DATA MINING

LAB MID

SUBMITTED BY:

NAME: MAHNOOR ASIF

REG NO: FA22-BCS-150

SEC: BCS-5D

DATE: 11 NOV, 2024

Table of Content

CHAPTER 1:

PURPOSE OF PROJ AND LAB GOALS:

Purpose of Project:

The project examines the effects of social media platforms on mental health and overall well-being. Social media enables users to share personal information, interact with others, and engage with content. However, excessive use can contribute to mental health issues such as anxiety, depression, and social comparison, particularly among younger individuals and regions with high internet access. Through analyzing survey data, the study seeks to identify vulnerable groups, including young females, students, and professionals, and to understand the risk factors associated with content consumption, online time, and emotional responses.

Lab Goals:

- The goals for this midterm lab is to use rapid miner for complete model/process creation as well as evaluation focusing on type of data in dataset and what model to apply on that dataset.
- The goals is to preprocess , analyze and visualize dataset to find insights which are helpful to make decisions.
- To clean dataset by handling missing values, duplicates, outliers, inconsistent data etc to help model gain better results or accuracy
- After preprocessing, the goal is to train two supervised learning models to find insights from dataset .
- The lab involves selecting appropriate models, setting model parameters, training them, evaluating their performance and visualizing results.
- The main focus is to find out factors related to social media which have strong impact on mental health

Dataset description:

Attribute	Type of Data	Data Type	Description
Age	Categorical (Ordinal)	Ordinal	Age group of the respondent (11-18, 18-24, 24-34, 34-54)
Gender	Categorical (Nominal)	Nominal	Gender of the respondent (Male, Female)
Province	Categorical (Nominal)	Nominal	Province of the respondent (Punjab, Sindh, KPK, Azad Kashmir, Balochistan, Overseas)
Occupation	Categorical (Nominal)	Nominal	Occupation of the respondent (Student, Employed, Unemployed, Retired, Housewife)
Content Type Engagement	Categorical (Nominal)	Nominal	Type of content respondent engages with (Entertainment, News, Educational, Professional)
Social Media Usage (hours/day)	Categorical (Ordinal)	Ordinal	Hours spent on social media daily (1-2 hours, 3-4 hours, etc.)
Social Media Platforms Used	Categorical (Nominal)	Nominal	Platforms used by the respondent (Instagram, Twitter, TikTok, Snapchat, YouTube, Facebook)
Emotional Response	Categorical (Nominal)	Nominal	Whether the respondent experiences emotional responses (YES, NO)
Impact on Sleep	Categorical (Ordinal)	Ordinal	Impact of social media on sleep quality (Worsened, No impact, Improved)
Fear of Missing Out (FOMO)	Categorical (Ordinal)	Ordinal	Frequency of experiencing FOMO from social media (Frequently, Occasionally, Rarely, Never)
Cyberbullying	Categorical (Nominal)	Nominal	Whether the respondent experienced cyberbullying (YES, NO, Prefer not to say)
Social Comparison	Categorical (Ordinal)	Ordinal	Frequency of comparing life to others on social media (Frequently, Occasionally, Rarely, Never)
Addiction	Ordinal (Scale)	Ordinal	Rating of social media addiction (1 to 5)
No of platforms	Categorical (Ordinal)	Ordinal	Number of platforms actively used (1, 2, 3, 4, 5, More than 5)

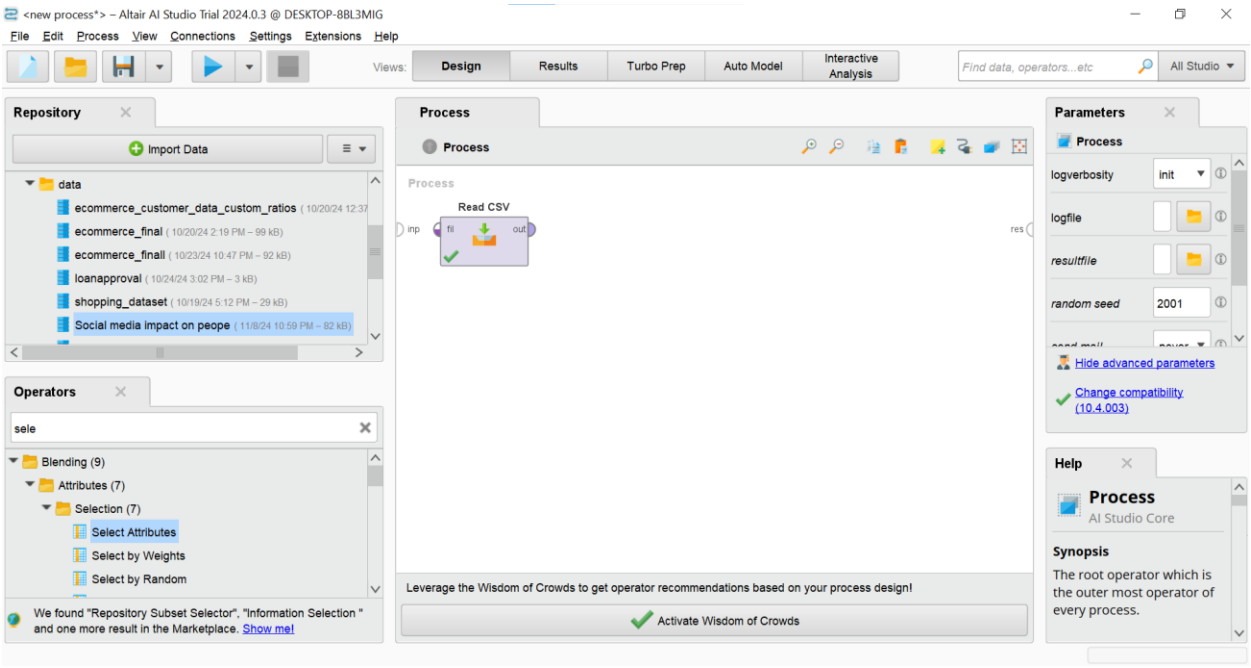
Attribute	Type of Data	Data Type	Description
Agreement on Social Media Enhancing Social Life	Ordinal (Scale)	Ordinal	Agreement with statement on social media enhancing social life (1 to 5)
Hours of Sleep Lost Due to Social Media Usage	Numeric (Continuous)	Continuous	Number of hours of sleep lost per week due to social media usage

Total records: 477

CHAPTER 2:

PREPROCESSING :

Reading dataset through csv:



Open in Turbo Prep Auto Model Interactive Analysis Filter (677 / 677 examples): all

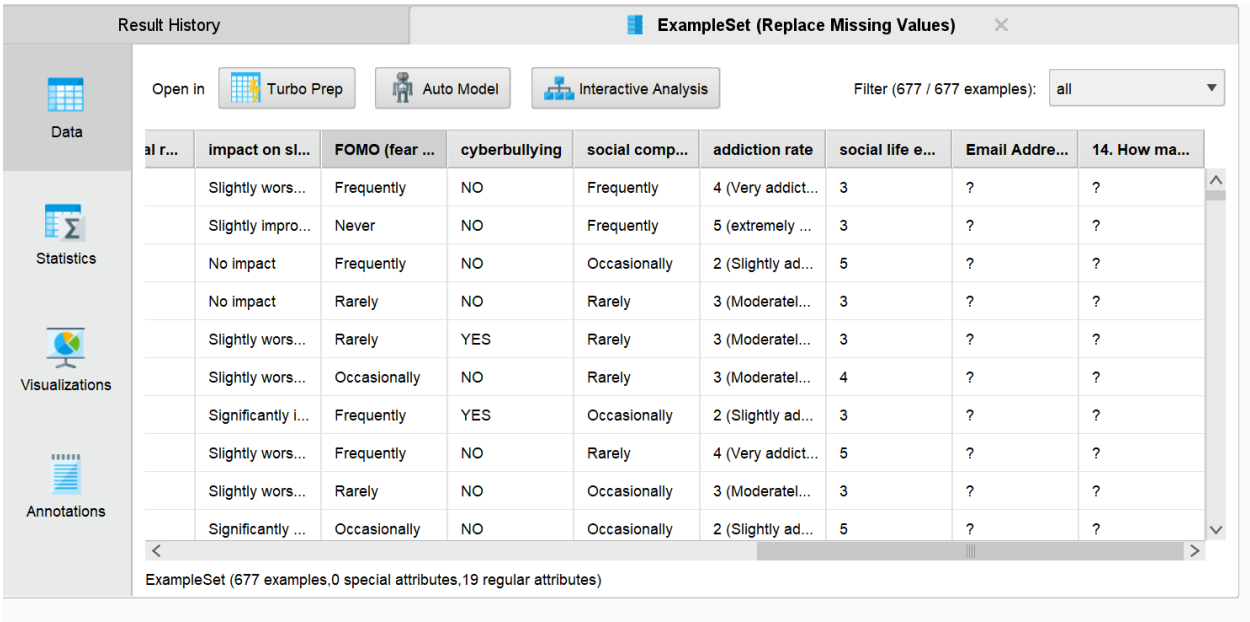
Row No.	Timestamp	Age	Gender	province	Occupation	type of cont...	hours	platfo
1	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Instag
2	?	18-24	Female	Rawalpindi	Student	Entertainmen...	7-9 hour	TikTo
3	?	18-24	Male	Taxila	Student	Entertainmen...	1-2 hour	Instag
4	?	18-24	Female	Islamabad	Student	News and cu...	1-2 hour	Instag
5	?	18-24	Female	Taxila	Student	Entertainmen...	3-4 hour	Youtu
6	?	18-24	Male	Rawalpindi	Student	Educational ...	5-6 hour	Instag
7	?	18-24	Male	punjab	Student	Entertainmen...	1-2 hour	Instag
8	?	18-24	Male	Rawalpindi	Student	Entertainmen...	7-9 hour	Instag
9	?	18-24	Female	Wah	Student	Entertainmen...	3-4 hour	TikTo
10	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Youtu

ExampleSet (677 examples,0 special attributes,19 regular attributes)

Open in Turbo Prep Auto Model Interactive Analysis Filter (677 / 677 examples): all

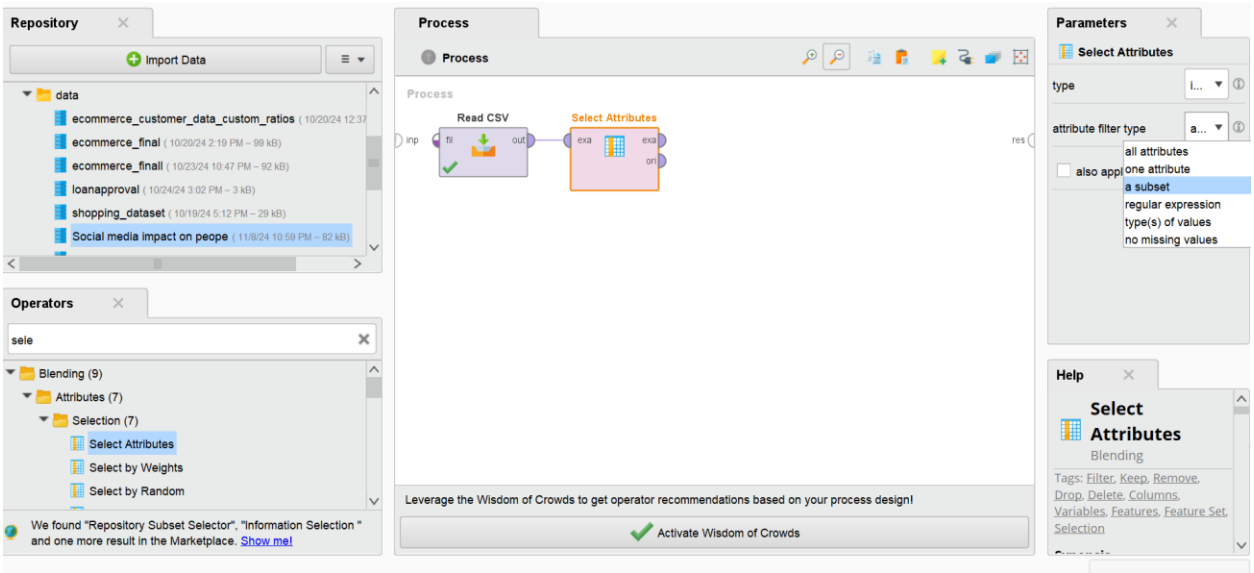
emotional r...	impact on sl...	FOMO (fear ...	cyberbullying	social comp...	addiction rate	no of platfor...	social life e...	sleep
YES	Slightly wors...	Frequently	NO	Frequently	4 (Very addict...	2	3	1.5 ho
YES	Slightly impro...	Never	NO	Frequently	5 (extremely ...	4	3	6
NO	No impact	Frequently	NO	Occasionally	2 (Slightly ad...	4	5	24 ho
YES	No impact	Rarely	NO	Rarely	3 (Moderatel...	2	3	0
YES	Slightly wors...	Rarely	YES	Rarely	3 (Moderatel...	2	3	21
NO	Slightly wors...	Occasionally	NO	Rarely	3 (Moderatel...	more than 5	4	5
NO	Significantly i...	Frequently	YES	Occasionally	2 (Slightly ad...	2	3	5
YES	Slightly wors...	Frequently	NO	Rarely	4 (Very addict...	3	5	26
NO	Slightly wors...	Rarely	NO	Occasionally	3 (Moderatel...	3	3	2
NO	Significantly ...	Occasionally	NO	Occasionally	2 (Slightly ad...	2	5	8

