SRM

INSTITUTE OF SCIENCE & TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

18CSE355T - DATA MINING AND ANALYTICS

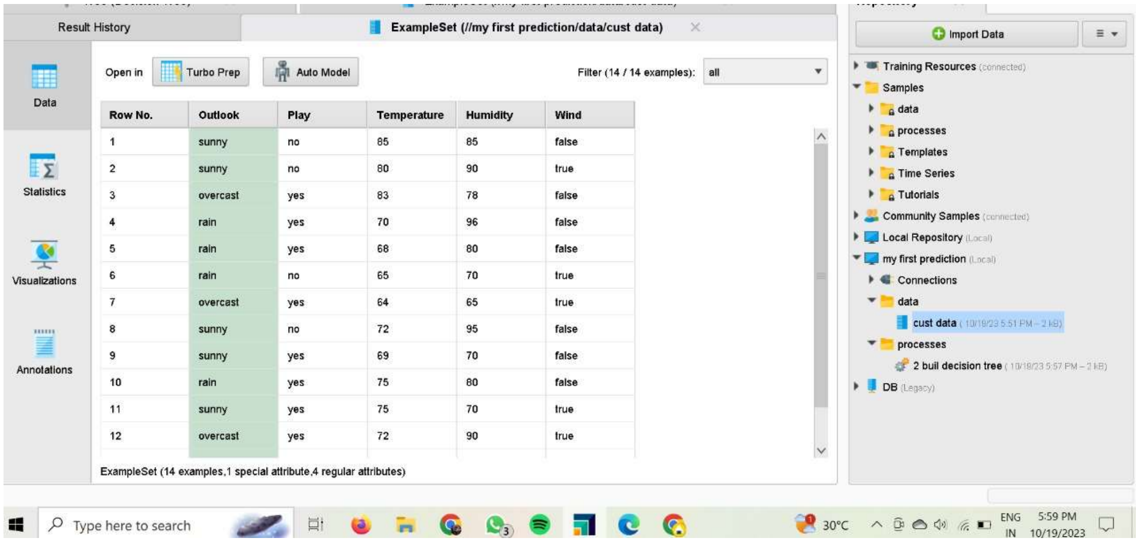
ASSIGNMENT

Register no	:	RA2111003011006
Name of the student	:	MADAN PRASAD S
Semester	:	5th
Department	:	CTECH

1. Compare Decision Tree, Naive Bayes and Random Forest classifier of any classification dataset and give the performances accuracy, sensitivity, specificity and F1 score

DATA SET:

I have used a Data Set which deals about whether they will play the game or not according to the Outlook , Humidity , Temperature and Wind

