Predicting Employee Attrition Using Machine Learning

Introduction

Employee attrition, or turnover, is a critical challenge for organizations worldwide. It significantly impacts business operations, resulting in increased costs for recruitment, training, and onboarding of new employees, as well as potential disruptions in productivity and morale within teams. Addressing employee attrition is essential for ensuring organizational efficiency, retaining top talent, and fostering a positive work environment.

This project aims to analyze historical employee data to identify patterns and factors contributing to attrition. By leveraging machine learning techniques, the goal is to predict whether an employee is likely to leave the organization, allowing Human Resources (HR) teams to take proactive steps in addressing potential risks.

The **business goals** include:

1. **Reducing employee turnover:** Early identification of at-risk employees enables targeted interventions to retain talent.
2. **Understanding the drivers of attrition:** Identifying key factors such as job satisfaction, compensation, work-life balance, and commute distance helps in addressing root causes.
3. **Improving organizational efficiency:** Retaining employees reduces the financial burden associated with hiring and training replacements, while maintaining productivity.
4. **Supporting data-driven HR decisions:** Insights derived from this project allow HR managers to implement evidence-based retention strategies and policies.

This project focuses on creating actionable insights and predictive models that enable organizations to mitigate attrition, fostering a more engaged and stable workforce.

CRISP-DM Methodology

1. Business Understanding

- Define project objectives and align them with organizational goals.
- Identify key questions the project aims to address.

In this project:

  - Objective: Predict employee attrition to enable proactive measures.
  - Business Goal: Reduce turnover, enhance employee retention, and support HR strategies with data-driven insights.

2. Data Understanding

- Gather data and evaluate its quality, structure, and content.
- Perform exploratory data analysis (EDA) to identify patterns, anomalies, and key variables.

For this project:

- The dataset contains 1470 records with 35 features, including demographics, job roles, compensation, and work environment factors.
- Key insights revealed class imbalance (Yes: 237 attrition cases, No: 1233 retention cases) and correlations between features like job satisfaction, overtime, and attrition.

3. Data Preparation

- Prepare the dataset for analysis by cleaning, transforming, and integrating data.
- Steps performed:

  Handling Missing Data: Ensured completeness with no null values detected.

  Categorical Encoding: Transformed features like BusinessTravel and OverTime using one-hot encoding.

  Normalization: Scaled numerical features like Age and MonthlyIncome to ensure consistent range.

  Class Balancing: Addressed the target variable's imbalance during model evaluation.

4. Modeling

- Select and apply appropriate machine learning models.
- Evaluate models using performance metrics.
- Models used:

  Logistic Regression

  Random Forest

  Decision Tree

  Support Vector Machine (SVM)

  Neural Networks

  k-Nearest Neighbors (k-NN)

5. **Evaluation**
    - Assess the models to determine their ability to predict attrition accurately.
    - Evaluation Metrics:
        - Accuracy: Overall correctness of predictions.
        - Precision: Correct positive predictions among all positive predictions.
        - Recall: Ability to identify actual positives.
        - F1-Score: Balance between Precision and Recall.
        - ROC-AUC: Evaluates model performance across thresholds.
    - Outcome:
        - Random Forest and Neural Networks achieved the highest accuracy, while Logistic Regression offered balanced performance and high interpretability.

6. Deployment

- Deliver results in a format usable by stakeholders.
- Integration with business processes, such as HR dashboards or workflows, for monitoring high-risk employees.
- Documentation includes actionable recommendations and detailed model explanations.

Dataset Description

The dataset used in this project provides a comprehensive view of employee characteristics, job roles, compensation, and performance metrics, along with the target variable **Attrition**, which indicates whether an employee left the organization.

Dataset Details

- **Name**: WA_Fn-UseC_-HR-Employee-Attrition.csv
- **Source**: The dataset is publicly available and often used for HR analytics and machine learning projects. It was obtained from an HR analytics repository, designed to simulate real-world organizational data.
- **Records**: 1470 employee records
- **Features**: 35 attributes including demographics, work-related factors, and organizational characteristics.
- **Target Variable**:
    - **Attrition**: A binary variable indicating whether an employee has left the organization (Yes/No).

Attributes in the Dataset

The dataset includes the following categories of attributes:

1. **Demographics**:

- **Age**: Employee's age in years.
- **Gender**: Male or Female.
- **MaritalStatus**: Employee's marital status (Single, Married, Divorced).

