Employee's Attrition

line 1: Rand Aljathlani
line 2: *Management Information Systems Department*
line 3: *King Saud University*
line 4: Riyadh, Saudi Arabia

line 1: Leena Alateeq
line 2: *Management Information Systems Department*
line 3: *King Saud University*
line 4: Riyadh, Saudi Arabia

line 1: Alhanouf Alhayan
line 2: *Management Information Systems Department*
line 3: *King Saud University*
line 4: Riyadh, Saudi Arabia

line 1: Rand Althaqib
line 2: *Management Information Systems Department*
line 3: *King Saud University*
line 4: Riyadh, Saudi Arabia

line 1: Aljowhara Alshammari
line 2: *Management Information Systems Department*
line 3: King Saud University
line 4: Riyadh, Saudi Arabia

line 1: Noufa Alsalhi
line 2: *Management Information Systems Department*
line 3: *King Saud University*
line 4: Riyadh, Saudi Arabia

***Abstract*—This study explores the application of data mining techniques in analyzing employee attrition data to gain insights into workforce management challenges using IBM dataset. The goal is to build a predictive model using RapidMiner to identify key factors contributing to employee attrition. In this project, we employed multiple models, including Decision Tree, Random Forest, and Logistic Regression, to analyze the data and predict attrition.**

I. INTRODUCTION

The dataset we are working on contains a range of features, including demographic information, job role, and satisfaction ratings. We will apply classification techniques to develop our predictive model, which will help human resources departments take proactive steps to manage employee retention.

Employee attrition refers to the process of employees leaving an organization. While it poses challenges, understanding attrition can lead to several benefits for companies. By analyzing the reasons behind employee departures, organizations can improve their recruitment and retention strategies, reduce costs associated with hiring, and foster a more engaged workforce. Additionally, attrition allows for the introduction of fresh talent and ideas, enhancing overall productivity and morale.

Our main goal is to explore the factors influencing employee attrition and their relationship with attrition rates. Understanding these elements is essential for organizations aiming to enhance workforce stability and productivity. This study focuses on identifying the key drivers of attrition, enabling companies to develop effective strategies for talent retention.

In our analysis, we utilized several data mining algorithms, including Decision Tree, Random Forest, and Logistic Regression, to build predictive models that highlight the critical factors impacting employee attrition. These models offer actionable insights that can help organizations mitigate attrition and retain valuable employees

II. LITERATURE REVIEW

A. The study "Analyzing Employee Attrition Using Decision Tree Algorithms" investigates the factors contributing to employee turnover in knowledge-based organizations, emphasizing the importance of minimizing attrition to maintain competitive advantage. Utilizing data from 309 employees at a Nigerian higher institution, the research applies decision tree algorithms through tools like WEKA and See5 to identify predictive models for attrition. The study highlights that employee-related attributes, such as demographic information and job-related factors, are crucial in predicting turnover. It also discusses the significance of data mining techniques in human resource management for developing effective retention strategies and enhancing decision-making processes. The findings suggest that understanding the reasons behind employee departures can aid organizations in implementing policies to reduce turnover and retain valuable talent. [2]
Keywords: Employee attrition, Turnover, Decision tree algorithms, Data mining

B. The study "Predicting Employee Attrition Using Tree-Based Models" aims to develop binary classification models to forecast employee turnover based on firm cultural and management attributes. Utilizing a dataset of resumes submitted through Glassdoor, the research employs decision tree, random forest, and gradient boosted tree models to assess the likelihood of employees leaving during job transitions. The findings indicate that random forest and decision tree methods are the most effective in predicting attrition, with factors such as compensation, company culture, and senior management performance significantly influencing employees' decisions to leave. The study provides valuable insights for human resources professionals to understand attrition drivers and apply tailored models for retention strategies. [3]
Keywords: Binary classification, Random Forest, Gradient boosted tree, Compensation

