Data Mining Project

Nouf Mansour

Amal Almutairi

Noor Balfas

Muna Saleh Dini

THE DATASET LINK

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Main Task:

Predicting Cardiovascular Disease

How we accomplish this?

Create a predictive model that uses demographic and health data to assess if a person has cardiovascular disease.

Does the individual have cardiovascular disease? (1 = Yes, 0 = No).

Before starting implementing the tasks..

We adjusted the dataset and made some errors like: Duplicating id , Missing values in age and weight , Value range error for Height and selected a random sample to focus our analysis

Task 1:

1. Understand the data: Import the data selected ".........csv" file into RapidMiner. While importing the data, you should properly choose the types of attributes. Identify the type of each attribute (binary, nominal, ordinal, numeric) and report it. Get statistics about the data to identify the values of the summarizing properties for each attribute, including frequency, location, and spread [e.g., value ranges of the attributes, frequency of values, medians, means, variances, or standard deviation].

1.1 Identify the type of each attribute:

Attribute Name	Type	Description
Age	Numeric	Described as numerical value , presented by days
Weight	Numeric	Described as numerical value ,Weight by kilogram
ID	Nominal	Identifier , unique for everyone
Gender	Numeric	Gender (1=female, 2=male)
Height	Numeric	Described as numerical value ,Height by centimeters
Ap_hi	Numeric	Systolic blood pressure
Ap_lo	Numeric	Diastllic blood pressure
Cholesterol	Ordinal	Cholesterol level (1=normal , 2= above normal , 3= High)
Gluc	Ordinal	Glucose level (1=normal , 2= above normal , 3= High)
Smoke	Binary	Smoking status (0= doesn't smoke , 1= smoker)
Aloc	Binary	Alcohol user (0= no , 1= yes)
Active	Binary	Physical status (unactive = 0 , active =1)
Cardio	Binary	Heart disease (0= safe , 1= not safe)

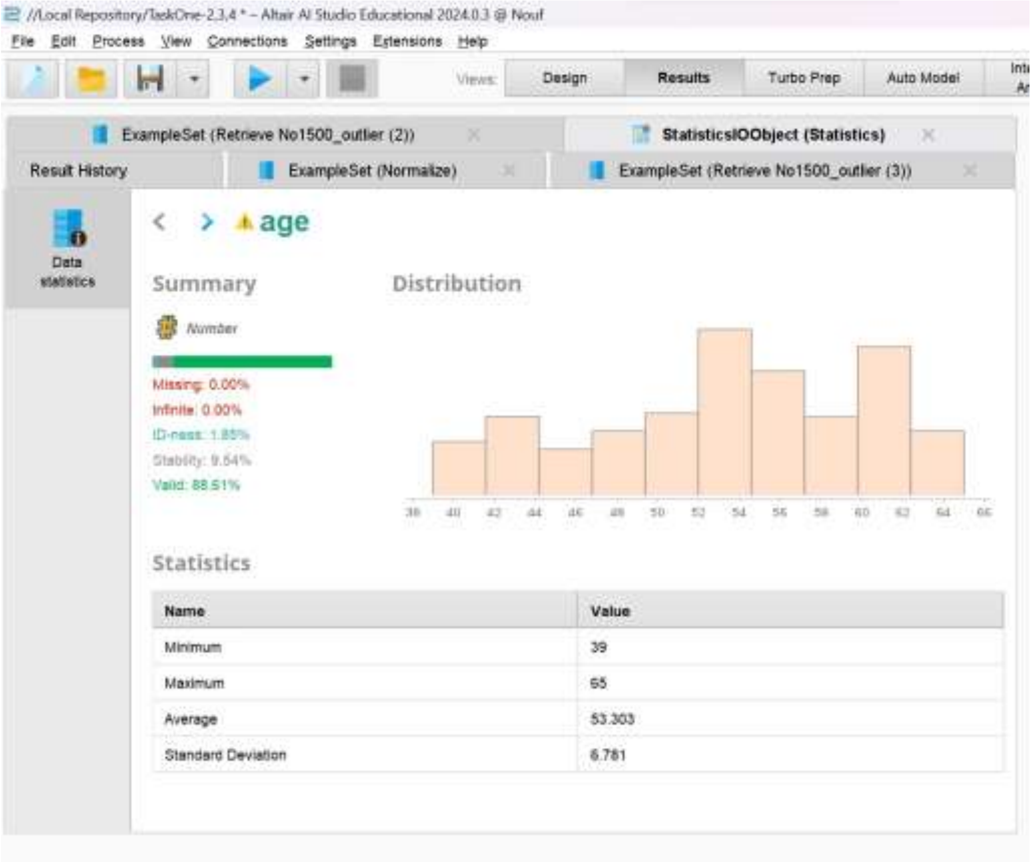
1.2 Properties of attributes

	Column	Description
1	**Age**	Represents the age of the individual in days. It is a numerical attribute used to calculate age in years or categorize individuals into age groups
2	**Weight**	Represents the weight of the individual in kilograms. It is a continuous variable that can be used for calculating Body Mass Index (BMI).
3	**ID**	A unique identifier for each individual in the dataset. It is non-numerical and used only for distinguishing individuals.
4	**Gender**	Represents the gender of the individual
5	**Height**	Represents the height of the individual in centimeters. This numerical attribute can also be used for BMI calculation.
6	**Ap_hi**	Represents the systolic blood pressure (the higher value in a blood pressure reading). It is a continuous variable.
7	**Ap_lo**	Represents the diastolic blood pressure (the lower value in a blood pressure reading). It is a continuous variable.
8	**Cholesterol**	Indicates the cholesterol level: ○ **1 = Normal** ○ **2 = Above Normal** ○ **3 = High**
9	**Gluc**	Indicates the glucose level: ○ **1 = Normal** ○ **2 = Above Normal** ○ **3 = High**
10	**Smoke**	Indicates smoking status: ○ **0 = Does not smoke** ○ **1 = Smoker**
11	**Alco**	Indicates alcohol usage: ○ **0 = Does not consume alcohol** ○ **1 = Consumes alcohol**

12	**Active**	Represents physical activity level: ○ **0 = Not active** ○ **1 = Active**
13	**Cardio**	Indicates the presence of cardiovascular disease: ○ **0 = No (safe)** ○ **1 = Yes (not safe)**

Data Exploratory Analysis and Visualization

1.3 Statistics about some of the data



//Local Repository/TaskOne-2.3.4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data statistics
Distribution
38 40 42 44 46 48 50 52 54 56 58 60 62 64 66

age

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 1.85%
Stability: 9.54%
Valid: 88.61%

Statistics

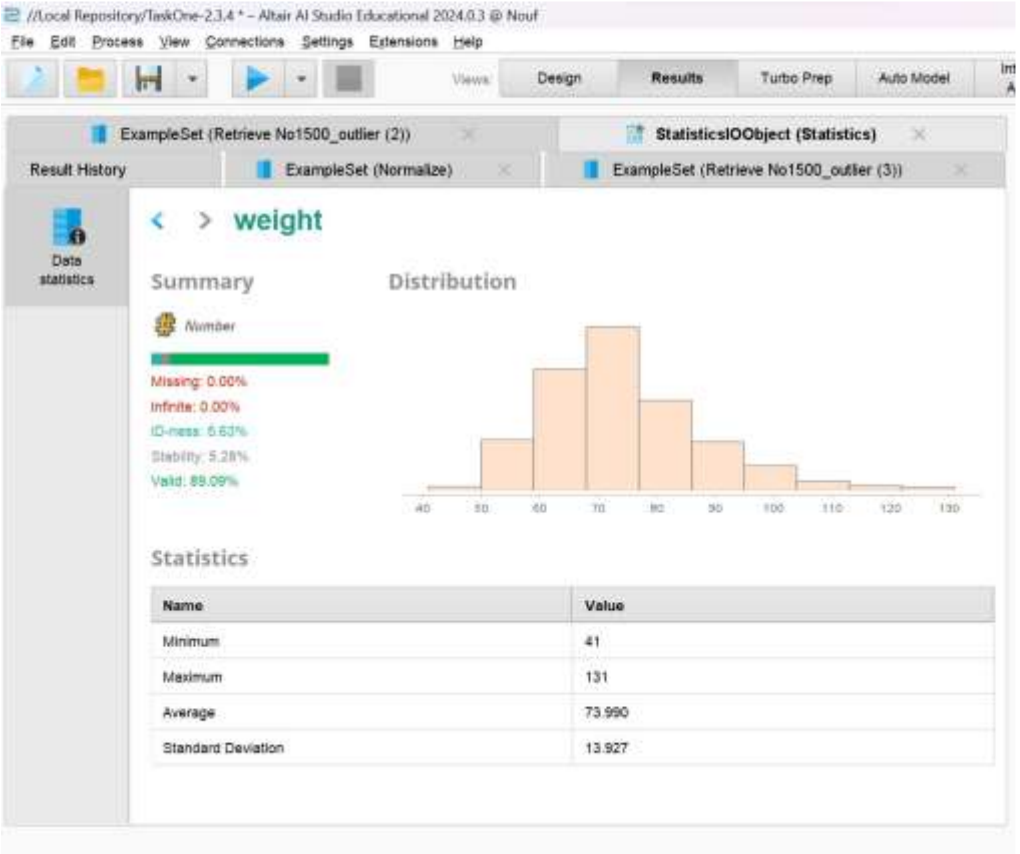
Name	Value
Minimum	39
Maximum	65
Average	53.303
Standard Deviation	6.781

Brief explaintion About Age :

age (Age Distribution) Description:

The plot shows the age distribution of the dataset's participants, with an average age of 53 and a range of 39 to 65.3.

Significance: An important risk factor for cardiovascular disease is age. Since older people are typically at higher risk, this feature is essential.



//Local Repository/TaskOne-2.3.4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data statistics
Distribution
40 50 60 70 80 90 100 110 120 130

weight

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 6.63%
Stability: 5.28%
Valid: 89.09%

Statistics

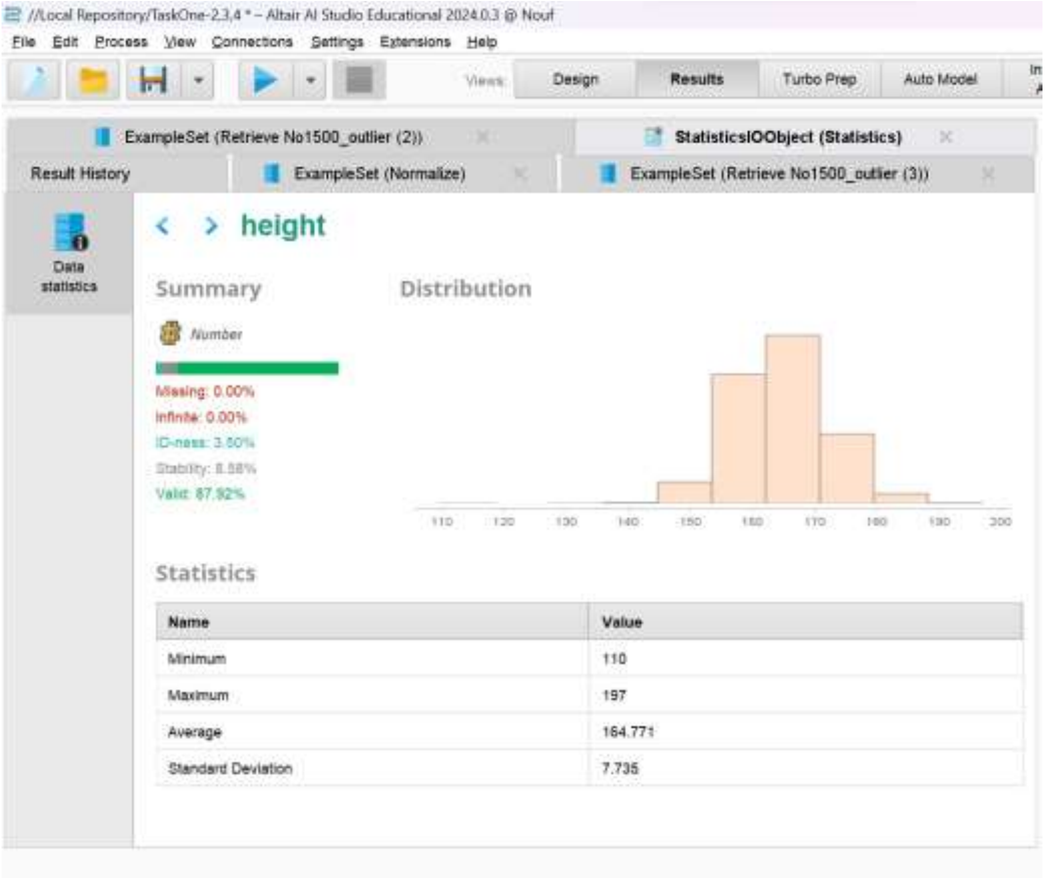
Name	Value
Minimum	41
Maximum	131
Average	73.990
Standard Deviation	13.927

Brief explaintion About Weight:

distribution of (weight)

The plot shows the weight distribution, with an average of 73.99 kg and a range of 41 to 131 kg.

Relevance: Body mass index (BMI), a recognized risk factor for cardiovascular health problems, is directly correlated with weight.


//Local Repository/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File
Edit
Process
View
Connections
Settings
Extensions
Help
Views:
Design
Results
Turbo Prep
Auto Model
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data statistics
Distribution
110
120
130
140
150
160
170
180
190
200

height

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 3.50%
Stability: 8.58%
Valid: 87.92%

Statistics

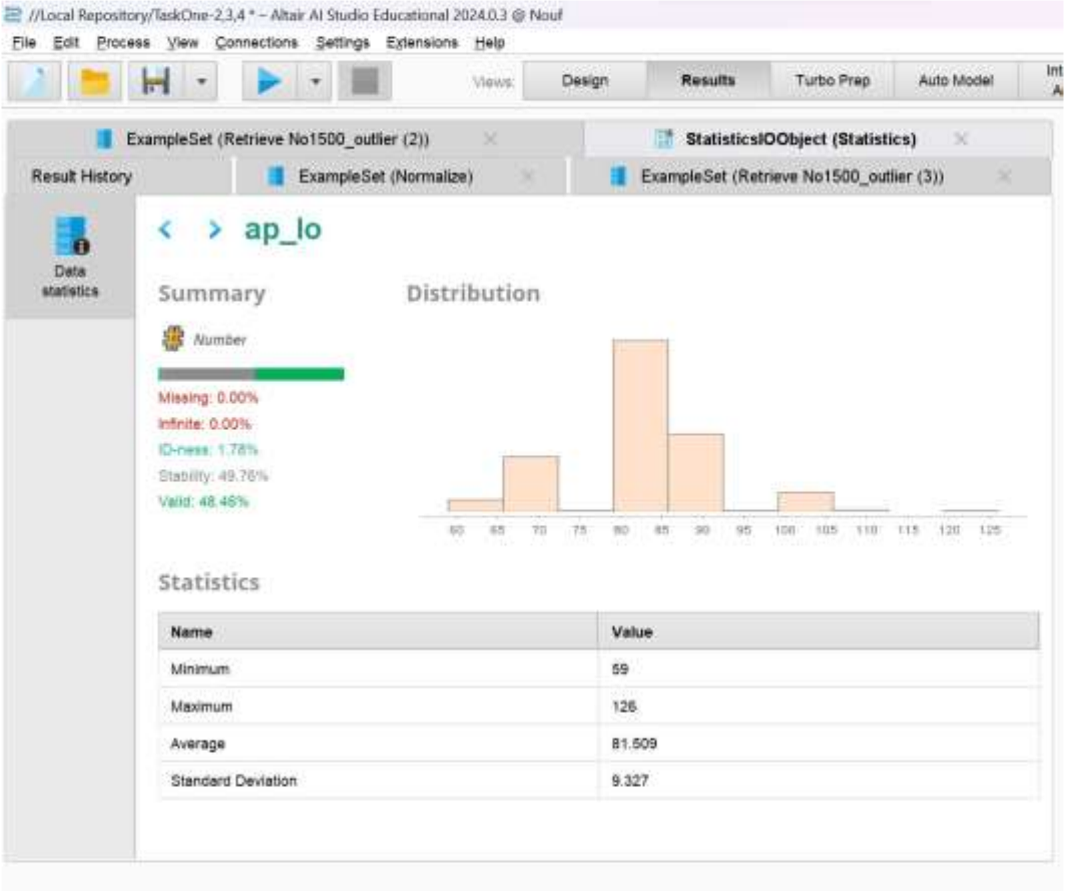
Name	Value
Minimum	110
Maximum	197
Average	164.771
Standard Deviation	7.735

Brief explaintion About Height: height

(Height Distribution)

This plot displays the range of individual heights, with an average of 164.77 cm and a range of 110 to 197 cm.

Relevance: Although height by itself may not be a reliable indicator of cardiovascular disease, it may serve as a stand-in when paired with weight (e.g., BMI).


//Local Repository/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File
Edit
Process
View
Connections
Settings
Extensions
Help
Views:
Design
Results
Turbo Prep
Auto Model
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data statistics
Distribution
60
65
70
75
80
85
90
95
100
105
110
115
120
125

ap_lo

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 1.78%
Stability: 49.76%
Valid: 48.46%

Statistics

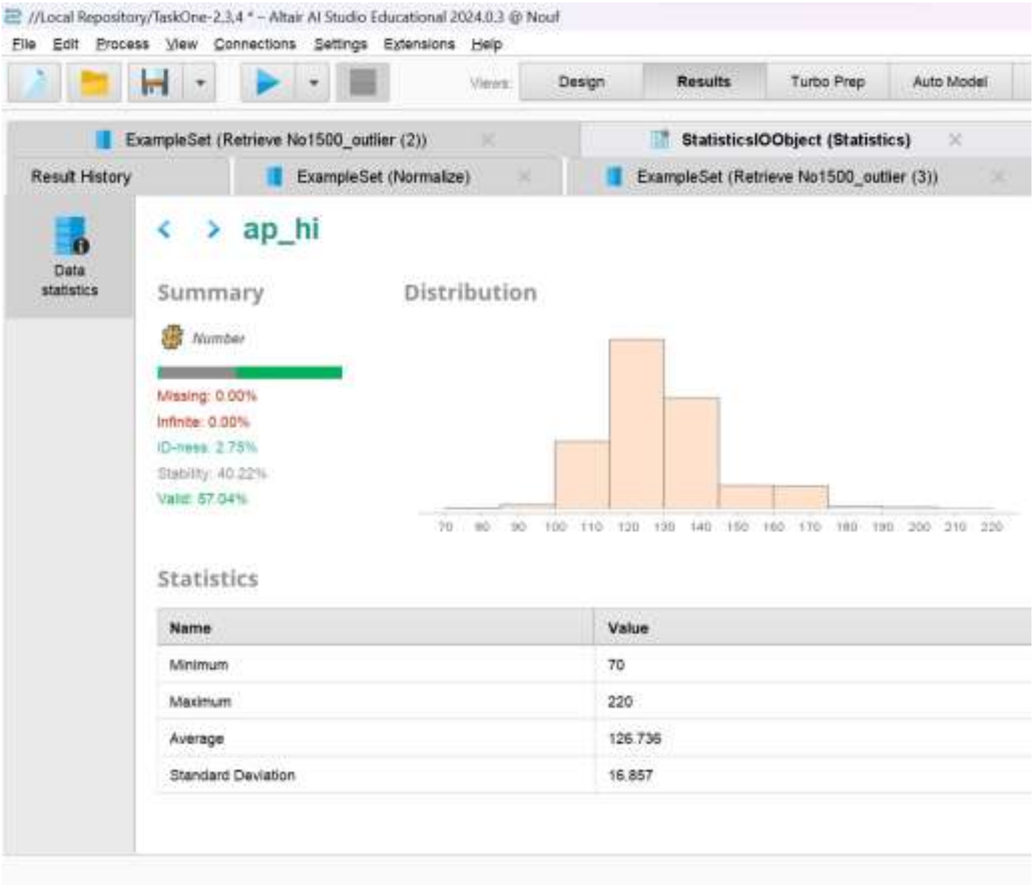
Name	Value
Minimum	59
Maximum	126
Average	81.509
Standard Deviation	9.327

Brief explaintion About ap_lo:

Diastolic Blood Pressure Distribution (ap_lo):

This plot displays the dataset's diastolic blood pressure distribution (ap_lo). It has an average of 81 and ranges from 59 to 126.51.

Relavance : One of the main markers of cardiovascular problems is high or low blood pressure.


//Local Repository/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data statistics
Distribution
70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220

ap_hi

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.75%
Stability: 40.22%
Valid: 57.04%

Statistics

Name	Value
Minimum	70
Maximum	220
Average	126.736
Standard Deviation	16.857

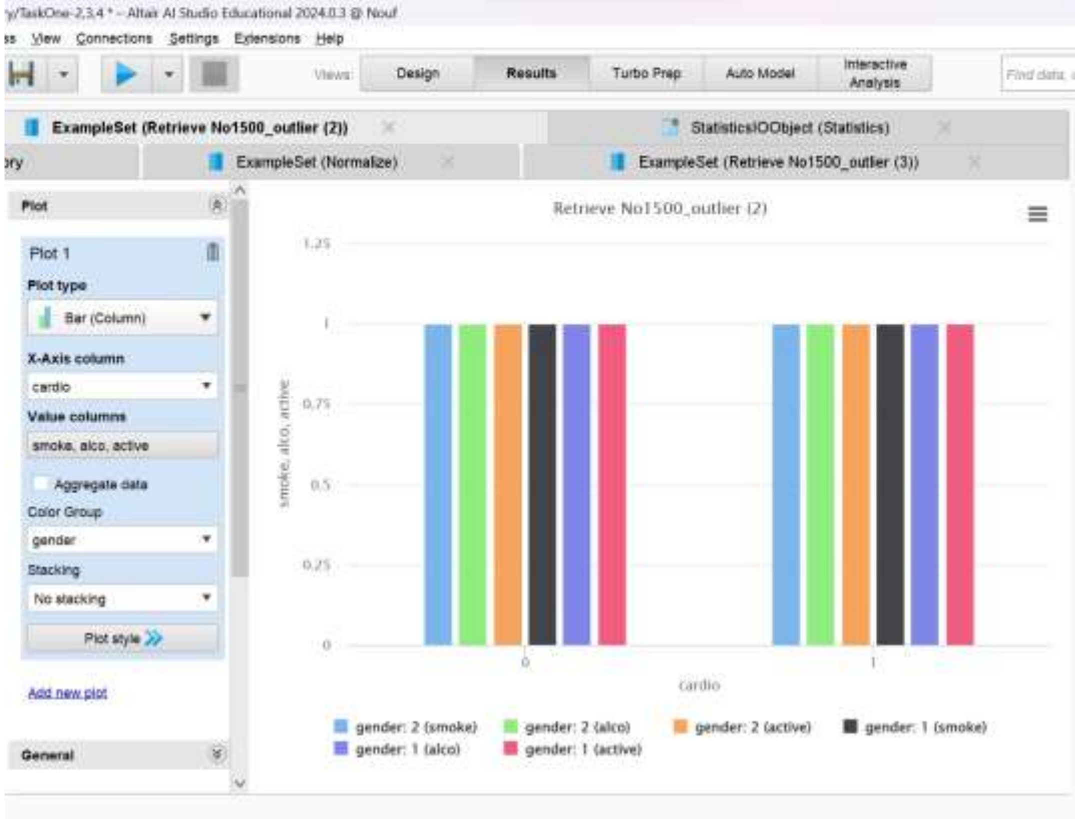
Brief explaintion About ap_high:

Systolic Blood Pressure Distribution (ap_hi)

This plot shows the average systolic blood pressure (ap_hi), which ranges from 70 to 220.
Relevance: Systolic blood pressure is essential for detecting hypertension, just like diastolic blood pressure.