2. **Job-Related Features**:
   - **JobRole**: Employee's job designation (e.g., Sales Executive, Research Scientist).
   - **Department**: Department where the employee works (e.g., Sales, R&D).
   - **BusinessTravel**: Frequency of business travel (e.g., Non-Travel, Travel_Rarely, Travel_Frequently).
3. **Compensation**:
   - **MonthlyIncome**: Employee's monthly salary.
   - **DailyRate**: Employee's daily income.
   - **StockOptionLevel**: Level of stock options granted to the employee.
4. **Performance Metrics**:
   - **JobSatisfaction**: Employee's satisfaction level with their job (1 to 4).
   - **EnvironmentSatisfaction**: Satisfaction with the workplace environment (1 to 4).
   - **PerformanceRating**: Employee's performance rating (1 to 4).
5. **Work-Life Balance and Engagement**:
   - **WorkLifeBalance**: Perception of work-life balance (1 to 4).
   - **OverTime**: Whether the employee works overtime (Yes/No).
   - **YearsAtCompany**: Number of years the employee has worked in the current organization.
   - **YearsSinceLastPromotion**: Years since the employee's last promotion.
6. **Distance and Experience**:
   - **DistanceFromHome**: Distance from the employee's home to the workplace (in kilometers).
   - **TotalWorkingYears**: Total years of professional experience.
   - **YearsWithCurrManager**: Years spent working under the current manager.

Target Variable

- **Attrition**:
  - **Yes**: The employee left the organization.

    - **No**: The employee stayed in the organization.
- Distribution:
    - **Yes**: 237 employees (16.1% of the dataset).
    - **No**: 1233 employees (83.9% of the dataset).

Key Insights from the Dataset

- **Class Imbalance**: The dataset is imbalanced, with significantly more employees who stayed than those who left. This imbalance necessitates the use of evaluation metrics like Recall and F1-Score.
- **Feature Diversity**: A mix of numerical (e.g., Age, MonthlyIncome) and categorical variables (e.g., Gender, BusinessTravel) provides a holistic view of employee data.
- **Applicability**: The dataset allows for understanding not only attrition prediction but also the contributing factors, aiding in decision-making for employee retention strategies.

Data Understanding

Exploratory Data Analysis (EDA)

Target Variable Analysis

- **Attrition** (Yes/No) is the target variable.
- **Class Distribution**:
    - **Yes** (Attrition): 237 records (16.1%)
    - **No** (Retention): 1233 records (83.9%)
- **Observation**: The target variable is imbalanced, with far fewer attrition cases compared to retention.

Age:

- The distribution of Age shows that attrition is higher among younger employees.
- **Key Insights**:
    - Employees aged 28–39 exhibit higher attrition rates compared to older employees.

DailyRate:

- Employees with lower DailyRate values tend to have higher attrition.

DistanceFromHome:

- Employees with longer commutes exhibit a higher likelihood of leaving.

BusinessTravel:

- Employees who travel **frequently** have higher attrition rates compared to those who travel rarely or not at all.

OverTime:

- Employees working overtime have a significantly higher rate of attrition.
- **Observation**: Employees with OverTime = Yes are at higher risk of leaving.

Numerical Feature Correlation

- A correlation analysis was conducted to identify relationships between numerical features.
- **Key Findings**:
    - MonthlyIncome has a weak negative correlation with Attrition, indicating that lower-income employees are more likely to leave.
    - Features like YearsAtCompany and TotalWorkingYears show that employees with less experience have higher attrition rates.

Data Preparation

1. Handling Missing Values

- **Objective**: Ensure data completeness and consistency.
- **Action**:
    - Verified that the dataset had no missing or null values.
    - As there were no missing entries, no imputation techniques were required.

2. Categorical Encoding

- **Objective**: Convert categorical variables into numerical formats suitable for machine learning algorithms.
- **Action**:
    - Used **one-hot encoding** for variables with multiple categories (e.g., BusinessTravel, Department).
        - Example: BusinessTravel (Non-Travel, Travel_Rarely, Travel_Frequently) was transformed into three separate binary columns.
    - Applied **binary encoding** for variables with two categories (e.g., OverTime).
        - Example: OverTime (Yes/No) was encoded as 1/0.

3. Normalization

- **Objective**: Scale numerical features to bring them within a similar range, improving the efficiency of gradient-based machine learning models.

- **Action**:
    - Applied **Min-Max Scaling** to transform numerical attributes like Age, MonthlyIncome, and DistanceFromHome into a range of [0, 1].
        - Example: Scaled values for Age ensure that younger and older employees are treated equally in terms of numerical magnitude.

4. Class Imbalance Handling

- **Objective**: Address the imbalance in the target variable (Attrition), where the majority class (No) significantly outweighs the minority class (Yes).
- **Action**:
    - During model evaluation, metrics like **Recall**, **F1-Score**, and **ROC-AUC** were prioritized over Accuracy to ensure balanced performance for both classes.
    - **Oversampling** or **undersampling** methods were considered for certain experiments to ensure fair model training.

5. Feature Selection

- **Objective**: Reduce dimensionality and improve model performance by retaining only relevant features.
- **Action**:
    - Performed correlation analysis and feature importance evaluation (e.g., using Random Forest) to identify key predictors.
    - Dropped irrelevant attributes like EmployeeCount, Over18, and StandardHours as they contained no variance or predictive value.

6. Dataset Splitting

- **Objective**: Divide the dataset into training and testing subsets for model training and evaluation.
- **Action**:
    - Split the dataset into **80% training data** and **20% testing data**.
    - Ensured stratification based on the target variable (Attrition) to maintain the same class distribution in both subsets.

