

ORANGE DATA MINING - PROJECT REPORT

Employee Attrition Prediction Using Machine Learning

Problem Statement

An organization wants to predict whether an employee is likely to leave the company based on factors such as age, department, job satisfaction, monthly income, years at the company, overtime, and work environment. By building a machine learning model in Orange Data Mining, the organization can identify employees who are at risk of leaving and take appropriate measures to improve employee retention and reduce turnover.

Software Required

- Orange Data Mining

Features

Feature	Description
Age	Employee age
MonthlySalary	Monthly salary
Experience	Years of experience
Overtime	Yes/No
JobSatisfaction	Low/Medium/High
WorkLifeBalance	Poor/Average/Good
TrainingHours	Training attended
EmployeeLeft	Target Variable
Target Variable Employee Left • Yes • No	

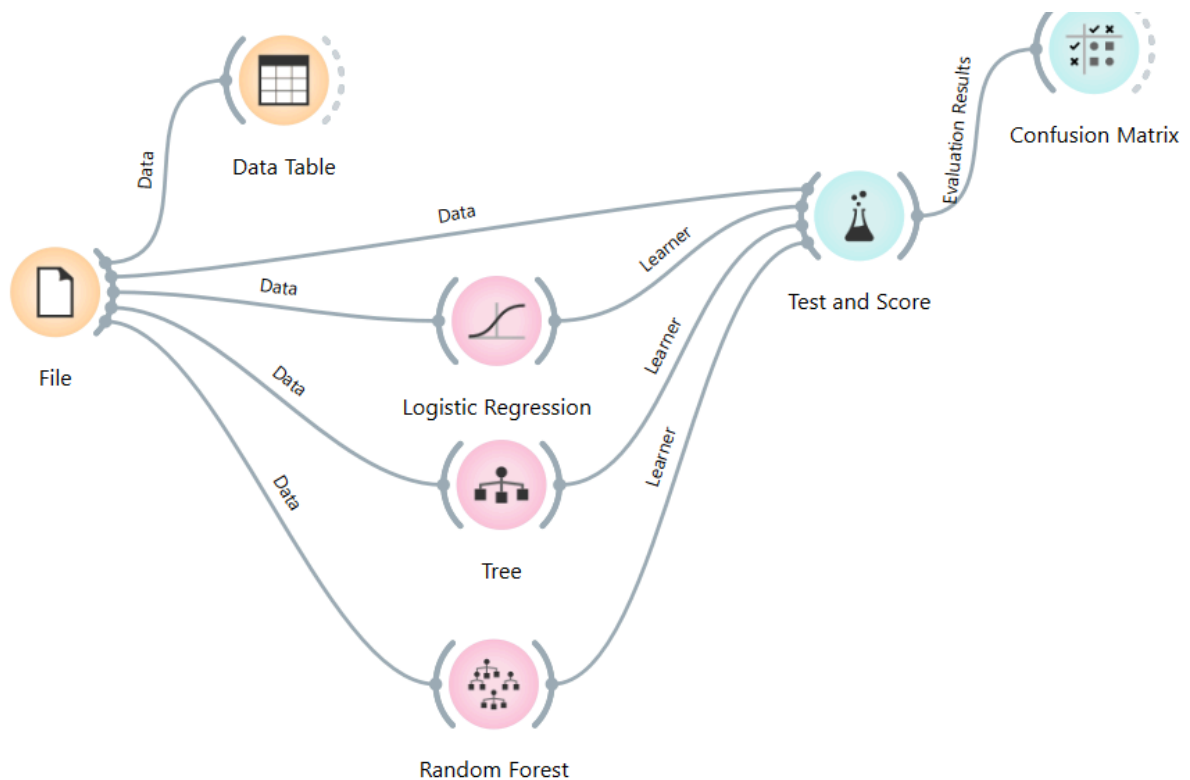
Step 1: Load Training Dataset

1. Open Orange Data Mining.
2. Add the File widget.
3. Click Browse and select Employee_Attrition_Train.csv.
4. Click Open to load the dataset.

Step 2: Check Variable Roles

Columns (Double click to edit)			
	Name	Role	Values
1	Age	feature	
2	MonthlySalary	feature	
3	Experience	feature	
4	Overtime	feature	No, Yes
5	JobSatisfaction	feature	High, Low, Medium
6	WorkLifeBalance	feature	Average, Good, Poor
7	TrainingHours	feature	
8	Attrition	target	No, Yes

ODM Workflow Step 4 - 6 :



Step 4: Add Machine Learning Models

- Random forest
- Logistic regression
- Tree

Step 5: Add Test & Score

Role of Test & Score Widget

Connection and function	Purpose
File → Test & Score	Provides the dataset for training and testing.
Model (Logistic Regression/Decision Tree/Random Forest) → Test & Score	Provides the machine learning algorithm (learner) to be trained and evaluated.
Display Metrics	Shows Accuracy (CA), Precision, Recall, F1-score, AUC, etc.
Select Best Model	Helps identify the best-performing model for deployment.