2. Analyze the data by using different visualization techniques provided by RapidMiner.

2.1 Explanation of plot one :


y/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
ss View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, o
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
ory
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Plot
Add new plot
General
Retrieve No1500_outlier (2)
1.25
1
0.75
0.5
0.25
0
smoke, alco, active
0
1
cardio
gender: 2 (smoke)
gender: 2 (alco)
gender: 2 (active)
gender: 1 (smoke)
gender: 1 (alco)
gender: 1 (active)

Plot 1

Plot type: Bar (Column)

X-Axis column: cardio

Value columns: smoke, alco, active

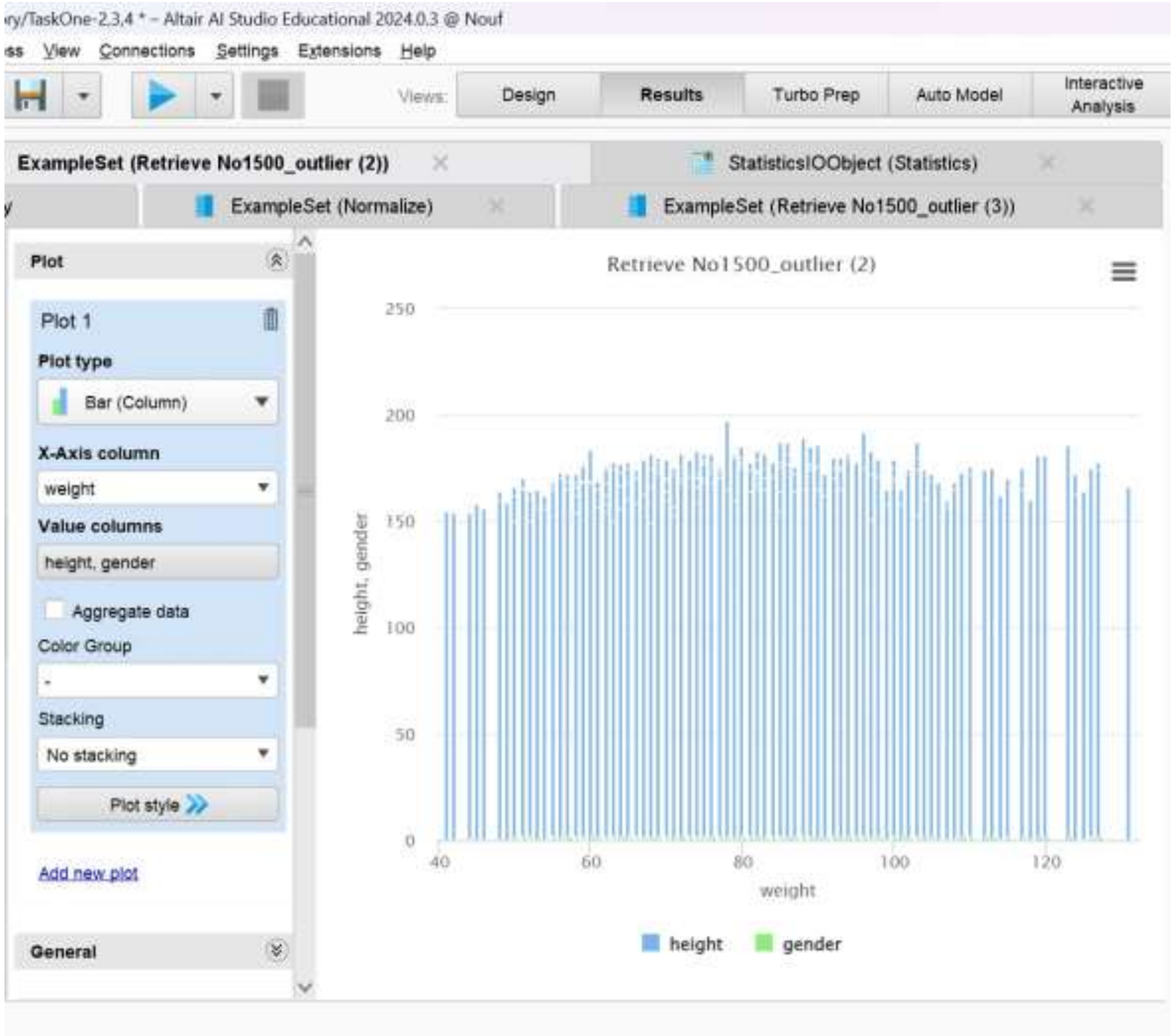
□ Aggregate data

Color Group: gender

Stacking: No stacking

Plot style

(Lifestyle Choices and Gender in Relation to Cardiovascular Disease)
The association between gender, physical activity, alcohol use, smoking, and cardiovascular disease is depicted in this plot.
Relevance: Modifiable risk factors for cardiovascular disease include lifestyle choices like drinking alcohol and smoking. Risk prediction may also be influenced by gender differences (1 = Female, 2= Male), you can tell that both male and female almost have equal Lifestyle, so the chance for having Cardiovascular Disease is possible for both of them if they don’t stop the bad habits like alco drinking , smoking.


ory/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
ss View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Plot
Plot 1
Add new plot
General
Retrieve No1500_outlier (2)
height, gender
0
50
100
150
200
250
40
60
80
100
120
weight
height
gender

Plot type: Bar (Column)

X-Axis column: weight

Value columns: height, gender

□ Aggregate data

Color Group: -

Stacking: No stacking

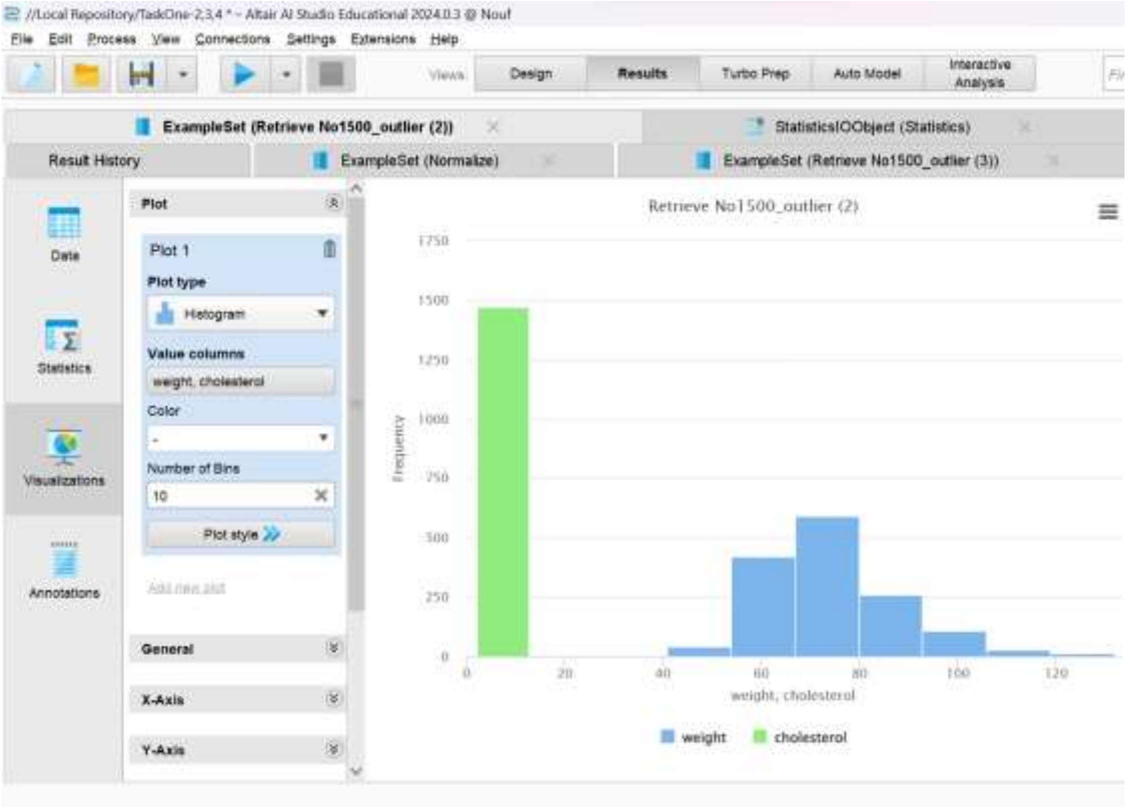
Plot style

2.2 Explanation of plot two:

(Weight and Height Overview)

This plot contrasts the distributions of height and weight by gender. It draws attention to individual differences in these parameters.

Relevance: These characteristics affect BMI, which is a measure of obesity, a major risk factor for cardiovascular disease.


//Local Repository/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
ExampleSet (Retrieve No1500_outlier (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve No1500_outlier (3))
Data
Statistics
Visualizations
Annotations
Plot
Plot 1
Add new plot
General
X-Axis
Y-Axis
Retrieve No1500_outlier (2)
Frequency
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100
120
weight, cholesterol
weight
cholesterol

Plot type: Histogram

Value columns: weight, cholesterol

Color: -

Number of Bins: 10

Plot style

2.3 Explanation of plot three:

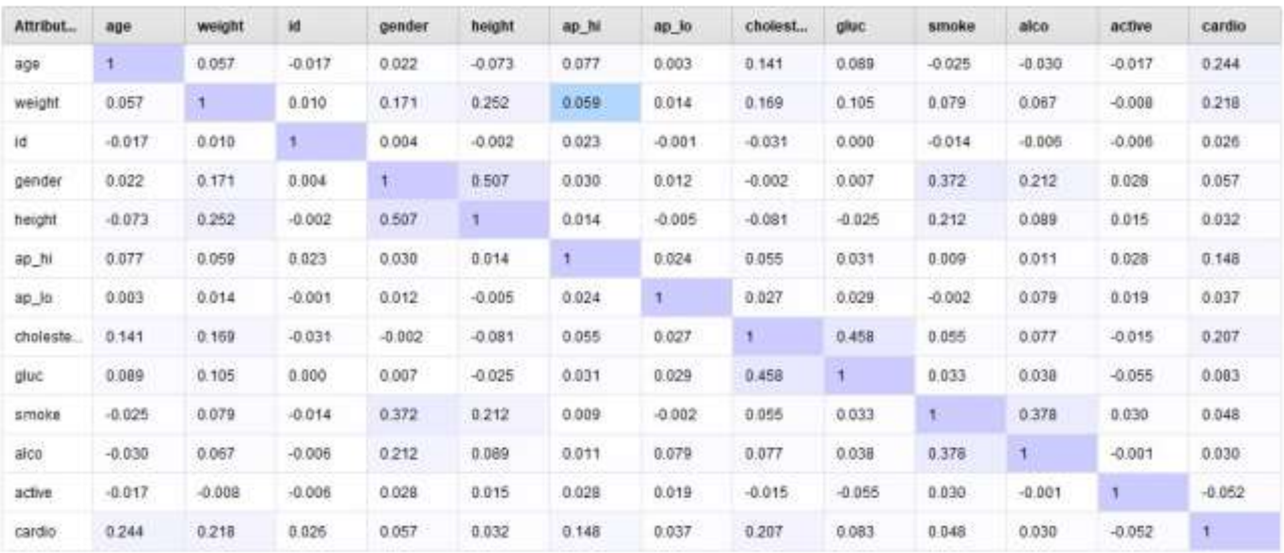
(Cholesterol and Weight)

The frequency of weight and cholesterol levels in the dataset is shown visually in this histogram.

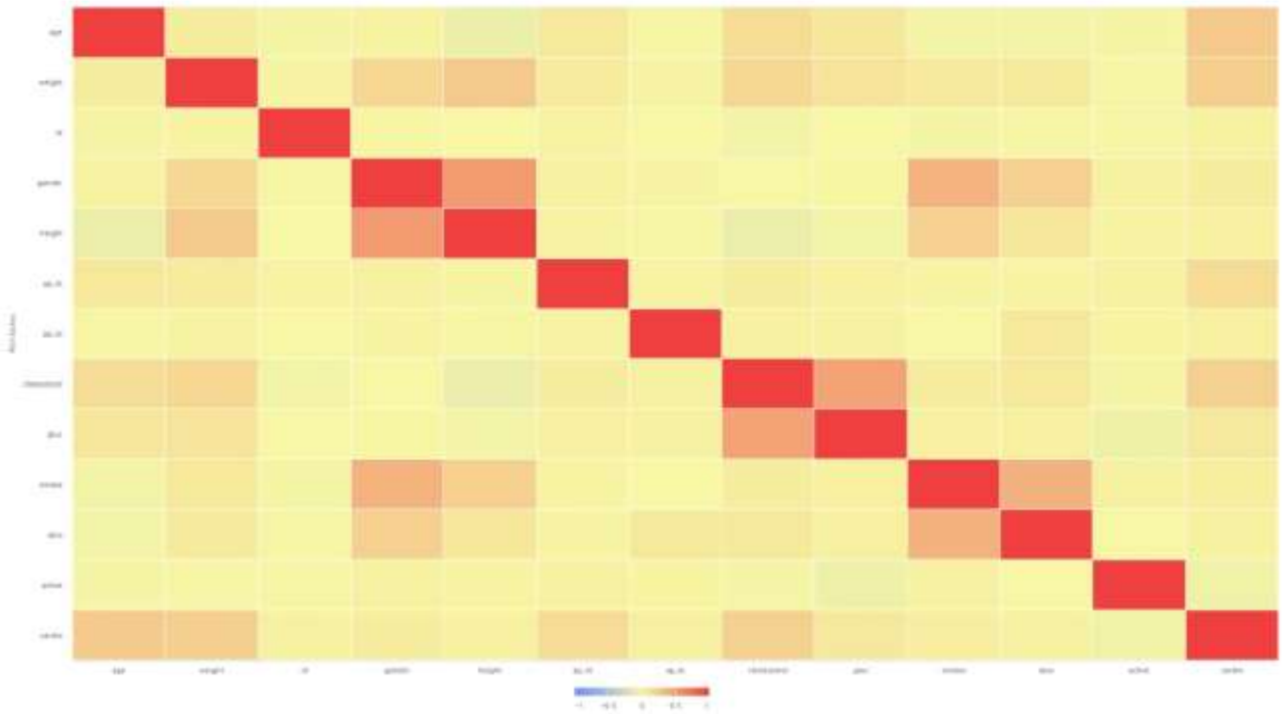
Relevance: One of the main cardiovascular risk factors, hyperlipidemia, can be identified by measuring cholesterol levels. The combination of weight and cholesterol yields important information.

3. Find and analyze the correlated attributes (you can use the “Correlation Matrix” operator or Scatter Plot chart).

3.1 Correlation matrix



Attribut...	age	weight	id	gender	height	ap_hi	ap_lo	cholest...	gluc	smoke	alco	active	cardio
age	1	0.057	-0.017	0.022	-0.073	0.077	0.003	0.141	0.089	-0.025	-0.030	-0.017	0.244
weight	0.057	1	0.010	0.171	0.252	0.059	0.014	0.169	0.105	0.079	0.067	-0.008	0.218
id	-0.017	0.010	1	0.004	-0.002	0.023	-0.001	-0.031	0.000	-0.014	-0.006	-0.006	0.026
gender	0.022	0.171	0.004	1	0.507	0.030	0.012	-0.002	0.007	0.372	0.212	0.028	0.057
height	-0.073	0.252	-0.002	0.507	1	0.014	-0.005	-0.081	-0.025	0.212	0.089	0.015	0.032
ap_hi	0.077	0.059	0.023	0.030	0.014	1	0.024	0.055	0.031	0.009	0.011	0.028	0.148
ap_lo	0.003	0.014	-0.001	0.012	-0.005	0.024	1	0.027	0.029	-0.002	0.079	0.019	0.037
choleste...	0.141	0.169	-0.031	-0.002	-0.081	0.055	0.027	1	0.458	0.055	0.077	-0.015	0.207
gluc	0.089	0.105	0.000	0.007	-0.025	0.031	0.029	0.458	1	0.033	0.038	-0.055	0.083
smoke	-0.025	0.079	-0.014	0.372	0.212	0.009	-0.002	0.055	0.033	1	0.378	0.030	0.048
alco	-0.030	0.067	-0.006	0.212	0.089	0.011	0.079	0.077	0.038	0.378	1	-0.001	0.030
active	-0.017	-0.008	-0.006	0.028	0.015	0.028	0.019	-0.015	-0.055	0.030	-0.001	1	-0.052
cardio	0.244	0.218	0.026	0.057	0.032	0.148	0.037	0.207	0.083	0.048	0.030	-0.052	1



Explanation:

Positive Correlations:

height and weight (0.252): Taller individuals tend to weigh more, which is a logical and expected trend.

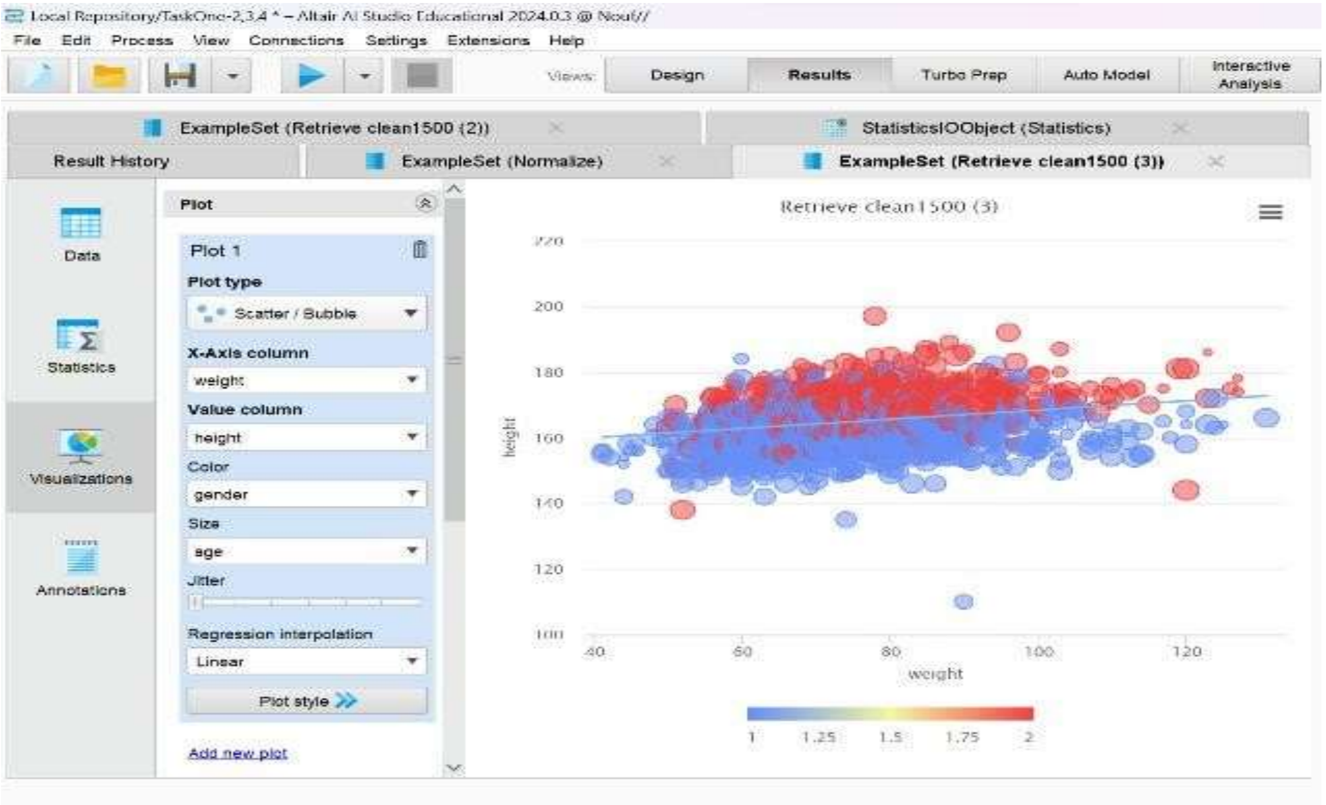
cholesterol and gluc (0.458): High cholesterol levels are moderately correlated with elevated glucose levels, indicating potential links to metabolic or cardiovascular health issues. **age and cardio (0.244):** Older age is moderately associated with cardiovascular conditions, reinforcing age as a risk factor.

weight and cardio (0.218): Higher weight is linked to cardiovascular issues, likely due to obesity-related risks.

Negative Correlations:

Attributes like smoke, alco, and active show very low correlations with most variables.

3.2 Scatter plot


Local Repository/TaskOne-2,3,4 * – Altair AI Studio Educational 2024.0.3 @ Nouf//
File Edit Process View Connections Settings Extensions Help
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
ExampleSet (Retrieve clean1500 (2))
StatisticsIOObject (Statistics)
Result History
ExampleSet (Normalize)
ExampleSet (Retrieve clean1500 (3))
Data
Statistics
Visualizations
Annotations
Plot
Plot 1
Plot style
Add new plot
Retrieve clean1500 (3)
height
220
200
180
160
140
120
100
40
60
80
100
120
weight
1
1.25
1.5
1.75
2

Plot type: Scatter / Bubble

X-Axis column: weight

Value column: height

Color: gender

Size: age

Jitter

Regression interpolation: Linear

Explanation:

There is a visible positive trend: as height increases, weight also increases.

This matches the correlation value of 0.252 from the matrix.

- Gender Based Differences:

Red points and blue points (representing different genders)

Red points (Female) are more concentrated in lower height and weight ranges.

Blue points (Male)are spread across a wider range of height and weight.

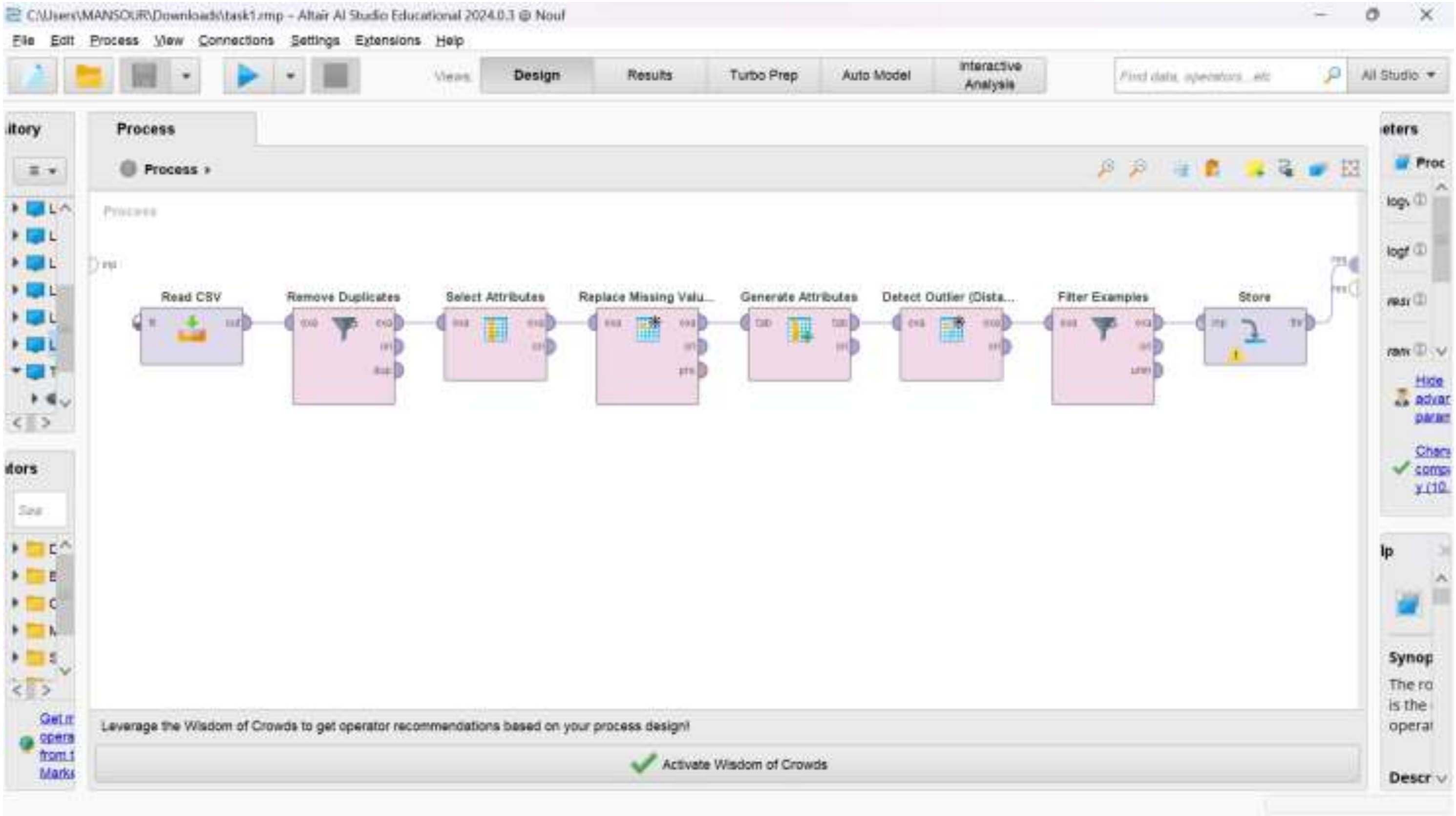
- Age Influence:

Larger-sized points (older individuals) are more distributed across the plot, indicating that age influences both weight and height variations.