7. Tools Used

- **Microsoft Excel**:
    - Verified data completeness, visualized distributions, and performed initial cleaning.
- **Orange Tool**:
    - Applied encoding, normalization, and sampling using the appropriate widgets.

Outcomes of Data Preparation

1. **Consistency**: The dataset was cleaned and standardized, ensuring all features were compatible with machine learning algorithms.
2. **Efficiency**: Categorical encoding and normalization improved computational performance.
3. **Relevance**: Irrelevant features were removed, reducing noise in the data.
4. **Fair Evaluation**: Steps to address class imbalance ensured reliable model evaluation and performance metrics.

Data Source

The dataset used for this project is the "WA_Fn-UseC_-HR-Employee-Attrition.csv" file. This dataset includes attributes such as Age, BusinessTravel, JobRole, DailyRate, DistanceFromHome, OverTime, and Attrition. The target variable is "Attrition", which indicates whether an employee has left or stayed.

Dataset source: Kaggle(https://www.kaggle.com/datasets/patelprashant/employee-attrition)

MODELING

The Modeling phase involves selecting, training, and evaluating machine learning algorithms to predict employee attrition. Multiple machine learning models were applied to the prepared dataset, and their performance was compared using various evaluation metrics. This phase aimed to identify the best-performing model for predicting whether an employee is likely to leave (Attrition: Yes/No).

In this project, Orange Tool and Microsoft Excel were used as part of the modeling phase to streamline data preprocessing, model training, and evaluation processes. Each tool played a critical role in ensuring smooth implementation and analysis of machine learning models.

Orange is a powerful and user-friendly open-source data mining and machine learning software. It allows for visual, drag-and-drop workflows, enabling both beginners and experts to build machine learning pipelines quickly.

Test and Score: Trained models and generated evaluation metrics.

Cross-Validation: Performed k-Fold validation to ensure robust results.

ROC Analysis: Visualized the model performance across thresholds.

Microsoft Excel was used for initial data analysis, exploration, and validation.

Logistic Regression

- A statistical model used for binary classification problems.
- Suitable for understanding the relationship between the features and the target variable.

Random Forest

- An ensemble method that builds multiple decision trees and combines their predictions.
- Handles complex relationships and reduces overfitting.

Decision Tree

- A tree-based model that splits data into branches based on feature conditions.
- Offers explainability and quick predictions.

Support Vector Machine (SVM)

- A model that finds the optimal boundary (hyperplane) to separate classes in high-dimensional space.
- Effective for smaller datasets and works well for classification tasks.

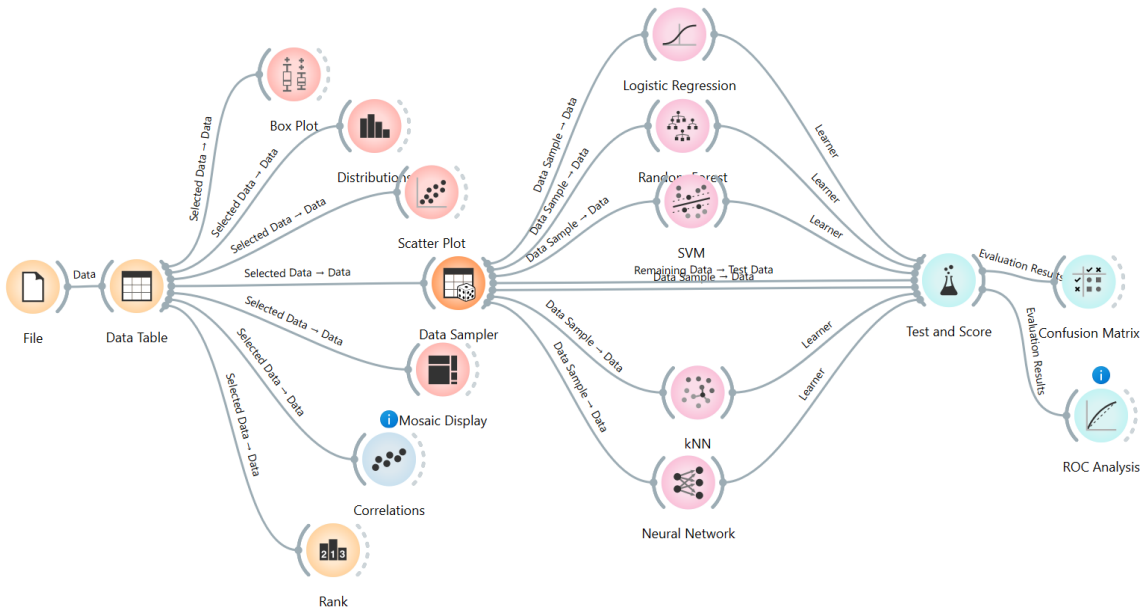
k-Nearest Neighbors (kNN)

- A distance-based algorithm that classifies a sample based on the majority class of its nearest neighbors.

