INSTITUTE OF TECHNOLOGICAL STUDIES OF BIZERTE
DEPARTMENT OF ELECTRICAL ENGINEERING

MACHINE LEARNING LAB REPORT

Chirine Bjaoui
chirneb@gmail.com
https://github.com/chirineb1

Abstract — **Comparative Analysis of Machine Learning Models for Rainfall Prediction**

Predicting rainfall accurately is essential for planning in sectors like agriculture, transportation, and disaster management. Traditional forecasting methods may not fully leverage historical data patterns. The challenge is to build a reliable classification model that predicts whether it will rain tomorrow, using weather data and machine learning techniques. This project addresses this problem by applying multiple supervised learning algorithms within the KNIME platform.

I. INTRODUCTION

This project aims to build and evaluate five machine learning classification models using KNIME on a weather dataset stored locally.

The models include:

Logistic Regression

Decision Tree

Random Forest

Gradient Boosting

Support Vector Classifier (SVC)

Evaluation is based on:

Accuracy

Precision

Recall

F1 Score

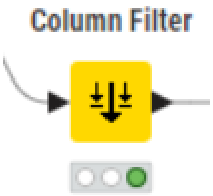
II. DATASET DESCRIPTION

The dataset used in this study is a weather-related dataset located locally on the system. The target variable is raintomorrow, indicating whether it will rain the next day.

1) *Data Preprocessing:*

Column Removal:

Removing unnecessary columns: [‘date’, ‘location’, ‘evaporation’, ‘sunshine’] using the **Column Filter** node.



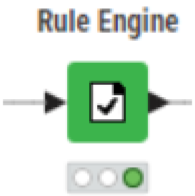
Missing Value Handling: Using the **Missing Value** node.

One-hot encoding:

Applied to categorical columns: ['raintoday', 'windgustdir', 'winddir9am', 'winddir3pm'] using **One to Many** node.

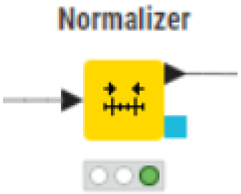
Label Encoding:

Label encode the 'raintomorrow' column (target variable) using a **Rule Engine** node.



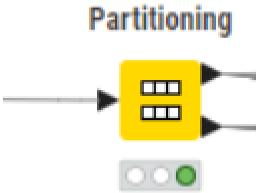
Normalization:

Applied to continuous features using the **Normalizer** node.



2) *Train-Test Split:*

Split the dataset using the **Partitioning** node with:



Training size: 80%

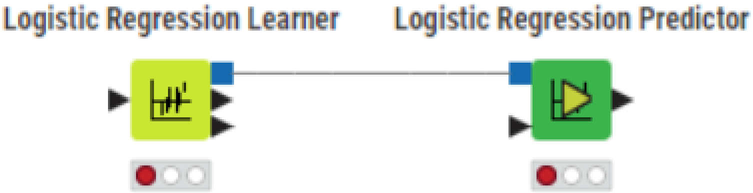
Testing size : 20%

Random seed: 10

3) *Model Configuration and Training:*

Each of the five models trained with specific hyperparameters:

Logistic Regression :

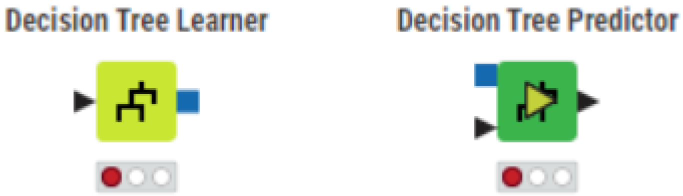


penalty=‘l2’

C=0.8

random_state=4

Decision Tree :



criterion='entropy'

min_samples_split=4

random_state=32

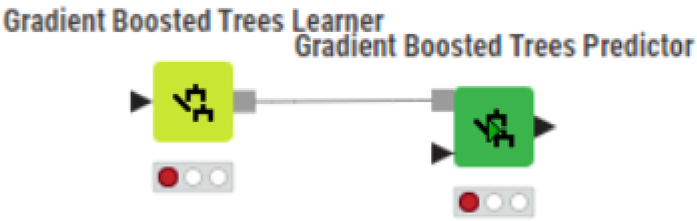
Random Forest :

n_estimators=128

criterion='entropy'

min_samples_split=8

Gradient Boosted classifier :