4. Propose a way of cleaning data: e.g., replacing missing values, doing normalization if needed, corrections or errors, etc. 4.1 Preprocessing

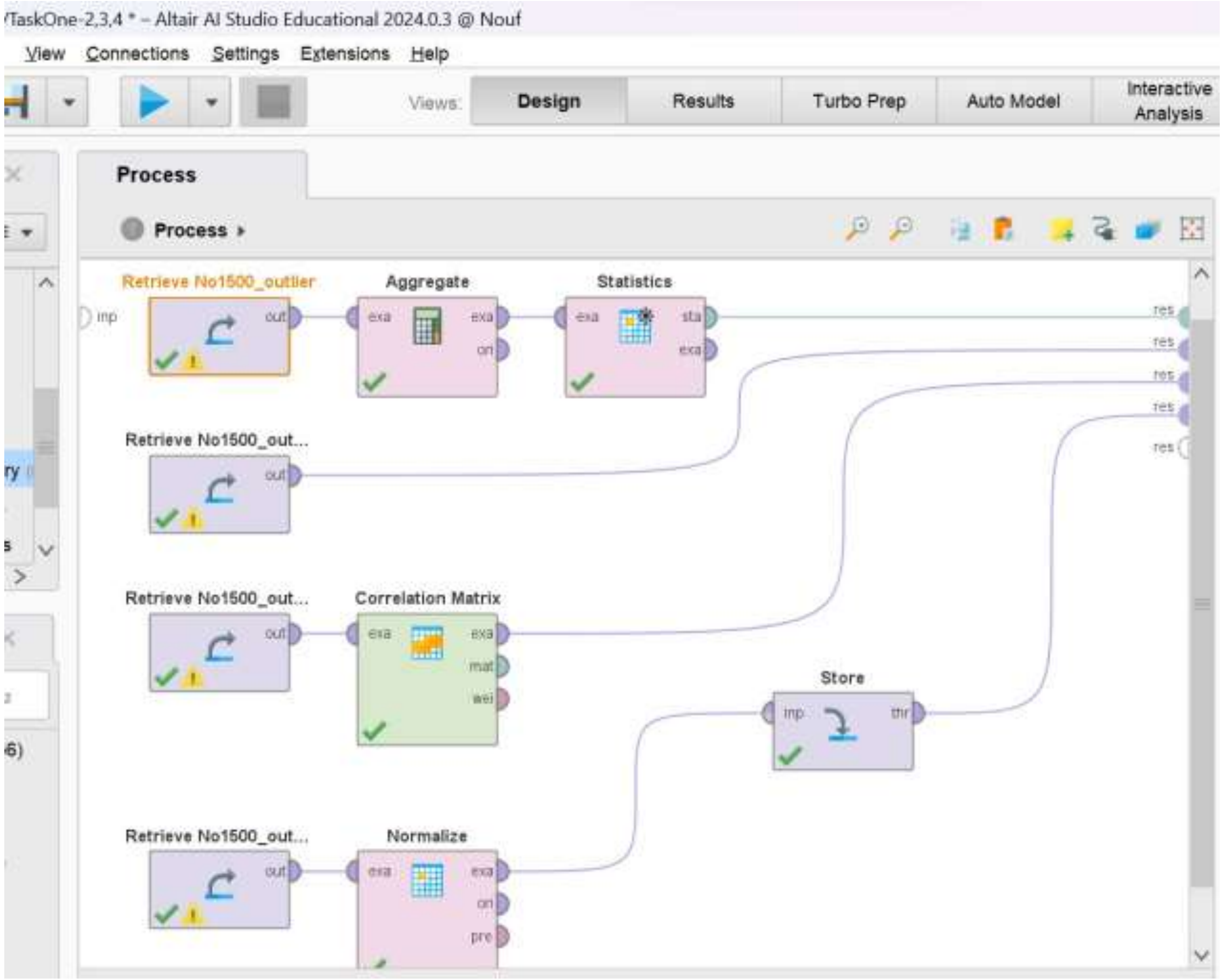
We have fixed those errors Duplicating id , Missing values in age and weight , Value range error for Height by and also did normalization but it will show on the final processes for task one down below :

Fixing the errors:


C:\Users\MANSOUR\Downloads\task1.rmp – Altair AI Studio Educational 2024.0.3 @ Nouf
File Edit Process View Connections Settings Extensions Help
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
All Studio
Process
Read CSV
Remove Duplicates
Select Attributes
Replace Missing Valu...
Generate Attributes
Detect Outlier (Dista...
Filter Examples
Store
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

Final Process for Task one:

statistics ,visualization , correlation , normalization of (Height , Weight, ap_lo, ap_hi) and it has been done by these processes



Task 2:

Evaluation Approach (Performance Measures)

Measure	Description	Definition of Value Function	Weight	Threshold
Accuracy	Percentage of test set tuples that are correctly classified	1 – Test MisClassification Rate	0.60	(>) 0.70
Simplicity	Model should be easy to understand.	Assuming that the ideal DT would have between 3(i.e. Ideal$_{\text{Bot}}$) through 7 (i.e. Ideal$_{\text{Top}}$) leaves, and the DT with 2 or less leaves (i.e. CutOff$_{\text{Bot}}$= 2) or 14 or more leaves (i.e. CutOff$_{\text{Top}}$= 14) is unacceptable;	0.05	3 ≤ Leaves < 14
Lift	For a given Target Event, it Measures the relative improvement that the model provides versus a random guess.	Based on the positive target event @ the 3rd decile	0.20	(>) 0
Stability	The model behaves similarly when applied to different datasets. Stability is binary, with 1 indicating a stable model and 0 indicating an unstable model as determined by the visual inspection of the noncumulative % Response Lift Chart for the given model	Based on visual inspection of Non-Cumulative Lift chart	0.15	1=>

1. Generate decision trees for the following parameter combinations and select the ‘best’ decision tree based on the criteria: Accuracy, Simplicity, Stability, and Lift (based on the positive target event @ the 3rd decile).

NumOfObsInModelSet =1471
TrainingPercent = 0.70
Test rows = 441
Train rows = 1030
MinObservationsPerLeaf = Ceiling(NumOfObsInModelSet * TrainingPercent * MinPercentOfTrainingObsPerLeaf)
Minimal Size for Split = 2* Leaf Size

DT No.	Splitting Criterion	MinPercentOfTrainingObsPerLeaf	Leaf Size	Minimal Size for Split	Confidence
1	Information Gain	3.50%	37	74	0.1
2	Information Gain	3.50%	37	74	0.2
3	Information Gain	6.00%	62	124	0.1
4	Information Gain	6.00%	62	124	0.2
5	Information Ratio	3.50%	37	74	0.1
6	Information Ratio	3.50%	37	74	0.2
7	Information Ratio	6.00%	62	124	0.1
8	Information Ratio	6.00%	62	124	0.2
9	Gini	3.50%	37	74	0.1
10	Gini	3.50%	37	74	0.2
11	Gini	6.00%	62	124	0.1
12	Gini	6.00%	62	124	0.2

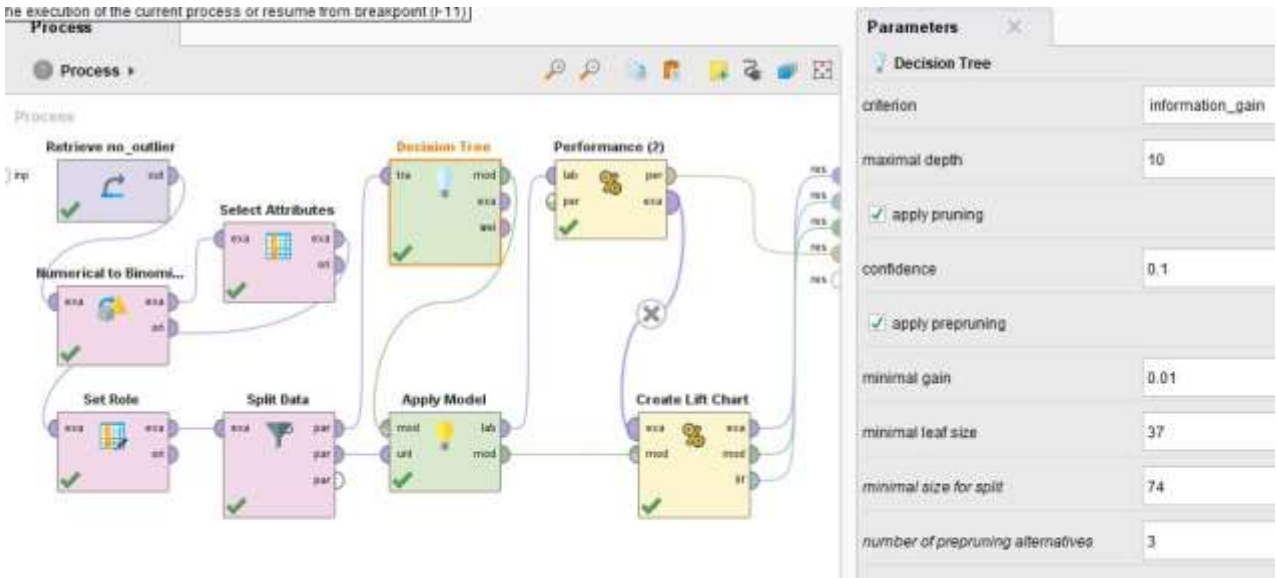
2. Specify combination function including weights that is to be used to obtain the Composite Performance Score for each DT:

$$\text{Overall\_Score} = w_{Accuracy}*\text{Score}_{Accuracy} + w_{Simplicity}*\text{Score}_{Simplicity} + w_{Lift}*\text{Score}_{Lift} + w_{Stability}*\text{Score}_{Stability}$$

Combination Function (overall score) =0.60*Accuracy+0.05*Simplicity+0.20*Lift+0.15*Stability 2.

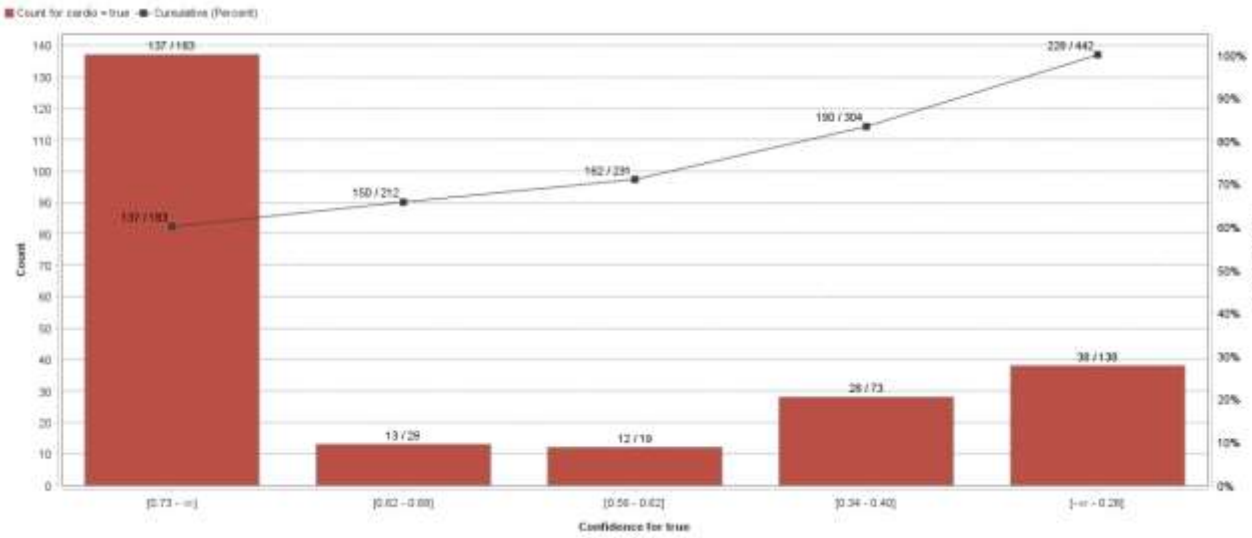
Summary of Results

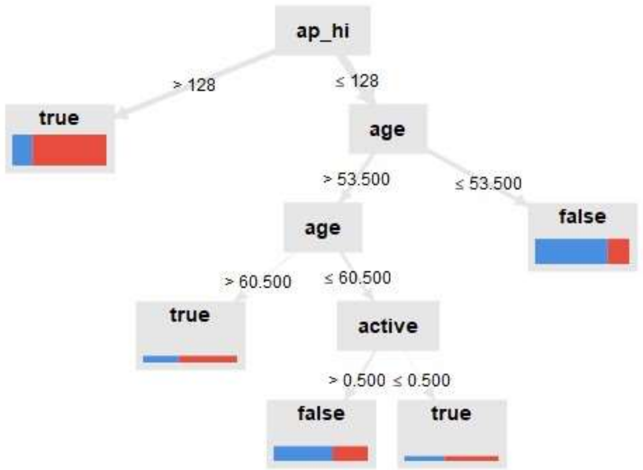
3. **Evidence of Experimentation** 1. DT 1



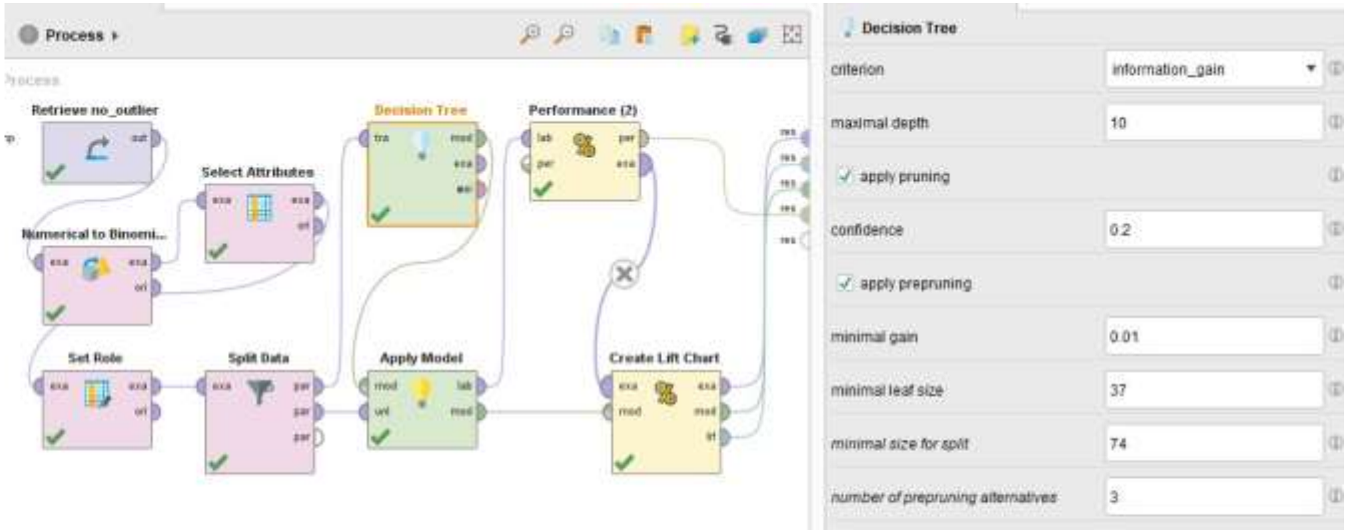
Decision Tree

criterion	information_gain
maximal depth	10
☑ apply pruning	
confidence	0.1
☑ apply prepruning	
minimal gain	0.01
minimal leaf size	37
minimal size for split	74
number of prepruning alternatives	3



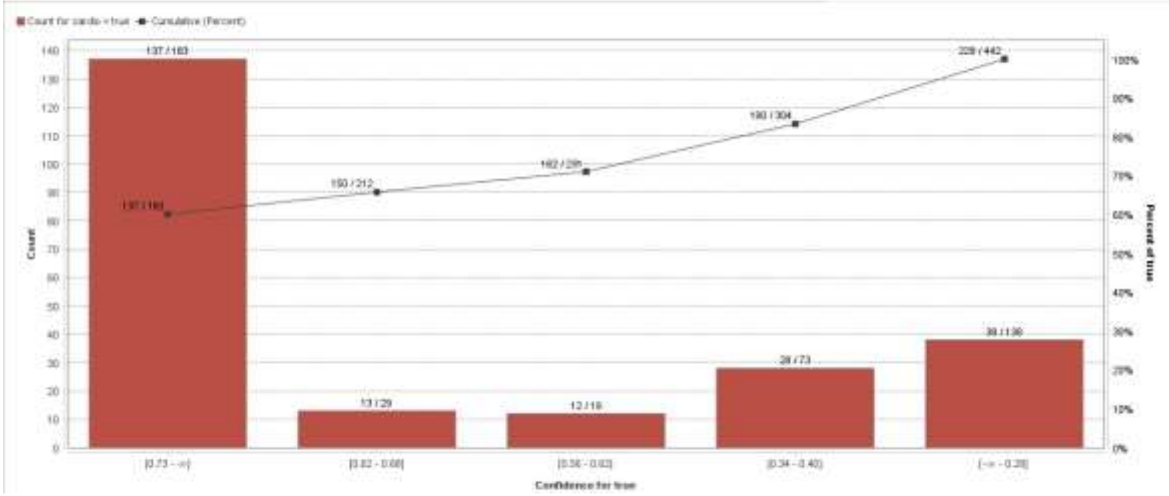


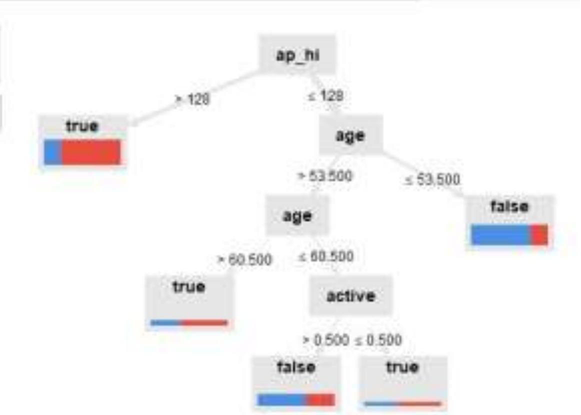
2. DT 2



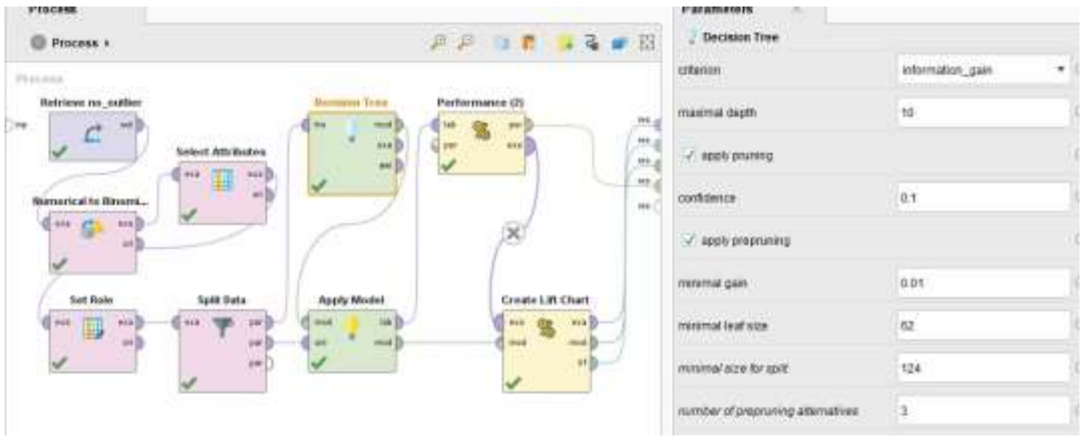
Decision Tree

criterion	information_gain
maximal depth	10
☑ apply pruning	
confidence	0.2
☑ apply prepruning	
minimal gain	0.01
minimal leaf size	37
minimal size for split	74
number of prepruning alternatives	3