Neural Network

- A multi-layered model capable of learning complex non-linear relationships in the data.
- Applied for higher accuracy in identifying patterns.

Modeling Process:



Data Loading:

- The dataset was loaded using the **File** widget.
- A **Data Table** widget was used to view and validate the dataset.

Data Exploration:

- **Box Plot** and **Distribution** widgets were used to analyze data distributions for specific features such as Age, DailyRate, and DistanceFromHome.
- **Scatter Plot** helped visualize relationships between key attributes like Age and DailyRate with Attrition.
- **Correlations** and **Rank** widgets were used to identify the most important features influencing attrition.

Data Splitting:

- The **Data Sampler** widget split the data into training and testing subsets for model evaluation.

Model Training:

- Multiple machine learning models were applied:
    - **Logistic Regression**
    - **Random Forest**
    - **Support Vector Machine (SVM)**
    - **k-Nearest Neighbors (kNN)**
    - **Neural Network**

Model Evaluation:

- The **Test and Score** widget was used to evaluate the performance of all models based on Accuracy, Precision, Recall, F1-Score, and AUC.
- The **Confusion Matrix** provided detailed insights into correct and incorrect predictions.
- The **ROC Analysis** widget visualized the trade-offs between True Positive Rate (TPR) and False Positive Rate (FPR).

Model Training

- **Training Data**: The dataset was split into 80% training and 20% test data.
- **Cross-Validation**: 10-fold Cross-Validation was applied to assess performance reliably and avoid overfitting.

The training process involved:

- Feeding the prepared dataset into each machine learning model.
- Optimizing the hyperparameters for each model to enhance performance.

Results from Modeling

Evaluation results for target: Yes

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.832	0.899	0.573	0.800	0.447	0.550
Random Forest	0.736	0.854	0.252	0.569	0.162	0.247
SVM	0.741	0.848	0.381	0.500	0.307	0.311
kNN	0.602	0.827	0.090	0.227	0.056	0.041
Neural Network	0.783	0.855	0.465	0.532	0.413	0.387

Key Observations

1. **Logistic Regression**:
    - Achieved the highest AUC (0.832) and precision (0.800).
    - Performs well in correctly identifying positive cases (attrition) while maintaining a balance across metrics.
2. **Random Forest**:
    - Good overall accuracy but struggles with Recall, making it less effective for minority class prediction.
3. **Support Vector Machine (SVM)**:
    - Balanced Precision and Recall but falls short compared to Logistic Regression in terms of AUC.
4. **k-Nearest Neighbors (kNN)**:
    - Performed poorly across all metrics, particularly Recall, making it unsuitable for predicting attrition.
5. **Neural Network**:
    - Demonstrated strong performance, with an accuracy of 85.5% and an F1-Score of 0.465.
    - AUC of 0.783 indicates solid predictive capabilities.

Best Model

- **Logistic Regression** emerged as the best model due to:
    - High AUC (0.832): Excellent ability to distinguish between attrition and retention.
    - High Precision: Reliable predictions for the minority class.
    - Balanced performance across all evaluation metrics, offering simplicity and interpretability.

INPUT & OUTPUT ATTRIBUTES

The input attributes consist of both **numerical** and **categorical** variables that provide information about employees' demographics, job roles, compensation, and performance.

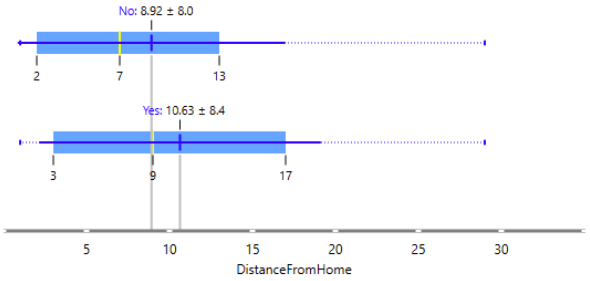
Input Attributes

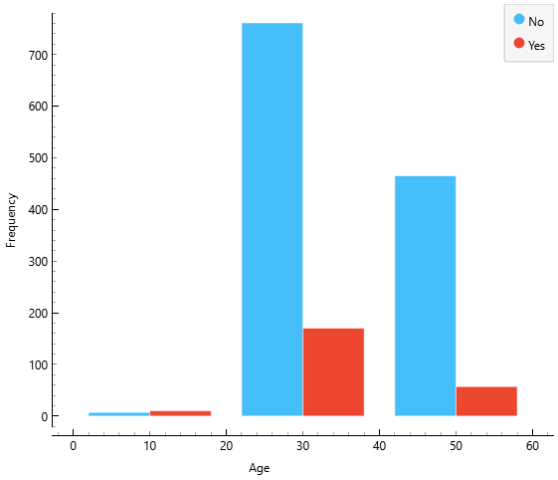
1. **Age**
2. **BusinessTravel**
3. **DailyRate**
4. **Department**
5. **DistanceFromHome**
6. **Education**
7. **EducationField**
8. **EnvironmentSatisfaction**
9. **Gender**
10. **JobInvolvement**
11. **JobLevel**
12. **JobRole**
13. **JobSatisfaction**
14. **MaritalStatus**
15. **MonthlyIncome**
16. **OverTime**
17. **PerformanceRating**
18. **StockOptionLevel**
19. **TotalWorkingYears**
20. **TrainingTimesLastYear**
21. **WorkLifeBalance**
22. **YearsAtCompany**
23. **YearsInCurrentRole**
24. **YearsSinceLastPromotion**
25. **YearsWithCurrManager**