C. The article "Analysis and Classification of Employee Attrition and Absenteeism in Industry: A Sequential Pattern Mining-Based Methodology" introduces a novel methodology called E(3A) CSPM to analyze and classify employee attrition and absenteeism using sequential pattern mining (SPM). This approach addresses the challenges in

predicting these phenomena, which can significantly impact productivity and costs. By applying SPM algorithms to four diverse public datasets, the methodology discovers frequent sequential patterns and rules that serve as features for both binary and multi-class classification tasks. The results demonstrate that E(3A) CSPM outperforms traditional methods in classifying and detecting employee attrition and absenteeism, providing valuable insights into the factors influencing these issues. [4]
Keywords: Absenteeism, Sequential pattern mining (SPM), Classification, Productivity

D. The document discusses the challenges of employee attrition in today's competitive job market and explores the application of Explainable Artificial Intelligence (XAI) in predicting and addressing this issue. It highlights how AI can analyze historical data and employee behavior to forecast attrition risks, enabling organizations to implement targeted retention strategies. The paper emphasizes the importance of interpretability in AI models, allowing HR professionals to understand predictions and make informed decisions. The methodology outlined in the study includes data collection, preprocessing, and the use of various classification algorithms to predict employee turnover. [5]
Keywords: retention strategies, interpretability, predictive modeling, human resources.

III. DATASET DESCRIPTION

The IBM dataset is a fictional dataset created to simulate employee attrition scenarios within an organization. It does not represent real-world data, but it is designed by IBM to mimic realistic patterns and relationships commonly found in human resources data.

Table 1. Description of all dataset attributes

Dataset description		Data type
age	Age of the Employee	integer
attrition	Whether the employee left the company (True, they left/False, they didn't)	Boolean
BusinessTravel	Describes the employee's travel frequency	String
DailyRate	The employee's daily rate of pay	Integer
Department	Department the employee belongs to	String
DistanceFromHome	Distance between the employee's home and the workplace	Integer
Education	Employee Education Level (1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor')	Integer
EducationField	The field of study (e.g., Life Sciences, Medical)	String
EmployeeCount	A constant value (always 1) and provides no variability	Integer
EmployeeNumber	A unique identifier for each employee	Integer
Environment Satisfaction	Satisfaction level with the work environment (1 'Low', 2 'Medium', 3 'High', 4 'Very High')	Integer
Gender	Gender of the employee (female, male)	String
HourlyRate	The hourly wage of the employee.	Integer
JobInvolvement	Degree of employee involvement in the job (1 'Low', 2 'Medium', 3 'High', 4 'Very High')	Integer
JobLevel	The employee's job level within the company (1 'Entry-level/ junior', 2 'Mid-level', 3 'Senior', 4 'Upper management', 5 'Executive')	Integer
JobRole	Employee Role in the company (e.g., Manager, Technician)	String
JobSatisfaction	Satisfaction level with the job (1'Low', 2 'Medium', 3 'High', 4 'Very High')	Integer
MaritalStatus	Marital status of the employee (single, married, divorced)	String
MonthlyIncome	The total monthly earnings, including bonuses and overtime, reflect the full income picture ($)	Integer
MonthlyRate	The monthly billing rate for the employee	Integer
NumCompaniesWorked	The number of companies the employee has previously worked for	Integer
Over18	Indicates whether the employee is over 18	boolean
OverTime	Indicates if the employee works overtime (Yes, they have/No, have not)	boolean
PercentSalaryHike	The percentage increase in the employee's salary compared to the previous year	Integer
Performance Rating	Employee's performance rating	Integer
Relationship Satisfaction	Satisfaction with workplace relationships (1 'Low', 2 'Medium', 3 'High', 4 'Very High')	Integer
StandardHours	The standard number of working hours	Integer
StockOption Level	Stock options provided to the employee (0 = None, 3 =Highest).	Integer
Total Working Years	Total years of professional experience.	Integer
Training TimesLast Year	The number of training sessions attended last year.	Integer
WorkLifeBalance	Employee's work-life balance satisfaction (1"Bad',2'Good',3'Better', 4 'Best)	Integer
YearsAtCompany	The number of years the employee has worked at the company.	Integer
YearsInCurrentRole	the number of years the employee has been in their current role.	Integer
YearsSinceLastPromotion	The number of years since the employee's last promotion.	Integer
Years WithCurrManager	The number of years the employee has worked with their current manager.	Integer