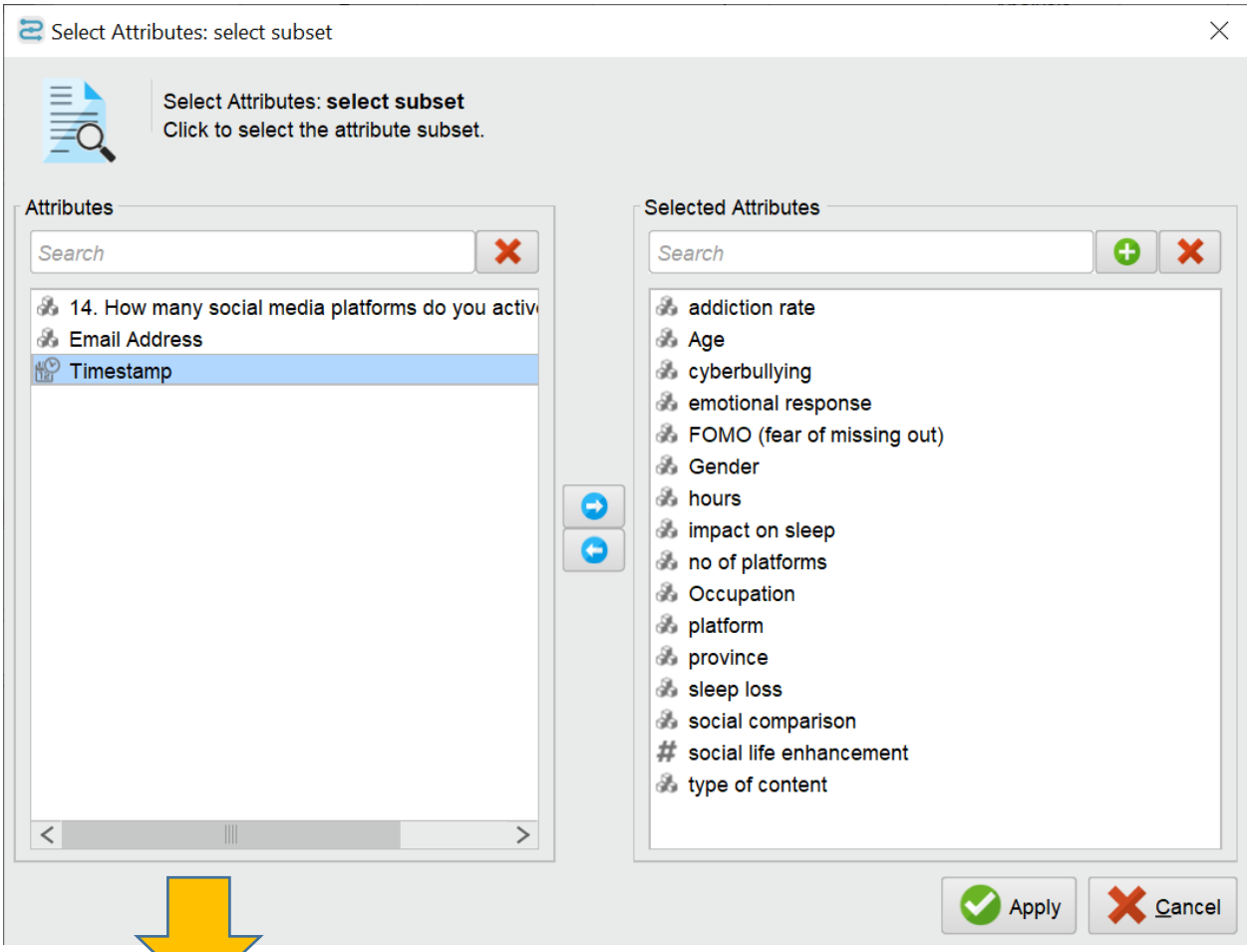
ExampleSet (677 examples,0 special attributes,19 regular attributes)



Operator 1: select attributes:

- This select operator will select necessary attributes which are helpful in analysis and remove those which are irrelevant or mostly empty.
- Here the attributes I didn't selected are shown below as they are empty and iirelevent according to my dataset.
- As this dataset is collected through survey form , that's why it has those irrelevant columns which should be removed before applying any model.



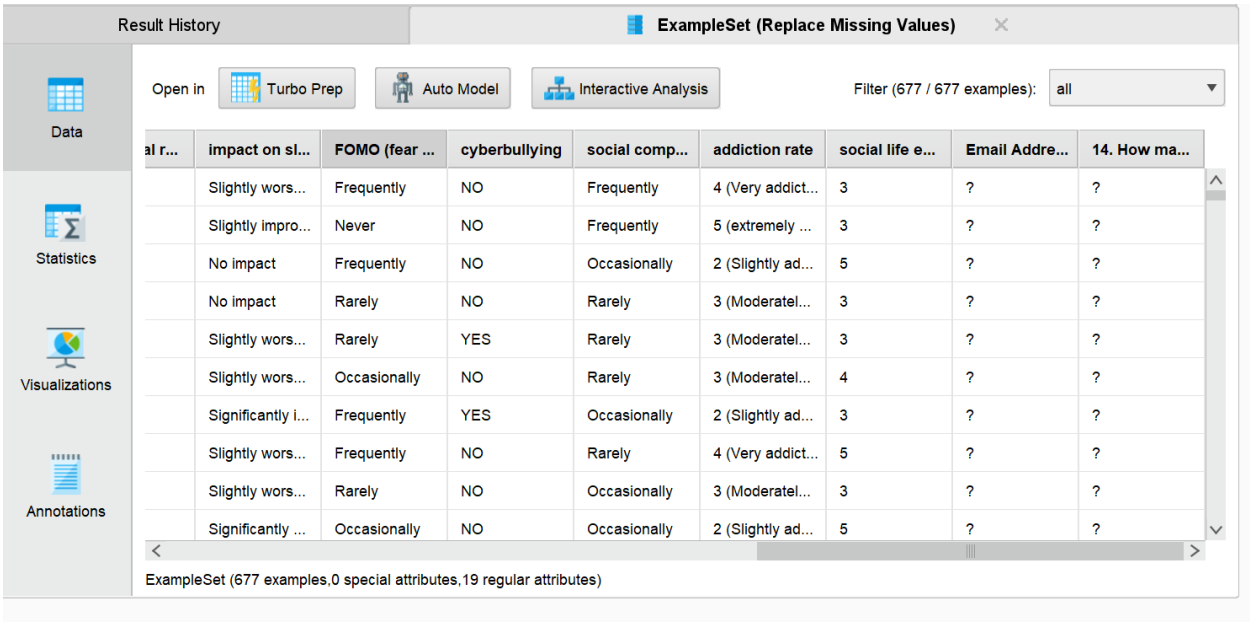


Open in Turbo Prep | Auto Model | Interactive Analysis

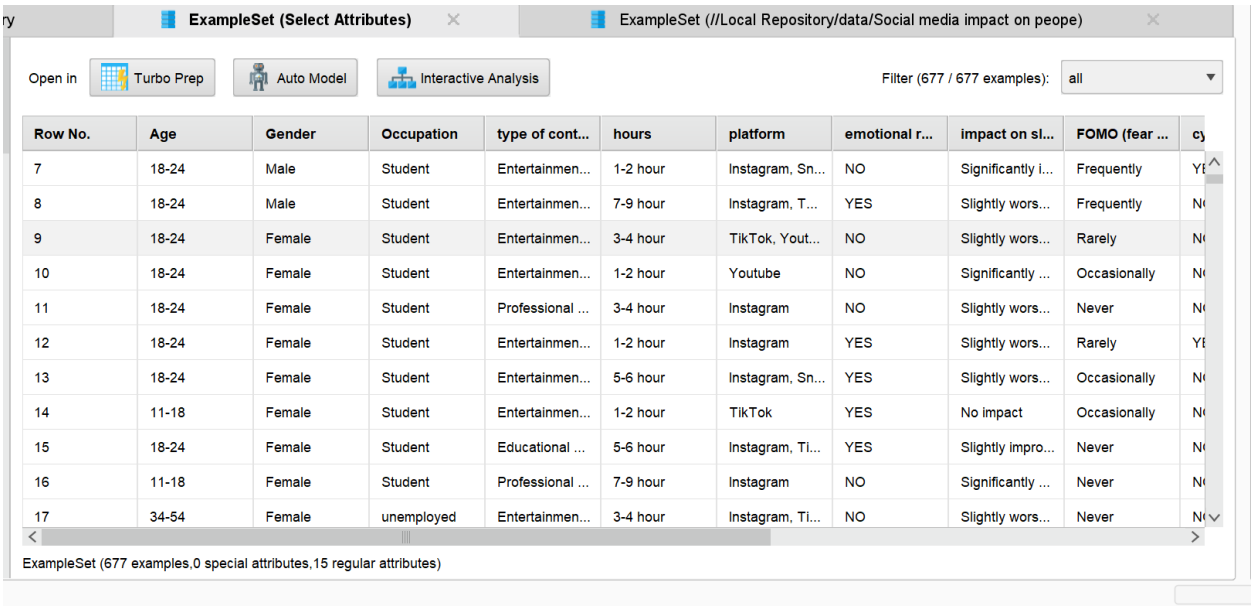
Filter (677 / 677 examples): all

Row No.	Timestamp	Age	Gender	province	Occupation	type of cont...	hours	platfc
1	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Instag
2	?	18-24	Female	Rawalpindi	Student	Entertainmen...	7-9 hour	TikTo
3	?	18-24	Male	Taxila	Student	Entertainmen...	1-2 hour	Instag
4	?	18-24	Female	Islamabad	Student	News and cu...	1-2 hour	Instag
5	?	18-24	Female	Taxila	Student	Entertainmen...	3-4 hour	Youtu
6	?	18-24	Male	Rawalpindi	Student	Educational ...	5-6 hour	Instag
7	?	18-24	Male	punjab	Student	Entertainmen...	1-2 hour	Instag
8	?	18-24	Male	Rawalpindi	Student	Entertainmen...	7-9 hour	Instag
9	?	18-24	Female	Wah	Student	Entertainmen...	3-4 hour	TikTo
10	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Youtu

ExampleSet (677 examples,0 special attributes,19 regular attributes)



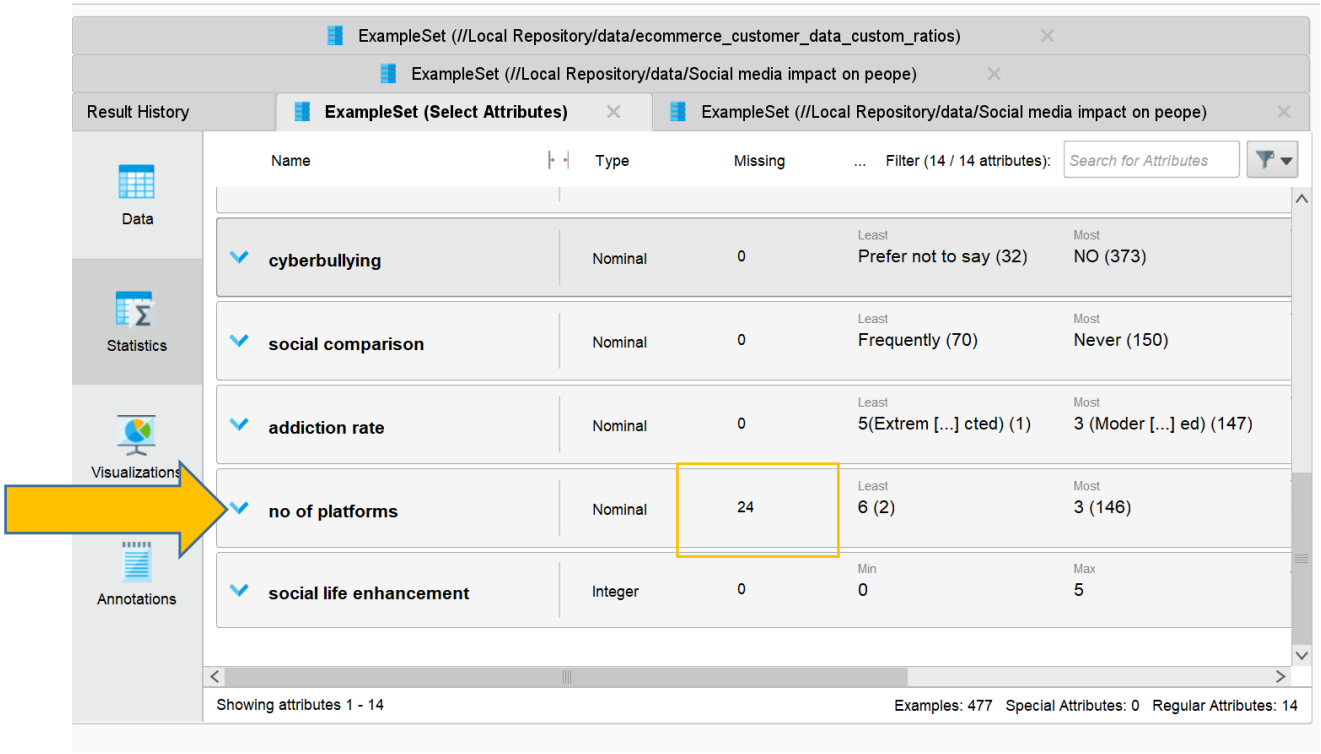
DATASET WHEN THEY ARE REMOVED:



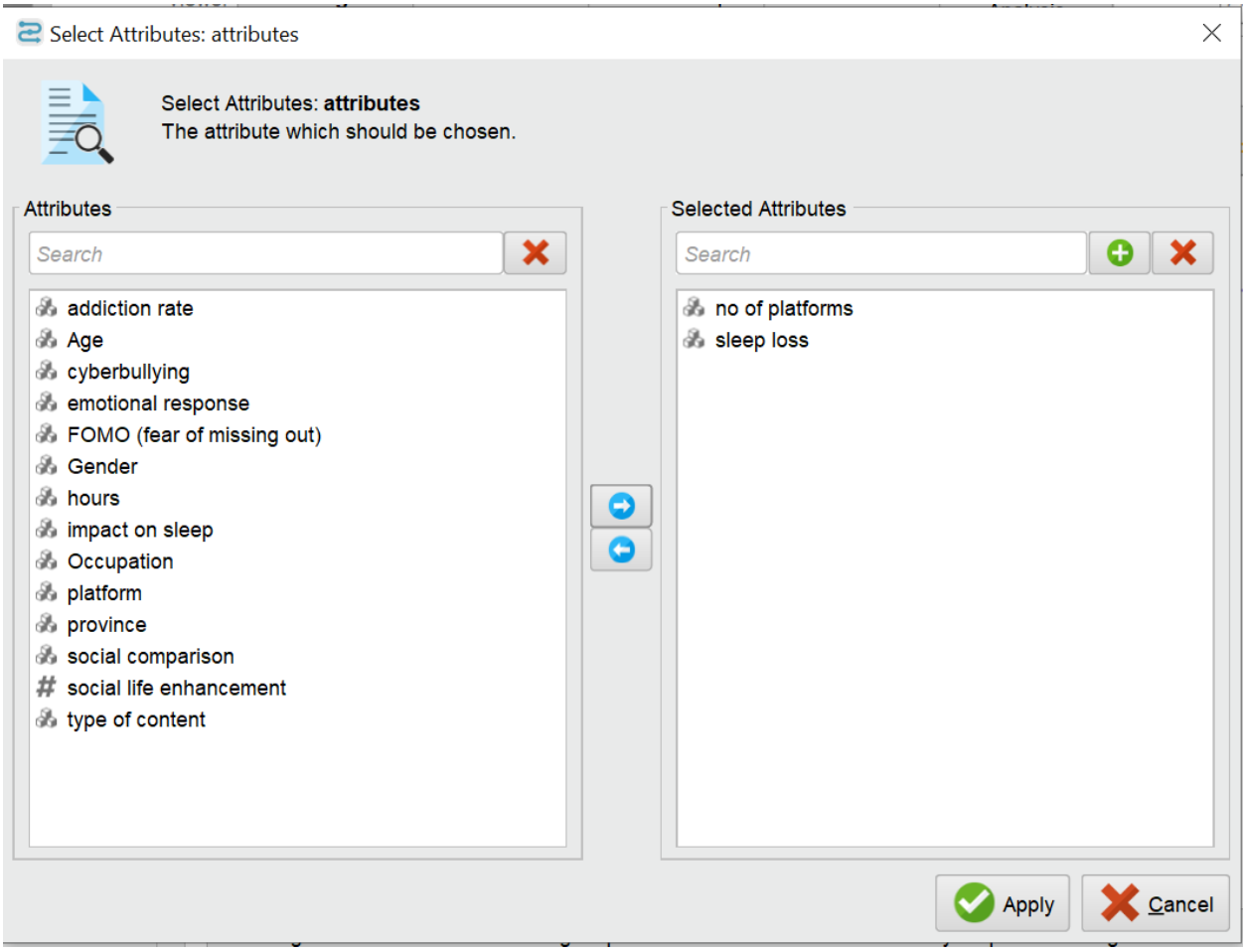
Operator 2: replace missing values

- Replacing missing rows with average or maximum or null value is more crutial because if whole row with missing values is removed , it can cause loss of important information.
- Here thr attribute platform has missing values so the operator replaced it by 3 which was most occurred in the column also known as mode.

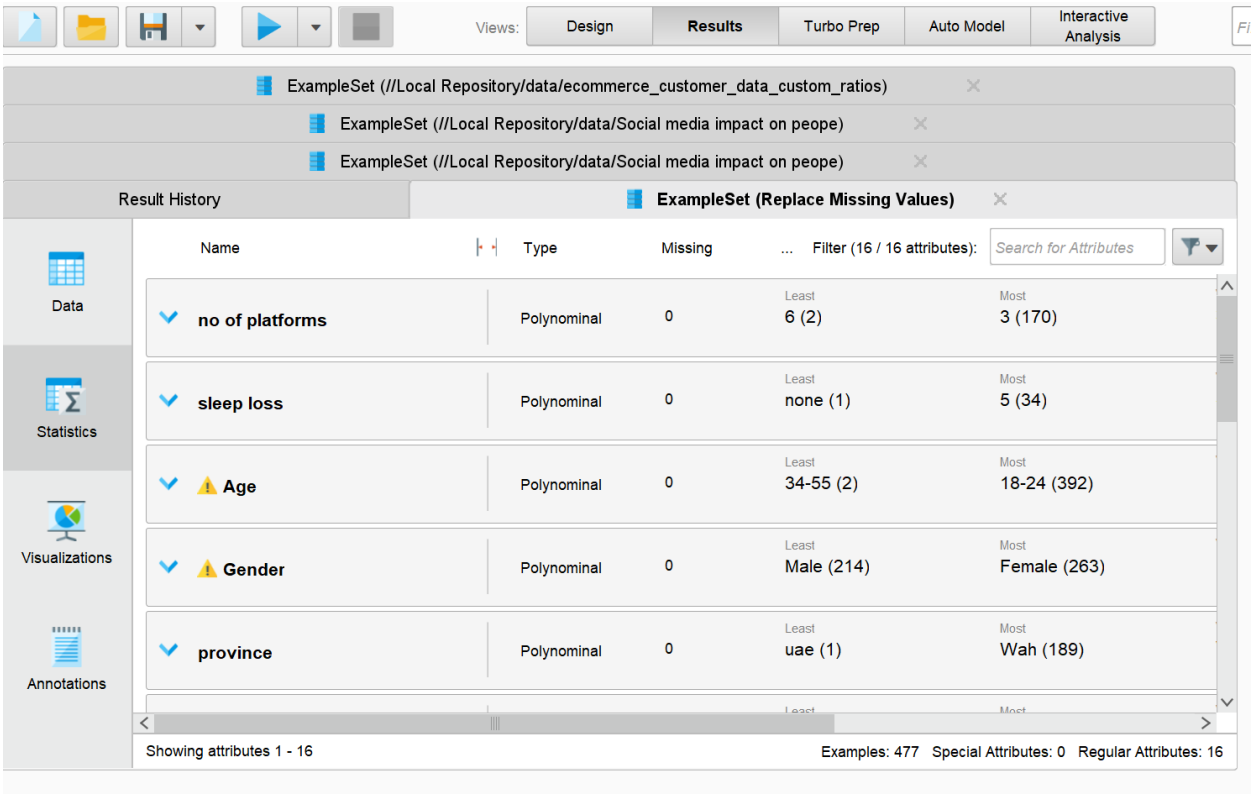
- In platform attribute, there are 24 cells which have missing values which are shown in statistics column.



- Sleep loss also had 1 missing value so it is also replaced.
- In attribute selection filter, I filtered attributes where I want to replace missing
- Those attributes are no of platform and sleep loss.

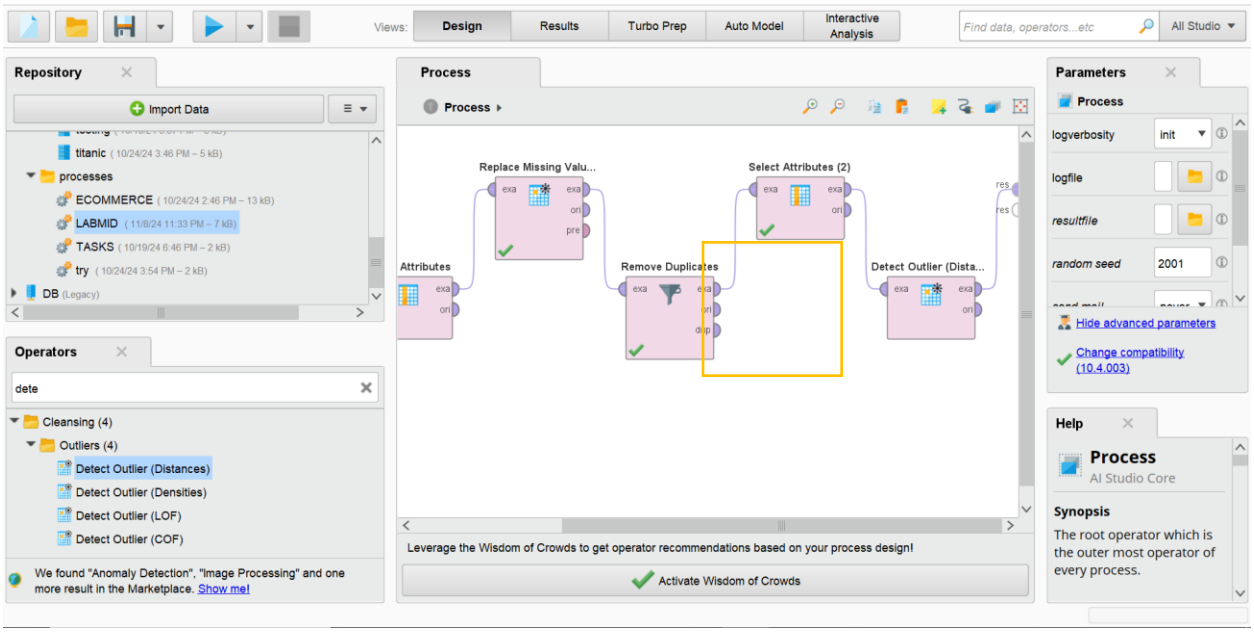


- After replacing missing values:



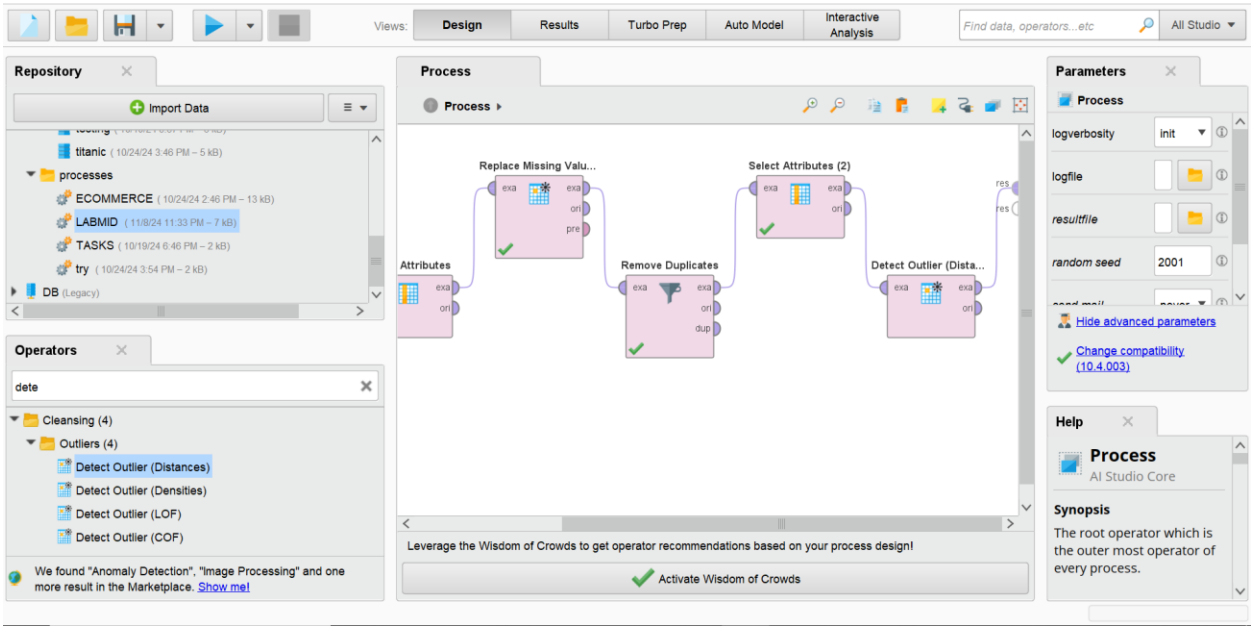
Operator 3: remove duplicates

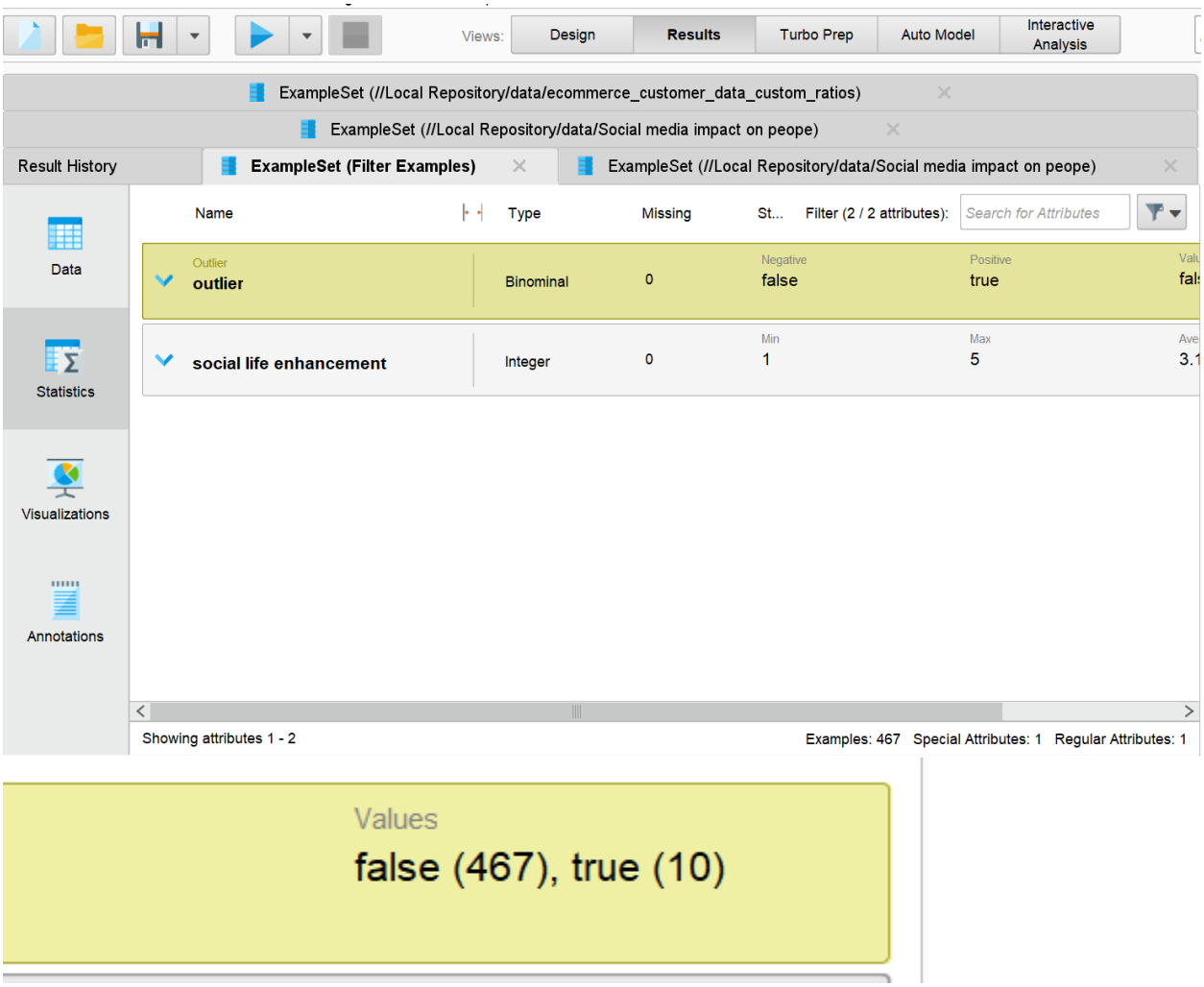
- If the same data appears multiple times, it can give you the wrong idea or lead to wrong conclusions.
- Without duplicates, your data is cleaner and easier to understand, making it easier for analysis.
- Removing duplicates are key for better accuracy . if not removed, it can degrade performance.
- Here the platform attribute has 24 missing values so they are replaced by max value in the column.



Operator 4: detect outlier:

- outliers can affect the accuracy of models, especially regression or classification models.
- Outliers ususally work on numerical columns .
- Here I had only one numerical column. (social life enhancement).
- So I selected this using select attributes and then connect detect outliers.

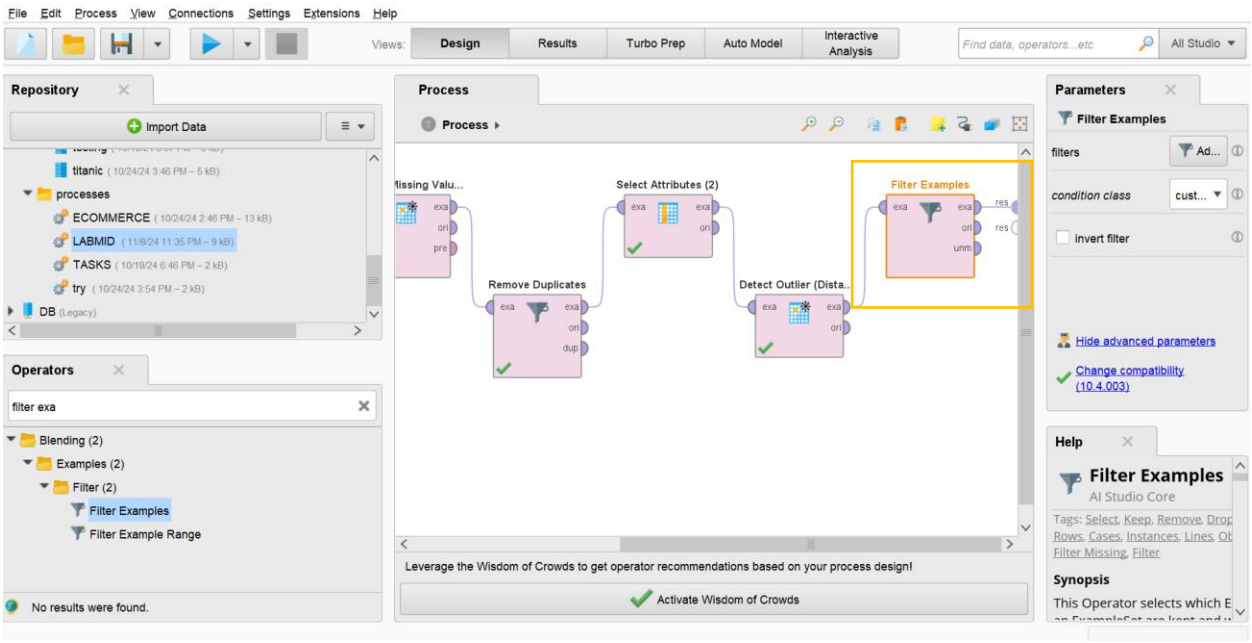


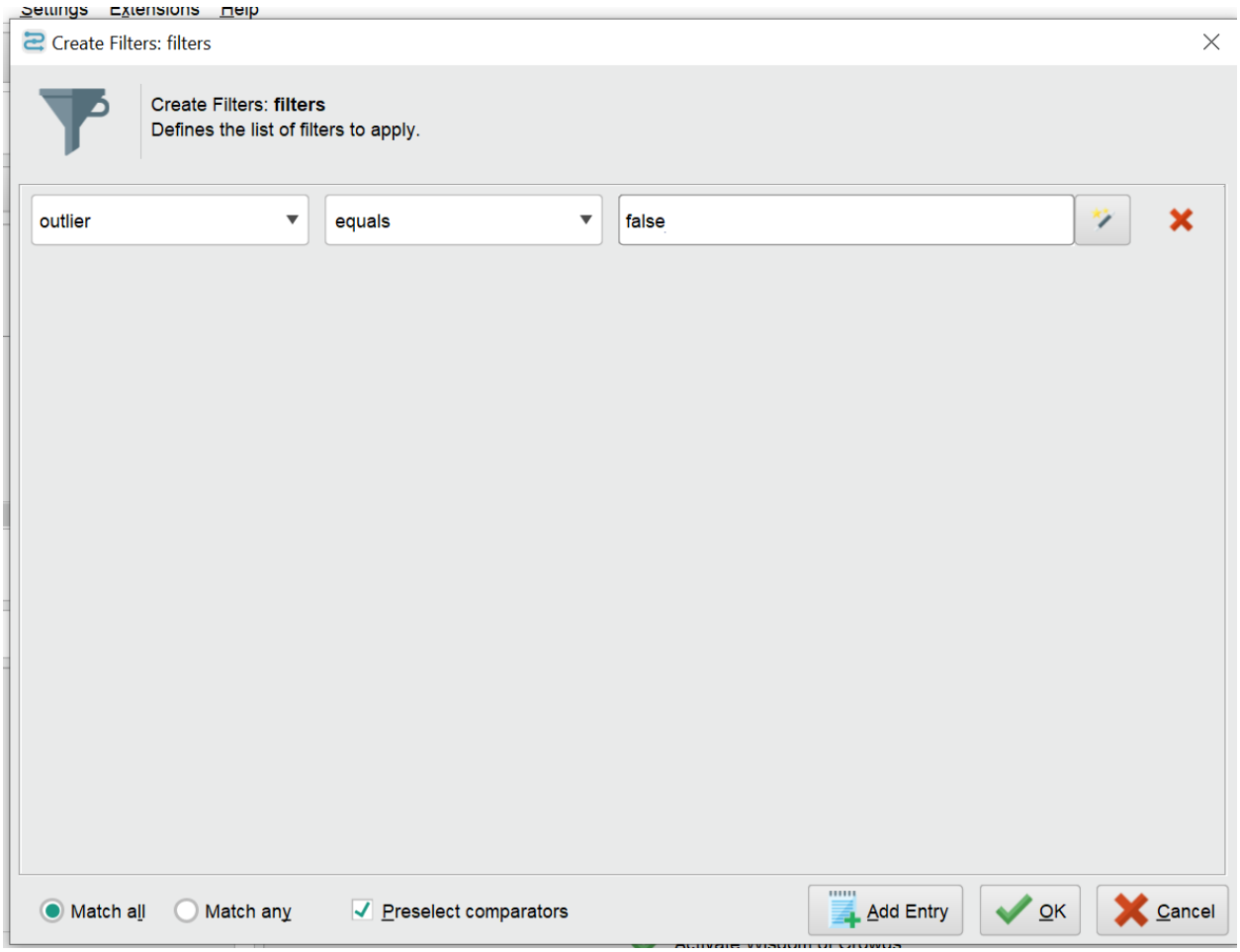


- It shows that it detected 10 values which are outliers so I will use filter example to remove those rows where outlier is true