Gradient Boosted Trees Learner
Gradient Boosted Trees Predictor

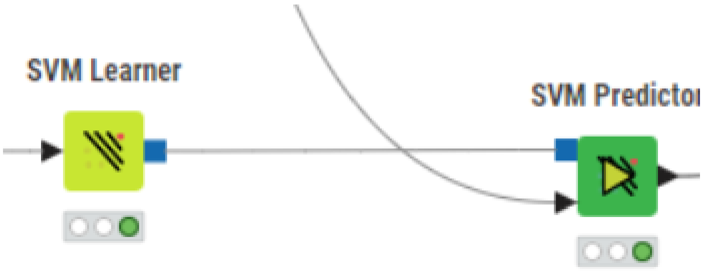
loss='exponential'

learning_rate=0.3

n_estimators=64

criteria='friedman_mse'

SVC :



kernel='sigmoid'

gamma='auto'

random_state=64

These settings were manually configured in the corresponding learner nodes.

4) *Model Training and Prediction:*

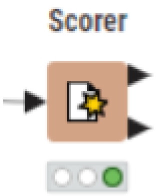
Each model follows this flow:

Learner node: Trained with x_train and y_train.

Predictor node: Used to predict y_pred from x_test.

Scorer node: Used to compute evaluation metrics.

scorer node :



5) *Evaluation Metrics:*

Each model's performance was evaluated using:

Confusion Matrix

Accuracy

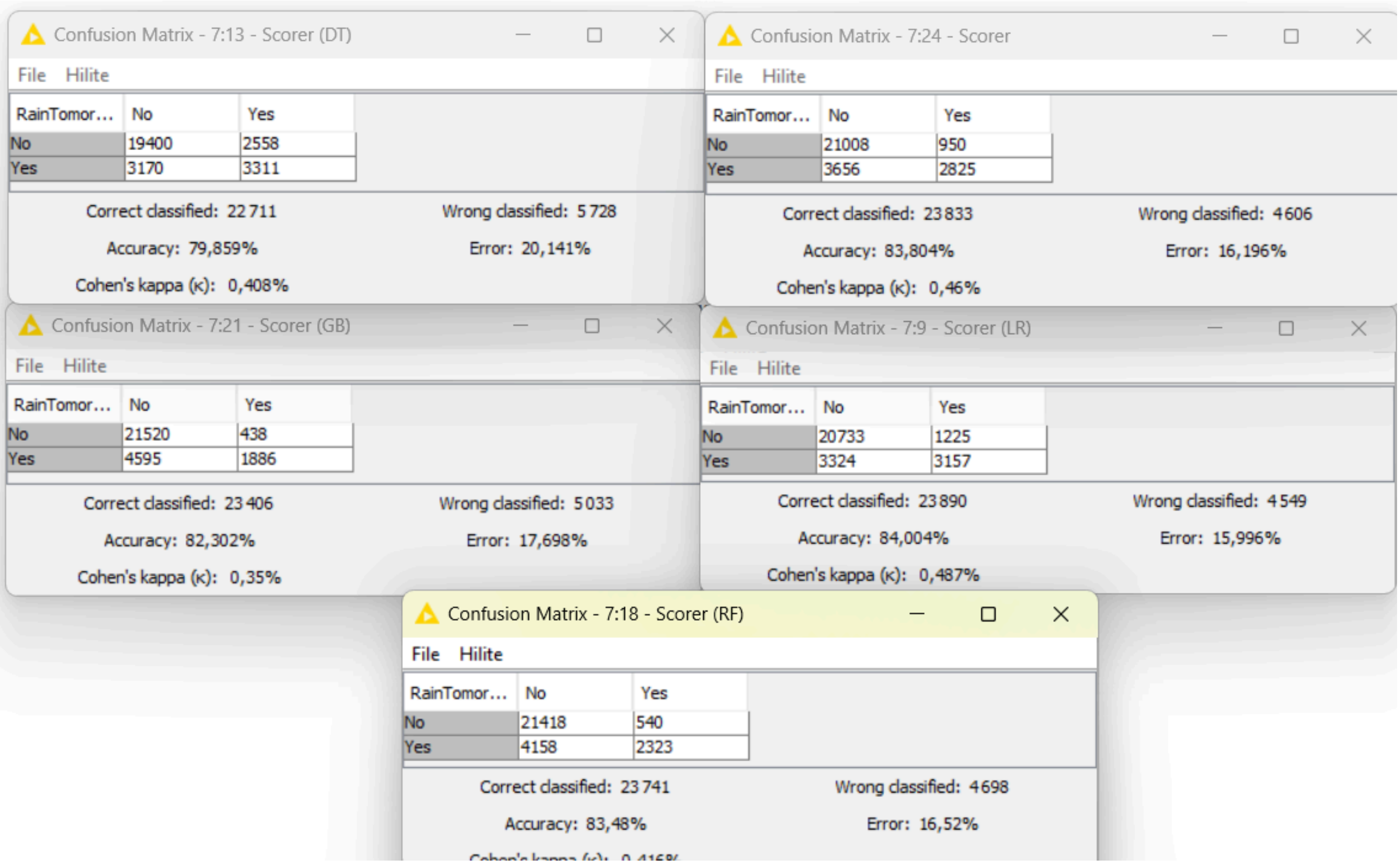
Precision

Recall

F1 Score

6) *Confusion Matrix:*

Each model's confusion matrix was analyzed to assess classification performance visually.



Confusion Matrix - 7:13 - Scorer (DT)

RainTomor...	No	Yes
No	19400	2558
Yes	3170	3311

Correct classified: 22 711 Wrong classified: 5 728

Accuracy: 79,859% Error: 20,141%

Cohen's kappa (κ): 0,408%

Confusion Matrix - 7:24 - Scorer

RainTomor...	No	Yes
No	21008	950
Yes	3656	2825

Correct classified: 23 833 Wrong classified: 4 606

Accuracy: 83,804% Error: 16,196%

Cohen's kappa (κ): 0,46%

Confusion Matrix - 7:21 - Scorer (GB)

RainTomor...	No	Yes
No	21520	438
Yes	4595	1886

Correct classified: 23 406 Wrong classified: 5 033

Accuracy: 82,302% Error: 17,698%

Cohen's kappa (κ): 0,35%

Confusion Matrix - 7:9 - Scorer (LR)

RainTomor...	No	Yes
No	20733	1225
Yes	3324	3157

Correct classified: 23 890 Wrong classified: 4 549

Accuracy: 84,004% Error: 15,996%

Cohen's kappa (κ): 0,487%

Confusion Matrix - 7:18 - Scorer (RF)

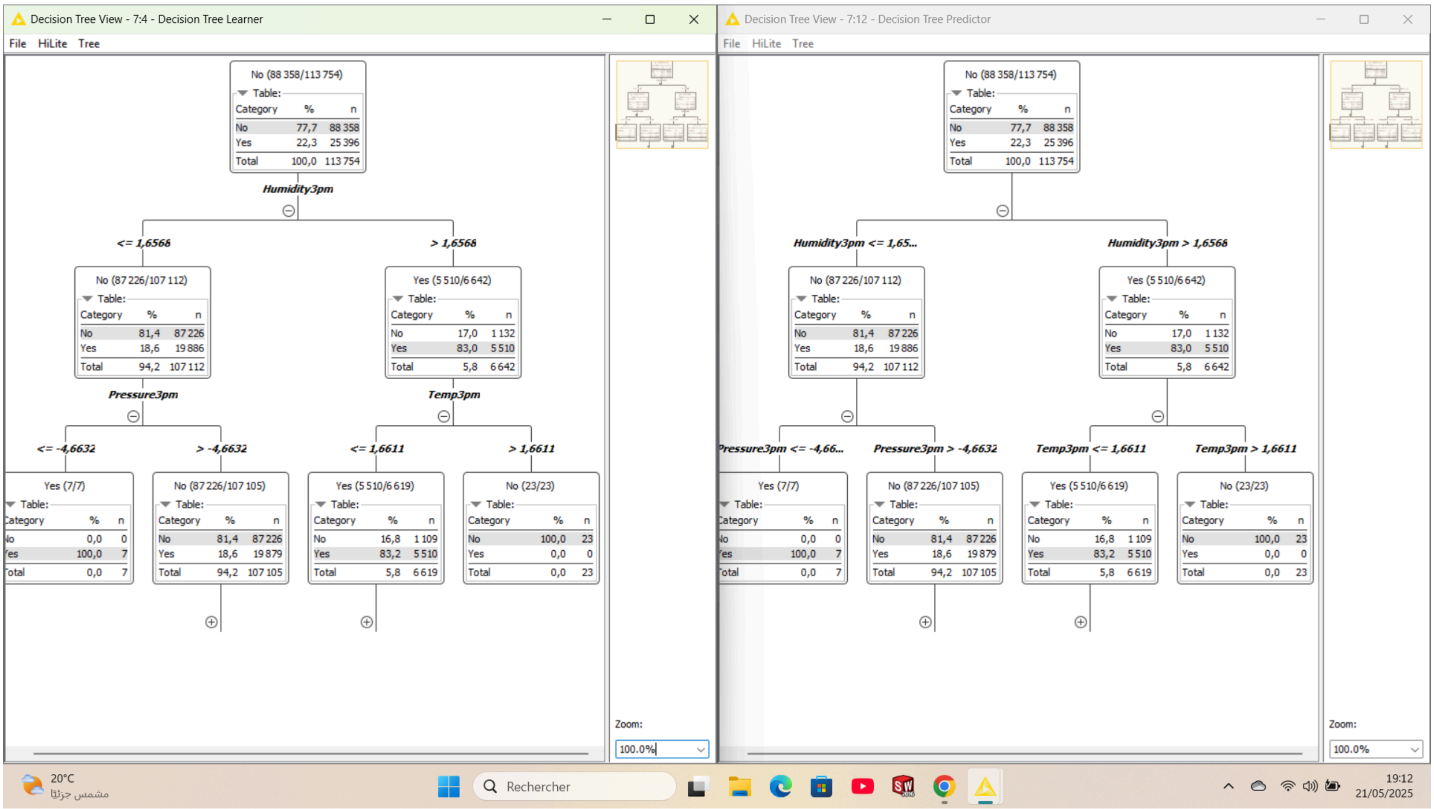
RainTomor...	No	Yes
No	21418	540
Yes	4158	2323