3. DT 3



Decision Tree

criterion: information_gain

maximal depth: 10

☑ apply pruning

confidence: 0.1

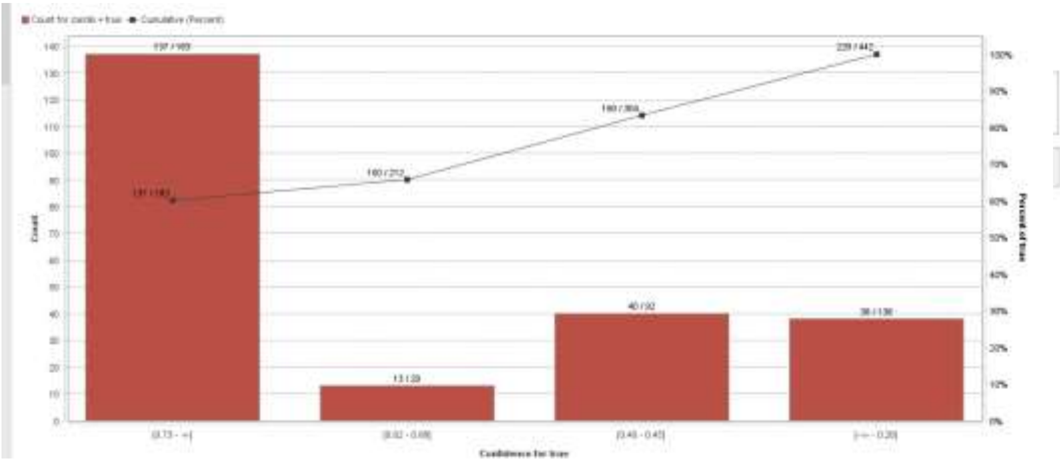
☑ apply prepruning

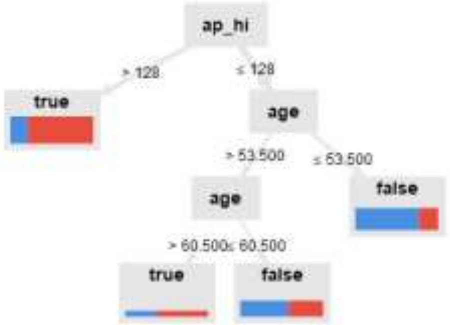
minimal gain: 0.01

minimal leaf size: 62

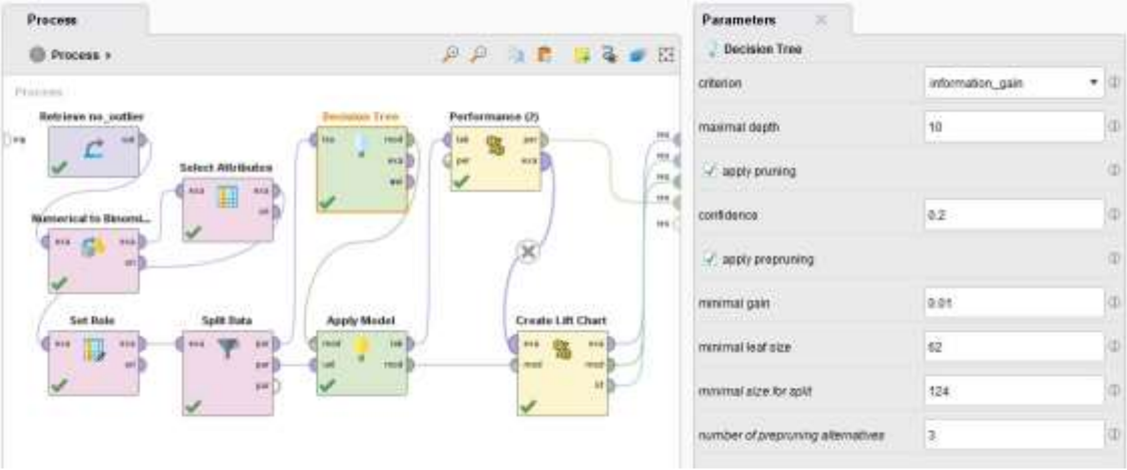
minimal size for split: 124

number of prepruning alternatives: 3





4. DT 4



Decision Tree

criterion: information_gain

maximal depth: 10

☑ apply pruning

confidence: 0.2

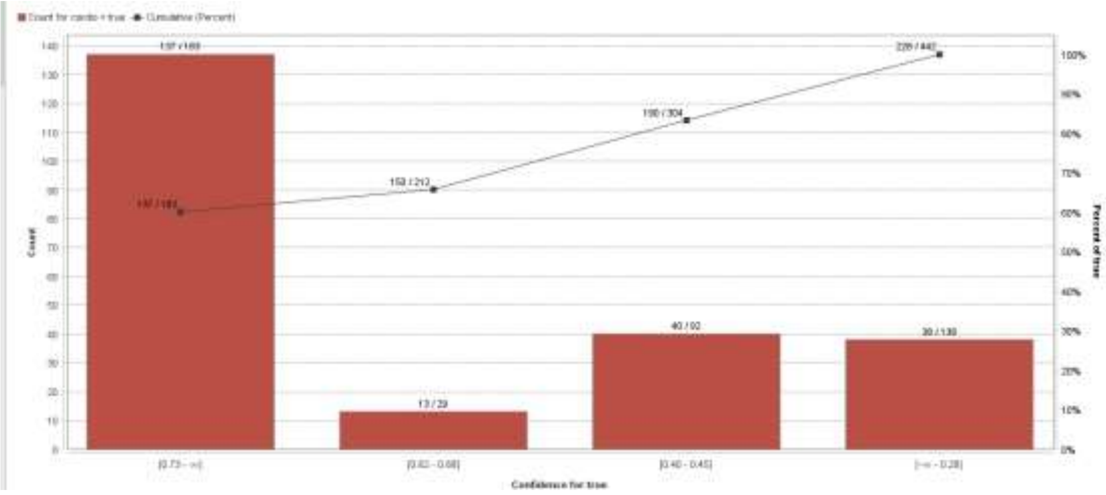
☑ apply prepruning

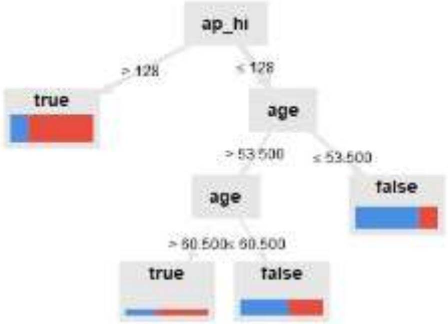
minimal gain: 0.01

minimal leaf size: 62

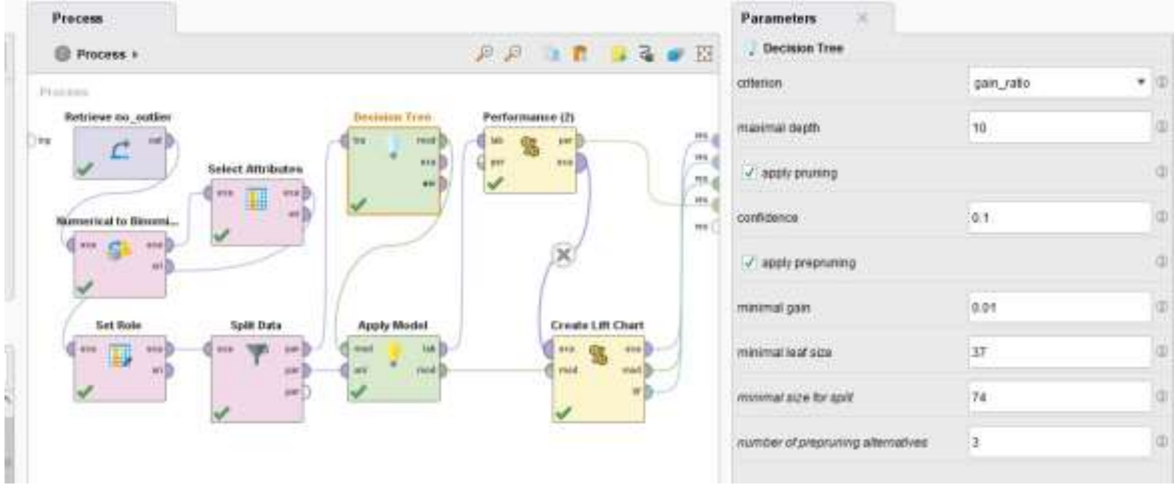
minimal size for split: 124

number of prepruning alternatives: 3





5. DT 5



Parameters

Decision Tree

criterion: gain_ratio

maximal depth: 10

☑ apply pruning

confidence: 0.1

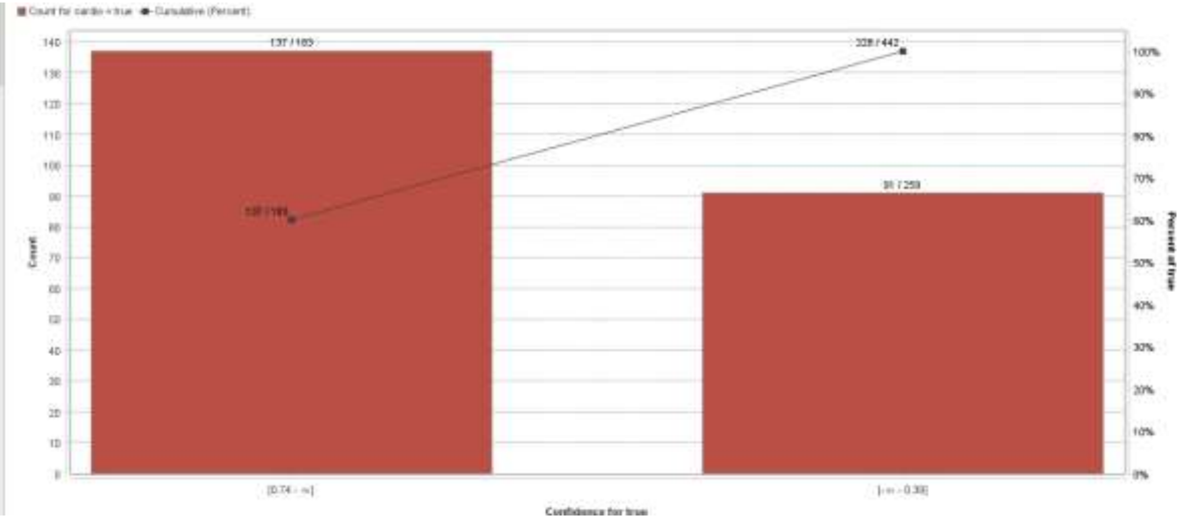
☑ apply prepruning

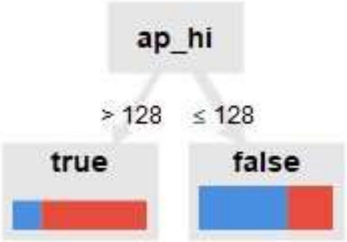
minimal gain: 0.01

minimal leaf size: 37

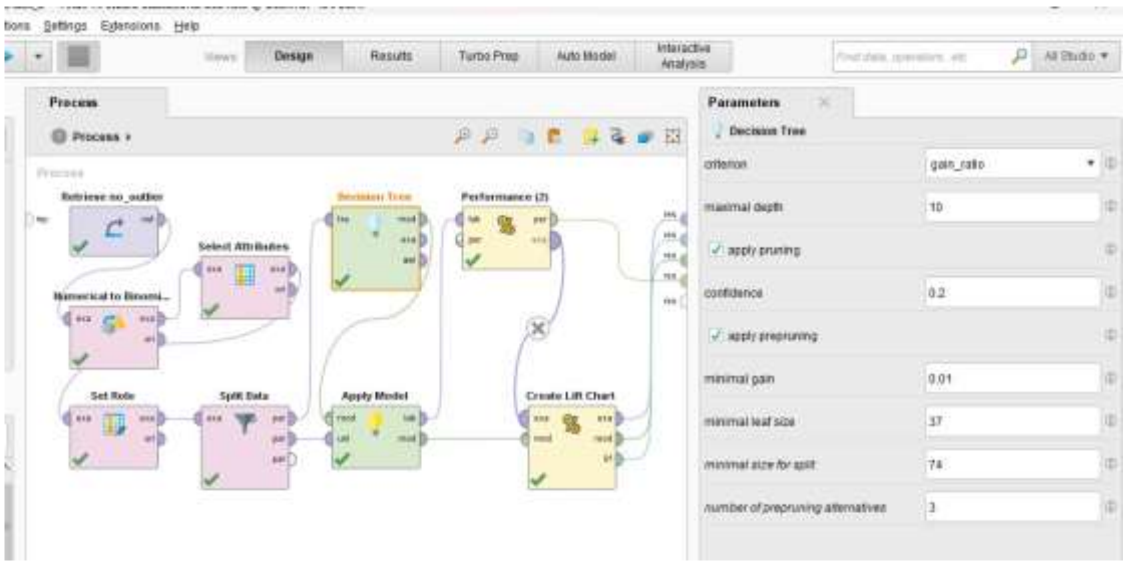
minimal size for split: 74

number of prepruning alternatives: 3





6. DT 6



Parameters

Decision Tree

criterion: gain_ratio

maximal depth: 10

☑ apply pruning

confidence: 0.2

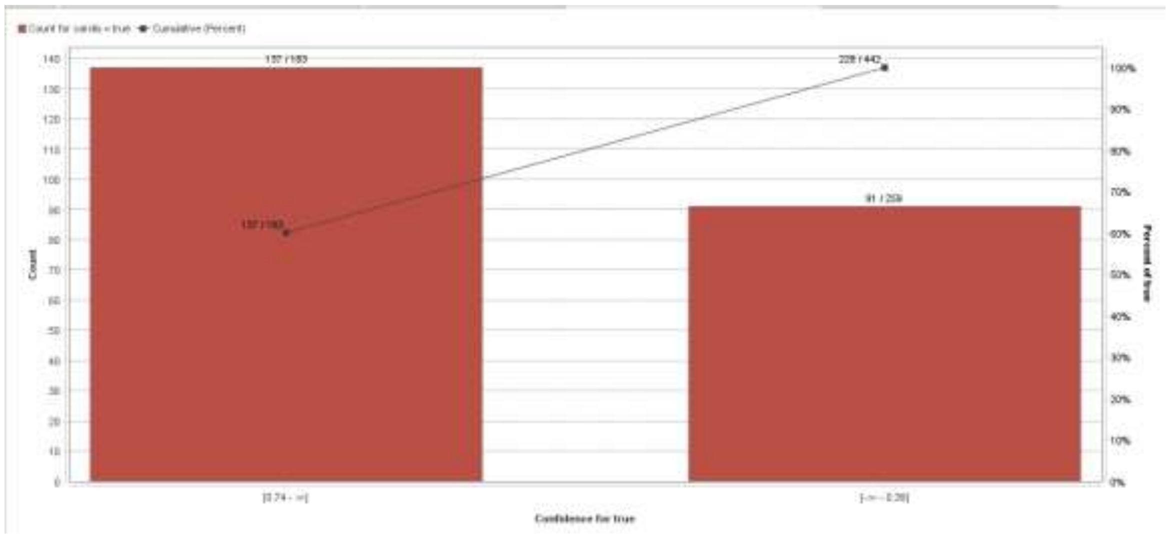
☑ apply prepruning

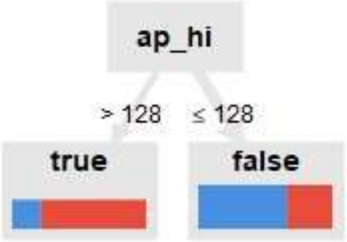
minimal gain: 0.01

minimal leaf size: 37

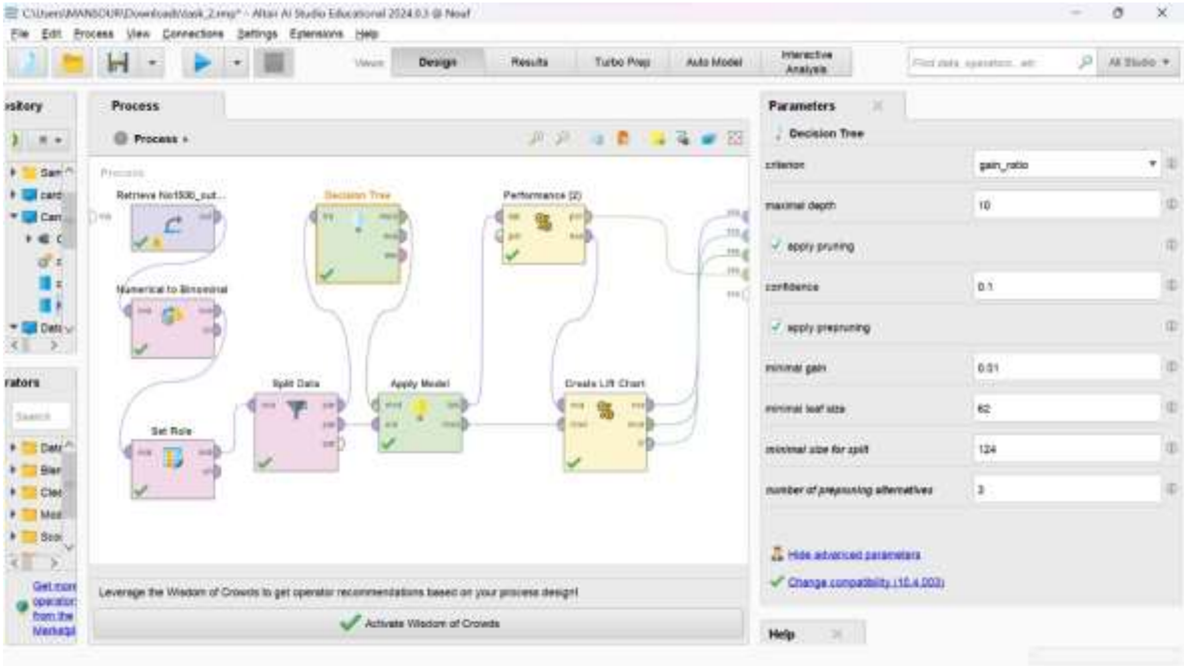
minimal size for split: 74

number of prepruning alternatives: 3



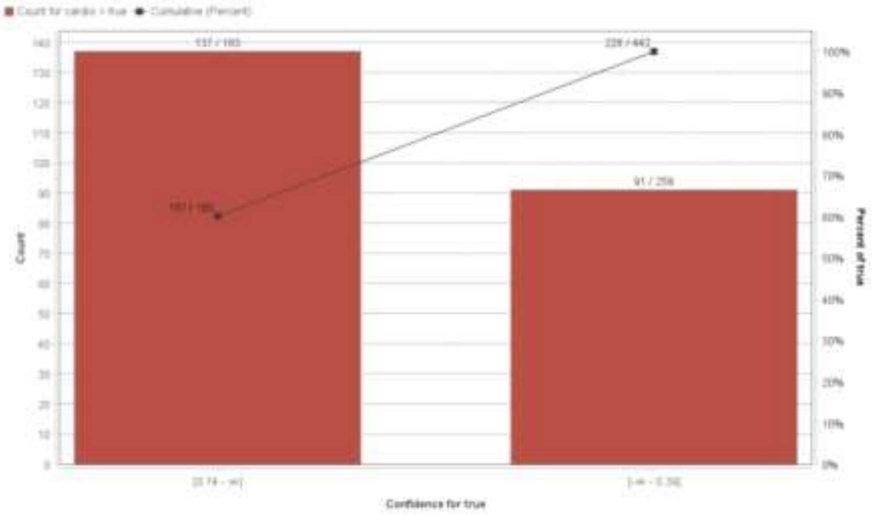


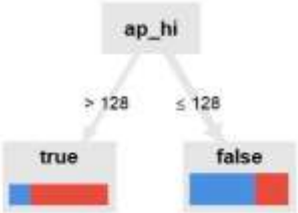
7. DT 7


C:\Users\MANSOUR\Downloads\task_2.rmp* – Altair AI Studio Educational 2024.0.3 @ Neaf
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
AI Studio
Process
Retrieve Nir1500_out...
Numerical to Binominal
Set Role
Split Data
Decision Tree
Apply Model
Performance (2)
Create Lift Chart
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Hide advanced parameters
Change compatibility (10.4.003)
Help

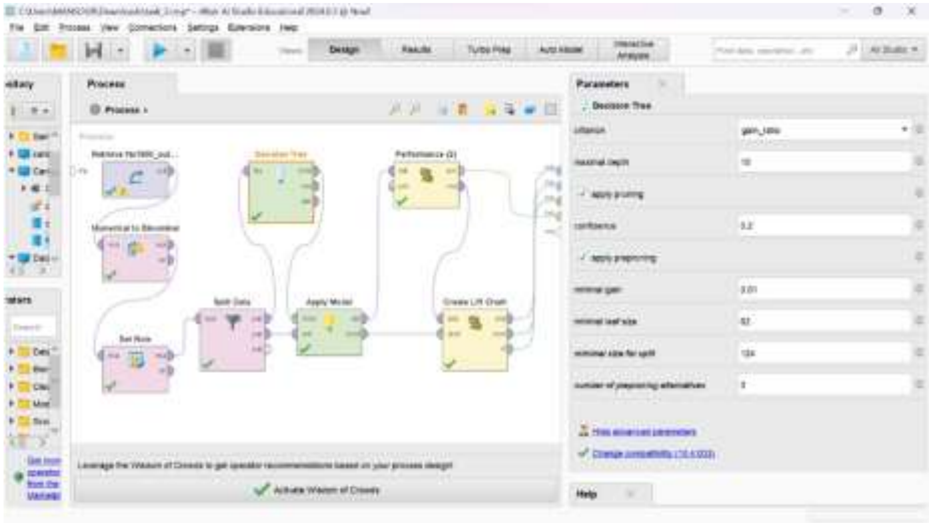
Parameters – Decision Tree

criterion: gain_ratio
maximal depth: 10
✓ apply pruning
confidence: 0.1
✓ apply prepruning
minimal gain: 0.01
minimal leaf size: 62
minimal size for split: 124
number of prepruning alternatives: 3



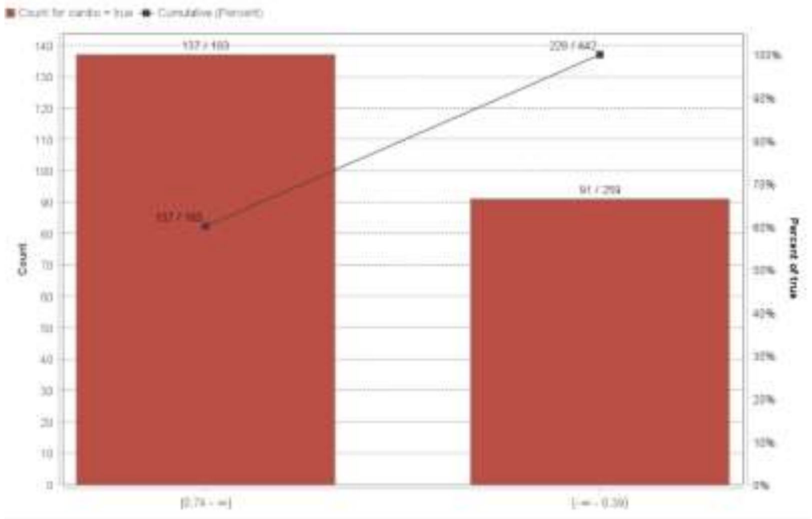


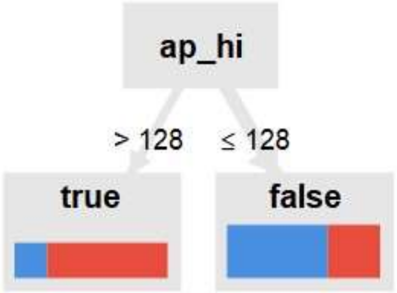
8. DT 8


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
AI Studio
Process
Retrieve Nir1500_out...
Numerical to Binominal
Set Role
Split Data
Decision Tree
Apply Model
Performance (2)
Create Lift Chart
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Hide advanced parameters
Change compatibility (10.4.003)
Help