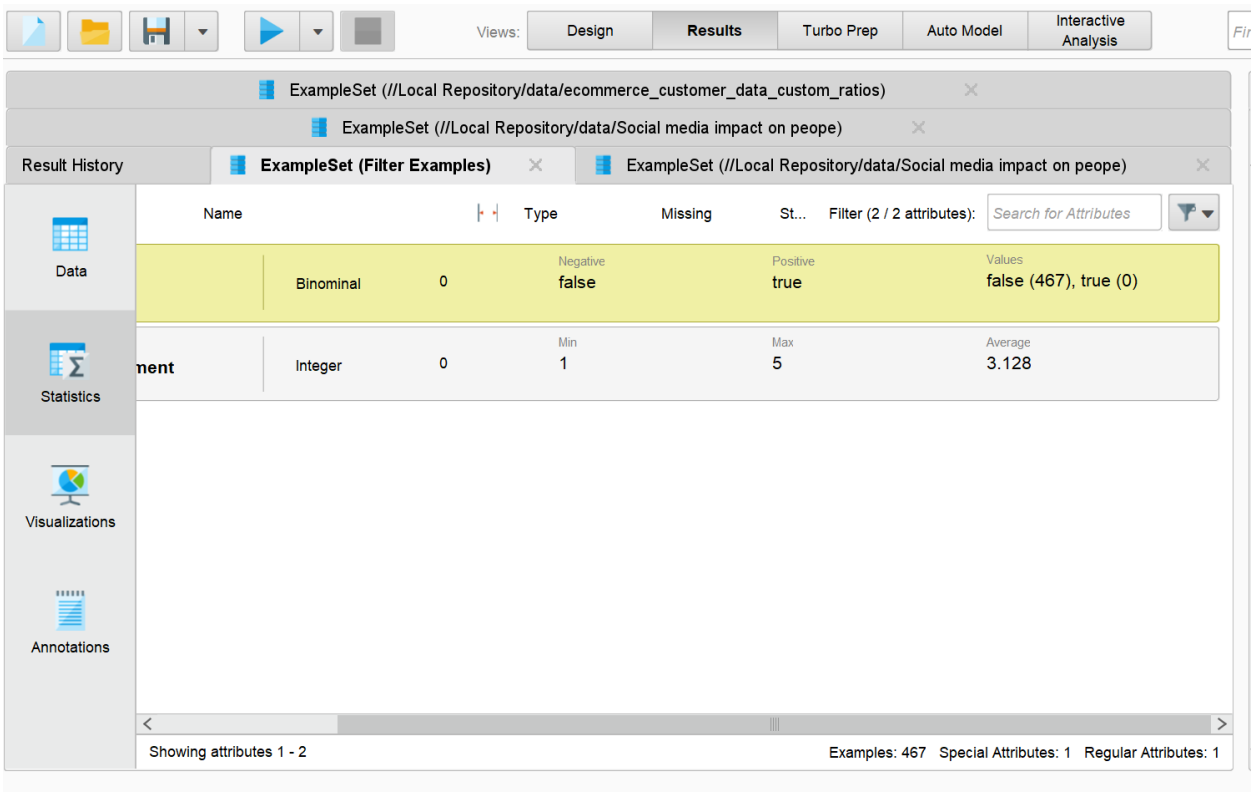
Operator 5: FILTER EXAMPLE:

- The purpose of this operator here is to remove those rows which has outlier values (outlier=true)
- It will select those rows only where outlier=false (means no outlier value)





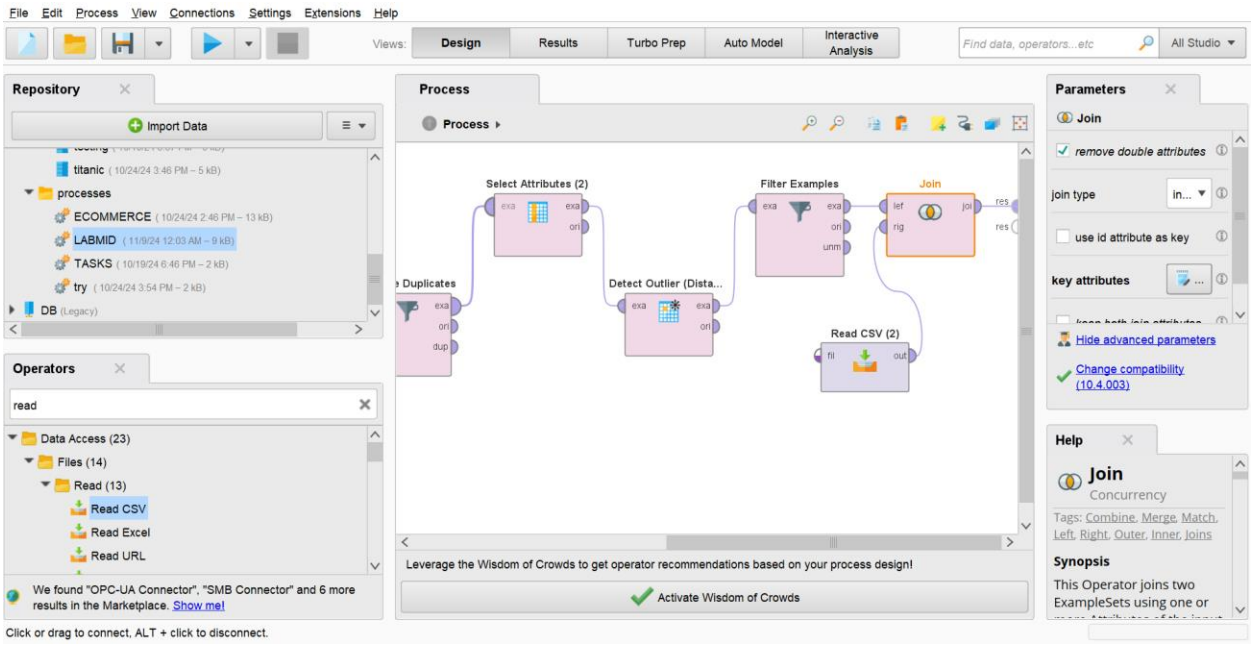
- Here in statistics ,we can see that true(0) means all outliers have been removed.



Operator 6: JOIN and GENERATE ATTRIBUTES :

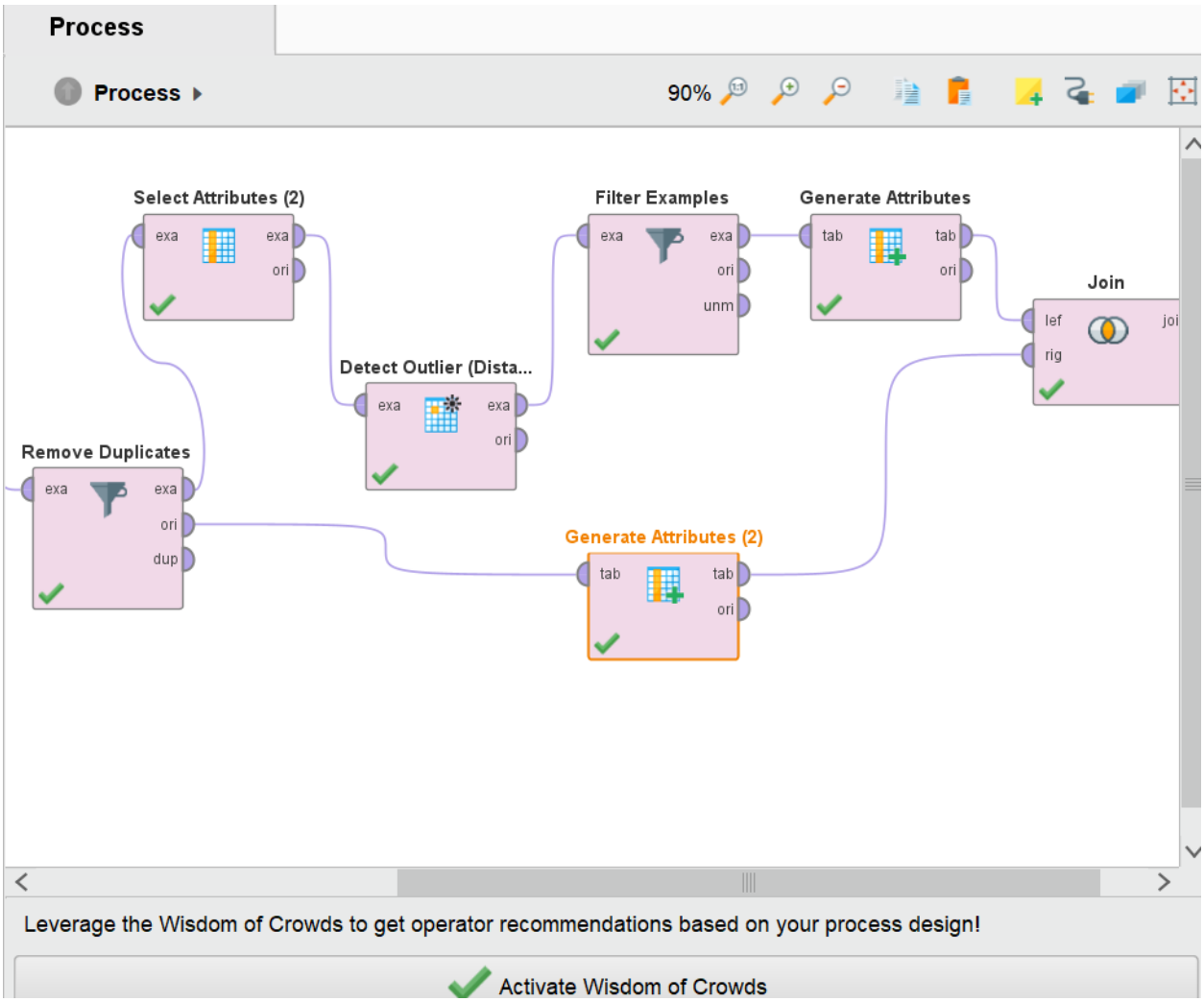
ISSUE WITH JOIN:

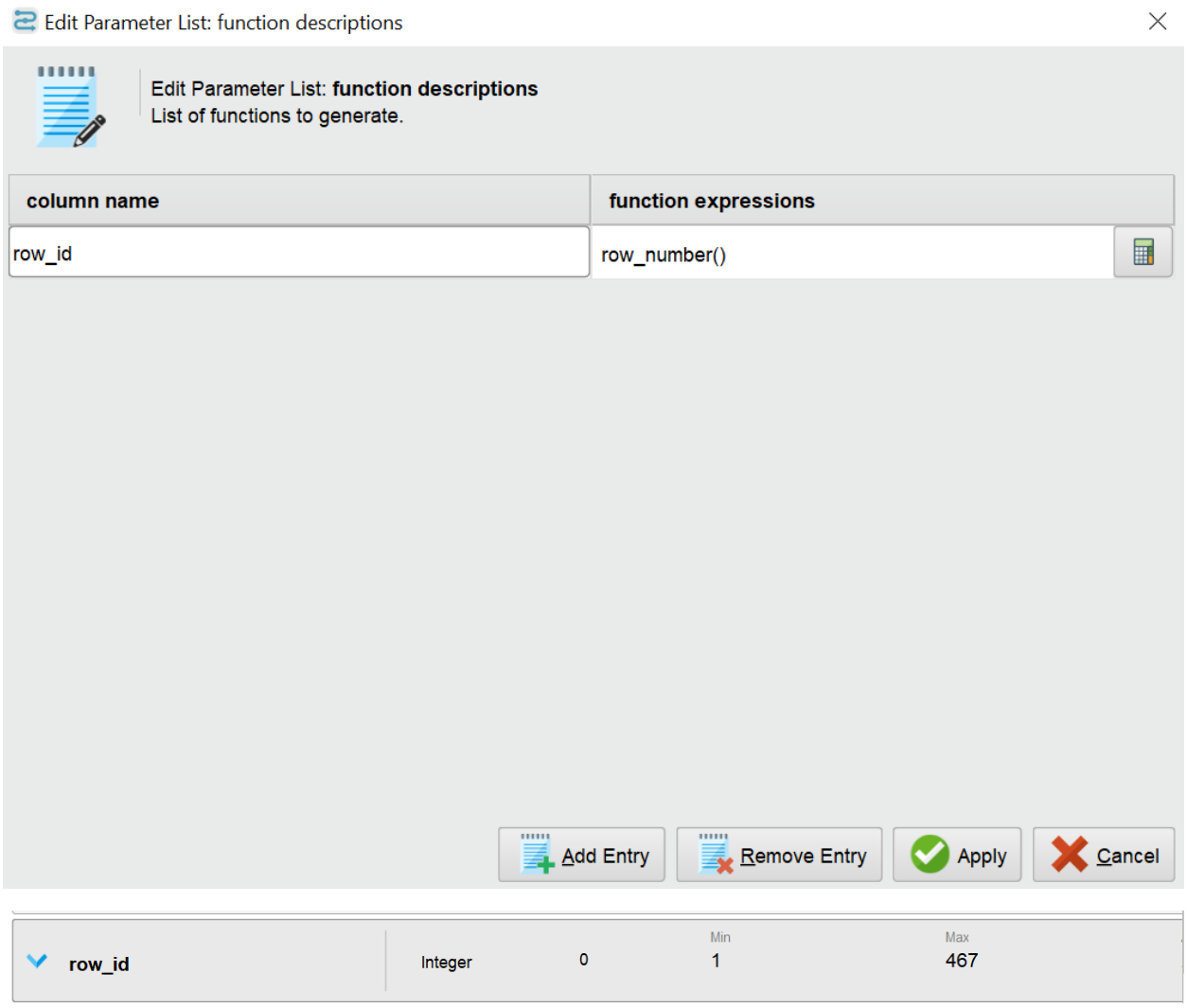
- **The issue with join operator here is that I had only 1 numeric column when I tried to combine the dataset(with outliers removed) to original dataset, it increases number of records.**
- **So that's why I generated row attribute with number of rows and on basis of that I joined both**
- **I did that bcz it has condition that it will join based on some common attribute so that's why .**



Correct join logic:

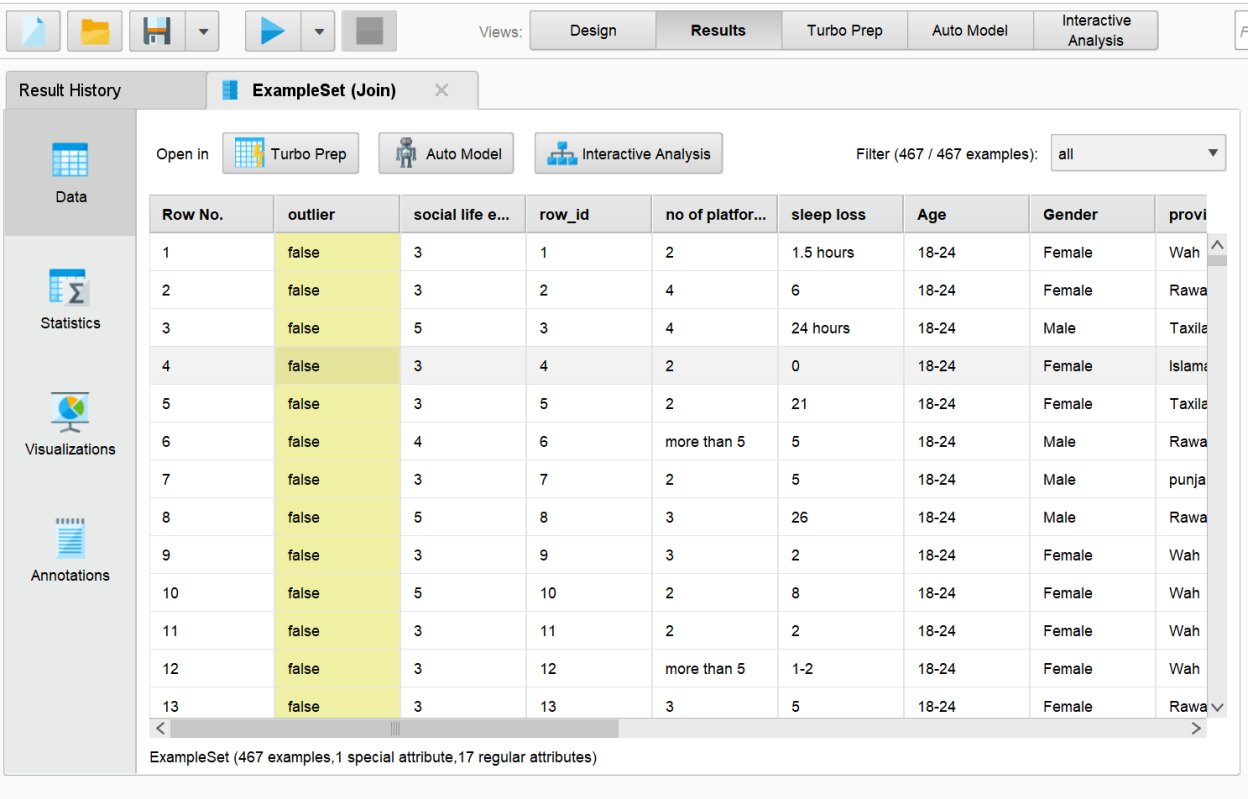
- The row number attribute generated in both datasets is common and will effectively join both datasets. Without multiplying number of rows.
- So it will join both datasets based on common attributes.


Process
Process
90%
Select Attributes (2)
exa
exa
ori
Filter Examples
exa
exa
ori
unm
Generate Attributes
tab
tab
ori
Join
lef
joi
rig
Detect Outlier (Dista...
exa
exa
ori
Remove Duplicates
exa
exa
ori
dup
Generate Attributes (2)
tab
tab
ori
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds



			Min	Max
row_id	Integer	0	1	467

FINAL:

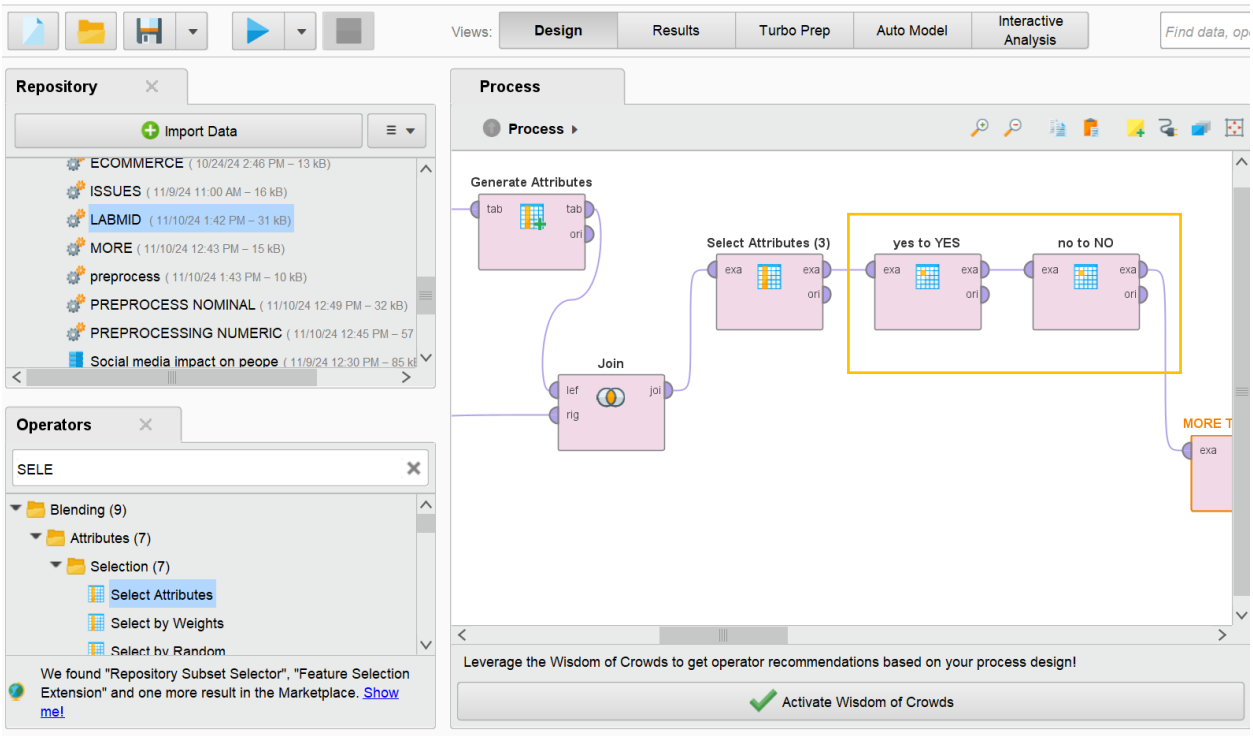


Row No.	outlier	social life e...	row_id	no of platfor...	sleep loss	Age	Gender	provi
1	false	3	1	2	1.5 hours	18-24	Female	Wah
2	false	3	2	4	6	18-24	Female	Rawa
3	false	5	3	4	24 hours	18-24	Male	Taxila
4	false	3	4	2	0	18-24	Female	Islama
5	false	3	5	2	21	18-24	Female	Taxila
6	false	4	6	more than 5	5	18-24	Male	Rawa
7	false	3	7	2	5	18-24	Male	punja
8	false	5	8	3	26	18-24	Male	Rawa
9	false	3	9	3	2	18-24	Female	Wah
10	false	5	10	2	8	18-24	Female	Wah
11	false	3	11	2	2	18-24	Female	Wah
12	false	3	12	more than 5	1-2	18-24	Female	Wah
13	false	3	13	3	5	18-24	Female	Rawa

OPERATOR 7: REPLACE OPERATOR :

- The purpose of replace operator here is that in emotional response column, which is going to be our class label, yes and no were in both capital or small (YES or yes)(No and no) as shown below.

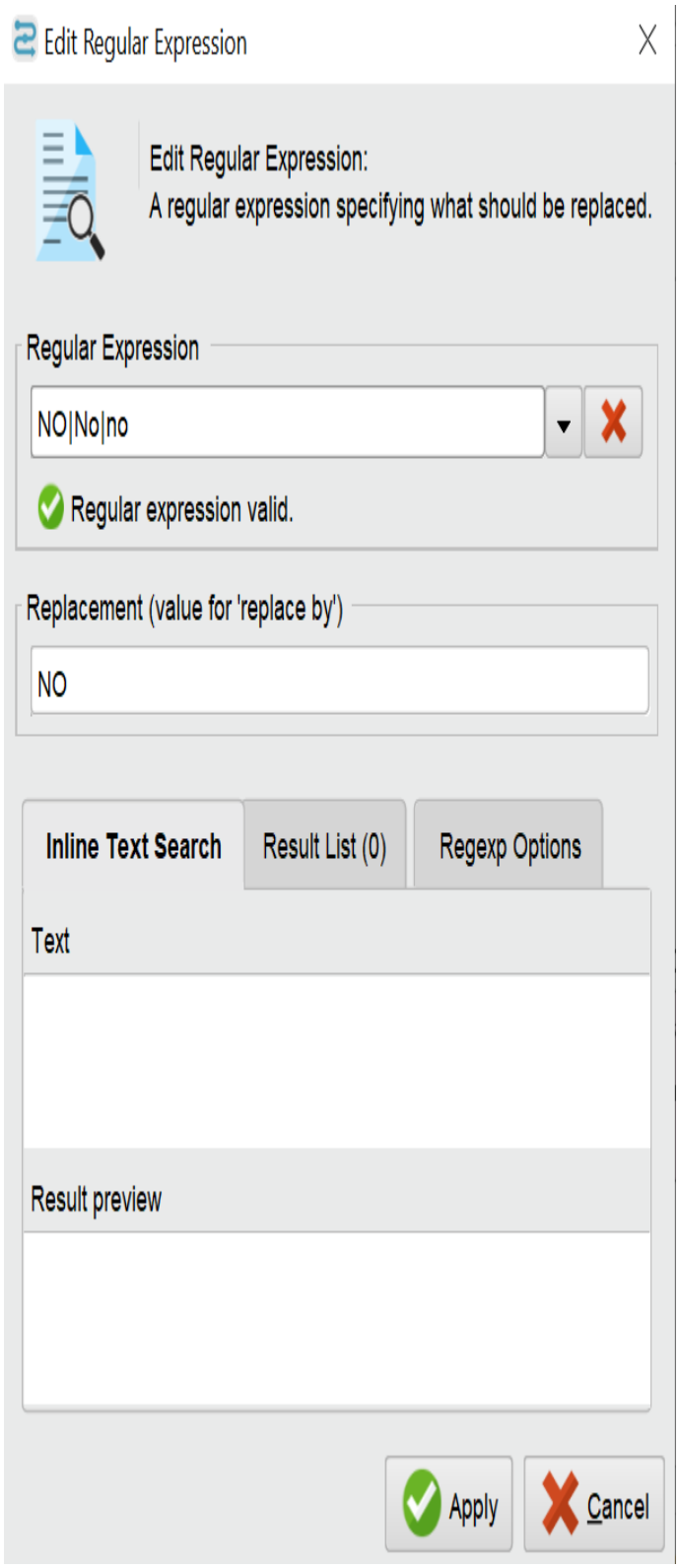
REPLACE OPERATOR (FOR EMOTIONAL RESPONSE):

