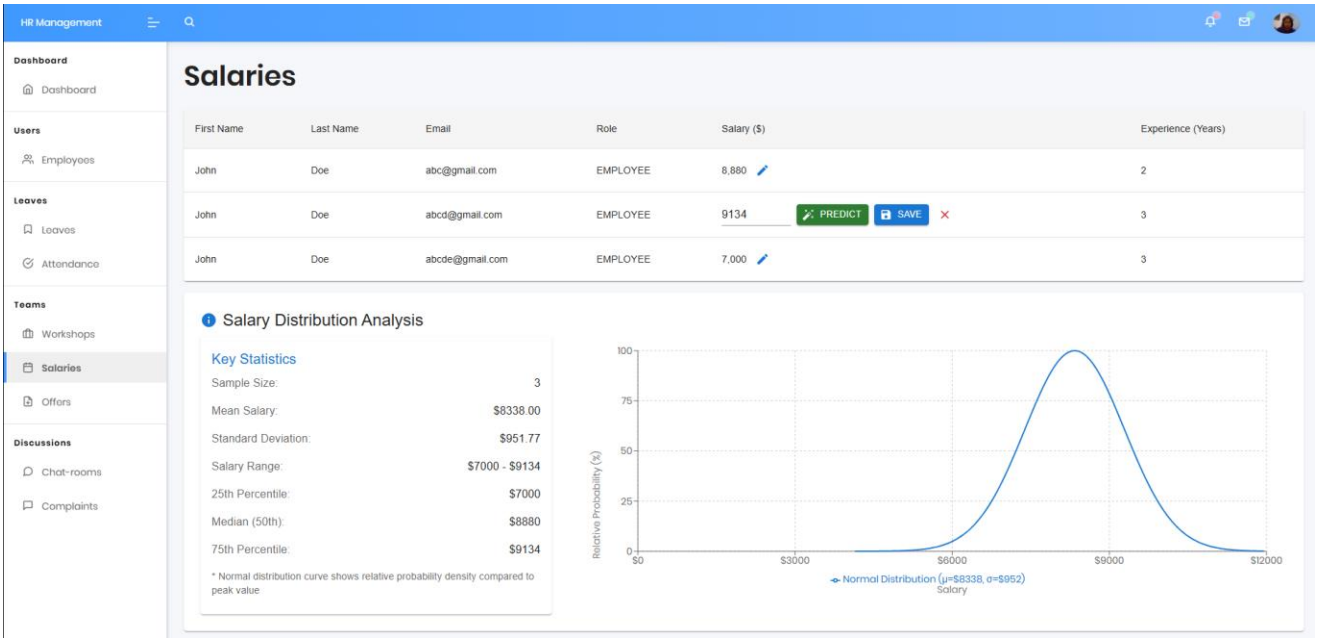
Model	ROC-AUC	Recall (Class 1)	Precision (Class 1)
Logistic Regression	**0.720**	**0.71**	0.36
SVM	0.694	0.69	0.33
Random Forest	0.687	0.46	0.54
KNN	0.568	0.15	0.78

Logistic Regression is the best model for this classification task.

