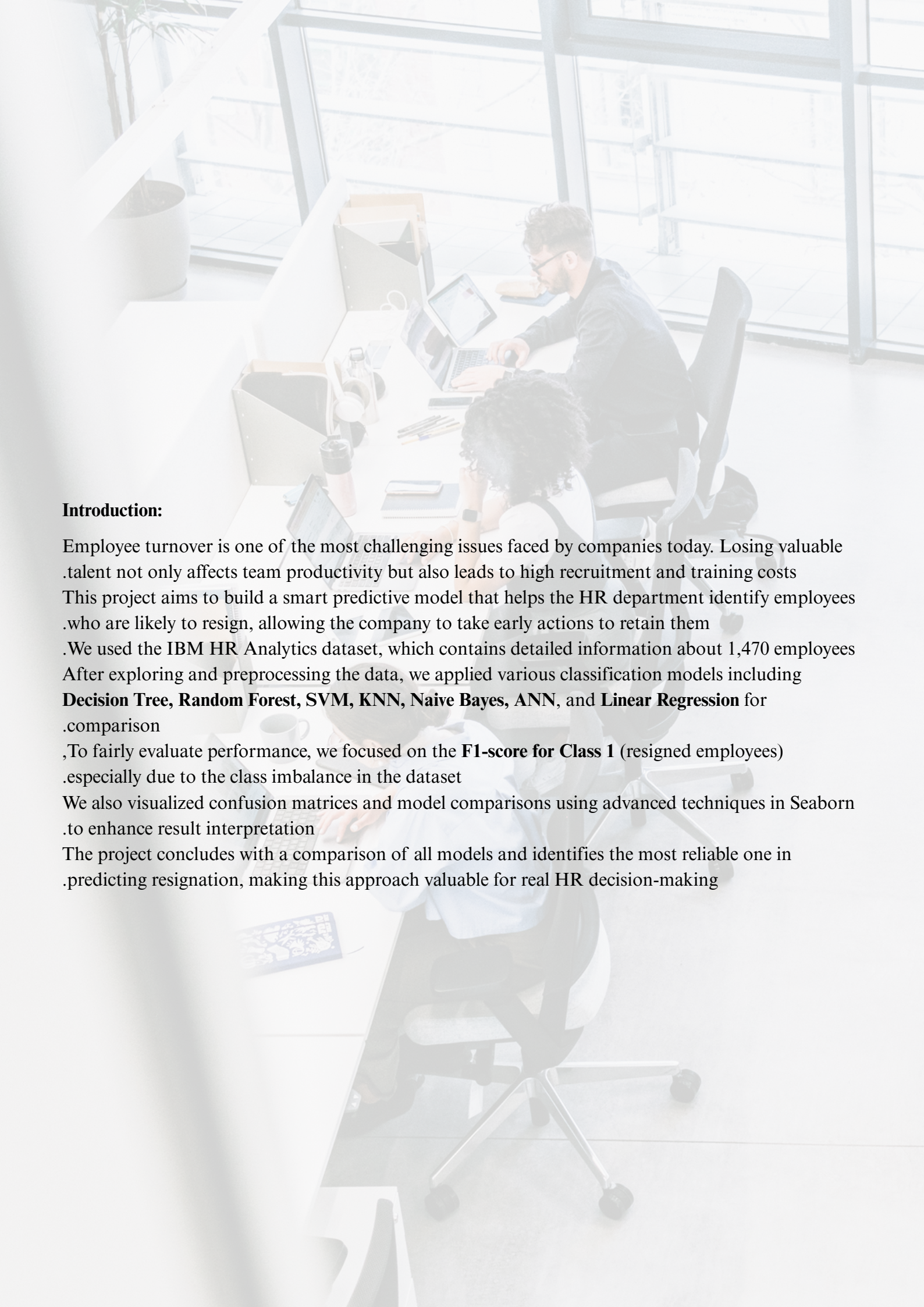
Data-Driven Prediction of Employee Resignation
A Machine Learning Approach



Introduction:

Employee turnover is one of the most challenging issues faced by companies today. Losing valuable .talent not only affects team productivity but also leads to high recruitment and training costs
This project aims to build a smart predictive model that helps the HR department identify employees .who are likely to resign, allowing the company to take early actions to retain them
.We used the IBM HR Analytics dataset, which contains detailed information about 1,470 employees
After exploring and preprocessing the data, we applied various classification models including **Decision Tree, Random Forest, SVM, KNN, Naive Bayes, ANN**, and **Linear Regression** for .comparison
,To fairly evaluate performance, we focused on the **F1-score for Class 1** (resigned employees) .especially due to the class imbalance in the dataset
We also visualized confusion matrices and model comparisons using advanced techniques in Seaborn .to enhance result interpretation
The project concludes with a comparison of all models and identifies the most reliable one in .predicting resignation, making this approach valuable for real HR decision-making

1. What is the name of your data?
IBM HR Analytics Employee Attrition & Performance

2. The source of the data (which database)?
Kaggle

3. Link to the original data:
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

4. Explain the data in words:
The dataset contains HR-related information for employees in a company, including attributes like age, department, overtime, job satisfaction, and whether or not the employee has left the company. The goal is to predict employee attrition based on these features.

