**Predicting Outcomes: A Study of Decision Trees and Logistic Regression in Orange Data Mining**

Abstract

In this report, we implemented and evaluated two machine learning models using Orange Data Mining. A Decision Tree model was used to classify whether Sara would go sailing based on weather conditions, company, and sailboat size. Additionally, a Logistic Regression model was applied to classify Iris flower species based on their petal and sepal measurements. The performance of both models was evaluated using confusion matrices and key metrics like accuracy, precision, recall, and F1-score. The report also includes calculations for class probabilities using Logistic Regression and a manual completion of confusion matrices for a wine classification task.

# Introduction

In this report, we implement two machine learning models using Orange Data Mining. Based on three input features, we applied a Decision Tree model to predict whether Sara would sail. We also implemented a Logistic Regression model to classify Iris flowers into three species. The performance of the models was evaluated using key metrics like accuracy, precision, recall, and F1-score. We also manually completed confusion matrices for a wine classification task.

# Methodology

We used Orange Data Mining to build two models. One is a Decision Tree to predict if Sara will sail. The other is a Logistic Regression model to classify Iris flowers. We evaluated both models using accuracy, precision, recall, and F1-score.

## **Decision Tree Model**

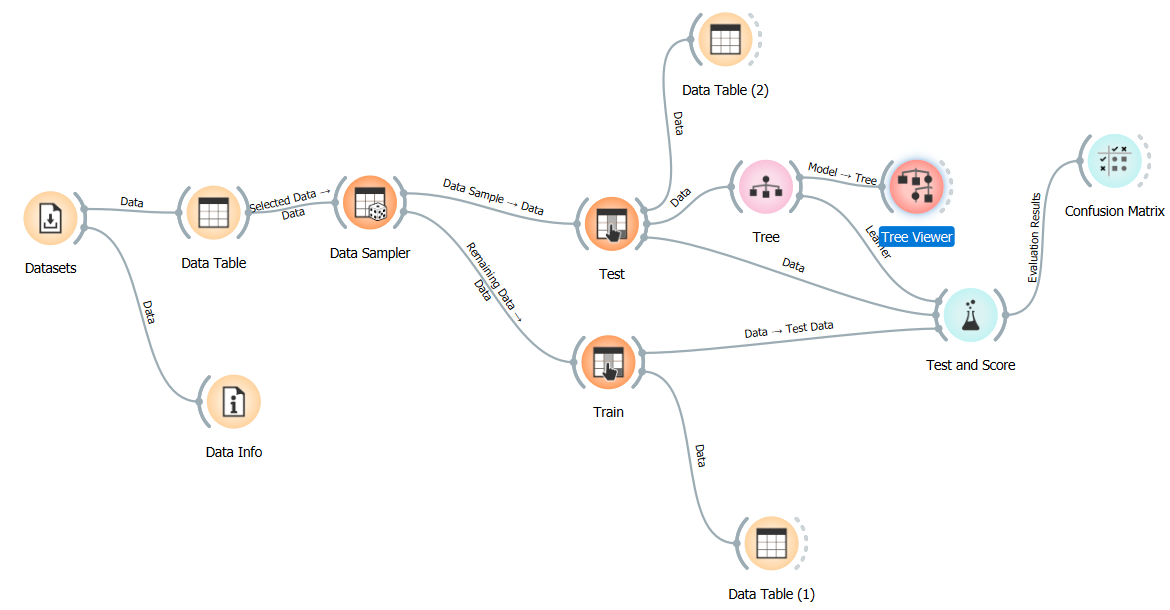


Figure 1: Decision tree evaluation flowchart.

We used the Sailing dataset to build a Decision Tree model to predict if Sara would sail. The data included weather, company, and sailboat size. We split the data into 75% training and 25% test sets. The tree stopped splitting when the majority class hit 90%, with a max depth of three. After training, we visualized the tree and made manual predictions for a few test examples. We evaluated the model using a confusion matrix, accuracy, precision, recall, and F1-score.

### Model Evaluation

The Binary Decision Tree model had a 60% accuracy on test data, with an AUC of 0.667, F1-score of 0.600, and MCC of 0.167. On training data, it performed better with 73.3% accuracy, an AUC of 0.804, F1-score of 0.733, and MCC of 0.464. The confusion matrix showed 3 correct and 2 incorrect predictions, with 2 correct “no” and 1 correct “yes” predictions. (Ishak, 2020)