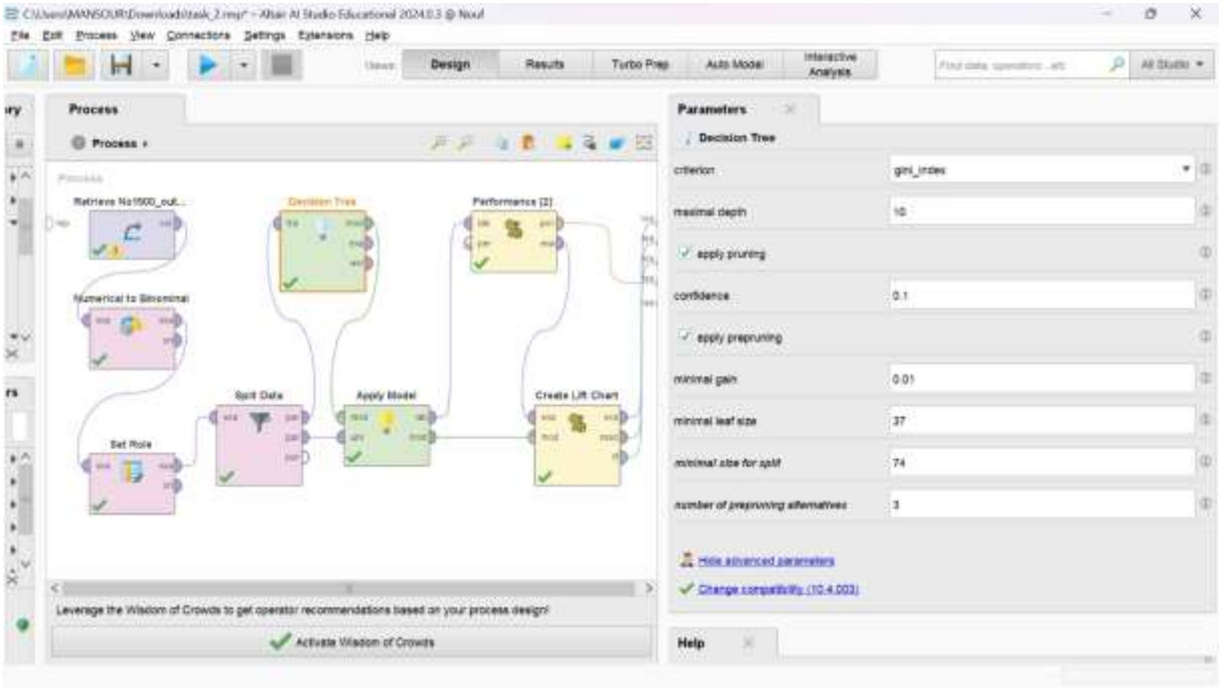
Parameters – Decision Tree

criterion: gain_ratio
maximal depth: 10
✓ apply pruning
confidence: 0.2
✓ apply prepruning
minimal gain: 0.01
minimal leaf size: 62
minimal size for split: 124
number of prepruning alternatives: 3





9. DT 9



C:\Users\MANSOUR\Downloads\task_2.rmp* – Altair AI Studio Educational 2024.0.3 @ Nouf

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

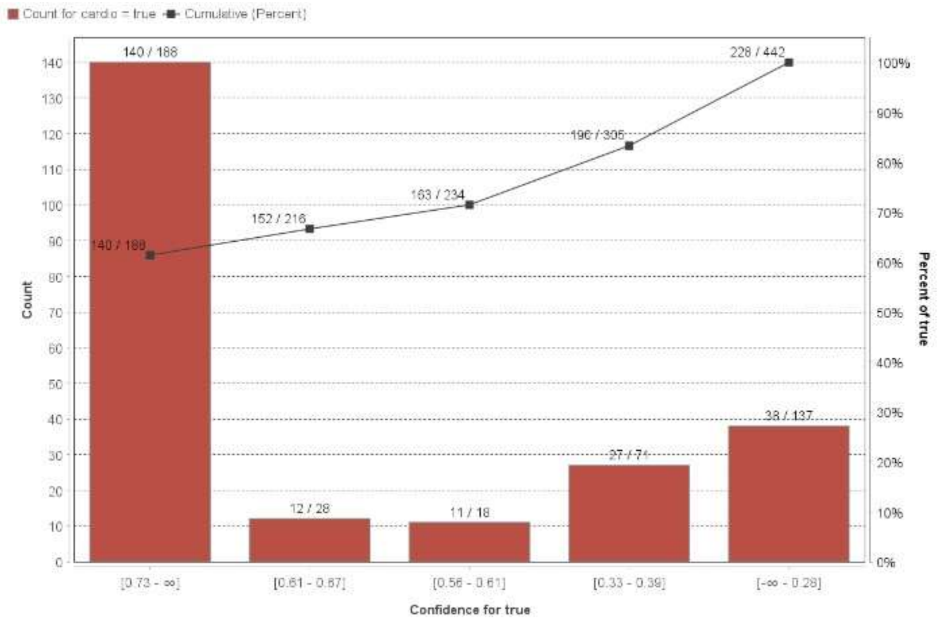
Activate Wisdom of Crowds

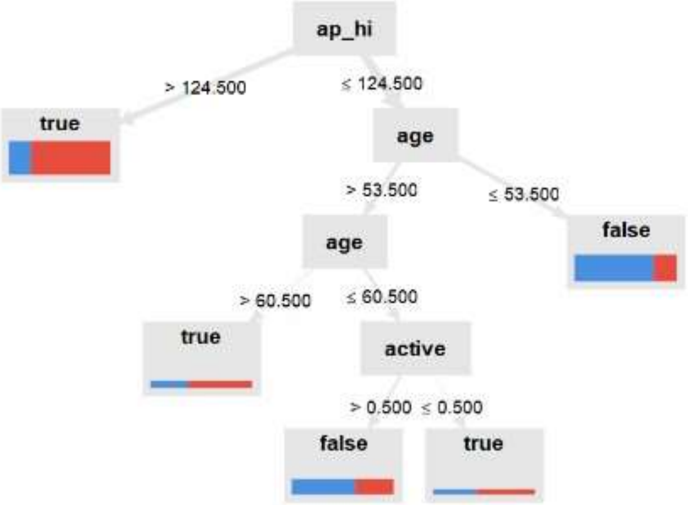
Parameters – Decision Tree

criterion	gini_index
maximal depth	10
✓ apply pruning	
confidence	0.1
✓ apply prepruning	
minimal gain	0.01
minimal leaf size	37
minimal size for split	74
number of prepruning alternatives	3

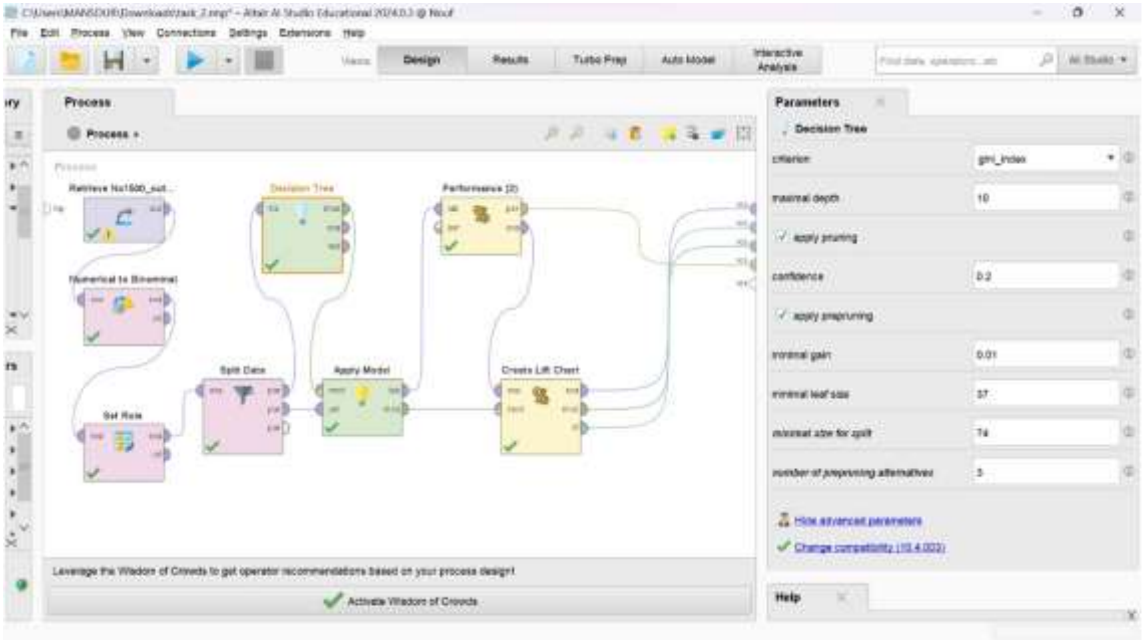
Hide advanced parameters

Change compatibility (10.4.003)





10. DT 10



C:\Users\MANSOUR\Downloads\task_2.rmp* – Altair AI Studio Educational 2024.0.3 @ Nouf

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

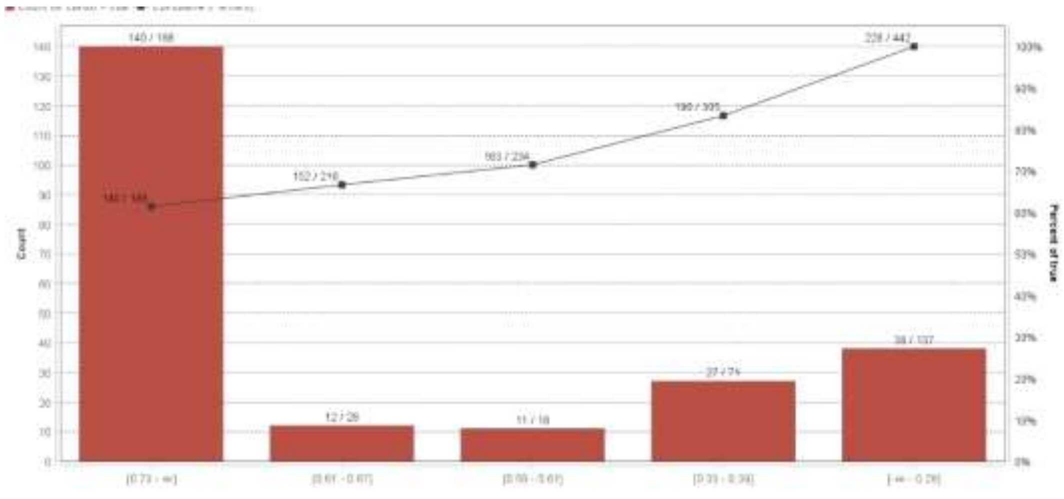
Activate Wisdom of Crowds

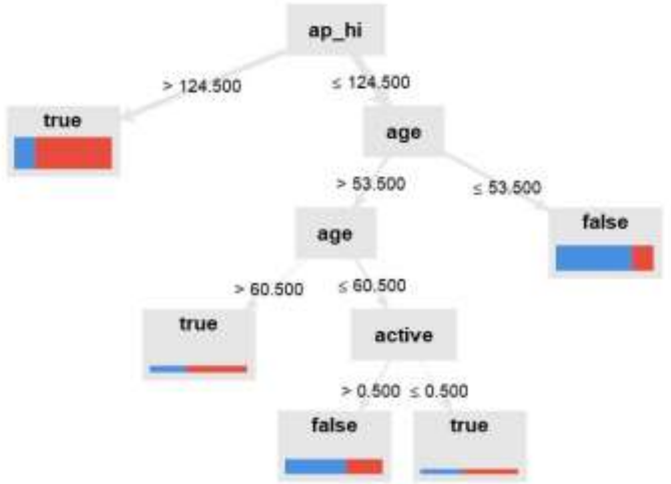
Parameters – Decision Tree

criterion	gini_index
maximal depth	10
✓ apply pruning	
confidence	0.2
✓ apply prepruning	
minimal gain	0.01
minimal leaf size	37
minimal size for split	74
number of prepruning alternatives	3

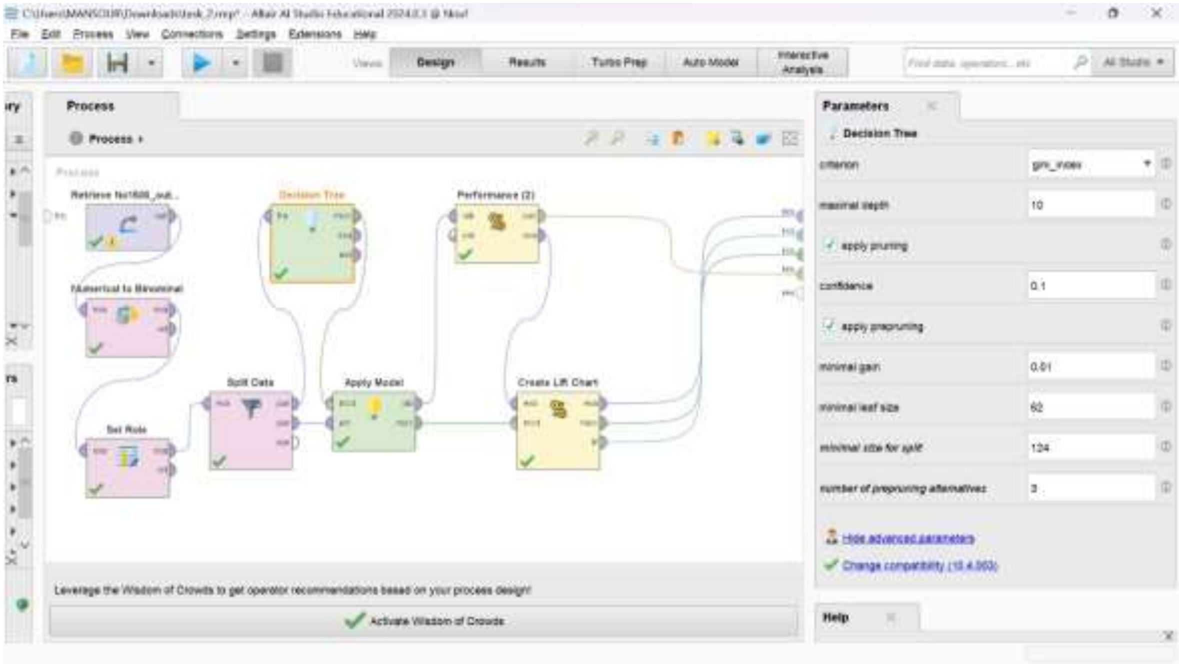
Hide advanced parameters

Change compatibility (10.4.003)





11.DT 11



Parameters

Decision Tree

criterion: gini_index

maximal depth: 10

☑ apply pruning

confidence: 0.1

☑ apply prepruning

minimal gain: 0.01

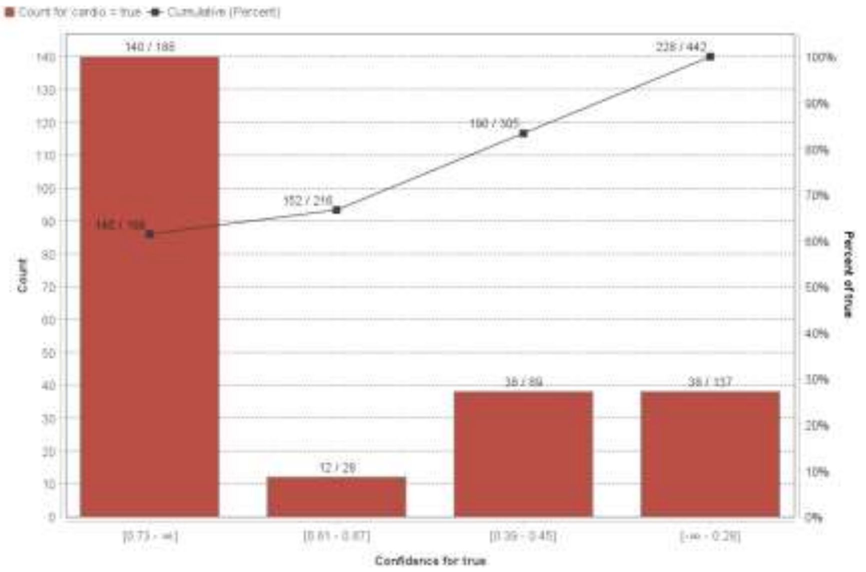
minimal leaf size: 62

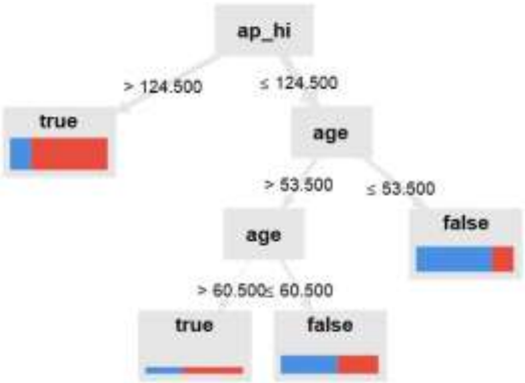
minimal size for split: 124

number of prepruning alternatives: 3

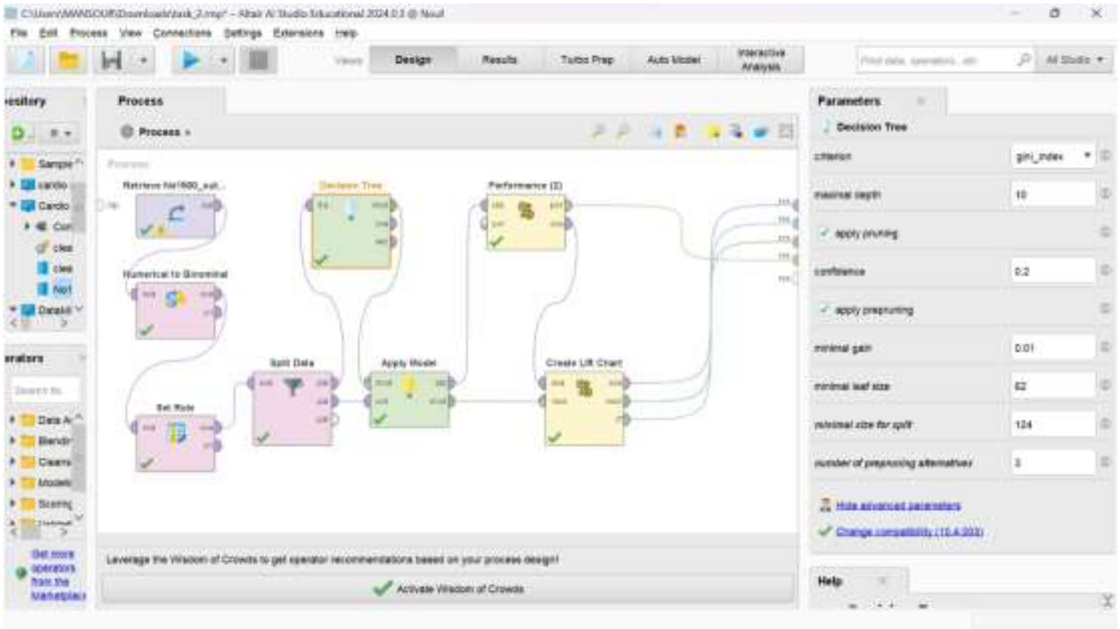
Hide advanced parameters

Change compatibility (10.4.003)





12.DT 12



Parameters

Decision Tree

criterion: gini_index

maximal depth: 10

☑ apply pruning

confidence: 0.2

☑ apply prepruning

minimal gain: 0.01

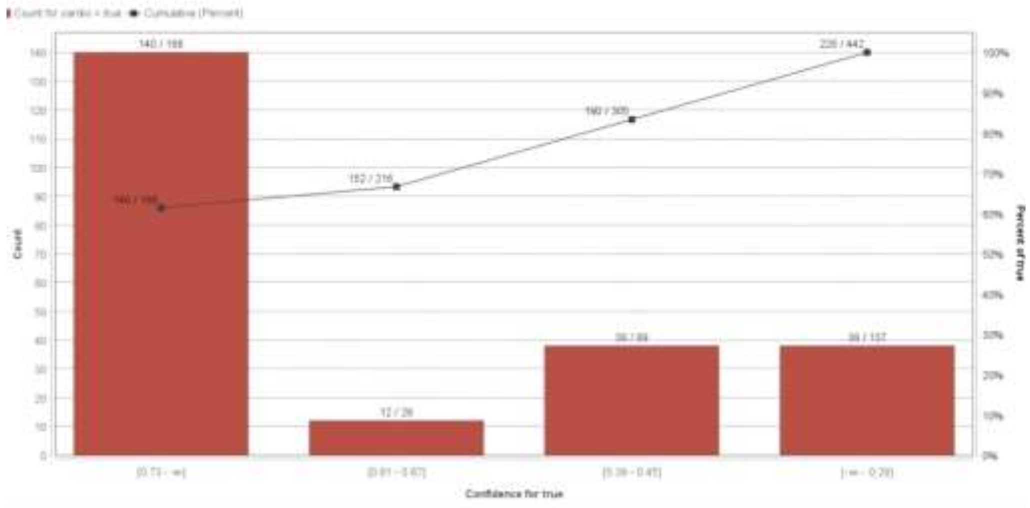
minimal leaf size: 62

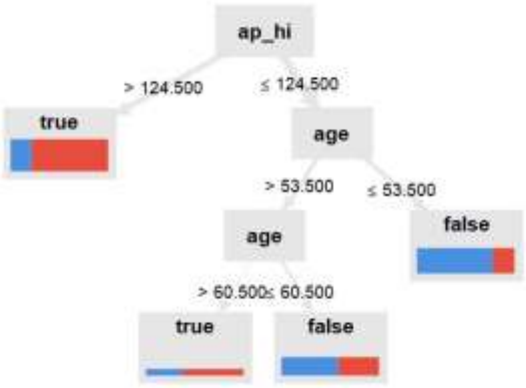
minimal size for split: 124

number of prepruning alternatives: 3

Hide advanced parameters

Change compatibility (10.4.003)



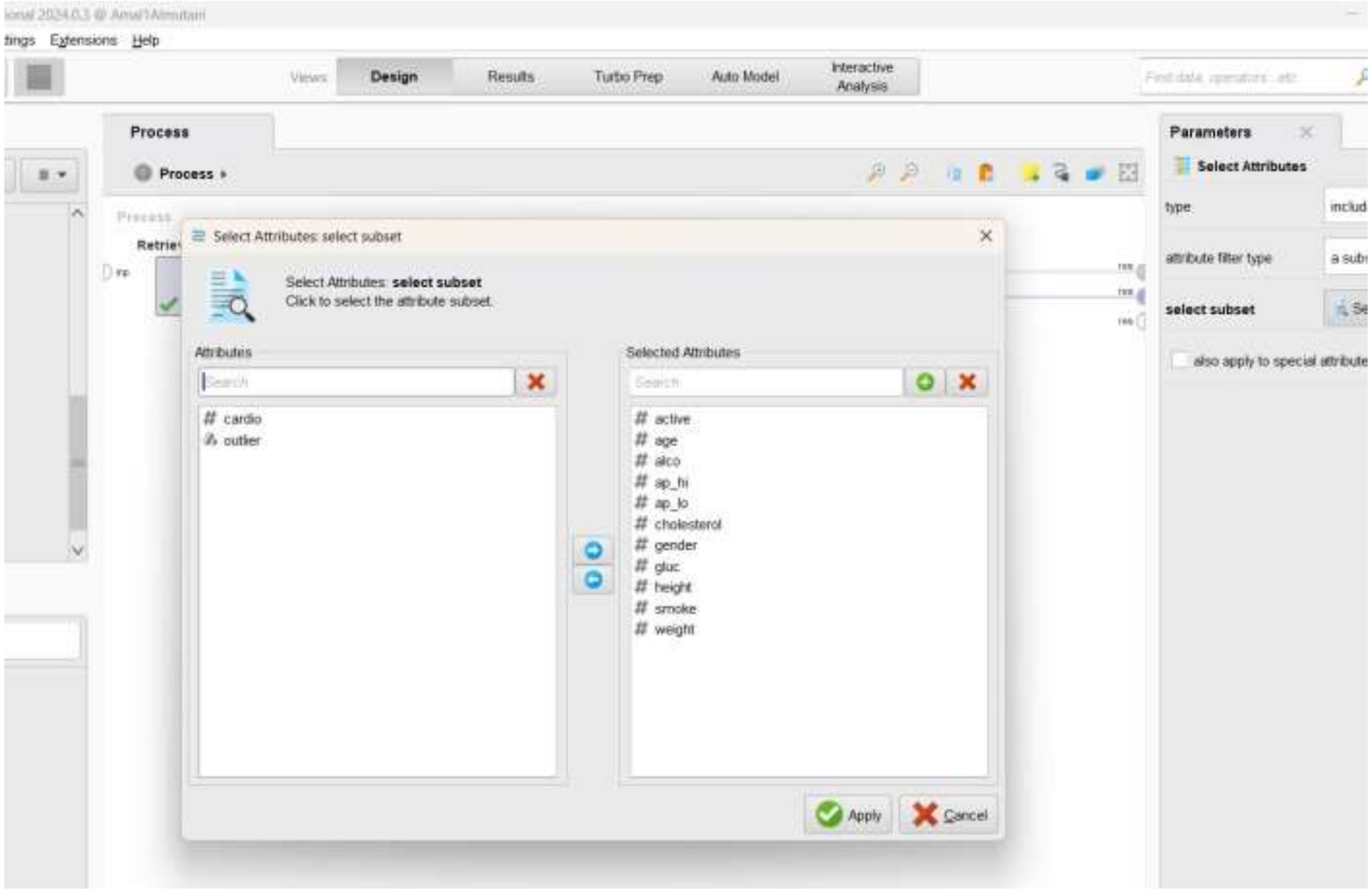


4. Try to find a combination different from the above combinations that gives the best results.

Task 3:

1. Use k-means to generate 4 segmentations with the parameter Number of Clusters described in the table below. You should ignore the following variables: ID variable, and the target variable.

we used the Select Attributes operator in RapidMiner to exclude the ID and target attributes from the analysis. The target attribute in our dataset, cardio, was excluded as it is not relevant to the clustering process



Select Attributes: select subset

Select Attributes: **select subset**
Click to select the attribute subset.