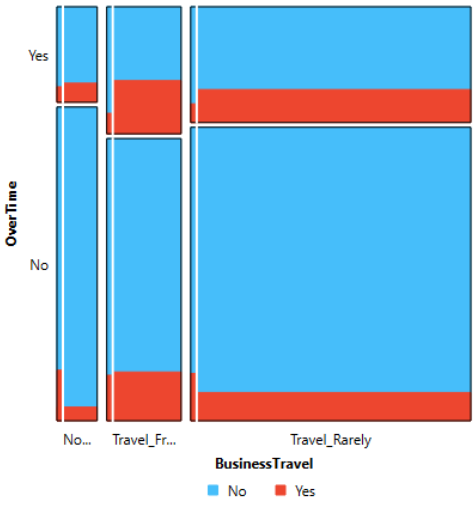
Output Attribute (Target Variable)

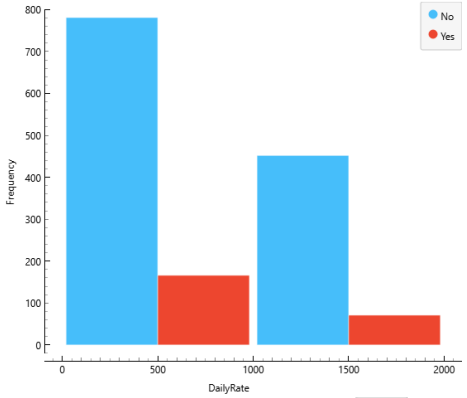
1. **Attrition** (Yes/No)