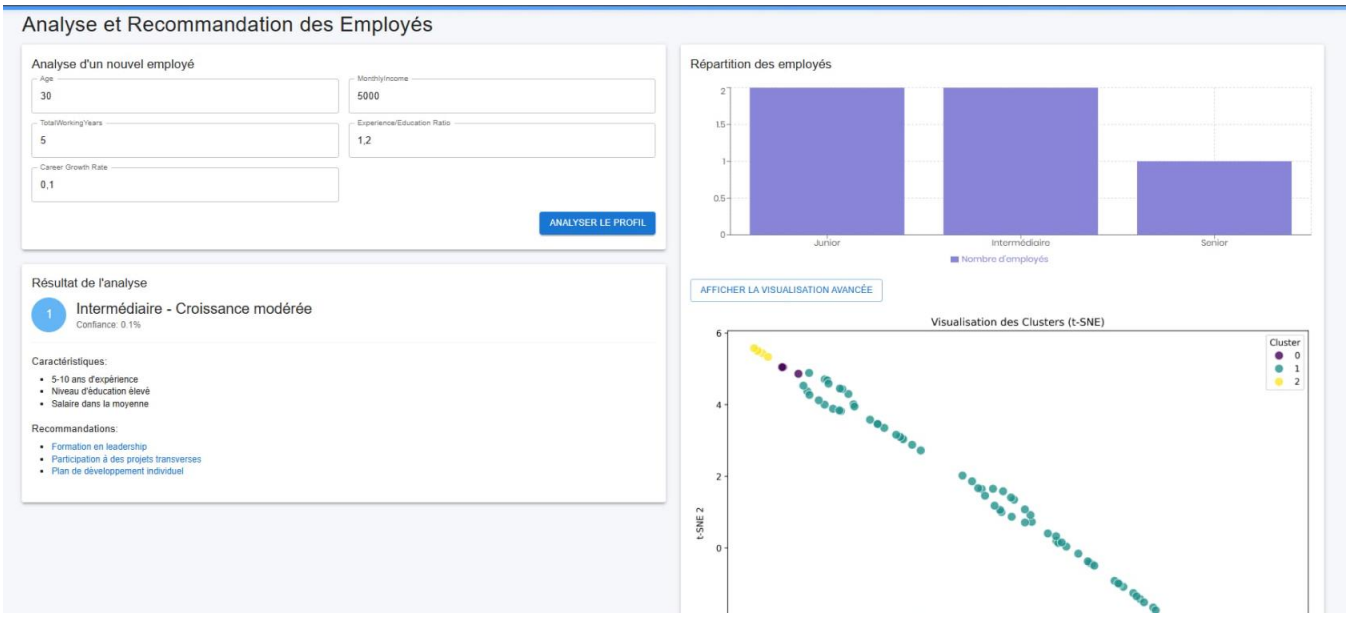
-Highest ROC-AUC , Balanced Recall (0.71) for Attrition , Reasonable Precision (0.36)