5. Is it a regression or classification problem?
Classification, the target is binary: "Yes" or "No" for employee attrition.

6. How many attributes?
35 columns (features)

7. How many samples?
1,470 (rows) employee records

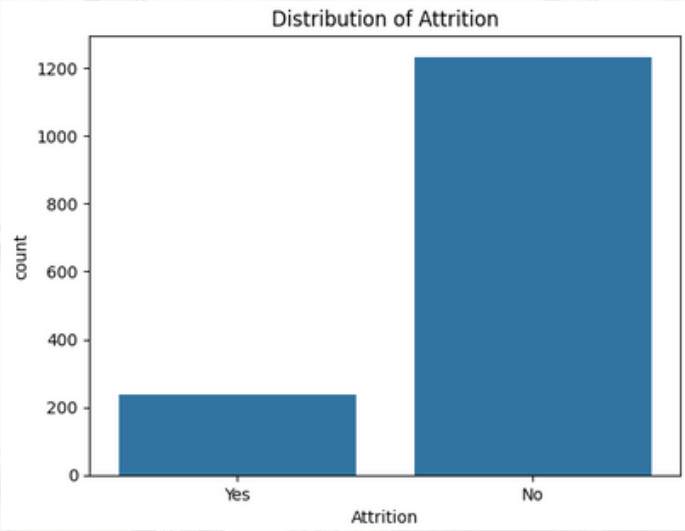
8. What are the properties of the data (statistics)?

- Age ranges from 18 to 60
- Monthly income ranges from 1,000 to over 19,000
- Most employees are in the Research & Development department
- Class imbalance: Only 16% of employees actually resigned

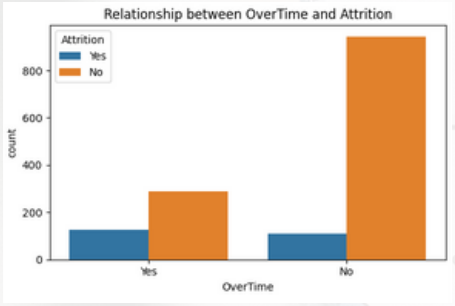
9. Are there any missing data? How did you fill in the missing values?
No missing values were found in the dataset.

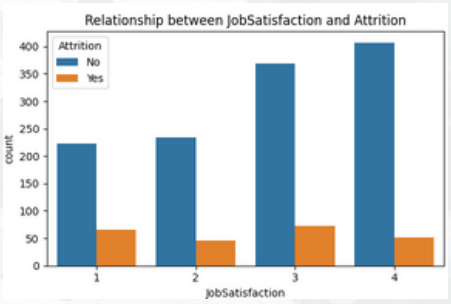
10. Visualize the data:

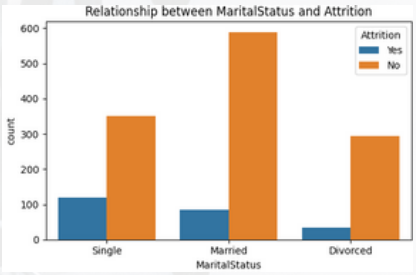
- Count plot of attrition ("Yes"/"No")



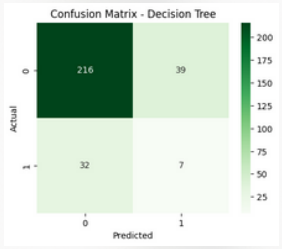
- **Count plots showing relationship between OverTime, MaritalStatus, JobSatisfaction and attrition**

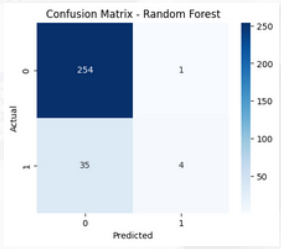


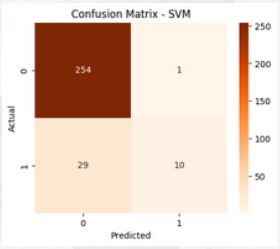


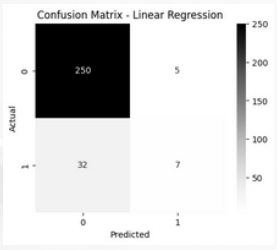


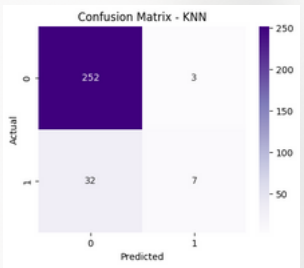
- **Heatmap of confusion matrix for each model**