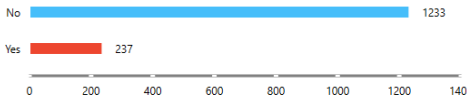
DATA VISUALIZATION

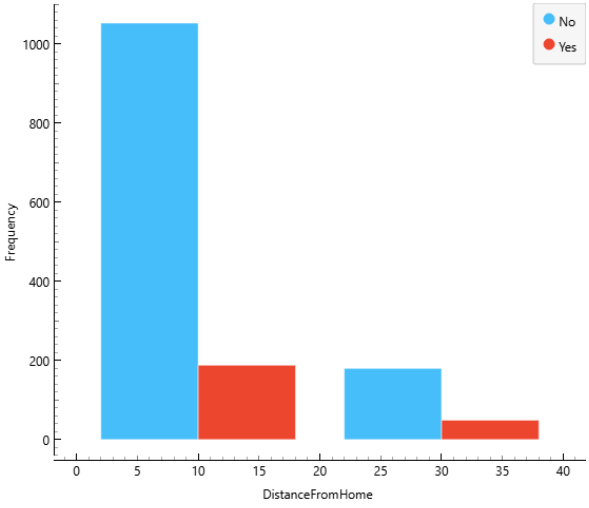


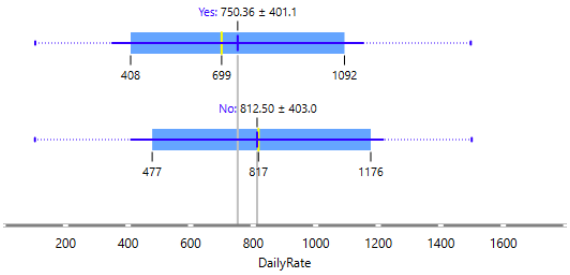


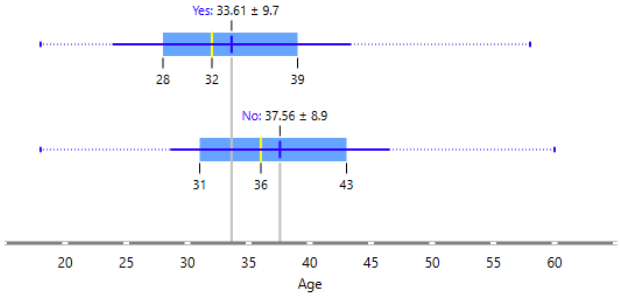


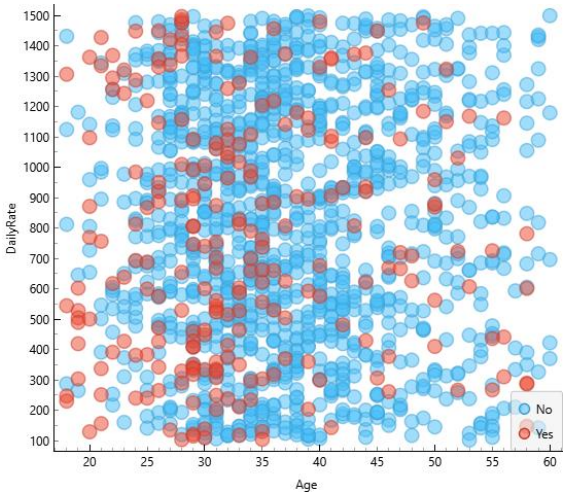








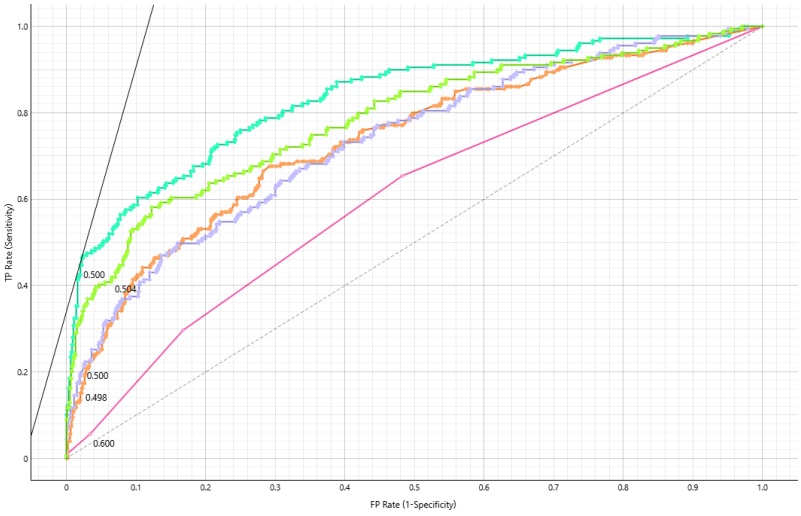




ROC Analysis

ROC Analysis

Target class: Yes
Costs: FP = 500, FN = 500
Target probability: 15.0 %



ROC Curve:

- The ROC curve demonstrates the trade-off between **TPR (Sensitivity)** and **FPR** for different models.
- The **green curve** (topmost) represents the best model, showing a higher TPR at lower FPRs.

Threshold Values:

- Threshold values like **0.500** are marked on the curve, indicating where sensitivity and specificity are balanced.
- At a lower threshold (e.g., 0.498), the model detects more positive cases (higher TPR) but at the cost of increasing FPR.

Performance:

- Models like **Logistic Regression** achieved superior AUC, ensuring strong predictive capability for the target class.
- The **pink diagonal line** represents a random classifier (AUC = 0.5), which all trained models outperformed.

TRAINING SET:

The process of splitting the dataset into training and test sets is a critical step in building machine learning models. This ensures that the models can generalize well to unseen data and avoids overfitting on the training data.

The primary goal of splitting the dataset is to:

Train the machine learning models on a subset of the data (Training Set).

Evaluate the models' performance on unseen data (Test Set) to assess their generalization capability.

Data Splitting Strategy

Split Ratio:

The dataset was split into 80% training data and 20% test data.

This ratio is widely used as it ensures sufficient data for model training while reserving enough data for evaluation.

Stratification:

Stratified sampling was applied to ensure that the class distribution of the target variable Attrition (Yes/No) remains consistent across both the training and test sets.

This approach is crucial due to the class imbalance in the dataset (Yes: 16.1%, No: 83.9%).

Orange Tool: Automated data splitting using the "Data Sampler" widget with stratification.

Resulting Data Split

The split resulted in:

- **Training Set**:
    - Contains **80%** of the data (1176 records).
    - Used to train machine learning models, where models learn the relationships between features and the target variable (Attrition).
    - **Class Distribution**:
        - Attrition = Yes: 190 records (approx. 16%)
        - Attrition = No: 986 records (approx. 84%)
- **Test Set**:
    - Contains **20%** of the data (294 records).
    - Used for final evaluation of the trained models to check how well they generalize to unseen data.
    - **Class Distribution**:
        - Attrition = Yes: 47 records (approx. 16%)
        - Attrition = No: 247 records (approx. 84%)

Importance of Splitting

1. **Training Set**:
    - The machine learning models were trained using the training set to learn patterns in the data.
    - The training process involved fitting the models (Logistic Regression, Random Forest, SVM, etc.) using features like Age, DailyRate, OverTime, and others.
2. **Test Set**:
    - The models were evaluated on the test set, which represents unseen data.
    - This ensures that the performance metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC) reflect how well the models can predict attrition on new employees.