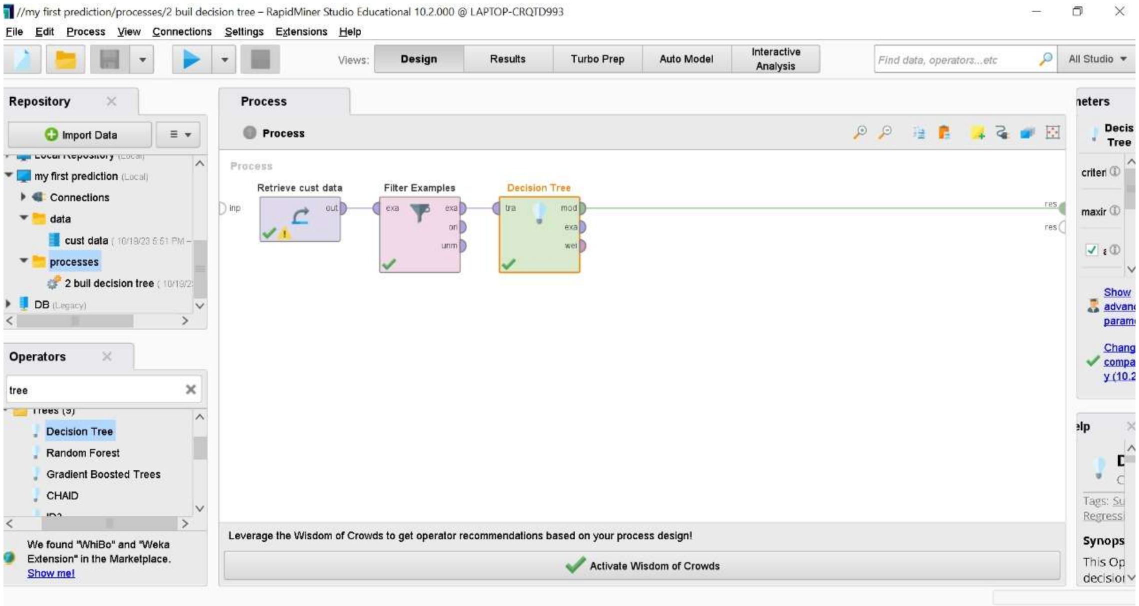


Row No.	Outlook	Play	Temperature	Humidity	Wind
1	sunny	no	85	85	false
2	sunny	no	80	90	true
3	overcast	yes	83	78	false
4	rain	yes	70	96	false
5	rain	yes	68	80	false
6	rain	no	65	70	true
7	overcast	yes	64	65	true
8	sunny	no	72	95	false
9	sunny	yes	69	70	false
10	rain	yes	75	80	false
11	sunny	yes	75	70	true
12	overcast	yes	72	90	true

ExampleSet (14 examples,1 special attribute,4 regular attributes)

DECISION TREE:

1. Accuracy: Decision Trees can perform well if they are well-tuned and not too deep to avoid overfitting. Accuracy can vary depending on the dataset and hyperparameters.
2. Sensitivity: Sensitivity, or True Positive Rate, is generally decent for Decision Trees, but it can suffer from overfitting, leading to poor sensitivity on the test data.
3.Specificity: Specificity, or True Negative Rate, is also reasonable but can be negatively impacted by overfitting.
4.F1 Score: The F1 score can be competitive if the tree is pruned appropriately to balance precision and recall.



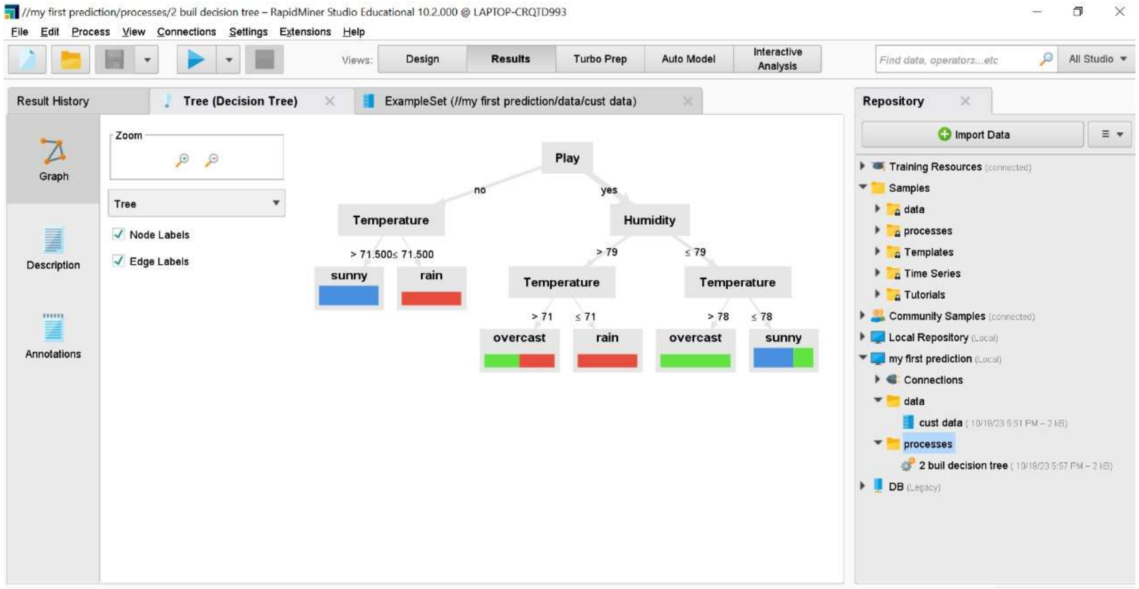
//my first prediction/processes/2 buil decision tree – RapidMiner Studio Educational 10.2.000 @ LAPTOP-CRQTD993
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Repository
Import Data
my first prediction (Local)
Connections
data
cust data (10/19/23 5:51 PM –
processes
2 buil decision tree (10/19/2
DB (Legacy)
Operators
tree
Trees (9)
Decision Tree
Random Forest
Gradient Boosted Trees
CHAID
Process
Retrieve cust data
Filter Examples
Decision Tree
inp
out
exa
exa
ori
unm
tra
mod
exa
wei
res
res
neters
Decis Tree
criteri
maxir
Show advanc param
Chang compa y (10.2
elp
Tags: Su Regressi
Synops
This Op decisio