IV. TECHNIQUE DESCRIPTION

In this project we will use **classification technique** as our primary data mining technique. Classification technique is a supervised data mining technique that assigns a class label to each record in a dataset based on several features.

Our target variable is **"Attrition"** which indicates whether the employee has left the company or not. The goal of classification is to build a model that accurately categorizes employees into two groups based on various features.

V. ALGORITHM DESCRIPTION

For this project, we selected **Decision Tree**, **Random Forest** and **Logistic Regression** as classification algorithms due to their effectiveness in predicting employee's attrition.

A. Decision Tree: is a predictive modeling tool that uses a tree-like graph of decisions and their possible consequences. It splits the dataset into branches based on feature values, leading to a decision (or classification) at the leaf nodes. We used a decision tree to identify key factors contributing to employee attrition. Each node represents a feature (e.g., monthly income, job role, years at the company), and each branch represents a decision rule based on that feature.

B. Random Forest: is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions (for classification tasks). It reduces overfitting compared to a single decision tree by averaging the results from multiple trees. Random forest can be employed to evaluate the importance of various features in predicting attrition. It provides insights into which attributes are most influential in determining whether employees leave.

C. Logistic Regression: is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular category, using a logistic function. The output is a value between 0 and 1, which can be interpreted as the probability of attrition. Logistic regression can be applied to estimate the relationship between employee attrition (dependent variable) and various independent variables (like age, monthly income, job satisfaction). It helps quantify the impact of each feature on the likelihood of attrition.

VI. PREPARATION METHODOLOGY

Using **Python**, we prepared the dataset for analysis by checking for missing values and duplicates, removing irrelevant features, encoding categorical variables, analyzing correlations, and selecting the most relevant attributes to optimize data quality and utility.

A. Check for Missing Values

To ensure the completeness of our dataset, we first checked for missing values using the isna() method, which verified that there are no missing values across any feature.

```
df.isna().sum().sum()
```

```
0
```

B. Check for Duplicates

Next, we checked for duplicate records in the dataset to ensure there were no redundancies. We used the duplicated() method and confirmed that no duplicate rows exist.

```
df.duplicated().sum()
```

```
0
```

Additionally, we specifically checked if there were duplicate employee numbers, which would indicate any duplication of employees in the dataset.

```
df['EmployeeNumber'].duplicated().sum()
```

```
0
```

C. Removing Irrelevant Features

Some features, including Over18, EmployeeCount, and StandardHours, were removed because they contained constant values across all rows, making them irrelevant for the analysis.

```
df['Over18'].value_counts()
```

```
Over18
Y    1470
Name: count, dtype: int64
```

```
df['EmployeeCount'].value_counts()
```

```
EmployeeCount
1    1470
Name: count, dtype: int64
```

```
df['StandardHours'].value_counts()
```

```
StandardHours
80    1470
Name: count, dtype: int64
```

Additionally, the EmployeeNumber feature was removed as it did not contribute any meaningful insights to the analysis.

```
df = df.drop(["Over18", "EmployeeCount", "StandardHours", 'EmployeeNumber'], axis=1)
```

D. Feature Encoding

Categorical features were encoded to convert them into numerical representations as follows:

1. Binary categorical features were encoded using **Label Encoding**:
   - Attrition (Target variable): 'No' = 0, : 'Yes' = 1
   - OverTime: 'No' = 0, 'Yes' = 1
   - Gender: 'Female' = 0, 'Male' = 1

```
from sklearn.preprocessing import LabelEncoder
```

```
lb = LabelEncoder()
```

```
binary_cols = ['Gender', 'OverTime', 'Attrition']

for col in binary_cols:
    df[col] = lb.fit_transform(df[col])
```

2. Ordinal categorical feature (BusinessTravel) was encoded **manually** to reflect its inherent order:
   - 'Non-Travel' = 0
   - 'Travel_Rarely' = 1
   - 'Travel_Frequently' = 2

```
df["BusinessTravel"] = df["BusinessTravel"].map({"Non-Travel": 0, "Travel_Rarely": 1, "Travel_Frequently": 2})
```

E. Correlation Analysis

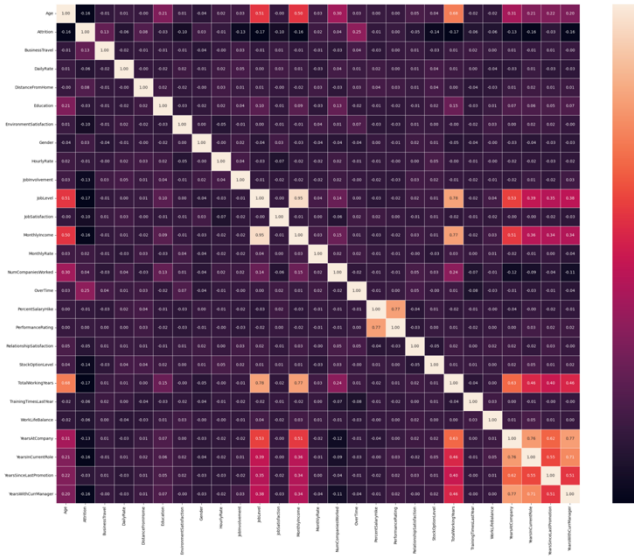
To identify relationships between features and the target variable (Attrition), we performed the following analyses:

1. Numeric Features (Pearson Correlation):

The Pearson correlation coefficients between numeric features and Attrition were computed. The top 16 features most correlated with the target were selected.

```
df_nums = df.select_dtypes(include=['number'])

f, ax = plt.subplots(figsize=(30, 23))
sns.heatmap(df_nums.corr(), annot=True, linewidths=.5, fmt='.2f');
```



```
df_nums_corr = df_nums.corr()['Attrition'].abs().sort_values(ascending=False)
top_features = df_nums_corr.head(17)
top_features
```

```
Attrition                 1.000000
OverTime                  0.246118
TotalWorkingYears         0.171063
JobLevel                  0.169105
YearsInCurrentRole        0.160545
MonthlyIncome             0.159840
Age                       0.159205
YearsWithCurrManager      0.156199
StockOptionLevel          0.137145
YearsAtCompany            0.134392
JobInvolvement            0.130016
BusinessTravel            0.127006
JobSatisfaction           0.103481
EnvironmentSatisfaction   0.103369
DistanceFromHome          0.077924
WorkLifeBalance           0.063939
TrainingTimesLastYear     0.059478
Name: Attrition, dtype: float64
```

2. Categorical Features (Chi-Square Test):

Using the Chi-Square test, we calculated the p-values for categorical features to evaluate their dependency on the target variable. A threshold value (0.022578) was determined based on the least significant numeric feature (TrainingTimesLastYear).

```
from scipy.stats import pointbiserialr
```

```
corr, p_val = pointbiserialr(df['Attrition'], df['TrainingTimesLastYear'])
print("TrainingTimesLastYear")
print(f"P-value: {p_val:.6f}")
```

```
TrainingTimesLastYear
P-value: 0.022578
```

```
from scipy.stats import chi2_contingency
```

```
df_obj=df.select_dtypes(include='object').columns.tolist()

df_obj
```

```
['Department', 'EducationField', 'JobRole', 'MaritalStatus']
```

```
def chi_square_test(df, feature, target='Attrition'):
    contingency_table = pd.crosstab(df[feature], df[target])
    _, p_value, _, _ = chi2_contingency(contingency_table)
    return p_value

results = {}
for feature in df_obj:
    p_value = chi_square_test(df, feature)
    results[feature] = {'p-value': f"{p_value:.6f}"}

results_df = pd.DataFrame(results).T
results_df
```

	p-value
Department	0.004526
EducationField	0.006774
JobRole	0.000000
MaritalStatus	0.000000