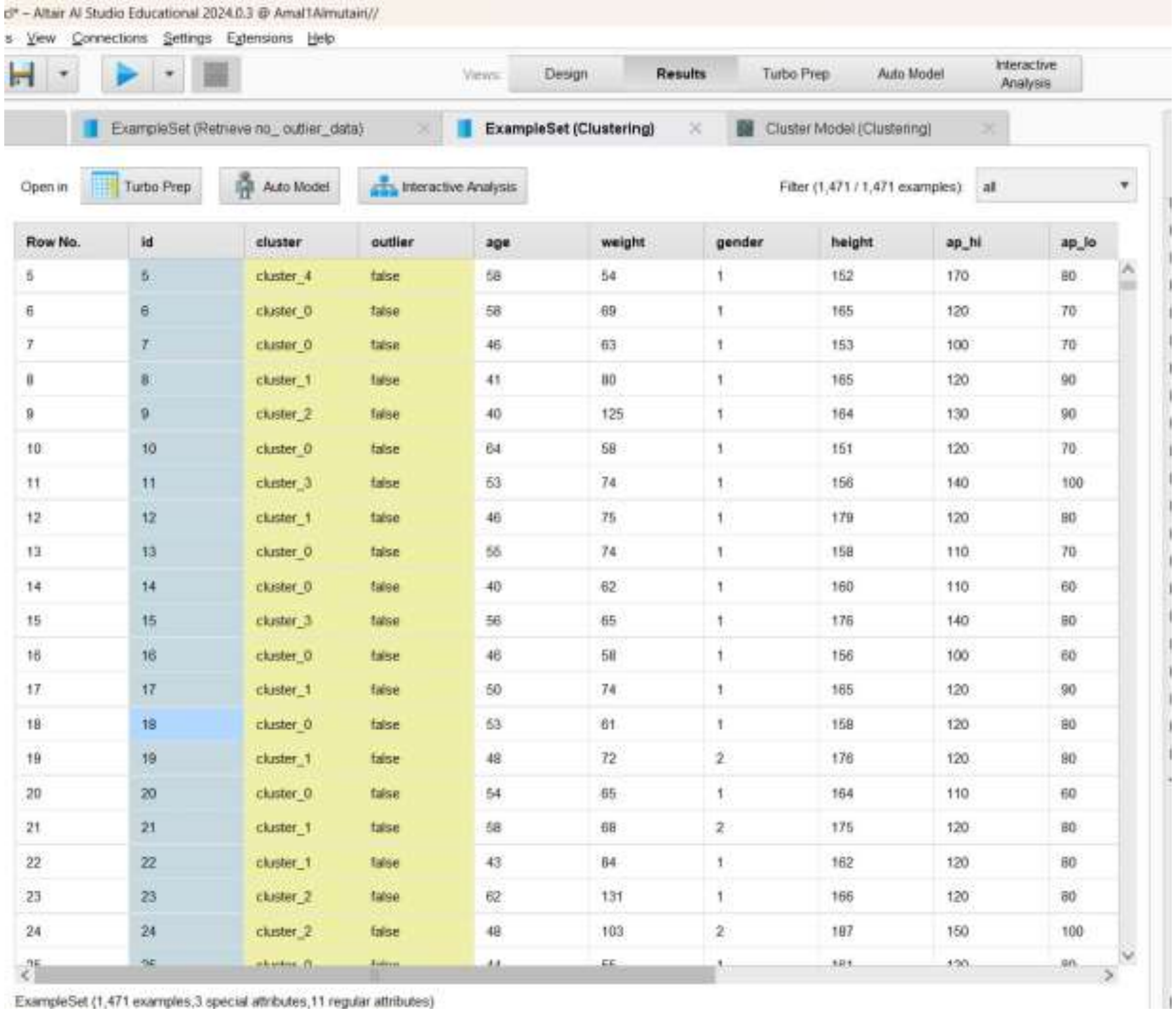
Attributes:
- cardio
- outlier

Selected Attributes:
- active
- age
- alco
- ap_hi
- ap_lo
- cholesterol
- gender
- gluc
- height
- smoke
- weight

Parameters — Select Attributes

type	includ
attribute filter type	a subs
select subset	Se

□ also apply to special attribute



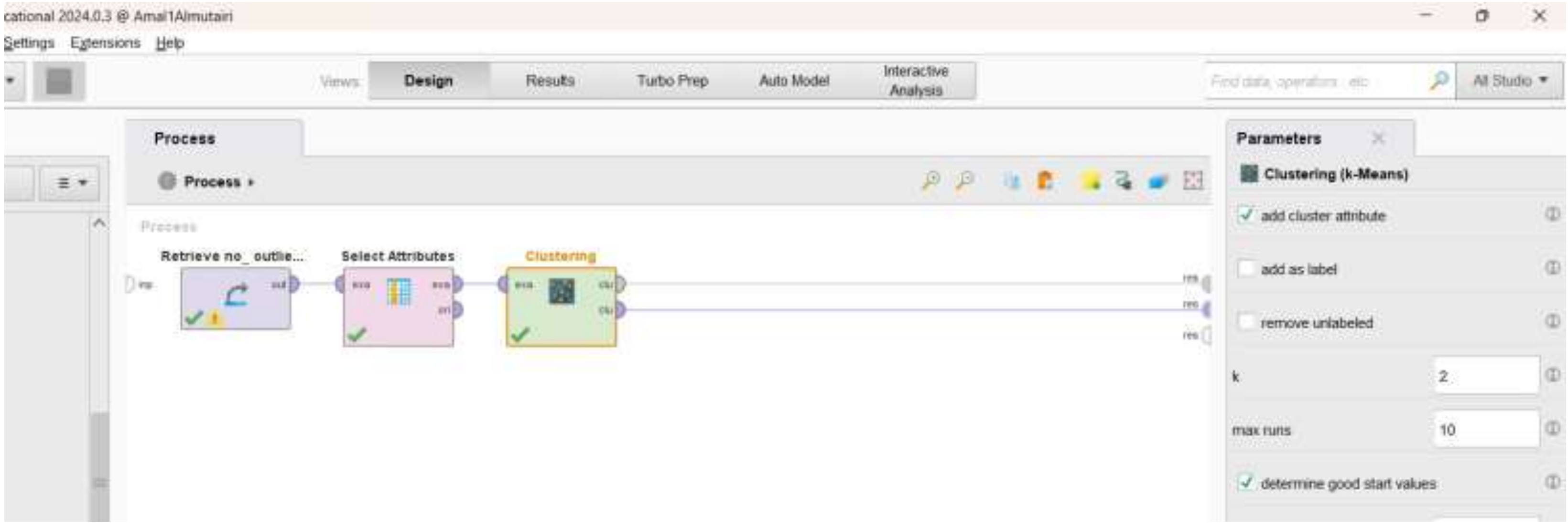
ExampleSet (Retrieve no_ outlier_data) | ExampleSet (Clustering) | Cluster Model (Clustering)

Filter (1,471 / 1,471 examples): all

Row No.	id	cluster	outlier	age	weight	gender	height	ap_hi	ap_lo
5	5	cluster_4	false	58	54	1	152	170	80
6	6	cluster_0	false	58	69	1	165	120	70
7	7	cluster_0	false	46	63	1	153	100	70
8	8	cluster_1	false	41	80	1	165	120	90
9	9	cluster_2	false	40	125	1	164	130	90
10	10	cluster_0	false	64	58	1	151	120	70
11	11	cluster_3	false	53	74	1	156	140	100
12	12	cluster_1	false	46	75	1	179	120	80
13	13	cluster_0	false	55	74	1	158	110	70
14	14	cluster_0	false	40	62	1	160	110	60
15	15	cluster_3	false	56	65	1	176	140	80
16	16	cluster_0	false	46	58	1	156	100	60
17	17	cluster_1	false	50	74	1	165	120	90
18	18	cluster_0	false	53	61	1	158	120	80
19	19	cluster_1	false	48	72	2	176	120	80
20	20	cluster_0	false	54	65	1	164	110	60
21	21	cluster_1	false	58	68	2	175	120	80
22	22	cluster_1	false	43	84	1	162	120	80
23	23	cluster_2	false	62	131	1	166	120	80
24	24	cluster_2	false	48	103	2	187	150	100

ExampleSet (1,471 examples,3 special attributes,11 regular attributes)

km_2 :



Parameters — Clustering (k-Means)

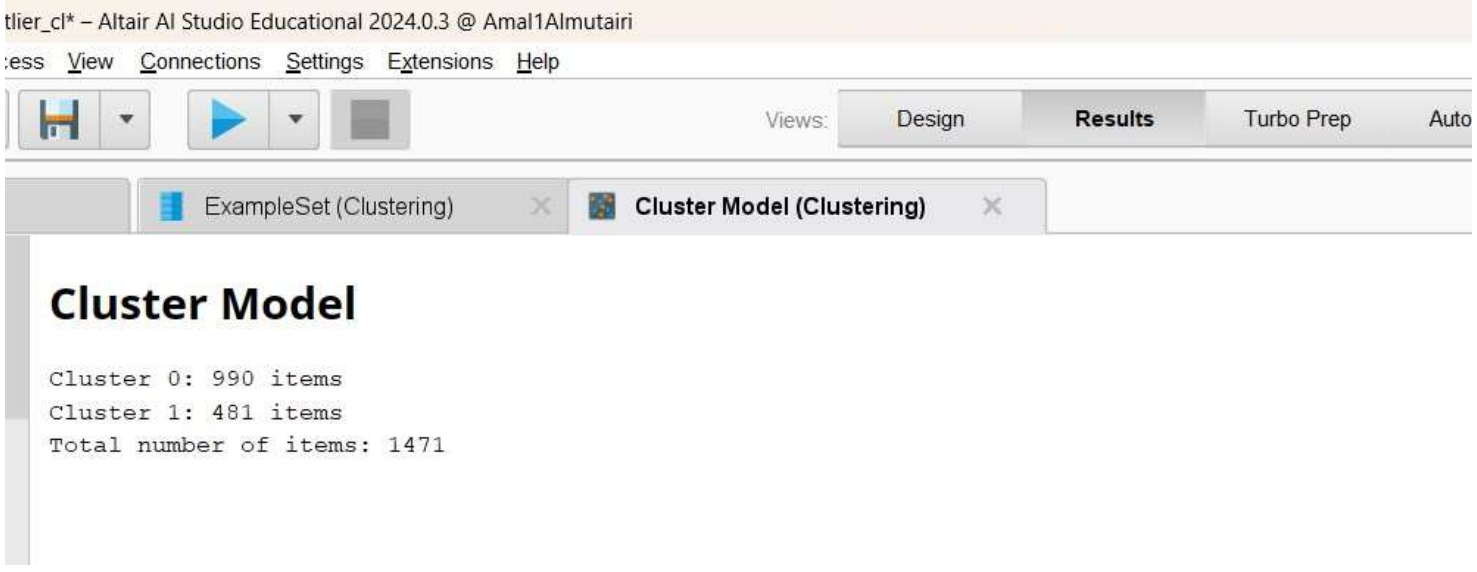
☑ add cluster attribute

□ add as label

□ remove unlabeled

k	2
max runs	10

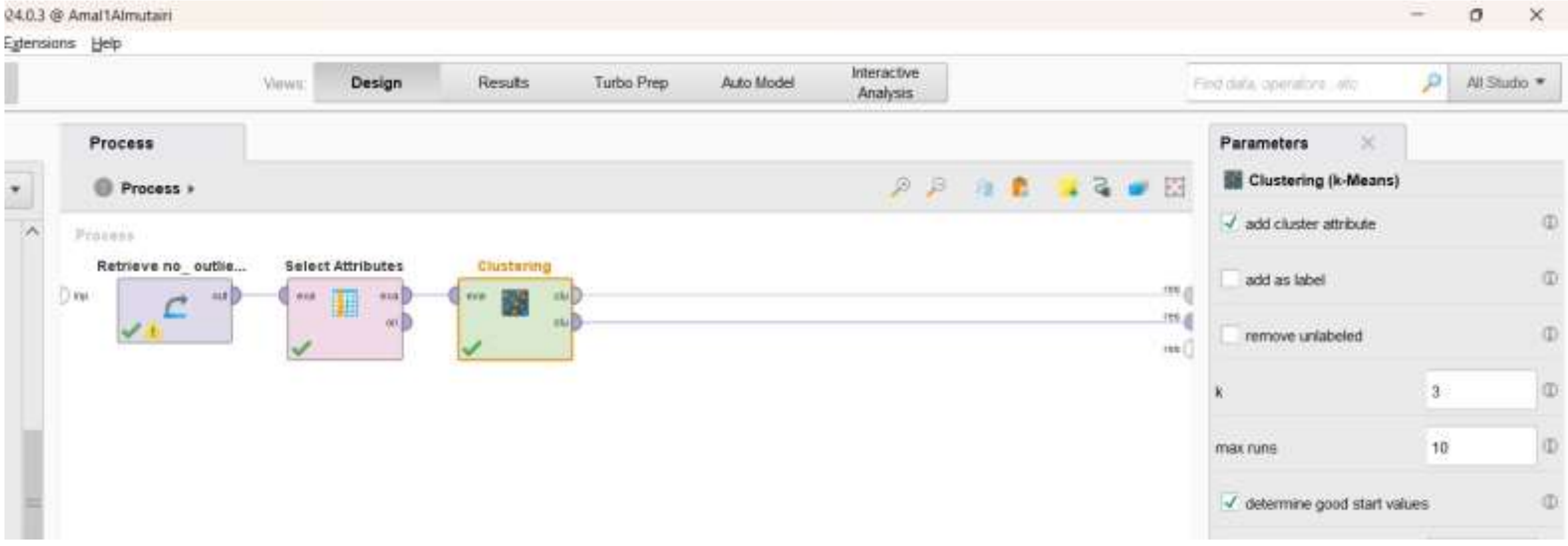
☑ determine good start values


tlier_cl* – Altair AI Studio Educational 2024.0.3 @ Amal1Almutairi
:ess View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto
ExampleSet (Clustering)
Cluster Model (Clustering)

Cluster Model

```
Cluster 0: 990 items
Cluster 1: 481 items
Total number of items: 1471
```

km_3 :



☑ add cluster attribute

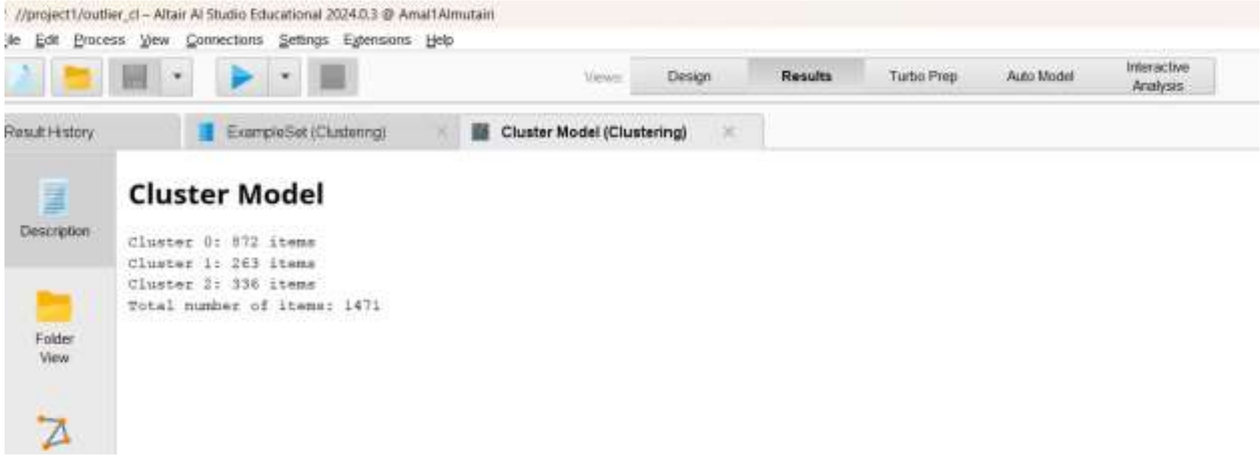
☐ add as label

☐ remove unlabeled

k 3

max runs 10

☑ determine good start values

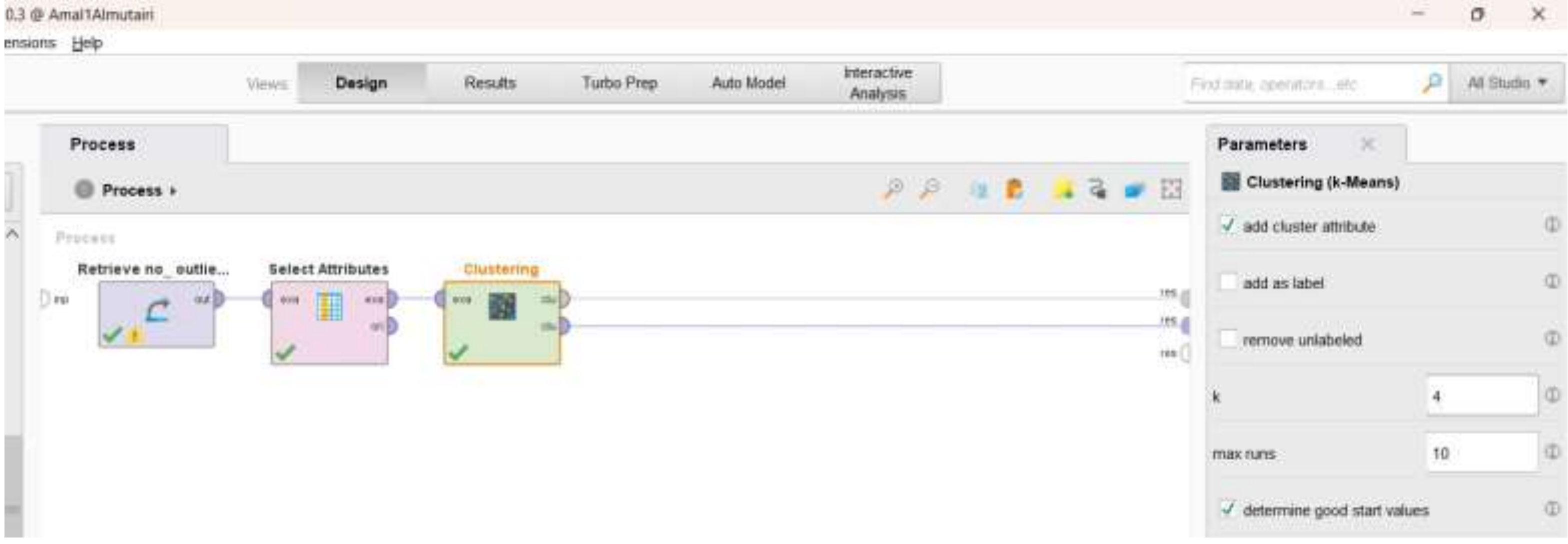

//project1/outlier_cl – Altair AI Studio Educational 2024.0.3 @ Amal1Almutairi
ile Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Result History
ExampleSet (Clustering)
Cluster Model (Clustering)
Description
Folder View

Cluster Model

```
Cluster 0: 872 items
Cluster 1: 263 items
Cluster 2: 336 items
Total number of items: 1471
```

km_4

:



☑ add cluster attribute

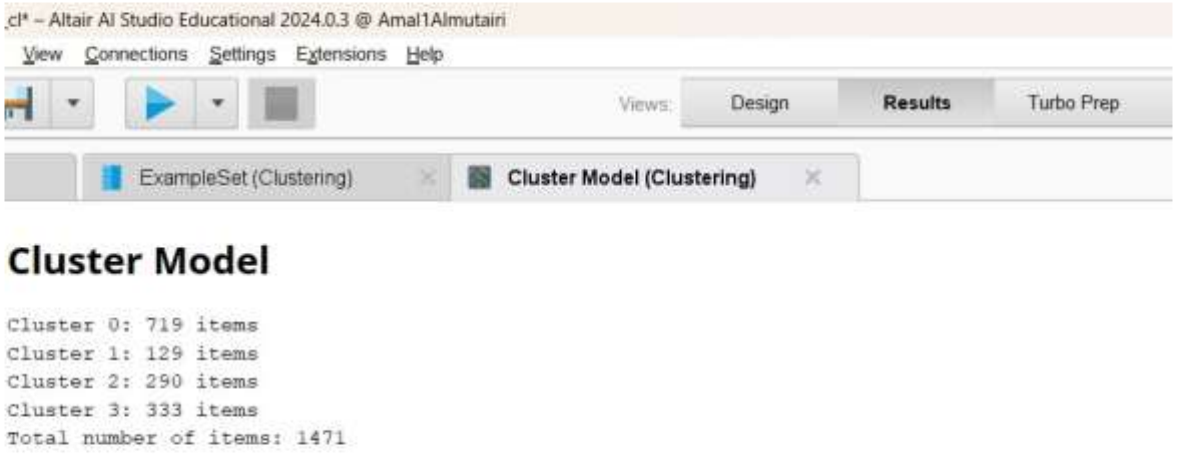
☐ add as label

☐ remove unlabeled

k 4

max runs 10

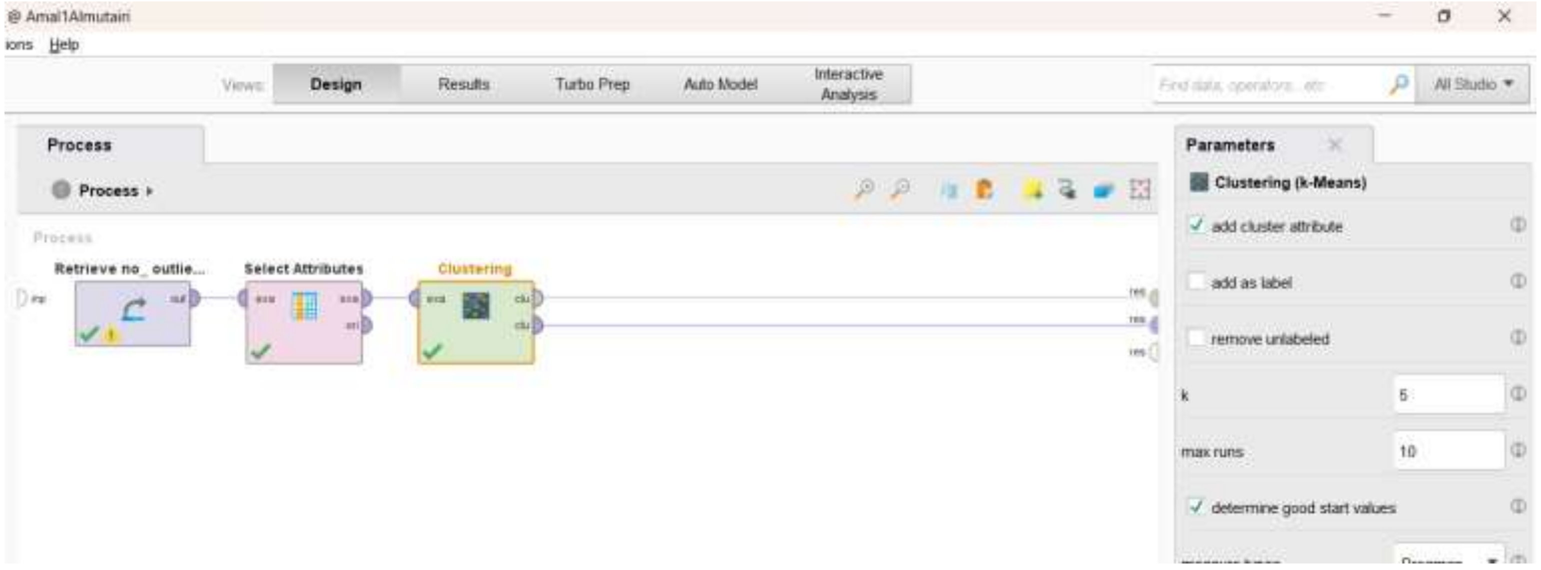
☑ determine good start values


_cl* – Altair AI Studio Educational 2024.0.3 @ Amal1Almutairi
View Connections Settings Extensions Help
Views: Design Results Turbo Prep
ExampleSet (Clustering)
Cluster Model (Clustering)