Correct classified: 23 741 Wrong classified: 4 698

Accuracy: 83,48% Error: 16,52%

Cohen's kappa (κ): 0,416%

Decision Tree Learner and Predictor Results



Decision Tree View - 7:4 - Decision Tree Learner

No (88 358/113 754)

Category	%	n
No	77,7	88 358
Yes	22,3	25 396
Total	100,0	113 754

Humidity3pm

<= 1,6568

No (87 226/107 112)

Category	%	n
No	81,4	87 226
Yes	18,6	19 886
Total	94,2	107 112

Pressure3pm

<= -4,6632

Yes (7/7)

Category	%	n
No	0,0	0
Yes	100,0	7
Total	0,0	7

> -4,6632

No (87 226/107 105)

Category	%	n
No	81,4	87 226
Yes	18,6	19 879
Total	94,2	107 105

> 1,6568

Yes (5 510/6 642)

Category	%	n
No	17,0	1 132
Yes	83,0	5 510
Total	5,8	6 642

Temp3pm

<= 1,6611

Yes (5 510/6 619)

Category	%	n
No	16,8	1 109
Yes	83,2	5 510
Total	5,8	6 619

> 1,6611

No (23/23)

Category	%	n
No	100,0	23
Yes	0,0	0
Total	0,0	23

Decision Tree View - 7:12 - Decision Tree Predictor

No (88 358/113 754)

Category	%	n
No	77,7	88 358
Yes	22,3	25 396
Total	100,0	113 754

Humidity3pm <= 1,65...

No (87 226/107 112)

Category	%	n
No	81,4	87 226
Yes	18,6	19 886
Total	94,2	107 112

Pressure3pm <= -4,66...

Yes (7/7)

Category	%	n
No	0,0	0
Yes	100,0	7
Total	0,0	7

Pressure3pm > -4,6632

No (87 226/107 105)

Category	%	n
No	81,4	87 226
Yes	18,6	19 879
Total	94,2	107 105

Humidity3pm > 1,6568

Yes (5 510/6 642)

Category	%	n
No	17,0	1 132
Yes	83,0	5 510
Total	5,8	6 642

Temp3pm <= 1,6611

Yes (5 510/6 619)