F. *Feature Selection*

Based on the correlation and Chi-Square test, the final dataset was reduced to 21 attributes, including the top numeric features and all significant categorical features.

```
df_num = top_features.index.tolist()
df_num
```

```
['Attrition',
 'OverTime',
 'TotalWorkingYears',
 'JobLevel',
 'YearsInCurrentRole',
 'MonthlyIncome',
 'Age',
 'YearsWithCurrManager',
 'StockOptionLevel',
 'YearsAtCompany',
 'JobInvolvement',
 'BusinessTravel',
 'JobSatisfaction',
 'EnvironmentSatisfaction',
 'DistanceFromHome',
 'WorkLifeBalance',
 'TrainingTimesLastYear']
```

```
data_cols = df_num + df_obj
data = df[data_cols]

data.head()
```

	Attrition	OverTime	TotalWorkingYears	JobLevel	YearsInCurrentRole	MonthlyIncome	Age	YearsWithCurrManager	StockOptionLevel	YearsAtCompany	JobIn
0	1	1	8	2	4	5993	41	5	0	6	
1	0	0	10	2	7	5130	49	7	1	10	
2	1	1	7	1	0	2090	37	0	0	0	
3	0	1	8	1	7	2909	33	0	0	8	
4	0	0	6	1	2	3468	27	2	1	2	

```
data.shape
```

```
(1470, 21)
```

```
data.to_csv('HR-Employee-Attrition-Updated.csv', index=False)
```

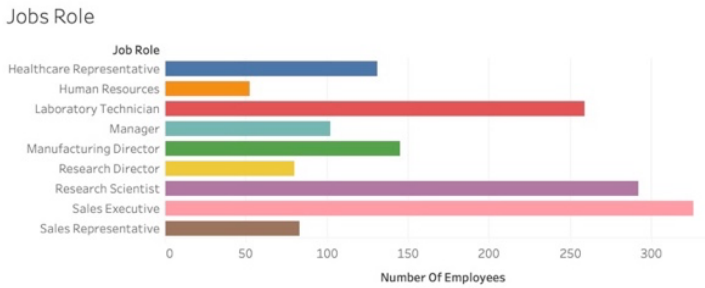
VII. VISUALIZATION METHODOLOGY

We have used tableau and python for data visualization part.

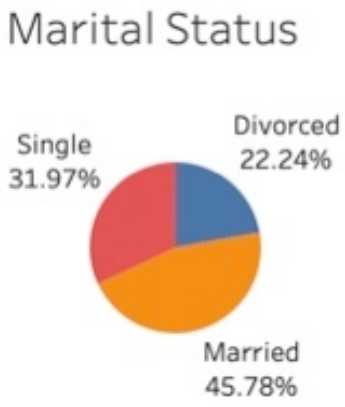
Tableau is a powerful and user-friendly data visualization tool that enables you to create interactive and shareable dashboards.

On the other hand, python is a versatile programming language widely used for data analysis and visualization through libraries such as matplotlib and seaborn.

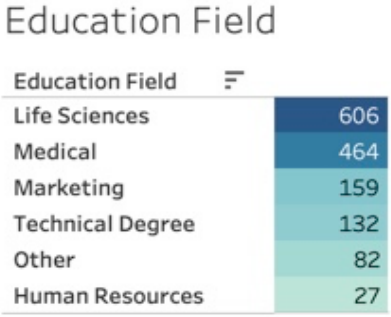
By combining these tools, we were able to present clear and simplified insights to support us in our project.