Cluster Model

```
Cluster 0: 719 items
Cluster 1: 129 items
Cluster 2: 290 items
Cluster 3: 333 items
Total number of items: 1471
```

km_5:



Clustering (k-Means)

☑ add cluster attribute

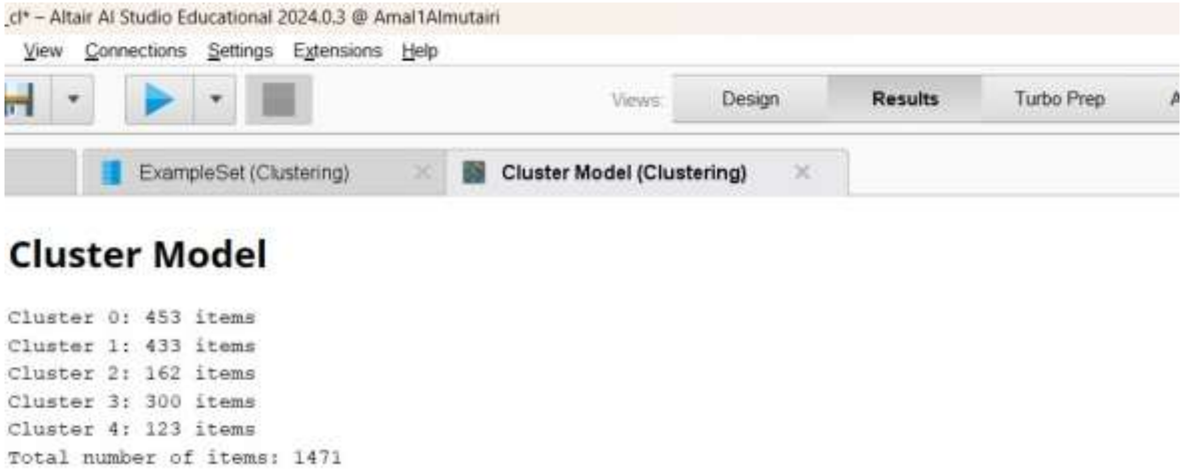
□ add as label

□ remove unlabeled

k 5

max runs 10

☑ determine good start values


_cl* – Altair AI Studio Educational 2024.0.3 @ Amal1Almutairi
View Connections Settings Extensions Help
Views: Design Results Turbo Prep
ExampleSet (Clustering)
Cluster Model (Clustering)

Cluster Model

```
Cluster 0: 453 items
Cluster 1: 433 items
Cluster 2: 162 items
Cluster 3: 300 items
Cluster 4: 123 items
Total number of items: 1471
```

s	Segmentation Label	Number of Clusters	Cluster Sizes	Size of Smallest Cluster	Outlier (Yes or No)
1.	km_2	2	Cluster 0: 990 Cluster 1: 481	Cluster 1: 481	NO
2.	km_3	3	Cluster 0: 872 Cluster 1: 263 Cluster 2: 336	Cluster 1: 263	NO
3.	km_4	4	Cluster 0: 719 Cluster 1: 129 Cluster 2: 290 Cluster 3: 333	Cluster 1: 129	Yes
4.	km_5	5	Cluster 0: 453 Cluster 1: 433 Cluster 2: 162 Cluster 3: 300 Cluster 4: 123	Cluster 4: 123	Yes

2. Describe the clusters in segmentation km_3 using the descriptive statistics from the results.

Cluster 0 : Contains 872 observations, the largest group.

- Age: The mean age is 52.525, making this group middle-aged.
- Weight: The mean weight is 67.655, the lowest among the groups, indicating that this group is lighter.
- Gender: The mean value for gender is 1.334, indicating that this group is mostly female.
- Height: The mean height is 164.347 cm, shorter compared to Group 1 but close to Group 2.
- Blood pressure:
  - Systolic (ap_hi): 117.366, the lowest among the groups, indicating better blood pressure health Diastolic (ap_lo): 77.462, also the lowest, indicating good cardiovascular health.
- Cholesterol and glucose: Both values are low (cholesterol: 1.258, glucose: 1.164), indicating healthier metabolic conditions.
- Smoking: Very low at 0.078.
- Alcohol consumption: Very low at 0.052.
- Physical activity: High at 0.776, indicating that this group is active.

Cluster 1 : Contains 263 observations, the smallest group.

- Age: Mean age 53.388, slightly older than Group 0.
- Weight: Mean weight 94.859, the heaviest of the groups, indicating a high BMI.
- Gender: The mean value for gender is 1.502, closer to 1.5, indicating that there are females and males, but the proportion of males is slightly higher than females.
- Height: Mean height 167.707 cm, the tallest of the groups.
- Blood pressure:
  - Systolic (ap_hi): 128.384, higher than group 0 but lower than group 2. Diastolic (ap_lo): 82.856, average.
- Cholesterol and glucose: slightly higher (cholesterol: 1.586, glucose: 1.308) compared to group 0, showing early signs of metabolic risk.
- Smoking: slightly higher at 0.148.
- Alcohol consumption: average at 0.091.
- Physical activity: high at 0.810, similar to group 2.

Cluster 2: Contains 336 observations, middle sized group.

- Age: The average age is 55.357, making this group the oldest.
- Weight: The mean weight is 73.765, higher than group 0 but lower than group 1.
- Gender : The mean gender value (1.381) shows that the group includes both males and females, with a higher proportion of females.
- Height: The mean height is 163.583 cm, the shortest among the groups.
- Blood pressure:
  - Systolic (ap_hi): 149.601, the highest, indicating a high prevalence of hypertension.
  - Diastolic (ap_lo): 90.926, also the highest, a potential indicator of cardiovascular risk.
- Cholesterol and glucose: the highest levels (cholesterol: 1.554, glucose: 1.310), reflecting a high metabolic risk.
- Smoking: slightly higher at 0.119 .
- Alcohol consumption: Low at 0.068.
- Physical activity: High at 0.812 , similar to the other groups.

3. If the domain expert was interested in evaluating outlier clusters, which segmentations (i.e., sets of clusters) would you provide to the expert for evaluation? Provide justification for your answer. You may assume that a cluster is an outlier if it contains less than 10% of the observations.

To evaluate outlier clusters, We will provide the following segmentation to the domain expert:

Segmentations with Outliers: km_4
(4 clusters):
Cluster 1 contains 129 observations, which is less than 10% of the total (1471 observations).
Justification: 10% of 1471 is approximately equal to 147 observations. Cluster 1 in km_4 contains less than this threshold, making it eligible to be an outlier.

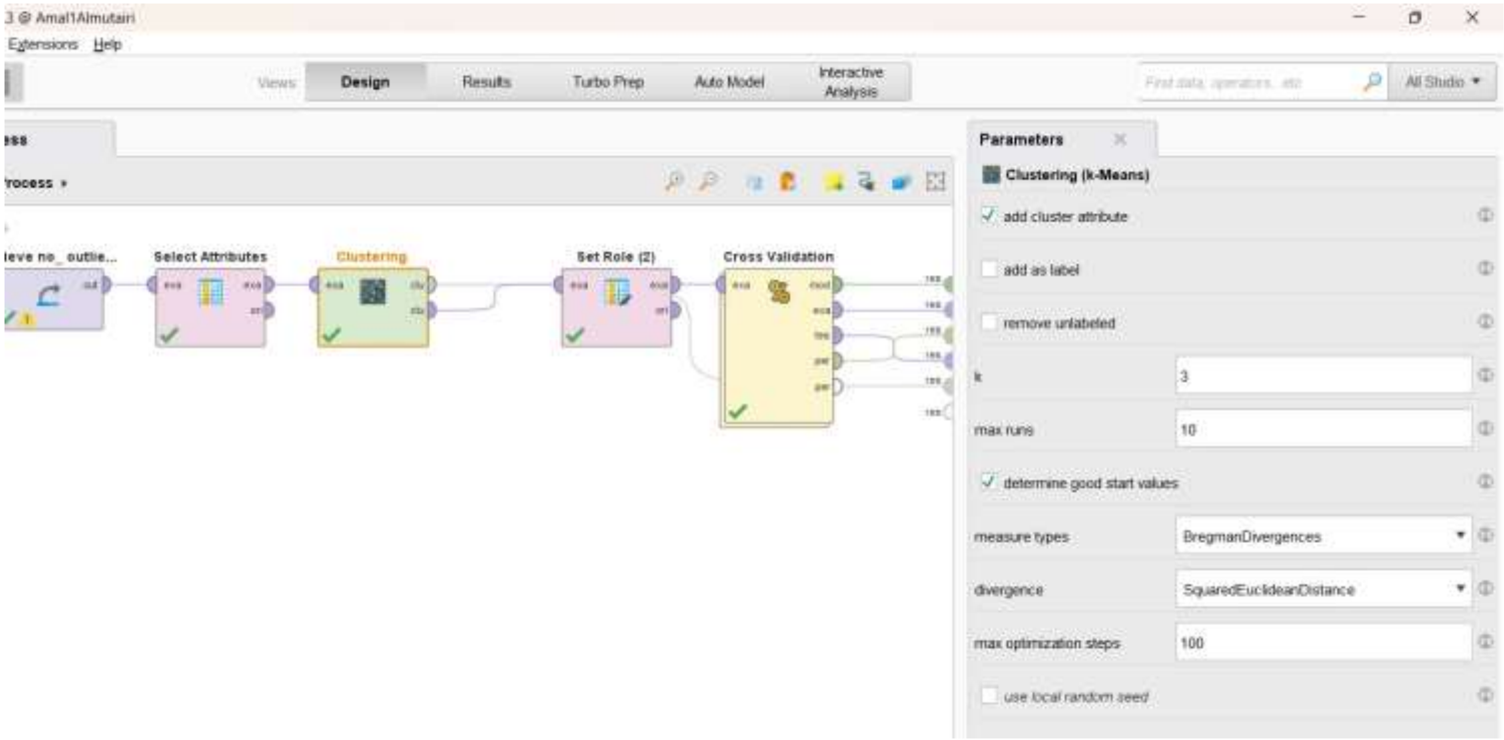
km_5 (5 clusters):
Cluster 4 contains 123 observations, which is less than 10% of the total (1471 observations).
Justification: Cluster 4 in km_5 also falls below the 10% threshold and is considered an outlier.

Excluded segmentation:
km_2 (2 clusters): Both clusters contain more than 10% of the total observations. km_3 (3 clusters): All clusters contain more than 10% of the total observations.

A domain expert should evaluate km_4 and km_5, because they contain clusters that meet the outlier criteria based on the 10% threshold.

4. For each cluster in km_3, generate 3 DTs (i.e., Information Gain, Gini, Gain Ratio) & select the DT that has the best test accuracy & good simplicity. Assume that the weights are: 0.60 for Accuracy & 0.40 for Simplicity. Also assume that there are no thresholds on these measures. Compare these segmentations based on the associated DTs. Note that for each of these DTs, the target variable is the Cluster ID.

We used the km_3 segmentation to generate 3 Decision Trees



Parameters

Clustering (k-Means)

☑ add cluster attribute

□ add as label

□ remove unlabeled

k: 3

max runs: 10

☑ determine good start values

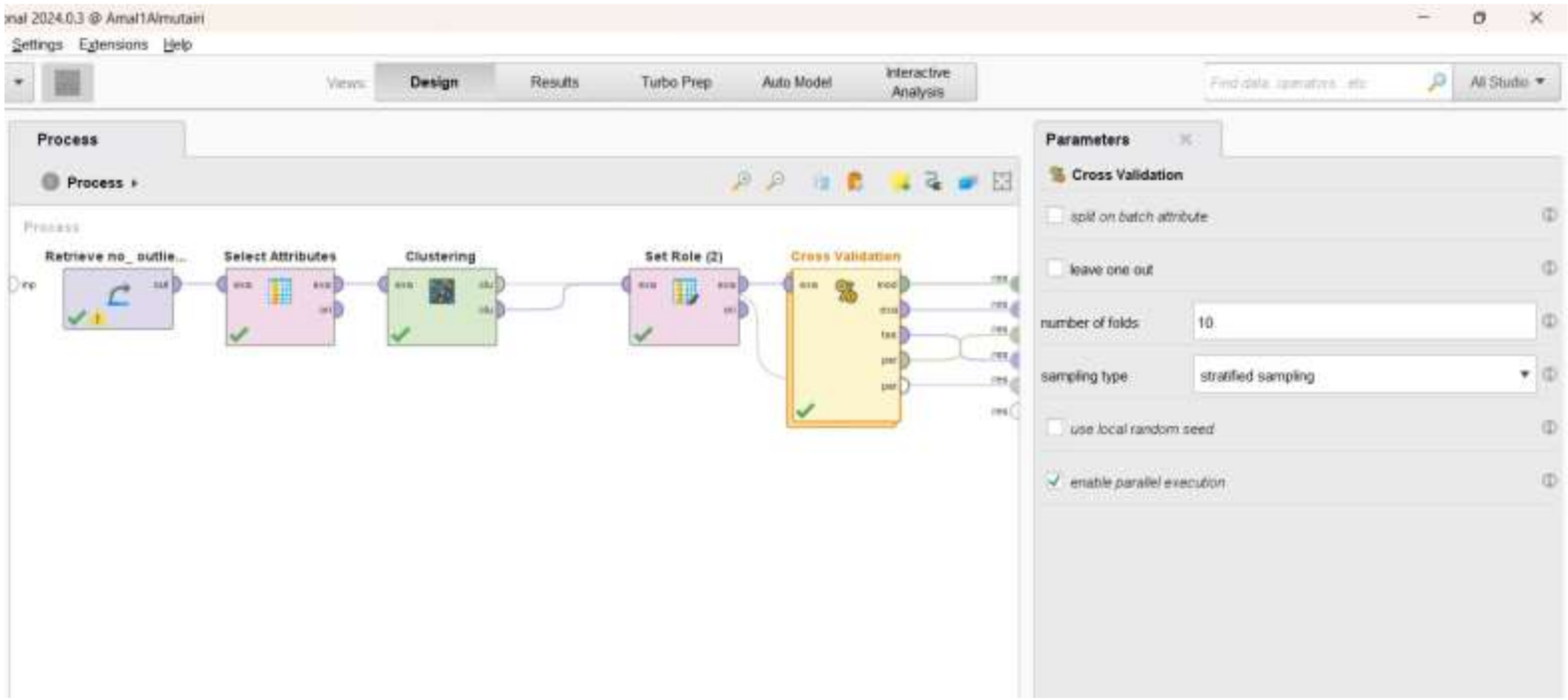
measure types: BregmanDivergences

divergence: SquaredEuclideanDistance

max optimization steps: 100

□ use local random seed

We specified 10 folds in Cross Validation and also used the Stratified method.



Parameters

Cross Validation

□ split on batch attribute

□ leave one out

number of folds: 10

sampling type: stratified sampling

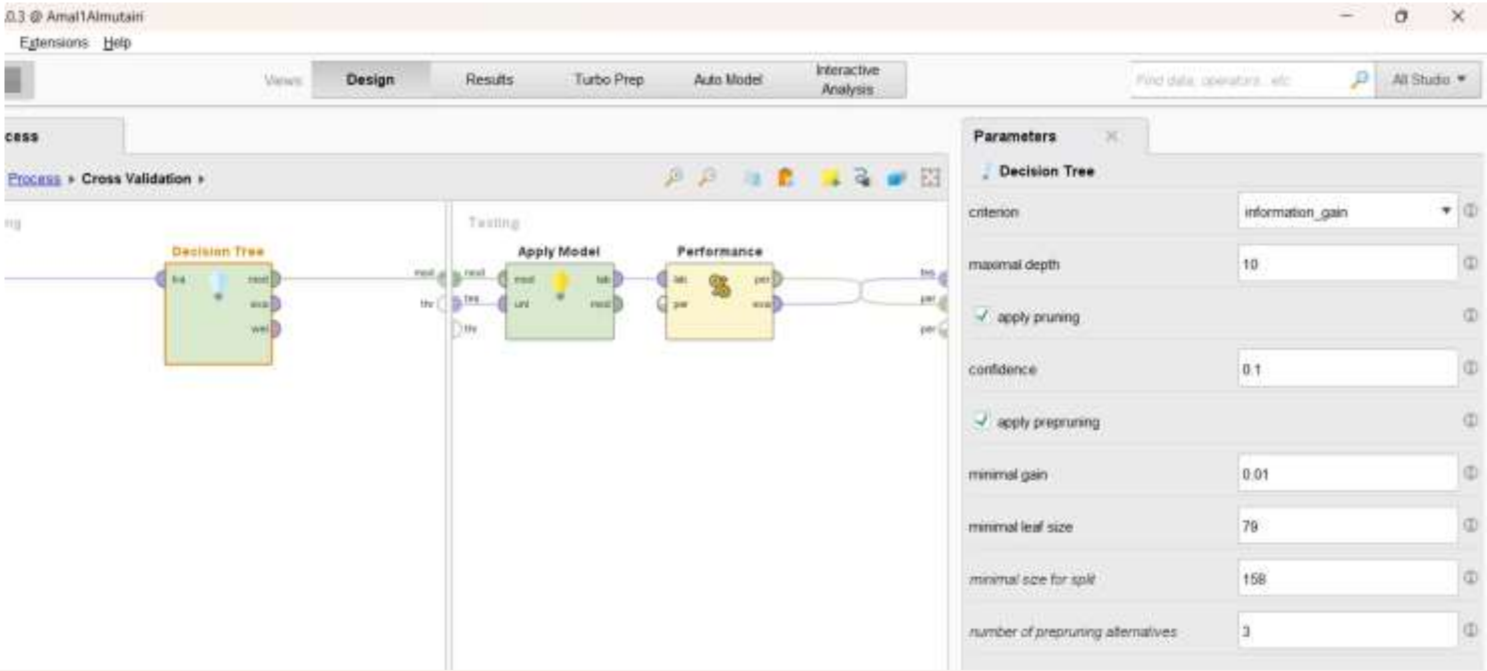
□ use local random seed

☑ enable parallel execution

We calculated the smallest cluster, which is Cluster 1 with a size of 263 and then multiplied it by 0.3 to get the number of observations per leaf, and multiplied the leaf size by 2 to get the minimal partition size. These values were applied to all the trees.

The Frist Decision Tree is Information Gain:

Confusion Matrix Table:


.0.3 @ Amal1Almutairi
Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
cess
Process ▸ Cross Validation ▸
Testing
Decision Tree
Apply Model
Performance