Category	%	n
No	16,8	1 109
Yes	83,2	5 510
Total	5,8	6 619

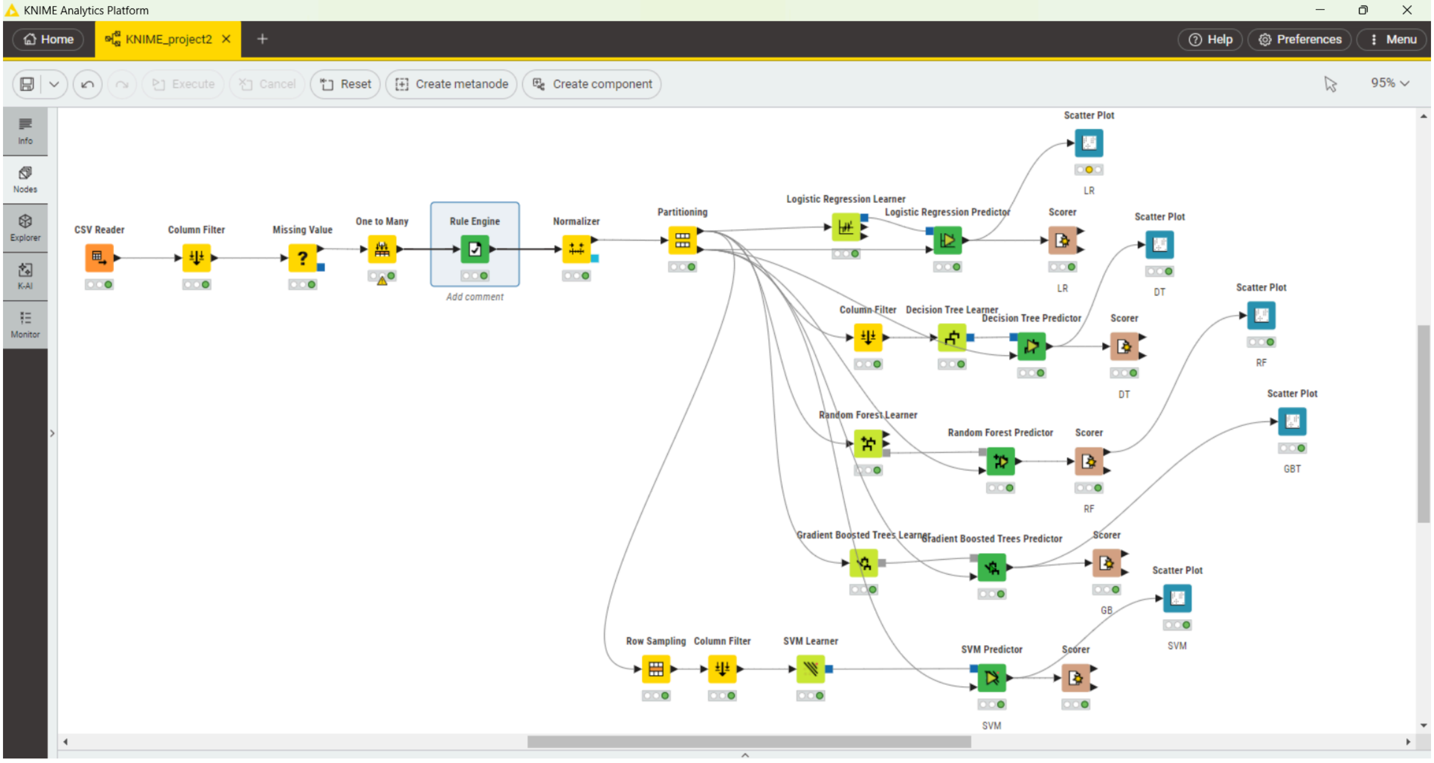
Temp3pm > 1,6611

No (23/23)

Category	%	n
No	100,0	23
Yes	0,0	0
Total	0,0	23

7) *KNIME Workflow Overview:*

The complete KNIME workflow, illustrated below, outlines all steps followed in the project — from data preprocessing to model training, prediction, evaluation, and visualization.



III. Observations

Best Performing Model: Logistic Regression achieved the highest accuracy (83.804%) and highest Cohen's Kappa score (0.46), indicating it was the most effective at predicting rainfall tomorrow.

Precision-Recall Tradeoff:

Gradient Boosting had the fewest false positives (438) but also the fewest true positives (1,886), suggesting it's very conservative about predicting rain.

Logistic Regression had more balanced results with 2,825 true positives while maintaining reasonable false positives (950).

IV. Conclusion

This KNIME workflow demonstrates how to apply multiple classification algorithms to a preprocessed dataset, evaluate performance using standard metrics, and visualize predictions.