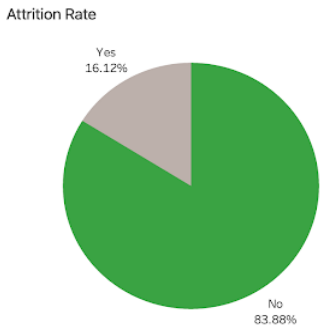
This chart shows the distribution of employees across different jobs roles.



This chart illustrates the distribution of employees according to their marital status, it highlights the proportion of single, married, and Divorced within the organization.



This chart displays the number of employees by education field, highlighting the diversity of educational backgrounds within the workforce.



This pie chart illustrates the attrition rate, or the rate at which employees are leaving the organization. The pie chart shows that 83.88% of employees did not leave (labeled as "No"), while 16.12% of employees did leave (labeled as "Yes").

The chart shows that most employees are aged 30-40, with fewer under 20 or over 50. This highlights a middle-aged workforce, useful for planning recruitment and retention strategies.

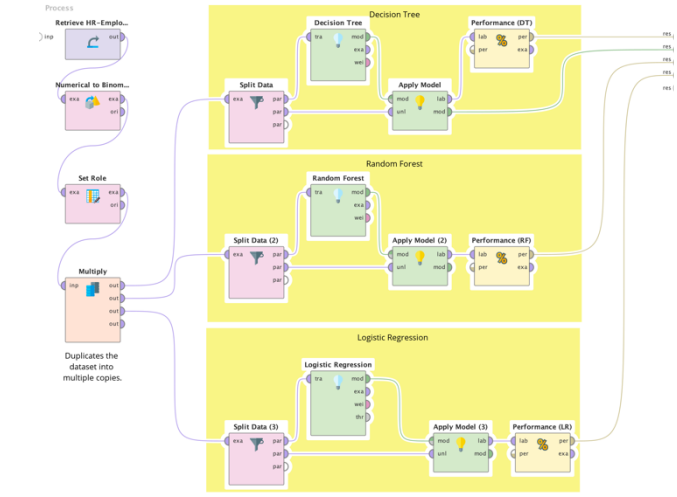
VIII. MODELING METHODOLOGY

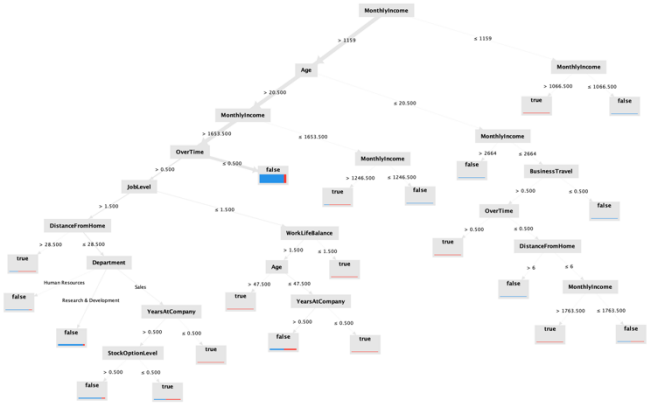
Our target variable is attrition, which contains two values, yes and no. The goal is to build a model that predicts the attrition based on different attributes.

We performed all the three models at the same time, we implemented two approaches for our models. The first method involved splitting the dataset into training and testing sets (70% training, 30% testing) to assess the model's performance on unseen data. The second method utilized cross-validation with 10 folds, which divides the data into multiple subsets and trains the model on different combinations of these subsets. This approach was employed to enhance the model's performance and address the issue of overfitting by ensuring the model generalizes well to new data.