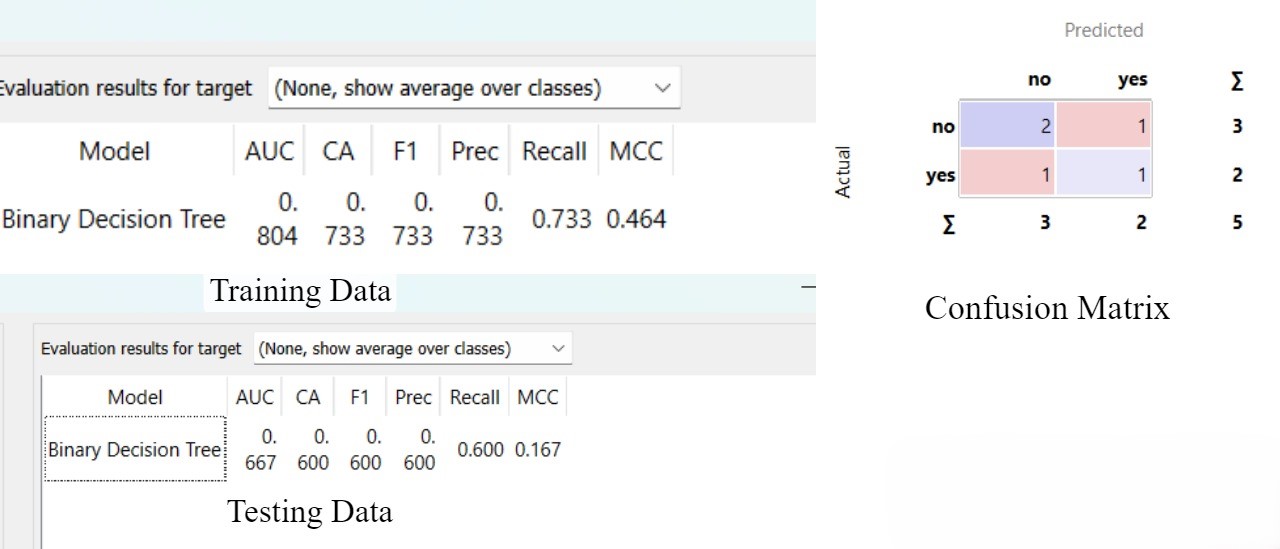


Figure 2: Model Evaluation Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test ID | Outlook | Company | Sailboat | Actual | Predicted |
| 1 | Sunny | Med | Big | Yes | Yes |
| 2 | Sunny | Med | Big | No | Yes |
| 3 | Rainy | No | Big | No | No |
| 4 | Rainy | Big | Big | Yes | No |
| 5 | Rainy | Med | Big | No | No |

### Insights

The decision tree provides insights into predicting whether Sara will sail based on the weather ("Outlook"), the company she has, and the size of the sailboat ("big" or "med").

* If it’s **rainy** and Sara has company, there’s an **83.3%** chance she won’t sail (**5 out of 6 times**).
* If it’s **sunny**, Sara is more likely to sail, with a **66.7%** chance (**6 out of 9 times**).
* On **sunny** days, if the boat is **big or medium**-sized and she has company, the chances of sailing increase to **71.4% (5 out of 7 instances)**.
* If Sara is alone or the boat is smaller, the likelihood of sailing decreases, though there is still some chance she will sail.

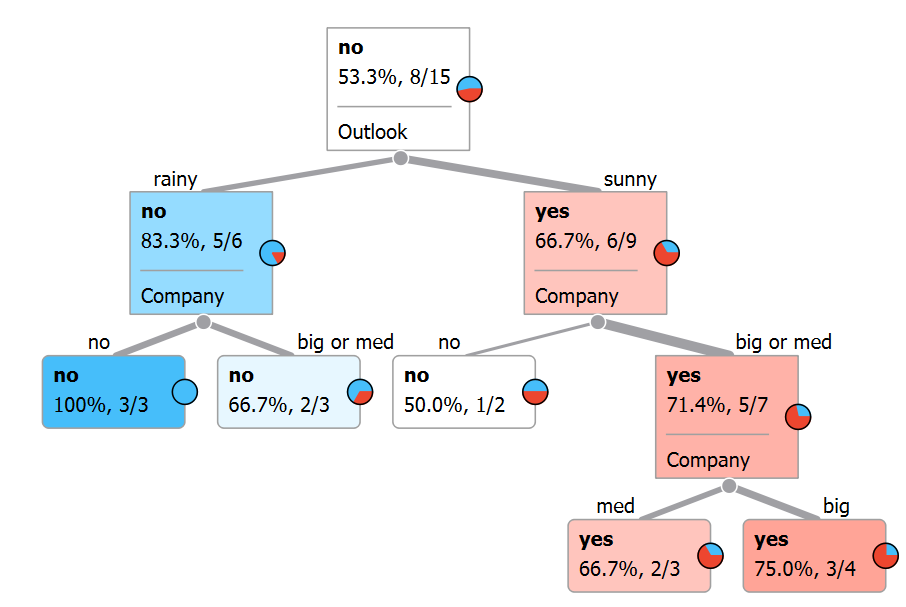


Figure 3: Decision Tree

## Logistic Regression Model

T, we used the Iris dataset to create a Logistic Regression model. The data included sepal and petal measurements for three Iris species. We split the data into 75% for training and 25% for testing. The model used L2 regularization. We calculated class probabilities for a few test examples manually. (Abdelmagid, 2020)