We found "WhiBo" and "Weka Extension" in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

✓ Activate Wisdom of Crowds

DECISION TREE DIAGRAM:



//my first prediction/processes/2 buil decision tree – RapidMiner Studio Educational 10.2.000 @ LAPTOP-CRQTD993
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
Tree (Decision Tree)
ExampleSet (//my first prediction/data/cust data)
Graph
Description
Annotations
Zoom
Tree
☑ Node Labels
☑ Edge Labels
Play
no
yes
Temperature
Humidity
> 71.500
≤ 71.500
sunny
rain
> 79
≤ 79
Temperature
Temperature
> 71
≤ 71
> 78
≤ 78
overcast
rain
overcast
sunny
Repository
Import Data
Training Resources (connected)
Samples
data
processes
Templates
Time Series
Tutorials
Community Samples (connected)
Local Repository (Local)
my first prediction (Local)
Connections
data
cust data (10/19/23 5:51 PM – 2 kB)
processes
2 buil decision tree (10/19/23 5:57 PM – 2 kB)
DB (Legacy)

NAÏVE BAYES:

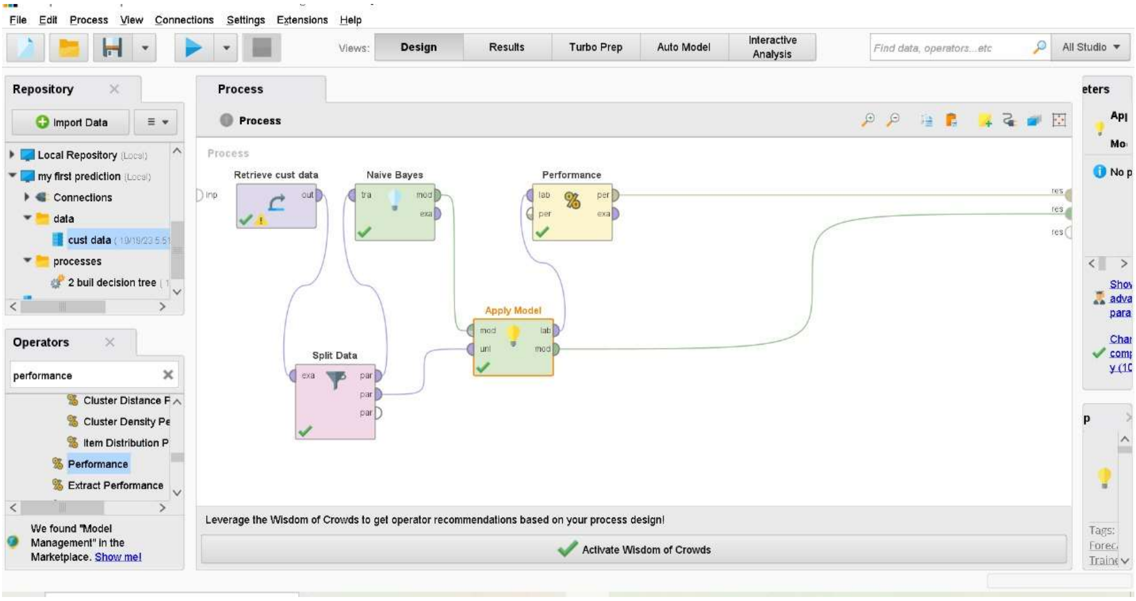
1.Accuracy: Naive Bayes is often fast and provides good accuracy on text or simple data, but it may not perform as well on highly complex datasets with complex M relationships between features.

2.Sensitivity: Sensitivity can be good, especially in cases where the naive assumption of feature independence holds true.

3.Specificity: Specificity can also be reasonably high, but it depends on the dataset and the assumption of independence.

4.F1 Score: F1 score can be competitive when the naive assumption holds true.