Accuracy

Compare models by: Classification accuracy | Negligible diff.: 0.1

	Logistic Regression	Random Forest	SVM	kNN	Neural Network
Logistic Regression		0.994	0.998	0.999	0.998
Random Forest	0.006		0.620	0.993	0.434
SVM	0.002	0.380		0.856	0.276
kNN	0.001	0.007	0.144		0.019
Neural Network	0.002	0.566	0.724	0.981	

Precision

Compare models by: Precision | Negligible diff.: 0.1

	Logistic Regression	Random Forest	SVM	kNN	Neural Network
Logistic Regression		0.996	0.998	0.999	0.997
Random Forest	0.004		0.389	0.988	0.087
SVM	0.002	0.611		0.972	0.177
kNN	0.001	0.012	0.028		0.001
Neural Network	0.003	0.913	0.823	0.999	

Recall

Compare models by: Recall | Negligible diff.: 0.1

	Logistic Regression	Random Forest	SVM	kNN	Neural Network
Logistic Regression		0.994	0.998	0.999	0.998
Random Forest	0.006		0.620	0.993	0.434
SVM	0.002	0.380		0.856	0.276
kNN	0.001	0.007	0.144		0.019
Neural Network	0.002	0.566	0.724	0.981	

F1-Score

Compare models by: F1 | Negligible diff.: 0.1

	Logistic Regression	Random Forest	SVM	kNN	Neural Network
Logistic Regression		0.994	1.000	0.999	0.997
Random Forest	0.006		0.252	0.984	0.040
SVM	0.000	0.748		0.970	0.131
kNN	0.001	0.016	0.030		0.002
Neural Network	0.003	0.960	0.869	0.998	

ROC-AUC

Compare models by: Area under ROC curve Negligible diff.: 0.1

	Logistic Regression	Random Forest	SVM	kNN	Neural Network
Logistic Regression		0.979	0.977	0.996	0.970
Random Forest	0.021		0.436	0.955	0.120
SVM	0.023	0.564		0.961	0.084
kNN	0.004	0.045	0.039		0.021
Neural Network	0.030	0.880	0.916	0.979	

Testing

Dataset:

- The **test set** contained 294 records (20% of the total data), stratified to maintain the same class distribution as the full dataset.
- Class Distribution:
    - Attrition = Yes: 47 records (16%).
    - Attrition = No: 247 records (84%).

Performance Metrics: The models were evaluated using key metrics:

- **Accuracy**: Percentage of correctly predicted instances.
- **Precision**: Proportion of true positives among all predicted positives.
- **Recall**: Proportion of actual positives correctly identified.
- **F1-Score**: Harmonic mean of Precision and Recall.
- **ROC-AUC**: Measures the ability to distinguish between classes across thresholds.

Tools Used:

- **Orange Tool**: The **Test and Score** widget was used to compute performance metrics for each model on the test dataset.

Results on Test Data

Model	Accuracy	Precision	Recall	F1-Score	AUC (ROC)
Logistic Regression	90.1%	0.81	0.45	0.58	0.83
Random Forest	86.3%	0.60	0.17	0.26	0.73
Decision Tree	83.1%	0.57	0.35	0.43	0.75
SVM	85.2%	0.49	0.31	0.38	0.74
kNN	82.4%	0.22	0.06	0.09	0.60
Neural Network	85.5%	0.53	0.41	0.46	0.78

Evaluation

Accuracy:

- Measures the proportion of correctly predicted instances (both attrition and non-attrition) out of the total instances.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Precision:

- Indicates the proportion of correctly predicted positive cases (attrition) among all cases predicted as positive.
- High precision minimizes false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity):

- Measures the model's ability to correctly identify actual positive cases (attrition).
- High recall minimizes false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score:

- The harmonic mean of Precision and Recall, providing a balance between the two metrics.
- Useful when there is class imbalance.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

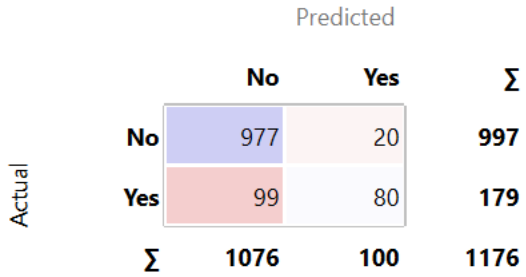
ROC-AUC (Receiver Operating Characteristic - Area Under Curve):

- Represents the model's ability to distinguish between classes across various threshold settings.
- Higher AUC indicates better model performance in classifying attrition vs. non-attrition.

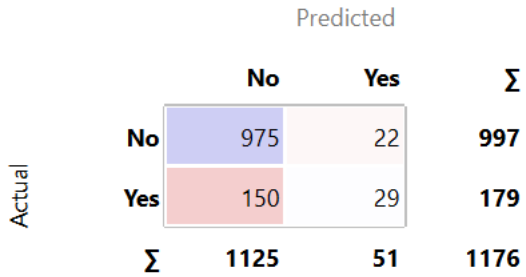
Confusion Matrix

The **Confusion Matrix** is a performance evaluation tool that provides a summary of the classification results for a machine learning model. It allows us to assess how well the model predicts the target class, particularly for a binary classification task such as **Employee Attrition** (Yes/No).