Parameters

Decision Tree

criterion: information_gain

maximal depth: 10

☑ apply pruning

confidence: 0.1

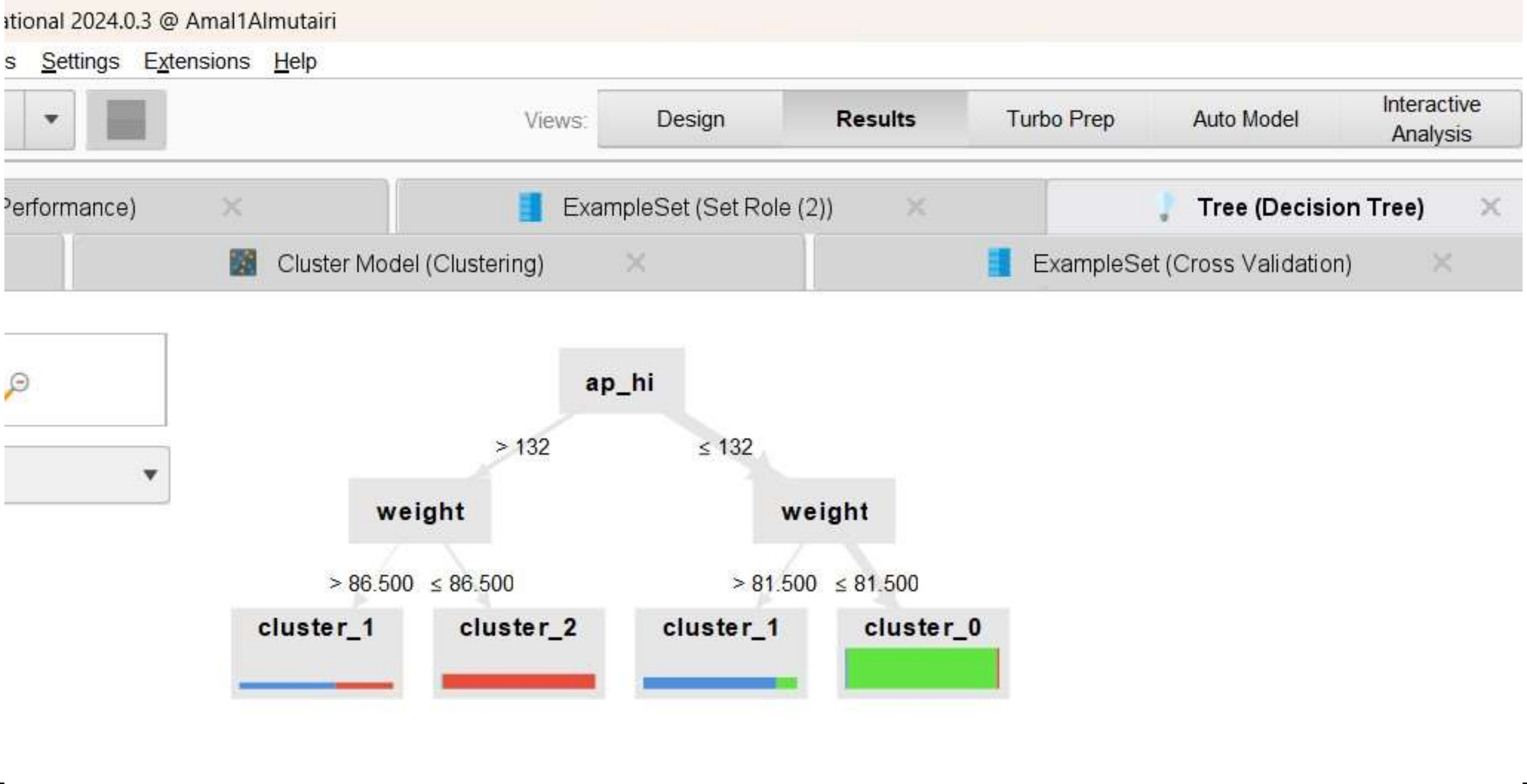
☑ apply prepruning

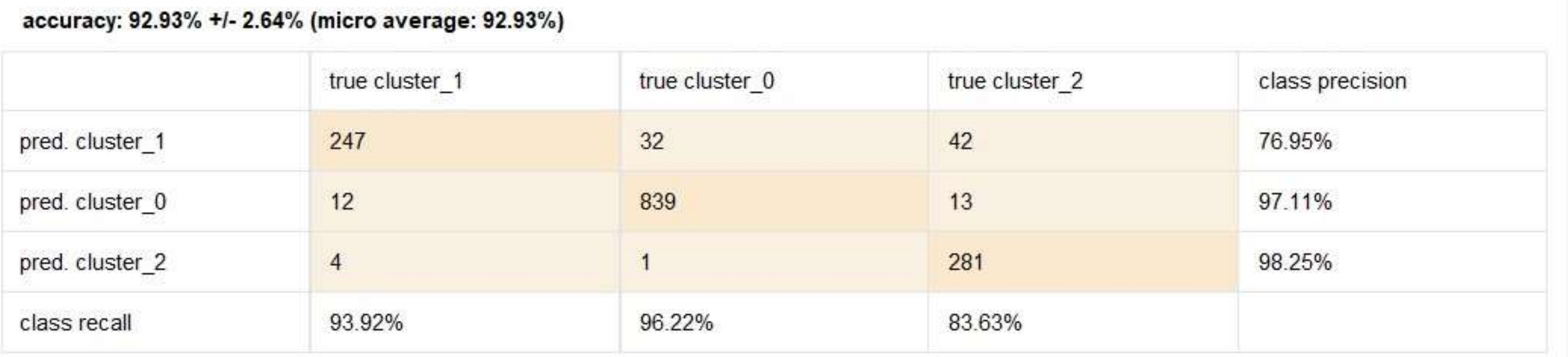
minimal gain: 0.01

minimal leaf size: 79

minimal size for split: 158

number of prepruning alternatives: 3

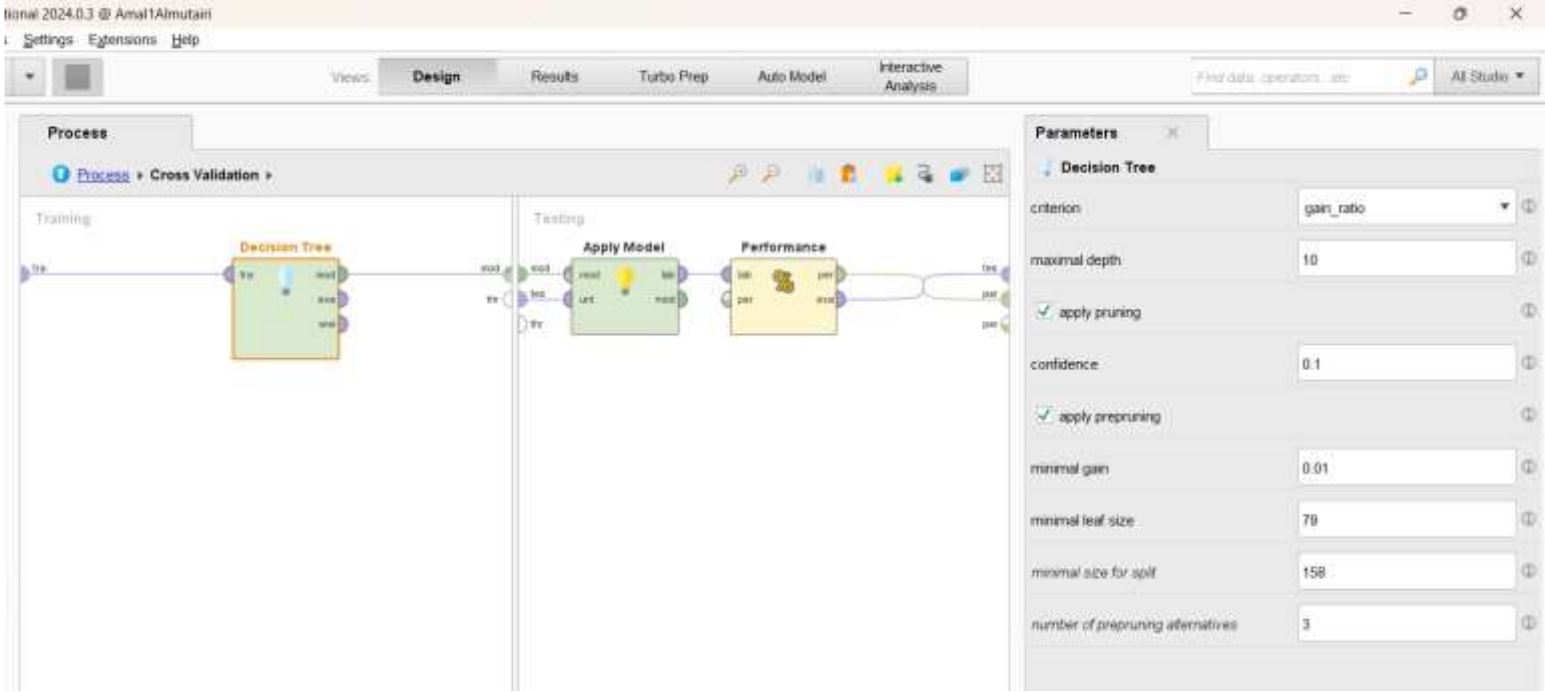

ational 2024.0.3 @ Amal1Almutairi
s Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Performance)
ExampleSet (Set Role (2))
Tree (Decision Tree)
Cluster Model (Clustering)
ExampleSet (Cross Validation)
ap_hi
> 132
≤ 132
weight
weight
> 86.500
≤ 86.500
> 81.500
≤ 81.500
cluster_1
cluster_2
cluster_1
cluster_0



accuracy: 92.93% +/- 2.64% (micro average: 92.93%)

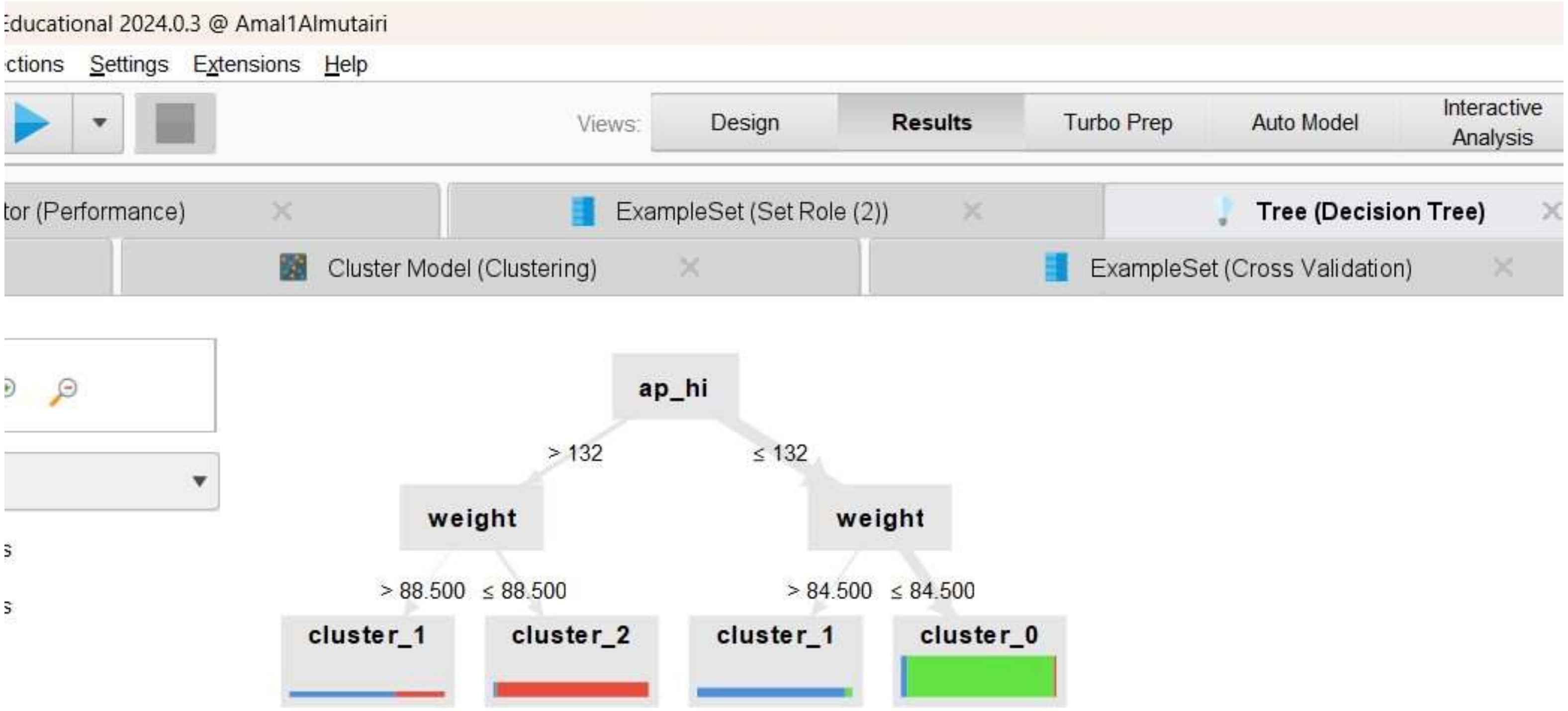
	true cluster_1	true cluster_0	true cluster_2	class precision
pred. cluster_1	247	32	42	76.95%
pred. cluster_0	12	839	13	97.11%
pred. cluster_2	4	1	281	98.25%
class recall	93.92%	96.22%	83.63%	

The Second Decision Tree is Gain Ratio:

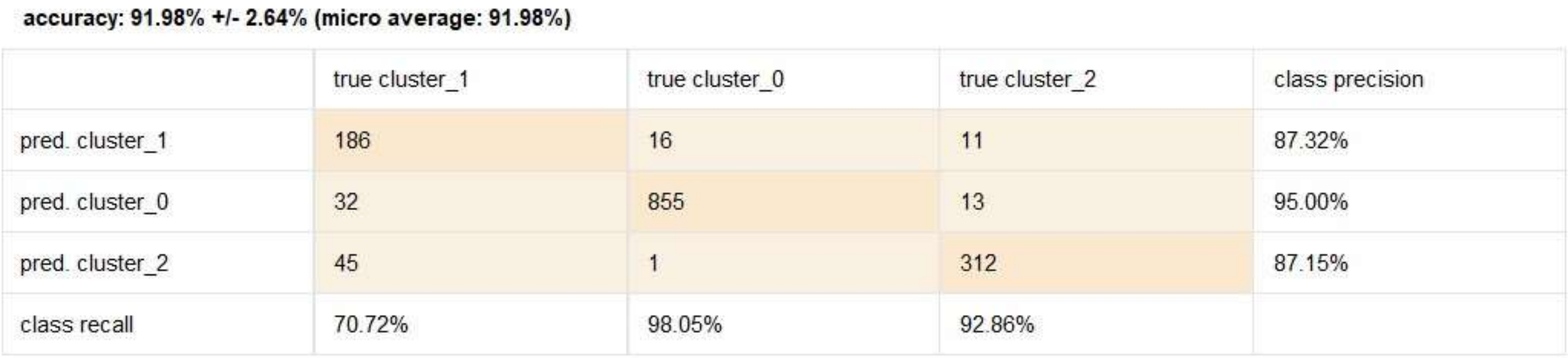


Decision Tree

criterion	gain_ratio
maximal depth	10
☑ apply pruning	
confidence	0.1
☑ apply prepruning	
minimal gain	0.01
minimal leaf size	79
minimal size for split	158
number of prepruning alternatives	3



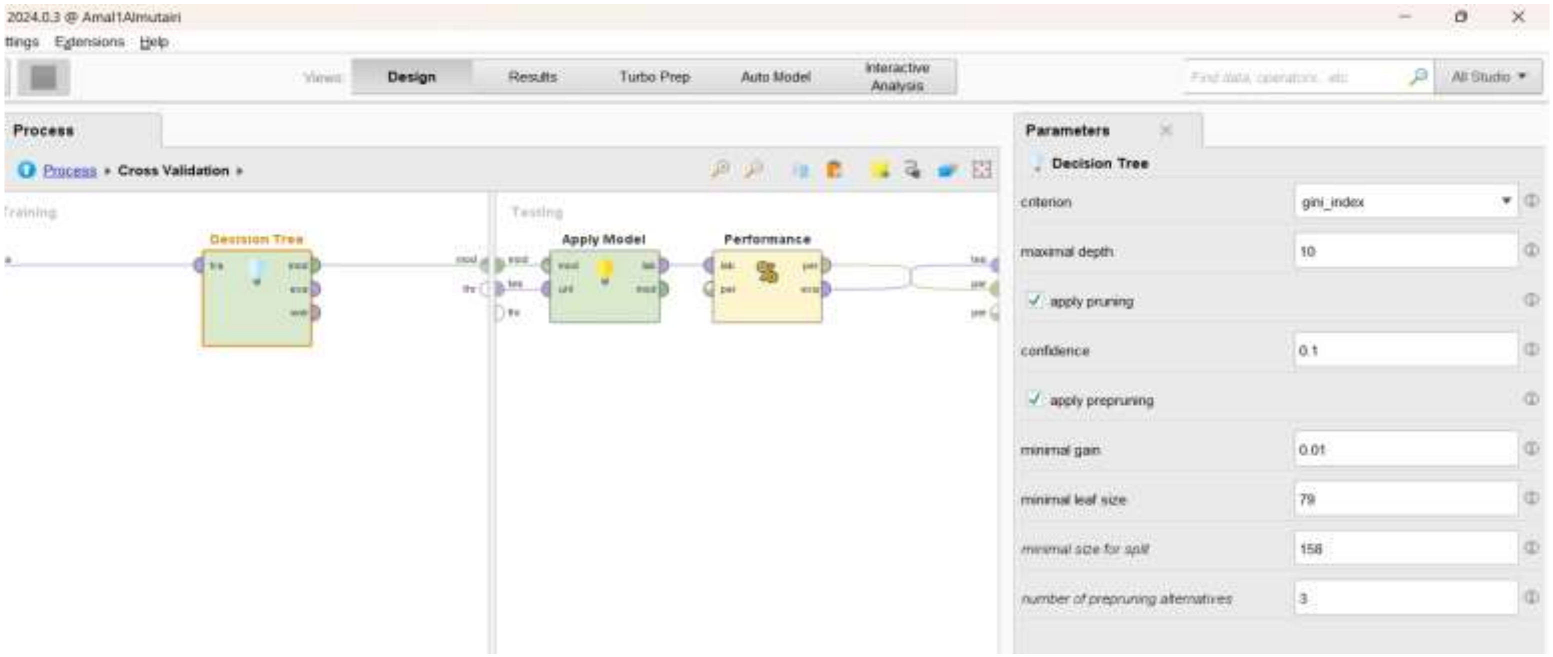
The Confusion Matrix Table:



accuracy: 91.98% +/- 2.64% (micro average: 91.98%)

	true cluster_1	true cluster_0	true cluster_2	class precision
pred. cluster_1	186	16	11	87.32%
pred. cluster_0	32	855	13	95.00%
pred. cluster_2	45	1	312	87.15%
class recall	70.72%	98.05%	92.86%	

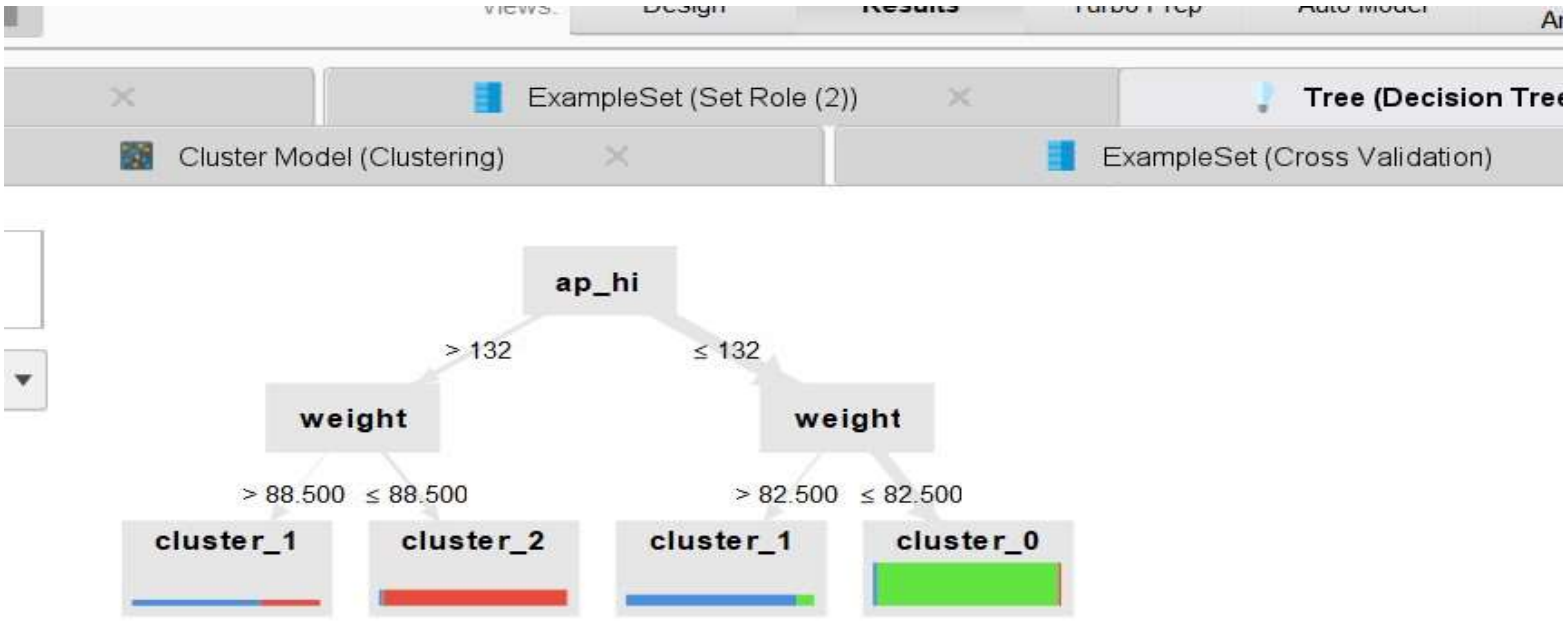
The third Decision Tree is Gini:


2024.0.3 @ Amal1Almutairi
ttings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators, etc
All Studio
Process
Process ▸ Cross Validation ▸
Training
Testing
Decision Tree
Apply Model
Performance

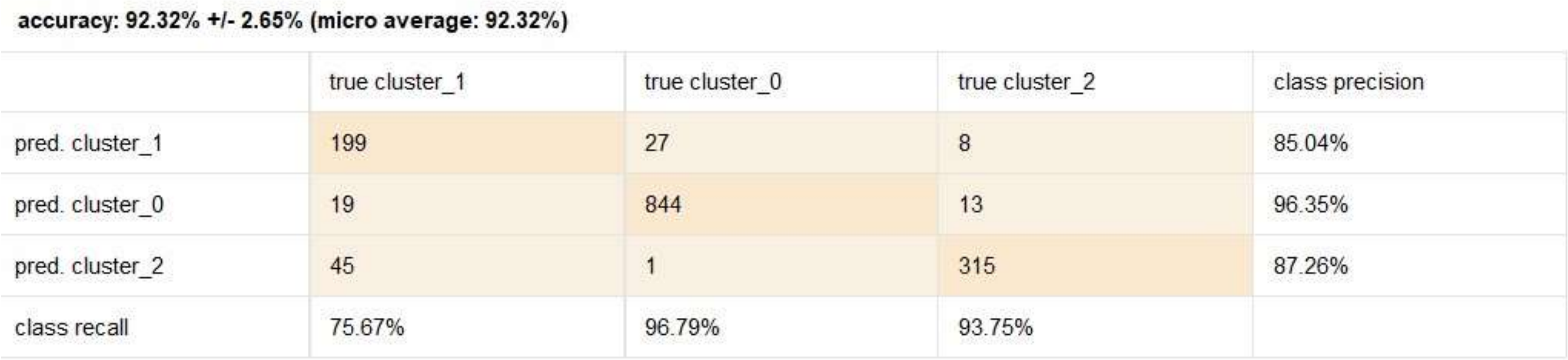
Parameters

Decision Tree

criterion	gini_index
maximal depth	10
☑ apply pruning	
confidence	0.1
☑ apply prepruning	
minimal gain	0.01
minimal leaf size	79
minimal size for split	158
number of prepruning alternatives	3



The Confusion Matrix Table:



accuracy: 92.32% +/- 2.65% (micro average: 92.32%)

	true cluster_1	true cluster_0	true cluster_2	class precision
pred. cluster_1	199	27	8	85.04%
pred. cluster_0	19	844	13	96.35%
pred. cluster_2	45	1	315	87.26%
class recall	75.67%	96.79%	93.75%	

combination function (overall score)=0.60*Accuracy+0.40*Simplicity

DT Label/Description	Performance Measures				Overall Score
	Accuracy		Simplicity		
	% Error	Score	No of Leaves	Score	
1-Information Gain	1-0.9293=0.0707	92.93%	4	1	0.60*0.9293+0.40*1=95758
2-Gain Ratio	1-0.9198=0.0802	91.98%	4	1	0.60*0.9198+0.40*1=0.95188
3-Gini	1-0.9232=0.0768	92.32%	4	1	0.60*0.9232+0.40*1=0.95392

We assumed in Task 2 that the ideal DT has 3–7 leaves (IdealBot = 3, IdealTop = 7). DTs with ≤2 or ≥14 leaves are considered unacceptable. Since all trees have 4 leaves, which falls within the ideal range, the Simplicity Score is 1 for all.

The formula used is:

ScoreSimplicity(NoOfLeaves) = 1 if IdealBot ≤ NoOfLeaves ≤ IdealTop.

The Information Gain decision tree achieves the highest overall score of 0.95758, making it the most suitable option as it offers the best balance between accuracy and simplicity.

5. Based on the results of the best DT, explain the produced clusters:

Importance of Variables :

Variable	Cluster_0	Cluster_1	Cluster_2	Explanation
ap_hi	**≤ 132**	**> 132**	**> 132**	The patients in **Cluster_0** have a systolic blood pressure (**ap_hi**) of ≤ 132, while those in **Cluster_1** and **Cluster_2** have higher blood pressure (> 132).
weight	**≤ 81.500**	**> 81.500 & ≤ 86.500**	**> 86.500**	**Cluster_0** consists of patients with lower weight (≤ 81.500), while Cluster_1 contains patients with moderate weight (between 81.500 and 86.500), and **Cluster_2** has patients with a weight greater than 86.500
Narrative on Comparison: This table explains how patients were grouped based on their ap_hi (systolic blood pressure) and weight. Cluster_0 includes patients with both lower blood pressure and weight, while Cluster_1 covers those with moderate blood pressure and moderate weight. Cluster_2 represents patients with high blood pressure and higher weight.				