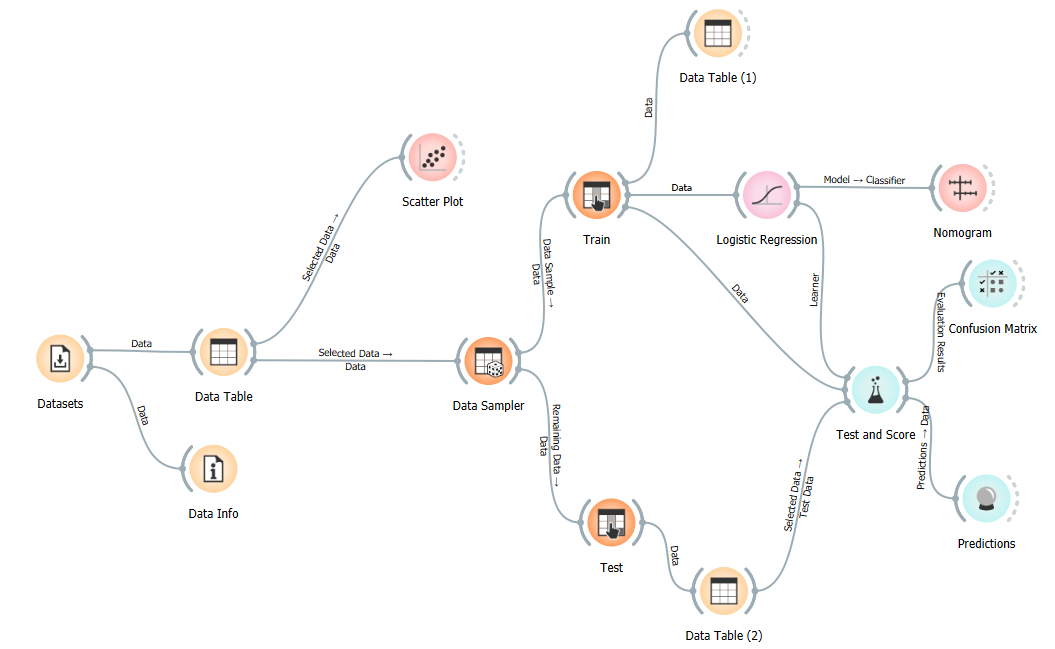


Figure : Logistic Regression evaluation flowchart.

### Model Evaluation

Both the logistic regression and decision tree models perfectly classified Iris flowers into Setosa, Versicolor, and Virginica. Sepal and petal measurements were key to their success. Petal length and width were especially important, as shown by the nomogram analysis. Overall, the models reliably predicted Iris species. (Fernandes, 2023)