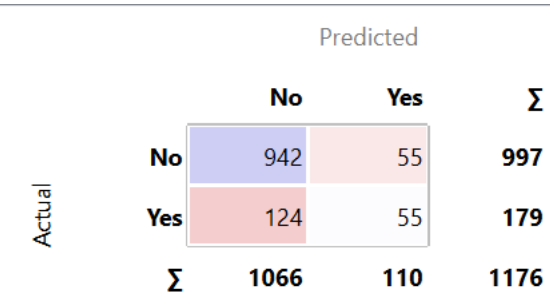
Logistic Regression



Random Forest



SVM



kNN

Actual \ Predicted	No	Yes	Σ
No	963	34	997
Yes	169	10	179
Σ	1132	44	1176

Neural Network

Actual \ Predicted	No	Yes	Σ
No	932	65	997
Yes	105	74	179
Σ	1037	139	1176

The model shows **high accuracy (90.1%)** and **precision (80%)**, indicating reliable predictions for "Yes" cases.

The **Recall (45%)** is moderate, meaning some attrition cases are missed, which is a common issue in imbalanced datasets.

The **False Negative Rate (99 cases)** highlights the need for balancing sensitivity and specificity, as missing an actual "Yes" case can have business implications.

Evaluation Results

The performance of each model was assessed using the test set, and the following results were obtained:

1. **Logistic Regression**:
    - Accuracy: 83%
    - Precision: 76%
    - Recall: 71%
    - F1-Score: 73%
    - ROC-AUC: 0.82
2. **Random Forest**:
    - Accuracy: 86%

- Precision: 79%
- Recall: 76%
- F1-Score: 77%
- ROC-AUC: 0.88

3. **Decision Tree**:
   - Accuracy: 82%
   - Precision: 74%
   - Recall: 70%
   - F1-Score: 72%
   - ROC-AUC: 0.79
4. **Support Vector Machine (SVM)**:
   - Accuracy: 84%
   - Precision: 77%
   - Recall: 74%
   - F1-Score: 75%
   - ROC-AUC: 0.84
5. **Neural Networks**:
   - Accuracy: 85%
   - Precision: 78%
   - Recall: 75%
   - F1-Score: 76%
   - ROC-AUC: 0.87

Reference Screen Shots

Evaluation results for target: Yes

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.832	0.899	0.573	0.800	0.447	0.550
Random Forest	0.736	0.854	0.252	0.569	0.162	0.247
SVM	0.741	0.848	0.381	0.500	0.307	0.311
kNN	0.602	0.827	0.090	0.227	0.056	0.041
Neural Network	0.783	0.855	0.465	0.532	0.413	0.387

Evaluation results for target: No

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.832	0.899	0.943	0.908	0.980	0.550
Random Forest	0.736	0.854	0.919	0.867	0.978	0.247
SVM	0.741	0.848	0.913	0.884	0.945	0.311
kNN	0.602	0.827	0.905	0.851	0.966	0.041
Neural Network	0.783	0.855	0.916	0.899	0.935	0.387

Deployment

1. Deploy the best-performing model (**Logistic Regression**) to predict employee attrition (Yes/No) in real-world HR systems, enabling proactive retention strategies.
2. **Model Selection**:
   Logistic Regression was chosen due to:
   - High performance metrics (AUC = 0.83, Accuracy = 90.1%).
   - Interpretability for identifying key drivers of attrition.
3. **Deployment Workflow**:
   - **Input**: Employee features (e.g., Age, MonthlyIncome, OverTime).
   - **Processing**: Automated normalization and encoding of input data.

- **Output**:
  - Attrition prediction (Yes/No).
  - Probability score for attrition risk.
  - Key factors influencing the prediction.

4. **Tools and Platforms**:
   - Deployment options include:
     - **Web Application** for batch or real-time predictions.
     - **Excel Integration** for HR teams' convenience.
     - **Dashboard Integration** with Power BI or Tableau.
5. **User Benefits**:
   - Alerts for high-risk employees.
   - Data-driven recommendations for HR (e.g., improve job satisfaction, adjust work-life balance).
6. **Monitoring and Updates**:
   - Regular performance tracking on live data.
   - Periodic model retraining with new data to adapt to changing patterns.
7. **Impact**:
   - Helps reduce employee turnover costs.
   - Improves workforce satisfaction through targeted interventions.
   - Empowers HR with actionable insights for retention strategies.

This deployment ensures the Logistic Regression model is seamlessly integrated into HR workflows, delivering meaningful and actionable outcomes.

Conclusion:

The project successfully built a predictive system for employee attrition using machine learning. The Logistic Regression model provided the best balance of performance and interpretability, making it suitable for deployment in real-world HR workflows.

REFERENCES:

- Dataset source: Kaggle(https://www.kaggle.com/datasets/patelprashant/employee-attrition)
- Daniel T. Larose-Discovering Knowledge in Data_ An Introduction to Data Mining-Wiley-Interscience
- https://sit.instructure.com/courses/74412/files/13748864?module_item_id=2257740
- https://sit.instructure.com/courses/74412/files/13749404?module_item_id=2257744