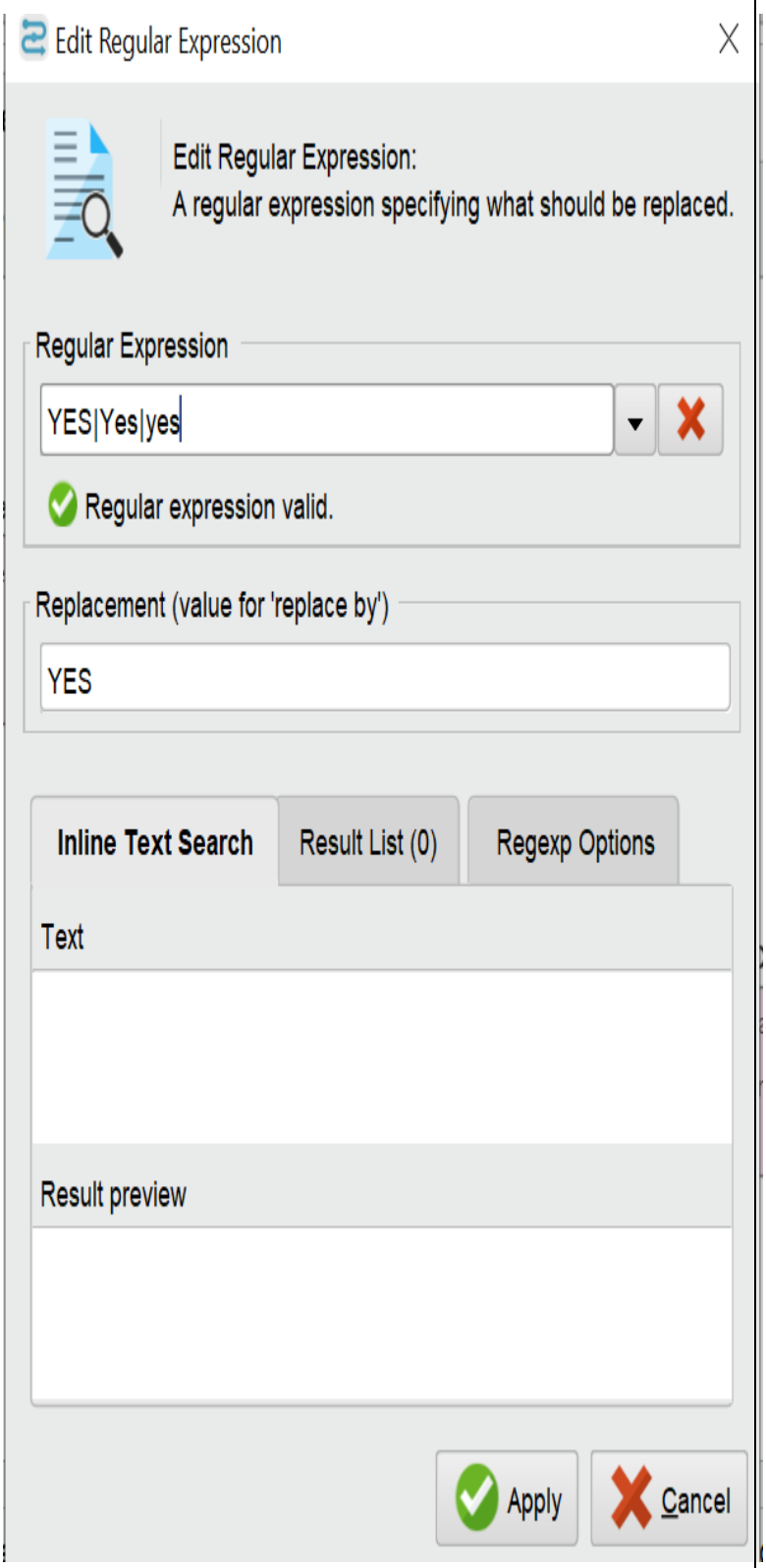


emotional response
Yes
Yes
No
Yes
Yes
No
No
No
No
No
No

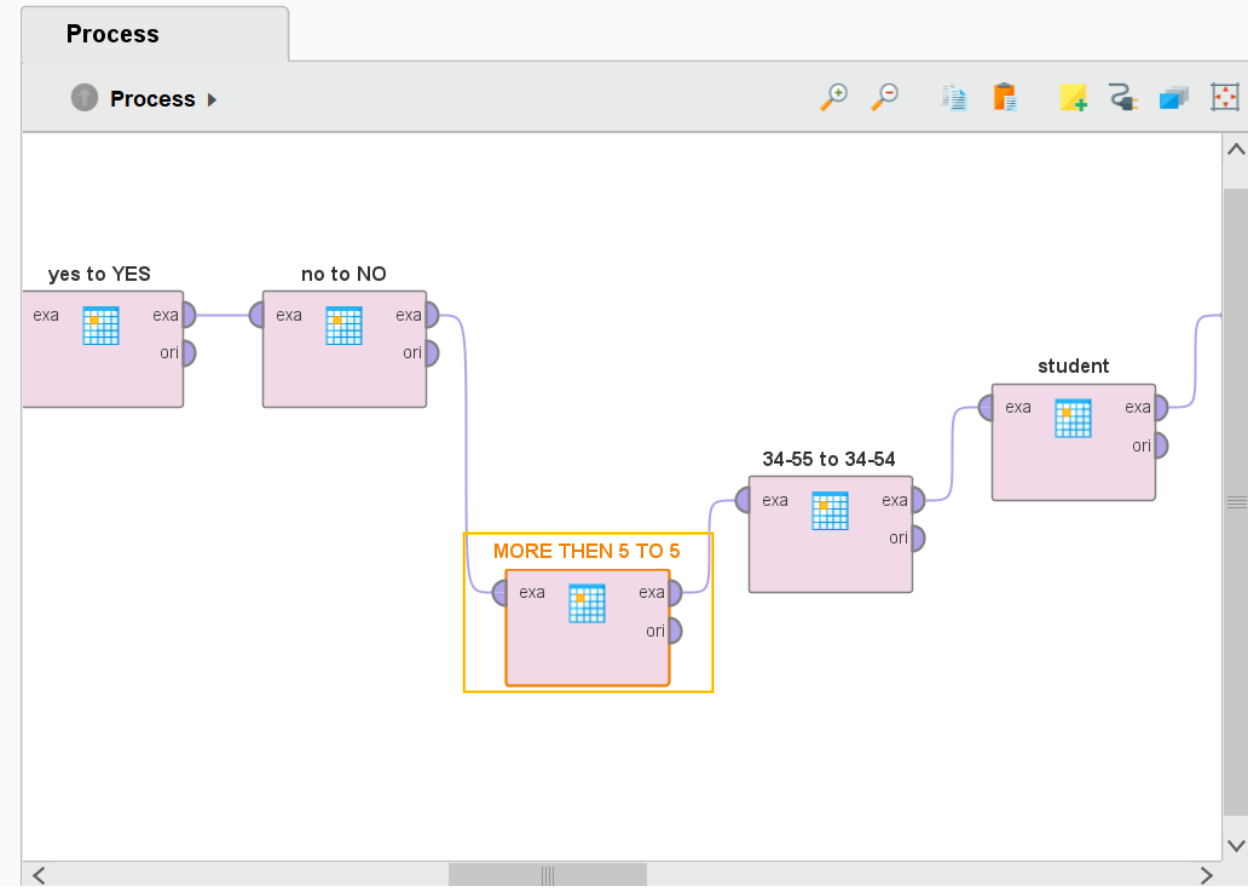
emotional response
NO
YES
NO
YES
YES
YES
YES
YES
NO
YES
YES

- So I modified them in single form
- When I don't do that, it will take 4 values in class labels
- After replacing values, there are only 2 class labels YES AND NO.


Edit Regular Expression
Edit Regular Expression:
A regular expression specifying what should be replaced.
Regular Expression
NO|No|no
Regular expression valid.
Replacement (value for 'replace by')
NO
Inline Text Search
Result List (0)
Regexp Options
Text
Result preview
Apply
Cancel


Edit Regular Expression
Edit Regular Expression:
A regular expression specifying what should be replaced.
Regular Expression
YES|Yes|yes
Regular expression valid.
Replacement (value for 'replace by')
YES
Inline Text Search
Result List (0)
Regexp Options
Text
Result preview
Apply
Cancel

REPLACE OPERATOR (FOR no of platforms):

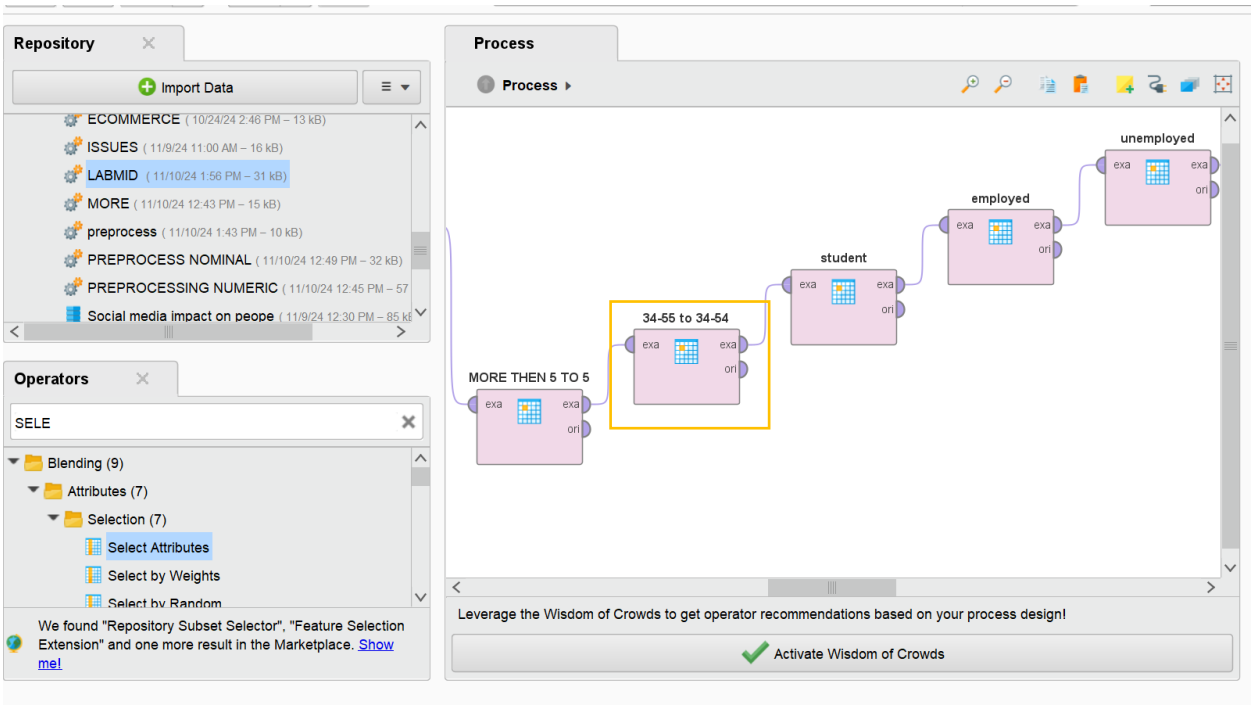


- similarly for number of platforms, where more then 5, I set them to figure 5 to gain better accuracy
- Accuracy will be better if all data is of same type.

no of platfor...
2
4
4
2
2
more than 5
2
3
3
2
2

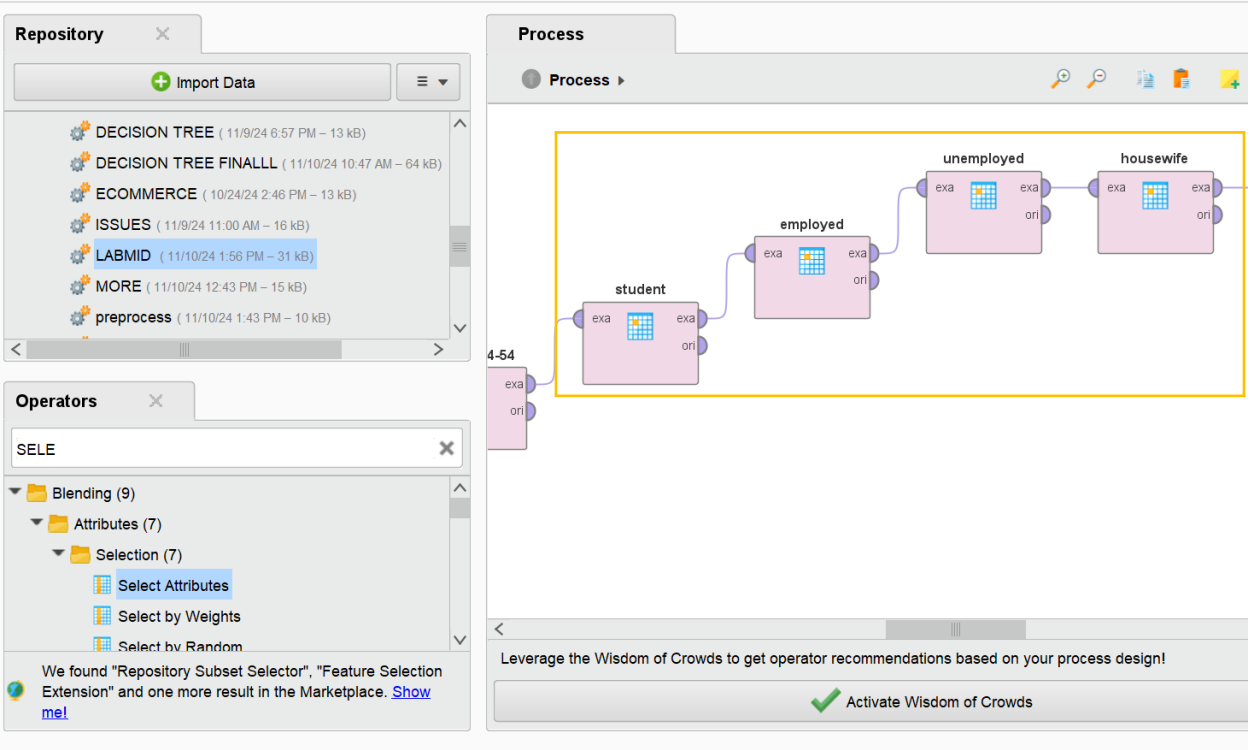
REPLACE OPERATOR (FOR AGE):