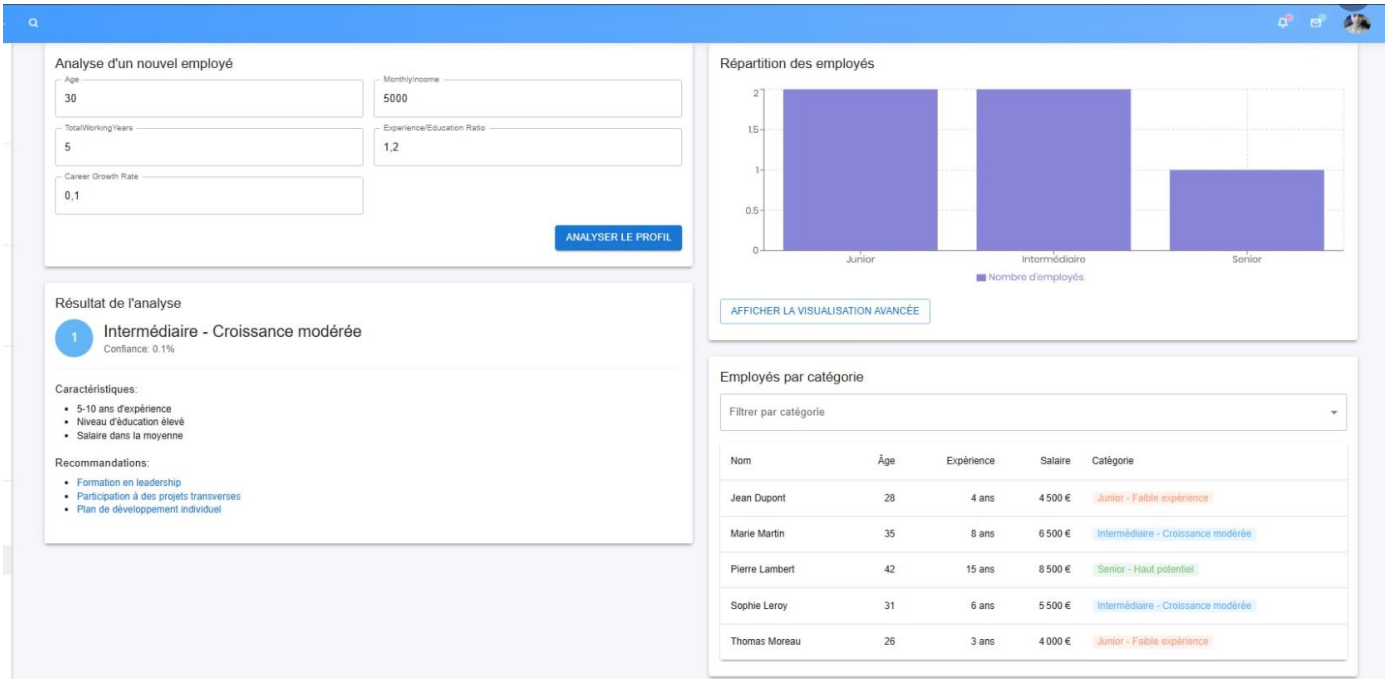
7.Deployment

DSO1(Regression):

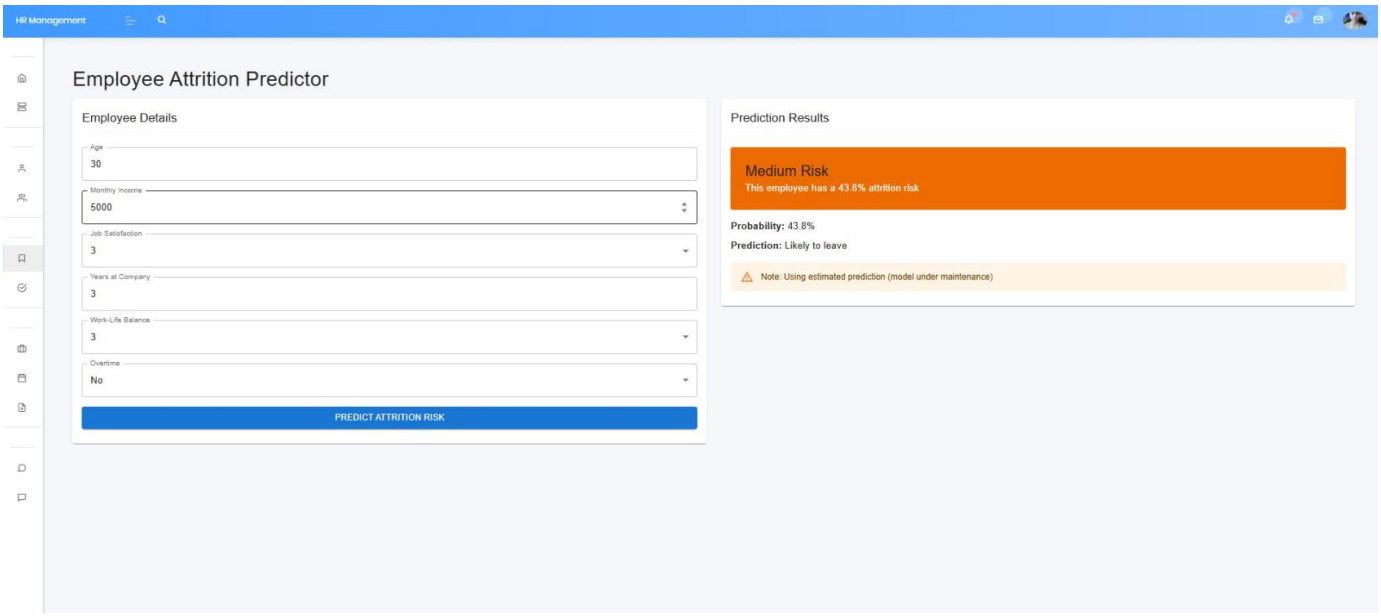


DSO2 (Clustering):





DSO3 (Classification):


HR Management
Employee Attrition Predictor
Employee Details
Age
30
Monthly Income
5000
Job Satisfaction
3
Years at Company
3
Work-Life Balance
3
Overtime
No
PREDICT ATTRITION RISK
Prediction Results
Medium Risk
This employee has a 43.8% attrition risk
Probability: 43.8%
Prediction: Likely to leave
Note: Using estimated prediction (model under maintenance)