The following steps are for split approach:

1. Import dataset into rapid miner.
2. Retrieve HR-Employee-Attrition-Updated dataset.
3. Use “Numerical to Binomial” to convert the target (Attrition) form numerical to binomial.
4. Use “Set Role” to specify the target attribute.
5. Use “Multiply” to duplicates our dataset into three copies.
6. For **all the three models using split approach** we used “split data” to split the data into 0.7 training and 0.3 testing after that we used “performance” to specify the accuracy of the model.





accuracy: 82.54%

	true false	true true	class precision
pred. false	351	58	85.82%
pred. true	19	13	40.62%
class recall	94.86%	18.31%	

accuracy: 85.71%

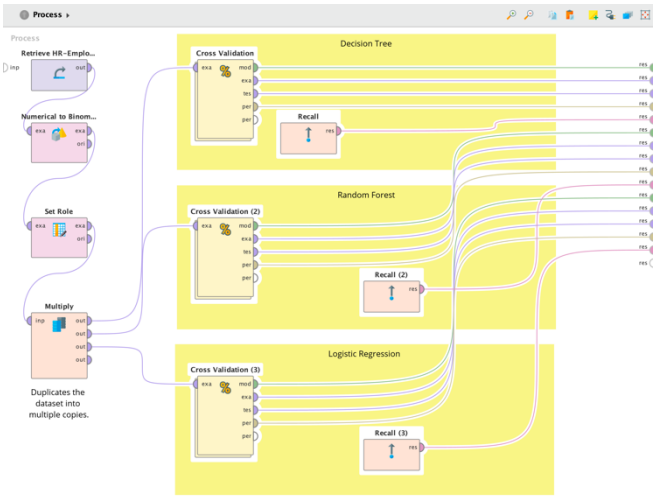
	true false	true true	class precision
pred. false	367	60	85.95%
pred. true	3	11	78.57%
class recall	99.19%	15.49%	

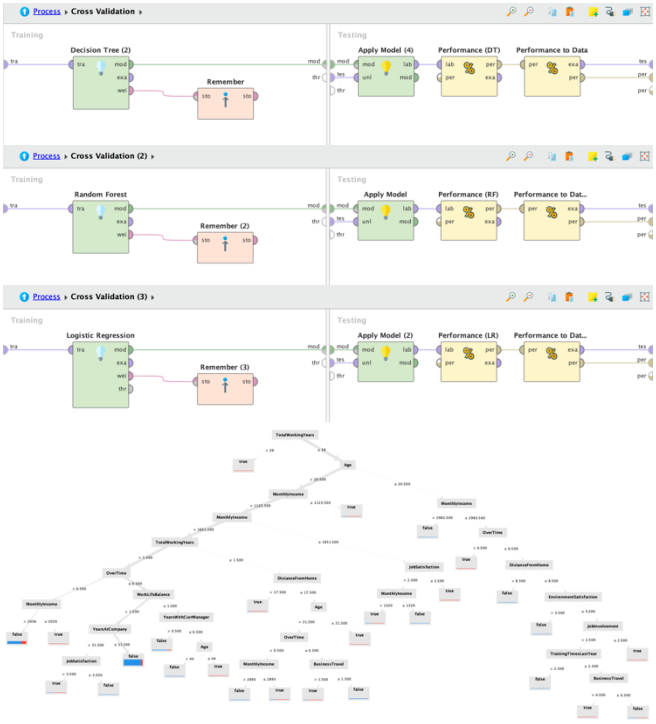
accuracy: 86.62%

	true false	true true	class precision
pred. false	340	29	92.14%
pred. true	30	42	58.33%
class recall	91.89%	59.15%	

The following steps are for CV approach:

1. Import dataset into RapidMiner.
2. Retrieve HR-Employee-attrition-updated dataset.
3. Use “numerical to binomial” to convert the target (attrition) from numerical to binomial.
4. Use “set role” to specify the target attribute.
5. Use “multiply” to duplicates our dataset into three copies.
6. Connect “multiply” with “cross-validation” for each model.
7. We used “recall” and “remember” to ensure the weights were optimized for the target attribute (Attrition).
8. “Performance DT, RF, LF” operator was used to assess the average performance for each model
9. “Performance to Data” operator was used to extract the Example Set containing accuracy metrics for each fold in the Cross-Validation (CV) process.