Step 6: Observe Results

Orange displays *(Actual values will depend on the dataset.)*

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.994	0.988	0.987	0.988	0.988	0.972
Tree	0.981	0.988	0.987	0.988	0.988	0.972
Random Forest	0.999	0.988	0.987	0.988	0.988	0.972

Step 7: Compare Models

Which model performed the best?

Random Forest has the highest Accuracy and F1 Score.

Step 8: Confusion Matrix

Learner : Logistic Regression

		Predicted		
		No	Yes	Σ
Actual	No	98.1 %	0.0 %	53
	Yes	1.9 %	100.0 %	27
Σ		54	26	80

TP :Correctly predicted employee will leave.

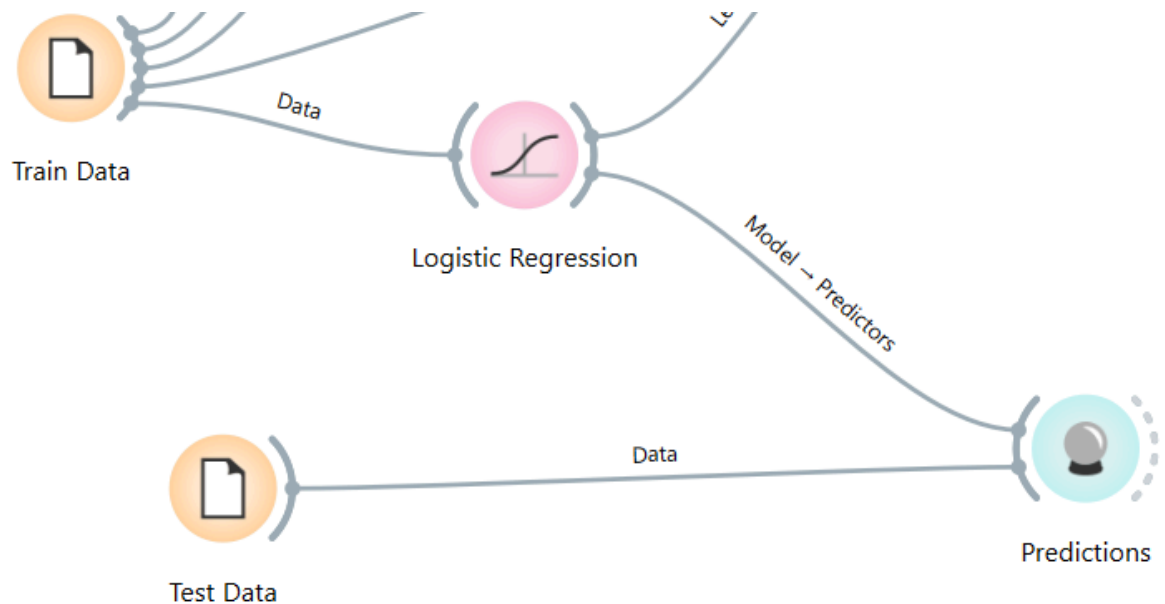
TN: Correctly predicted employee will stay.

FP: Predicted employee will leave but actually stayed.

FN: Predicted employee will stay but actually left.

Step 9: Prediction on New Employees

1. Open **Employee_Attrition_test_Predict.csv** using the **File** widget.
2. Open **Predictions** to view the results.



3.

Example Output:

Employee	Prediction
Employee 1	No
Employee 2	Yes
Employee 3	No

Result:

The AI model predicts whether a new employee is likely to leave the company.

AI predicts outcomes for new unseen employees.

Understanding the Algorithms

- **Random Forest:** Combines many decision trees to make more accurate and reliable predictions. It reduces overfitting and improves performance.
- **Linear Regression:** Predicts a continuous value by finding the best-fit straight line between input and output data.

- **Decision Tree:** Makes predictions by asking a series of yes/no questions based on the data. It is simple to understand and easy to interpret.

Why do the algorithms give different results?

The dataset contains

- Borderline employees
- Similar employees with different outcomes
- Some noisy records

Therefore, Different algorithms learn different patterns.

Evaluation Metrics

Accuracy : Higher Accuracy is better.

Precision: Higher Precision means fewer False Positives.

Recall : Higher Recall means fewer False Negatives.

F1 Score : Balances Precision and Recall. Useful when both are important.

Which Model Should We Choose?

Do **not** choose a model based only on **Accuracy**.

Compare all evaluation metrics.

Criteria	Logistic Regression	Decision Tree	Random Forest	kNN
Accuracy	Good	Better	Excellent	Very Good
Precision	Good	Better	Excellent	Very Good
Recall	Good	Better	Excellent	Very Good
F1 Score	Good	Better	Excellent	Very Good

Decision Rule for Students

1. Compare Accuracy.
 2. If Accuracy is similar, compare F1 Score.
 3. If F1 Score is similar, compare Precision and Recall.
 4. If performance is still similar, choose the simpler and more explainable model.
 5. If one model is clearly better across most metrics, deploy that model.
-

Key Takeaway

There is **no universally "best" algorithm**. The best model depends on:

- The characteristics of the dataset.
- The evaluation metrics (Accuracy, Precision, Recall, F1-score).
- The need for interpretability versus predictive performance.
- The application domain and deployment requirements.