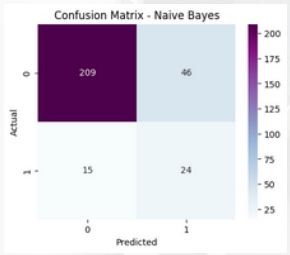


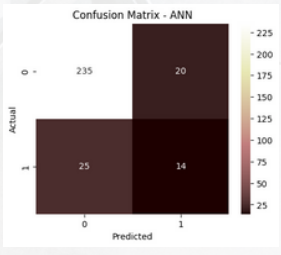




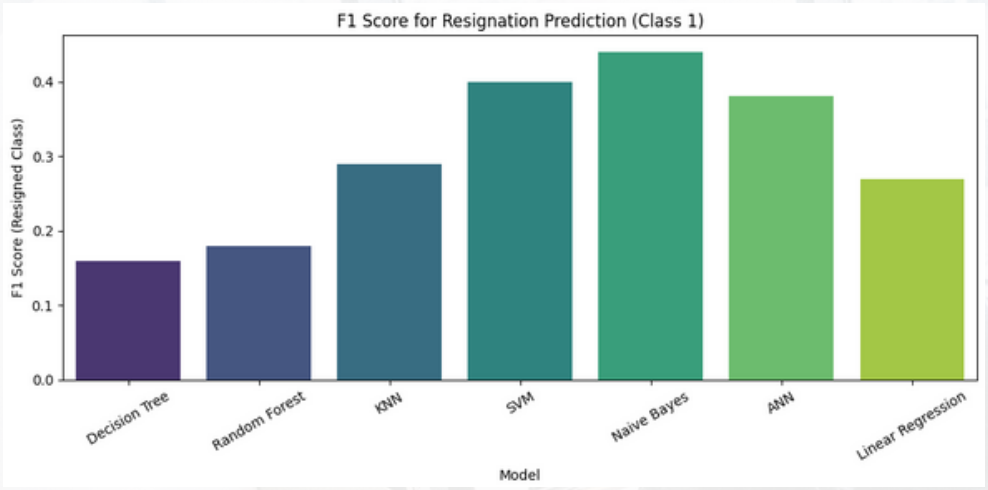








- **Bar chart comparing F1-Score for Class 1 across models**



11. Did you normalize or standardize any of your data? Why?

Yes, StandardScaler was applied for models like KNN and SVM because these models are sensitive to feature scales.

12. What type of preprocessing did you apply to your data? List everything and explain why.

- Removed uninformative columns: EmployeeNumber, EmployeeCount, Over18, StandardHours
- Applied One-Hot Encoding to convert categorical columns into numeric
- Scaled features where necessary
- Split the data into training and testing sets (80-20 split)

13. How did you divide the train and test data? What are the proportions?

Used train_test_split() from sklearn with test_size=0.2 , which gives 80% training and 20% testing.

14. Apply all the machine learning models you have learned in this course to your data and report the results. What is the best/worst performing model? Why?

Applied the following models:

- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Support Vector Machine
- Naive Bayes
- Artificial Neural Network
- Linear Regression (for comparison)

Best Performing: **Naive Bayes** (F1 Score: 0.44 for resigned class)

It had the best balance between precision and recall in predicting employees who resigned.

Worst Performing: **Decision Tree** (F1 Score: 0.16 for resigned class)

Very low recall and precision for class 1.

15. The accuracy of all models using tables and figures?

Yes

the accuracy and evaluation of all models were presented using both confusion matrices and a comparison bar chart of the F1-Score for Class 1 (resigned employees).

Note:These visualizations are shown above in question 10 for clarity and to avoid repetition.

16. If your ability to present the result is advanced...

Yes

all visualizations were created using the Seaborn library to ensure clarity, readability, and a professional presentation style. The Seaborn plots used for confusion matrices and F1-Score comparison are already included above in **question 10**.

What is the reason you picked up this data? What is the importance of your data in reality, and what is the importance of your best-performing model? Is there any insight you could share from the data and the model ?

I selected this dataset because employee attrition is a real-world problem that affects almost every organization. Understanding why employees leave and being able to predict it in advance is a major asset for Human Resources and business continuity.

This dataset contains rich employee-related features like age, department, job satisfaction, overtime, and more. These attributes gave me the opportunity to apply machine learning in a practical way to solve a meaningful business challenge.

After exploring and preparing the data, I applied all the classification algorithms we studied during the course. I focused on predicting employees who are likely to resign, which is class 1 in the dataset. Since the data was imbalanced, I chose to evaluate models based on their F1-score for class 1 instead of just accuracy.

Among all models, Naive Bayes gave the best balance between precision and recall when predicting resigned employees. It was able to detect patterns that other models missed, making it the most reliable for this task.

Through this project, I also learned the importance of data preprocessing and choosing the right evaluation metrics. I realized that accuracy alone is not enough, especially in imbalanced scenarios.

The visualizations helped me understand how certain factors like overtime and low job satisfaction are linked to employee attrition. These insights could be used by HR teams to create better retention strategies.

In conclusion, this project gave me hands-on experience in solving a real problem using machine learning, and I was able to identify a model that performs well and delivers meaningful insights.

GitHub Repository:

You can find the full project code and results in the following repository:

🔗 https://github.com/rubaTech/employee-attrition-prediction.git

This report was prepared by:

Ruba Aloufi

Student ID: 44104338