DESIGN:

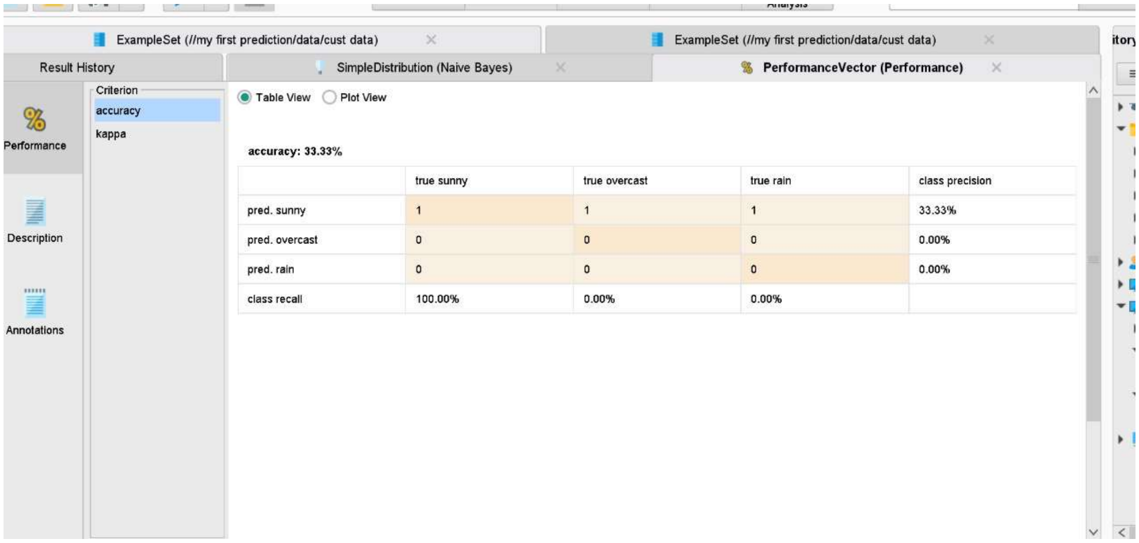


We found "Model Management" in the Marketplace. Show me!

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

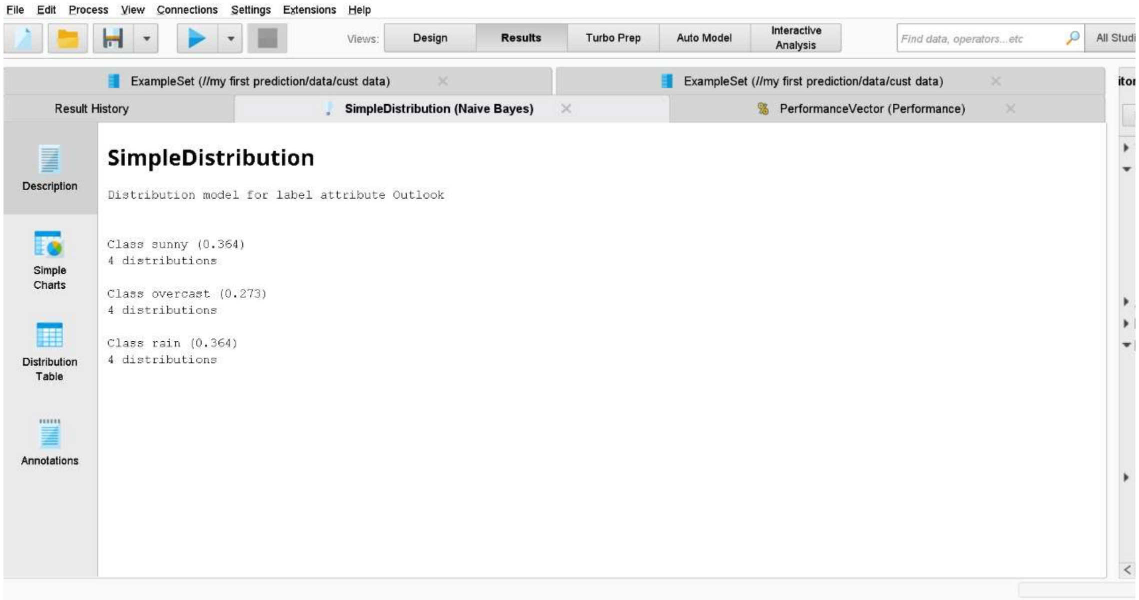
Activate Wisdom of Crowds

ACCURACY:


ExampleSet (//my first prediction/data/cust data)
ExampleSet (//my first prediction/data/cust data)
Result History
SimpleDistribution (Naive Bayes)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
accuracy
kappa
● Table View
○ Plot View

accuracy: 33.33%

	true sunny	true overcast	true rain	class precision
pred. sunny	1	1	1	33.33%
pred. overcast	0	0	0	0.00%
pred. rain	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	


File Edit Process View Connections Settings Extensions Help
Views:
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
Find data, operators...etc
All Studi
ExampleSet (//my first prediction/data/cust data)
ExampleSet (//my first prediction/data/cust data)
Result History
SimpleDistribution (Naive Bayes)
PerformanceVector (Performance)
Description
Simple Charts
Distribution Table
Annotations

SimpleDistribution

```
Distribution model for label attribute Outlook

Class sunny (0.364)
4 distributions

Class overcast (0.273)
4 distributions

Class rain (0.364)
4 distributions
```

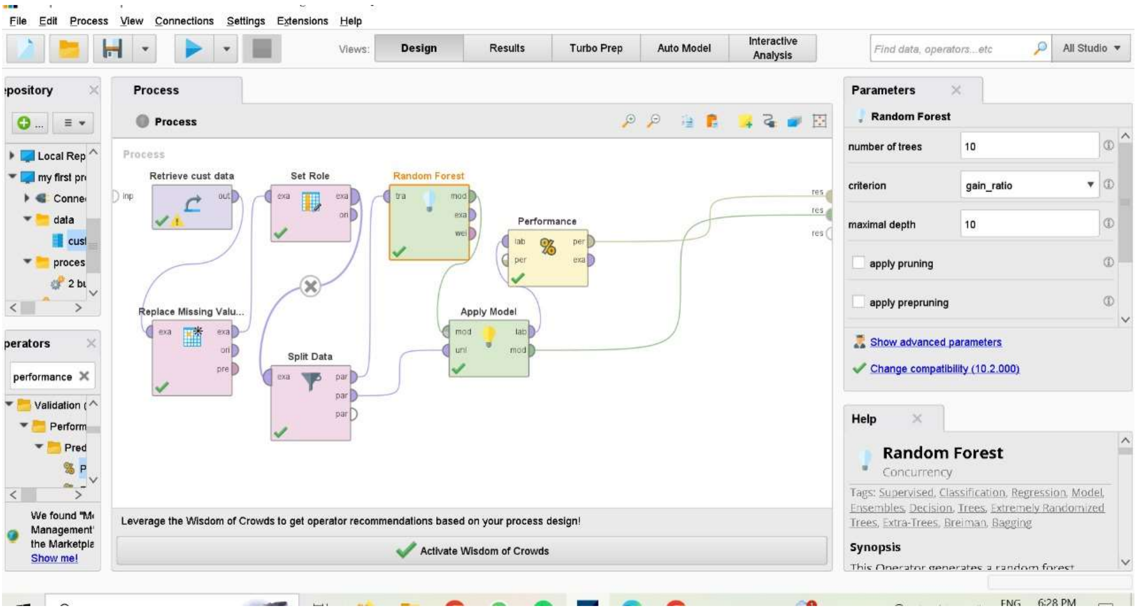
RANDOM FOREST:

Accuracy: Random Forest is known for its high accuracy because it combines multiple decision trees and reduces overfitting.

Sensitivity: Sensitivity is often good due to the ensemble nature of Random Forest, which helps in capturing different aspects of the data. Specificity: Specificity is also usually high because of the ensemble approach, which helps reduce false positives.

F1 Score: The F1 score is often competitive due to the balanced nature of the Random Forest

DESIGN:


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Process
Retrieve cust data
Set Role
Random Forest
Performance
Replace Missing Valu...
Split Data
Apply Model
Local Rep
my first pr
Conne
data
cust
proces
perators
performance
Validation
Perform
Pred
We found "M Management' the Marketpla Show me!
Show advanced parameters
Change compatibility (10.2.000)
Activate Wisdom of Crowds
Help
Random Forest
Concurrency
Tags: Supervised, Classification, Regression, Model, Ensembles, Decision, Trees, Extremely Randomized Trees, Extra-Trees, Breiman, Bagging
Synopsis
This Operator generates a random forest