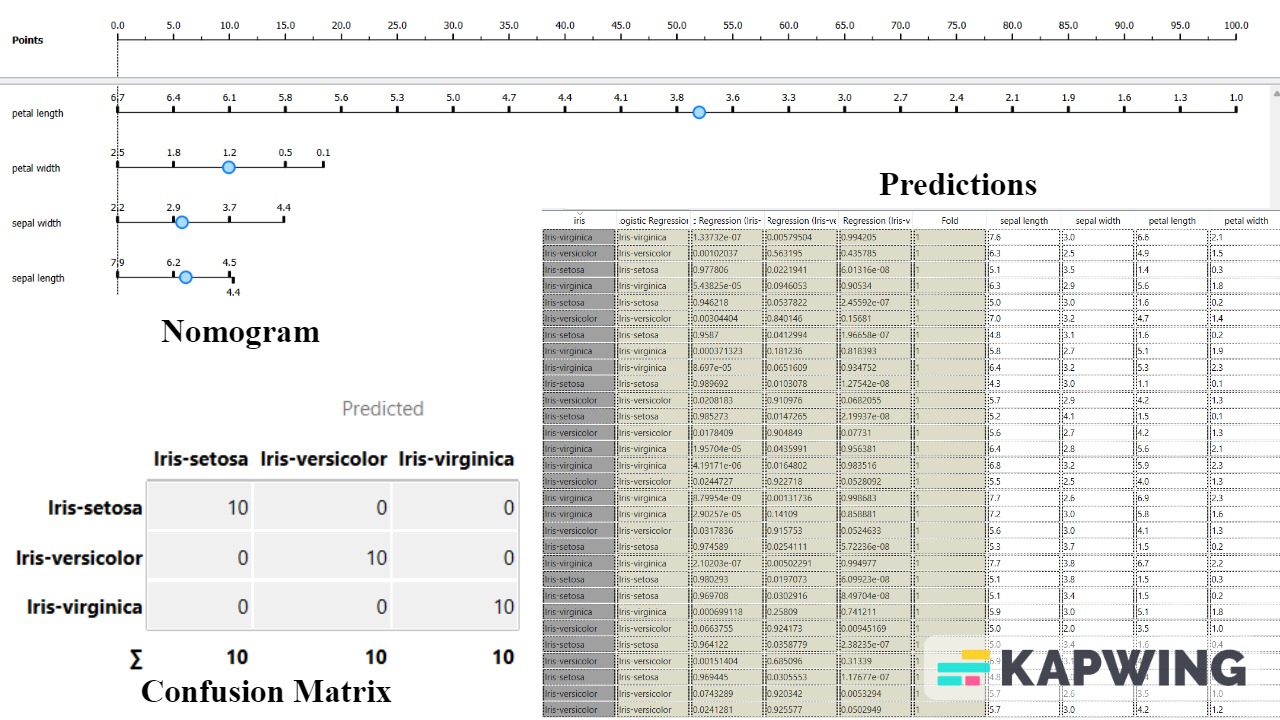


Figure : Model Evaluation Statistics

### Logistic Regression Model and Predictions

We built a logistic regression model to classify Iris flowers, with the following equations for each class:

* **Setosa:**  
  YSetosa=9.26545−0.398527×(Sepal Length)+0.91131×(Sepal Width)−2.36824×(Petal Length)−1.03474×(Petal Width)Y\_{Setosa} = 9.26545 - 0.398527 \times (Sepal \ Length) + 0.91131 \times (Sepal \ Width) - 2.36824 \times (Petal \ Length) - 1.03474 \times (Petal \ Width)YSetosa​=9.26545−0.398527×(Sepal Length)+0.91131×(Sepal Width)−2.36824×(Petal Length)−1.03474×(Petal Width)
* **Versicolor:**  
  YVersicolor=2.17695+0.415094× (Sepal Length) −0.233456×(Sepal Width)−0.147726×(Petal Length)−0.863045×(Petal Width)Y\_{Versicolor} = 2.17695 + 0.415094 \times (Sepal \ Length) - 0.233456 \times (Sepal \ Width) - 0.147726 \times (Petal \ Length) - 0.863045 \times (Petal \ Width)YVersicolor​=2.17695+0.415094×(Sepal Length)−0.233456×(Sepal Width)−0.147726×(Petal Length)−0.863045×(Petal Width)
* **Virginica:**  
  YVirginica=−11.4424−0.0165671× (Sepal Length) −0.677854×(Sepal Width)+2.51597×(Petal Length)+1.89779×(Petal Width)Y\_{Virginica} = -11.4424 - 0.0165671 \times (Sepal \ Length) - 0.677854 \times (Sepal \ Width) + 2.51597 \times (Petal \ Length) + 1.89779 \times (Petal \ Width)YVirginica​=−11.4424−0.0165671×(Sepal Length)−0.677854×(Sepal Width)+2.51597×(Petal Length)+1.89779×(Petal Width)

Using these equations, we calculated probabilities for the first four instances of the test data. The class with the highest probability was selected as the prediction, and all predictions matched the actual class, demonstrating the model's accuracy.

# Wine Classification Using Decision Tree and Logistic Regression

We used both Decision Tree and Logistic Regression models to classify wines. The Decision Tree model had an accuracy of 80%, with Type 3 wines achieving 100% recall and a 0.89 F1 score. The Logistic Regression model performed better overall, with a 90% accuracy. It showed perfect precision for Type 1 wines and 100% recall for Type 3 wines, indicating strong performance across all types. Overall, the Logistic Regression model was more effective, especially for Type 3 wines.

# References

Abdelmagid, A. a. Q. A., 2020. Utilizing the Educational Data Mining Techniques" Orange Technology" for Detecting Patterns and Predicting Academic Performance of University Students. *nf. Sci. Lett,* Volume 12(3), pp. 1415-1431..

Fernandes, A. F. F. D. R. E. a. N. W., 2023. Revista de Sociologia e Política. pp. 28, p.006..

Ishak, A. S. K. G. R. a. A. M., 2020. *Orange software usage in data mining classification method on the dataset lenses.* s.l., IOP Conference Series: Materials Science and Engineering.