accuracy: 83.13% +/- 1.31% (micro average: 83.13%)

	true false	true true	class precision
pred. false	1195	210	85.05%
pred. true	38	27	41.54%
class recall	96.92%	11.39%	

accuracy: 84.90% +/- 1.39% (micro average: 84.90%)

	true false	true true	class precision
pred. false	1219	208	85.42%
pred. true	14	29	67.44%
class recall	98.86%	12.24%	

accuracy: 85.78% +/- 1.68% (micro average: 85.78%)

	true false	true true	class precision
pred. false	1130	106	91.42%
pred. true	103	131	55.98%
class recall	91.65%	55.27%	

IX. THE BEST MODEL

After evaluating our models, we have decided to proceed with logistic regression for several reasons. It demonstrated higher performance than both the Random Forest and decision tree models, leading to better insights and outcomes. Given our primary objective of achieving optimal results using data mining techniques, logistic regression emerged as the best choice.

This model is particularly effective for binary classification, such as predicting employee attrition. It provides a clear interpretation of the relationships between independent variables and the likelihood of an employee leaving, allowing us to quantify the impact of factors like job satisfaction, monthly income, and tenure. Additionally, the probabilistic outputs of logistic regression enhance its utility in human resources decision-making, enabling organizations to implement targeted retention strategies based on the model's insights.

We specifically used recall as a key evaluation metric due to the imbalance in our dataset. This focus on recall ensures that we prioritize identifying employees at risk of attrition, thereby improving our overall retention strategies.

Table 2. Comparing models recall

Model	Recall	
	Split	CV
Decision Tree	18.31%	11.39%
Random Forest	15.49%	12.24%
Logistic Regression	59.15%	55.27%

X. INSIGHTS AND RECOMMENDATIONS

Based on the analysis and findings from our models, the following insights and strategic recommendations are proposed:

- Monthly income

Insight:

Employees with a monthly income of ≤ 1,066.5 are less likely to leave the organization, as they may find it challenging to secure better-paying opportunities elsewhere.

Recommendation:

Assess the potential for salary increases or performance-based incentives for those nearing the income threshold to help retain top talent.

- Working years

Insight:
Total working years > 39 increases the likelihood of attrition. This could be due to retirement, fatigue, or lack of challenges for highly experienced employees.

Recommendation:
Increase the retention of experienced employees by offering flexible roles (e.g., mentorship or advisory positions), recognition programs, and phased retirement options to keep them engaged and valued.

- Overtime work

Insight:
Employees who work overtime are more likely to leave the company. Specifically, 53.6% of employees who left worked overtime, compared to only 25% of those who stayed. This suggests that overtime may be contributing to burnout or dissatisfaction, leading to higher attrition rates.

Recommendation:
To reduce attrition, monitor and manage overtime effectively. Implement overtime limits, and explore options such as flexible schedules or additional support to reduce excessive work hours.

XI. CONCLUSION

This project successfully utilized data mining techniques to analyze employee attrition, with logistic regression emerging as the most effective model due to its high accuracy and interpretability. Tools like Python and Tableau played a crucial role in Data representation. The insights gained can help organizations implement targeted retention strategies. Future studies could explore other predictive models and incorporate real-time data to enhance accuracy. The success of this project is attributed to the teamwork and dedication of all involved.

XII. REFERENCES

[1] Pavan Subhash. (n.d.). IBM HR Analytics Employee Attrition & Performance. Retrieved from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[2] https://doi.org/10.1108/IJOA-10-2019-1903

[3] https://doi.org/10.1016/j.compind.2024.104106

[4] https://doi.org/10.3390/math11224677