Parameters – Random Forest

number of trees: 10

criterion: gain_ratio

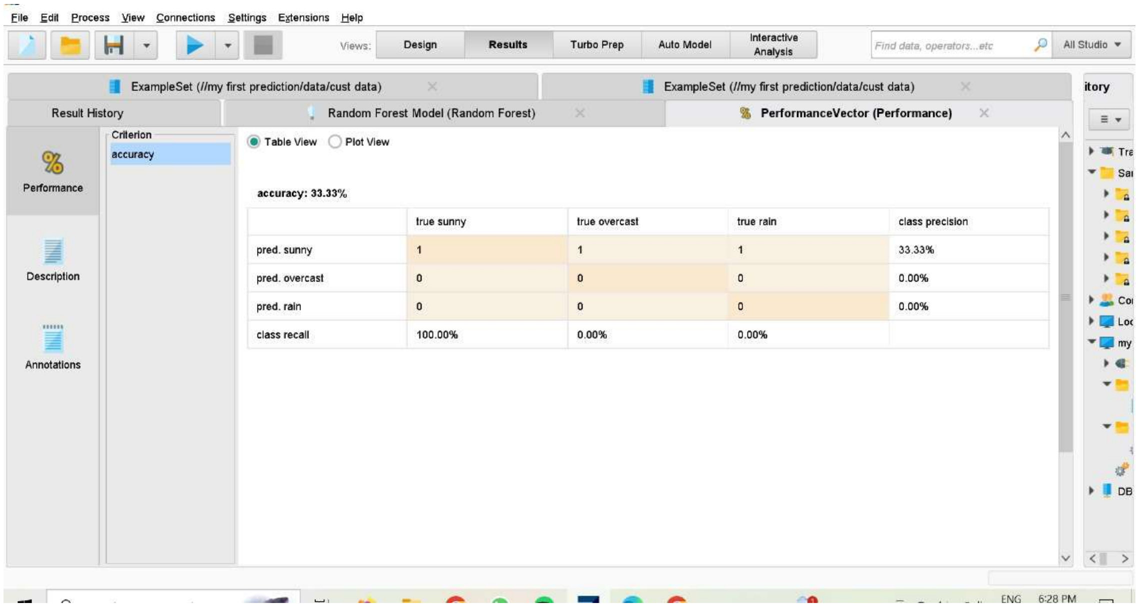
maximal depth: 10

□ apply pruning

□ apply prepruning

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

ACCURACY:


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
ExampleSet (//my first prediction/data/cust data)
ExampleSet (//my first prediction/data/cust data)
Result History
Random Forest Model (Random Forest)
PerformanceVector (Performance)
Performance
Description
Annotations
Criterion
accuracy
● Table View ○ Plot View

accuracy: 33.33%

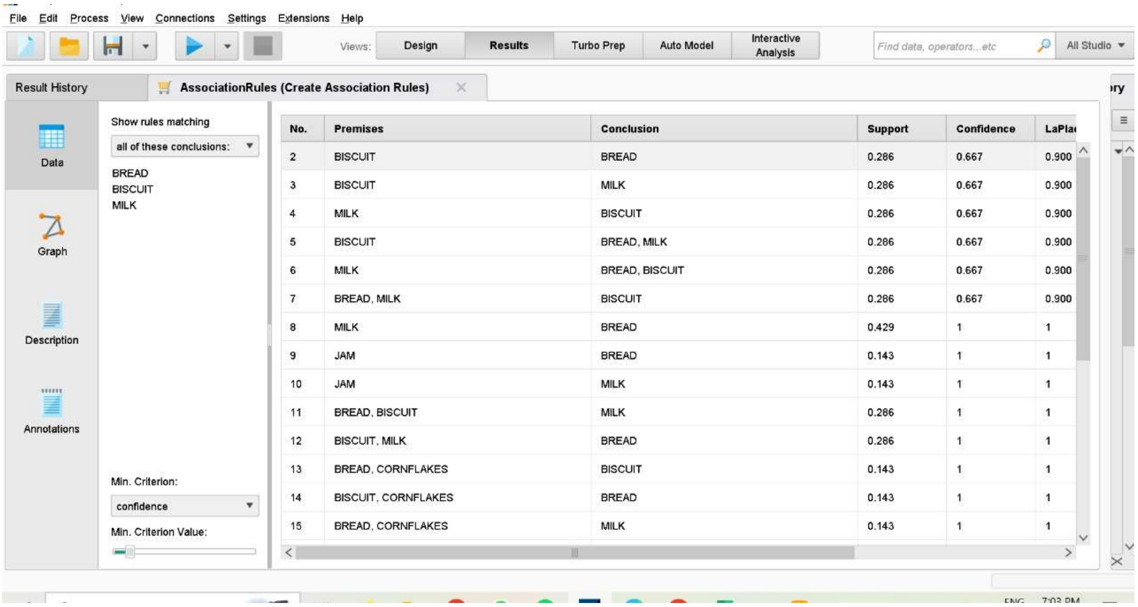
	true sunny	true overcast	true rain	class precision
pred. sunny	1	1	1	33.33%
pred. overcast	0	0	0	0.00%
pred. rain	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

2.

APRIORI ALGORITHM:

DATASET:

I have chose a dataset which is used for Market Basket Analysis which comprise about the Products (Premises and Conclusion) with their Support and Confidence.

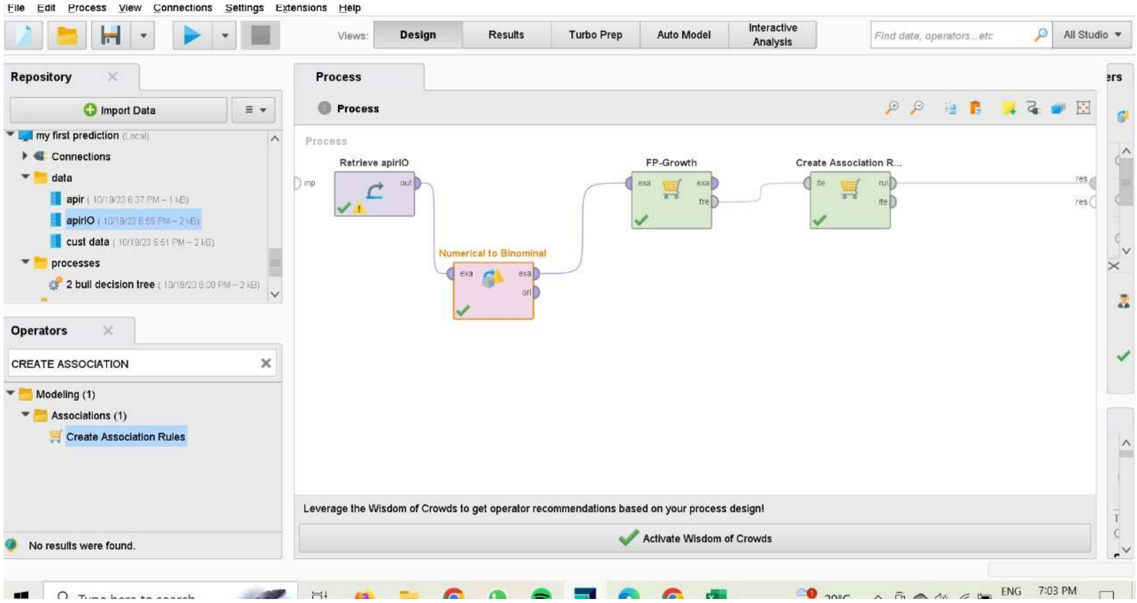


AssociationRules (Create Association Rules)

Show rules matching
all of these conclusions:
BREAD
BISCUIT
MILK

Min. Criterion:
confidence
Min. Criterion Value:

No.	Premises	Conclusion	Support	Confidence	LaPla
2	BISCUIT	BREAD	0.286	0.667	0.900
3	BISCUIT	MILK	0.286	0.667	0.900
4	MILK	BISCUIT	0.286	0.667	0.900
5	BISCUIT	BREAD, MILK	0.286	0.667	0.900
6	MILK	BREAD, BISCUIT	0.286	0.667	0.900
7	BREAD, MILK	BISCUIT	0.286	0.667	0.900
8	MILK	BREAD	0.429	1	1
9	JAM	BREAD	0.143	1	1
10	JAM	MILK	0.143	1	1
11	BREAD, BISCUIT	MILK	0.286	1	1
12	BISCUIT, MILK	BREAD	0.286	1	1
13	BREAD, CORNFLAKES	BISCUIT	0.143	1	1
14	BISCUIT, CORNFLAKES	BREAD	0.143	1	1
15	BREAD, CORNFLAKES	MILK	0.143	1	1



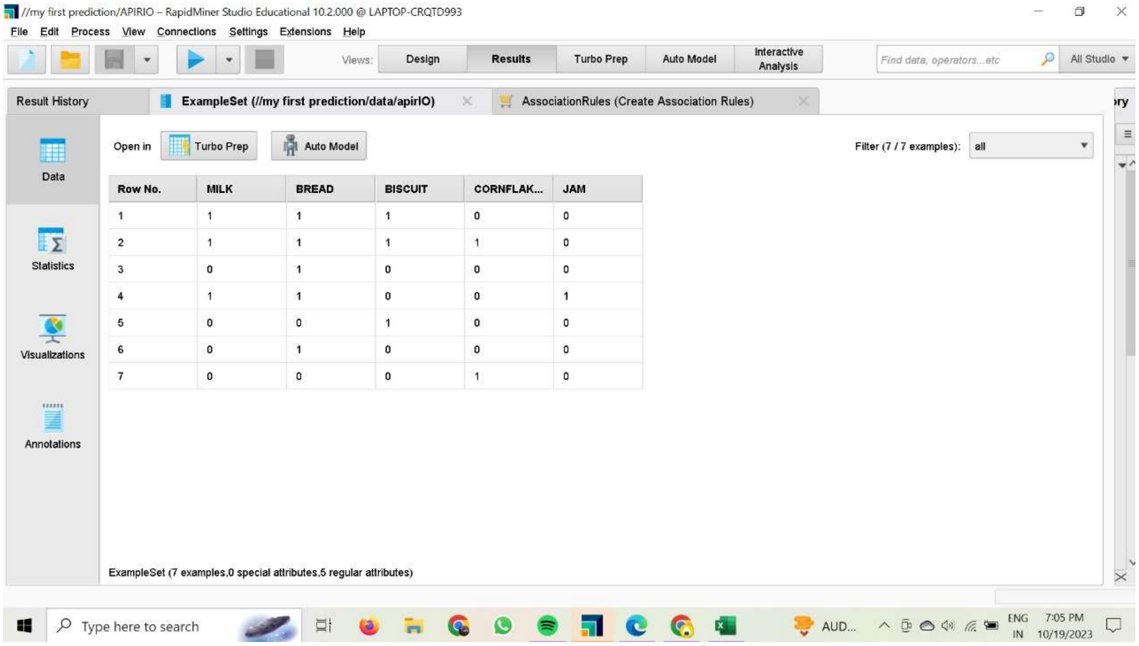
my first prediction (Local)
Connections
data
apir (10/19/23 6:37 PM – 1 kB)
apirIO (10/19/23 6:55 PM – 2 kB)
cust data (10/19/23 5:51 PM – 2 kB)
processes
2 buil decision tree (10/19/23 6:00 PM – 2 kB)

CREATE ASSOCIATION
Modeling (1)
Associations (1)
Create Association Rules

No results were found.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

ASSOCIATION RULES:


//my first prediction/APIRIO – RapidMiner Studio Educational 10.2.000 @ LAPTOP-CRQTD993
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
ExampleSet (//my first prediction/data/apirIO)
AssociationRules (Create Association Rules)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (7 / 7 examples): all
Type here to search
AUD...
ENG IN
7:05 PM 10/19/2023

Row No.	MILK	BREAD	BISCUIT	CORNFLAK...	JAM
1	1	1	1	0	0
2	1	1	1	1	0
3	0	1	0	0	0
4	1	1	0	0	1
5	0	0	1	0	0
6	0	1	0	0	0
7	0	0	0	1	0

ExampleSet (7 examples,0 special attributes,5 regular attributes)

CONCLUSION:-

RapidMiner is a versatile platform for implementing and evaluating various machine learning and data mining algorithms, including Decision Trees, Naive Bayes, Random Forest, and the Apriori algorithm. The choice of which algorithm to use depends on the nature of your data and the specific problem you're trying to solve. RapidMiner's user-friendly interface and comprehensive set of operators make it an excellent choice for data scientists and analysts to experiment with these algorithms and analyze their results effectively. Make sure to experiment with different algorithms and configurations to find the best solution for your particular use case.