- In age, I witnessed a value which has range 34-55 instead of 34-54



REPLACE OPERATOR (FOR OCCUPATION):

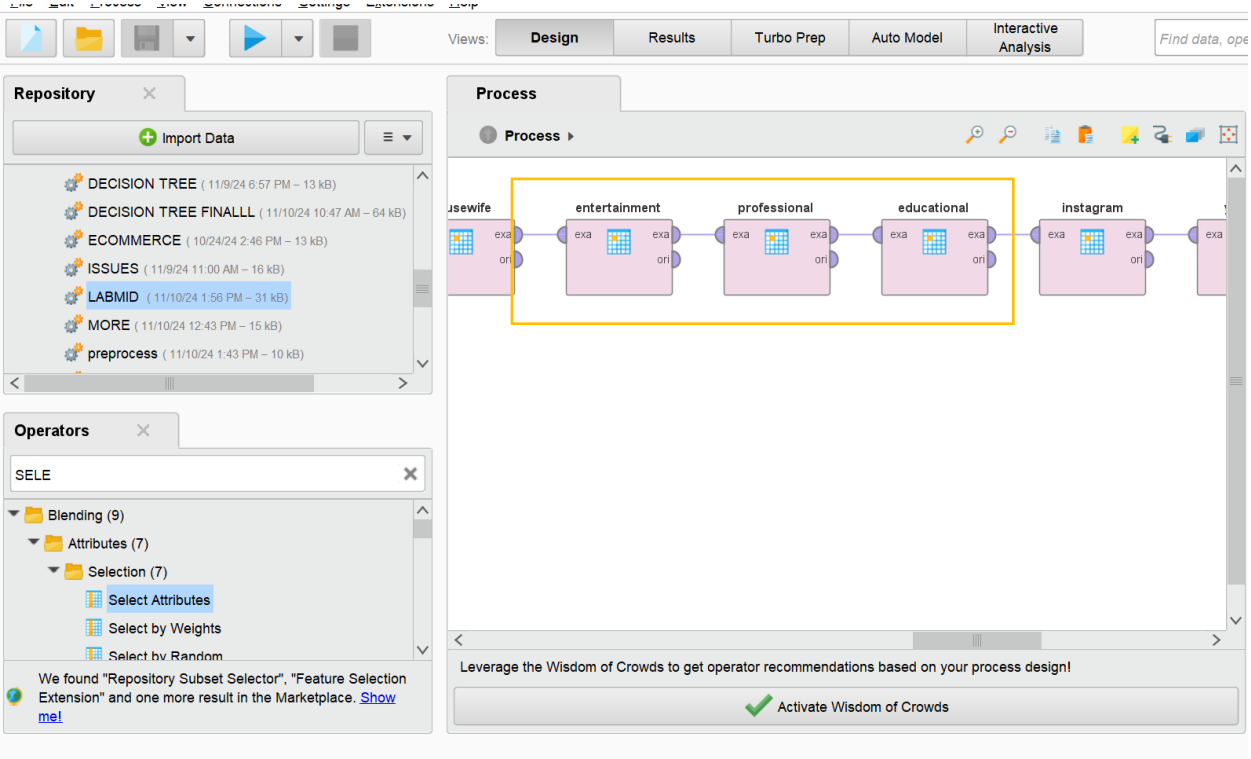
- There were again spelling issues and some irrelevant values in occupation
- Student wriiten as student
- Unemployed as unemployed
- Housewife was written as House wife.
- And some irrelevant like AJK, TAYLOR, HOME MAKEE when we had option to input values from user.
- So these replace operators correct those values.


Repository
Import Data
DECISION TREE (11/9/24 6:57 PM – 13 kB)
DECISION TREE FINALLL (11/10/24 10:47 AM – 64 kB)
ECOMMERCE (10/24/24 2:46 PM – 13 kB)
ISSUES (11/9/24 11:00 AM – 16 kB)
LABMID (11/10/24 1:56 PM – 31 kB)
MORE (11/10/24 12:43 PM – 15 kB)
preprocess (11/10/24 1:43 PM – 10 kB)
Operators
SELE
Blending (9)
Attributes (7)
Selection (7)
Select Attributes
Select by Weights
Select by Random
We found "Repository Subset Selector", "Feature Selection Extension" and one more result in the Marketplace. Show me!
Process
Process
4-54
student
employed
unemployed
housewife
exa
ori
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds

Occupation
Student
Student
unemployed
AJK
Student
Student
unemployed
unemployed
Student
unemployed
Student

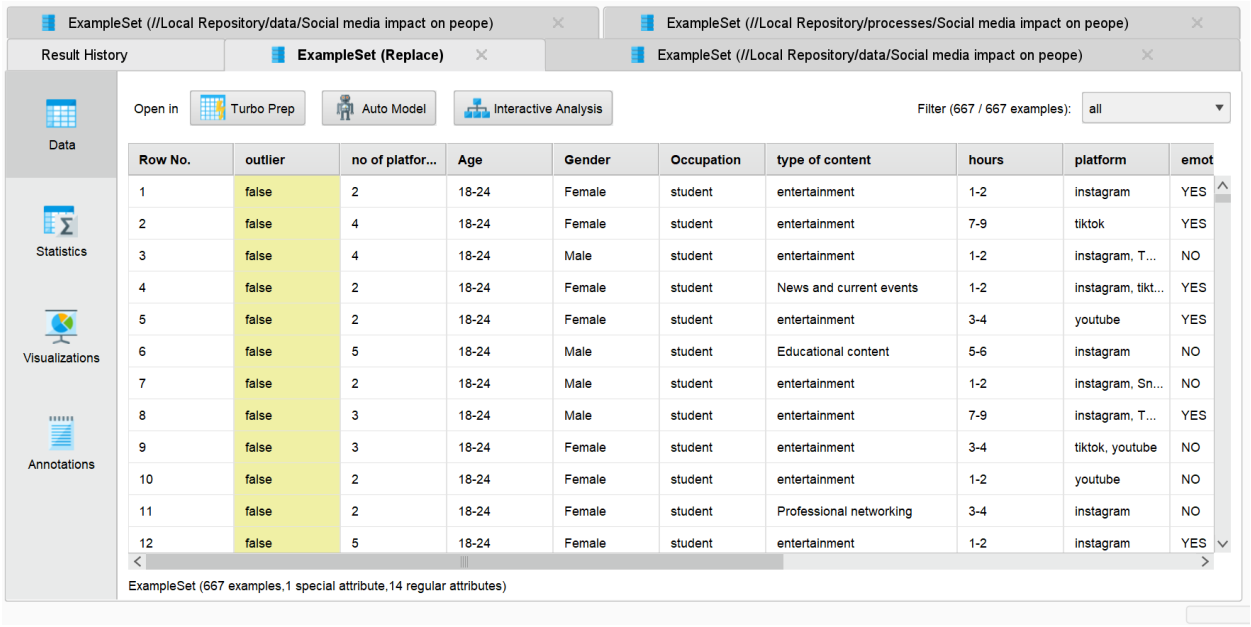
Occupation
Student
Employed
Student
Student
Unemployed
Student
Student
Employed
Student
Student
Retired

REPLACE OPERATOR (for type of content):



- And similarly I used other replace operators as well where data was uneven in columns .

FINAL TABLE AFTER PREPROCESSING:



ExampleSet (//Local Repository/data/Social media impact on peope) | ExampleSet (//Local Repository/processes/Social media impact on peope)

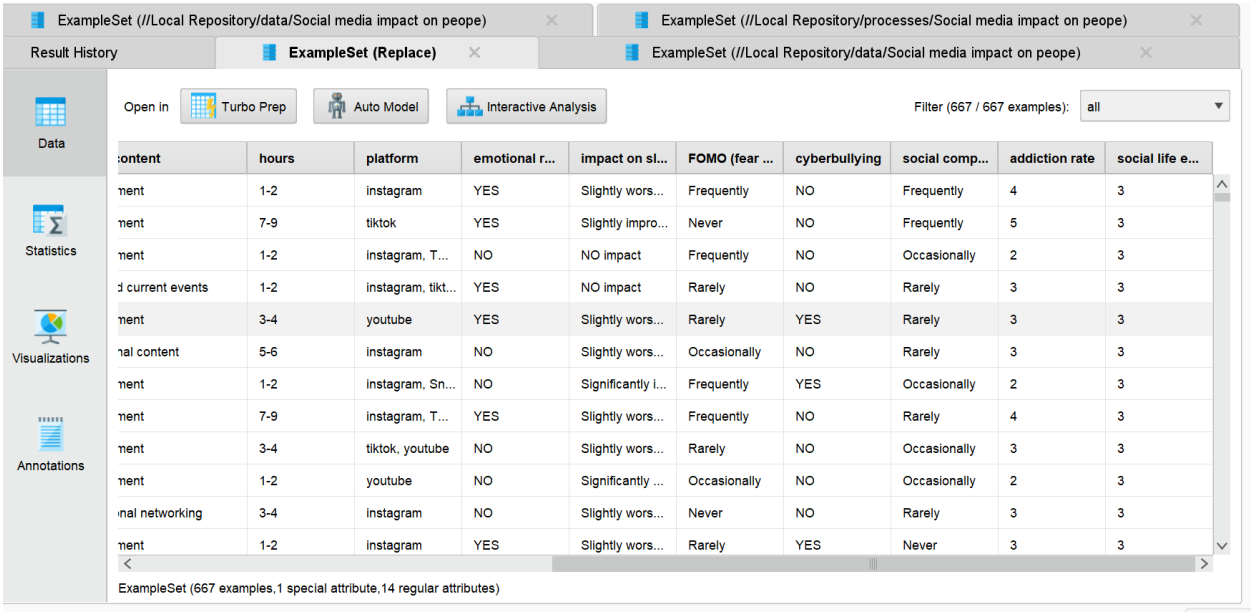
Result History | ExampleSet (Replace) | ExampleSet (//Local Repository/data/Social media impact on peope)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model | Interactive Analysis — Filter (667 / 667 examples): all

Row No.	outlier	no of platfor...	Age	Gender	Occupation	type of content	hours	platform	emot
1	false	2	18-24	Female	student	entertainment	1-2	instagram	YES
2	false	4	18-24	Female	student	entertainment	7-9	tiktok	YES
3	false	4	18-24	Male	student	entertainment	1-2	instagram, T...	NO
4	false	2	18-24	Female	student	News and current events	1-2	instagram, tikt...	YES
5	false	2	18-24	Female	student	entertainment	3-4	youtube	YES
6	false	5	18-24	Male	student	Educational content	5-6	instagram	NO
7	false	2	18-24	Male	student	entertainment	1-2	instagram, Sn...	NO
8	false	3	18-24	Male	student	entertainment	7-9	instagram, T...	YES
9	false	3	18-24	Female	student	entertainment	3-4	tiktok, youtube	NO
10	false	2	18-24	Female	student	entertainment	1-2	youtube	NO
11	false	2	18-24	Female	student	Professional networking	3-4	instagram	NO
12	false	5	18-24	Female	student	entertainment	1-2	instagram	YES

ExampleSet (667 examples,1 special attribute,14 regular attributes)



ExampleSet (//Local Repository/data/Social media impact on peope) | ExampleSet (//Local Repository/processes/Social media impact on peope)

Result History | ExampleSet (Replace) | ExampleSet (//Local Repository/data/Social media impact on peope)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model | Interactive Analysis — Filter (667 / 667 examples): all

ontent	hours	platform	emotional r...	impact on sl...	FOMO (fear ...	cyberbullying	social comp...	addiction rate	social life e...
ment	1-2	instagram	YES	Slightly wors...	Frequently	NO	Frequently	4	3
ment	7-9	tiktok	YES	Slightly impro...	Never	NO	Frequently	5	3
ment	1-2	instagram, T...	NO	NO impact	Frequently	NO	Occasionally	2	3
d current events	1-2	instagram, tikt...	YES	NO impact	Rarely	NO	Rarely	3	3
ment	3-4	youtube	YES	Slightly wors...	Rarely	YES	Rarely	3	3
nal content	5-6	instagram	NO	Slightly wors...	Occasionally	NO	Rarely	3	3
ment	1-2	instagram, Sn...	NO	Significantly i...	Frequently	YES	Occasionally	2	3
ment	7-9	instagram, T...	YES	Slightly wors...	Frequently	NO	Rarely	4	3
ment	3-4	tiktok, youtube	NO	Slightly wors...	Rarely	NO	Occasionally	3	3
ment	1-2	youtube	NO	Significantly ...	Occasionally	NO	Occasionally	2	3
nal networking	3-4	instagram	NO	Slightly wors...	Never	NO	Rarely	3	3
ment	1-2	instagram	YES	Slightly wors...	Rarely	YES	Never	3	3

ExampleSet (667 examples,1 special attribute,14 regular attributes)

CHAPTER 3:

MODELS IMPLEMENTATION AND EVALUATION :

MODEL 1: DECISION TREE:

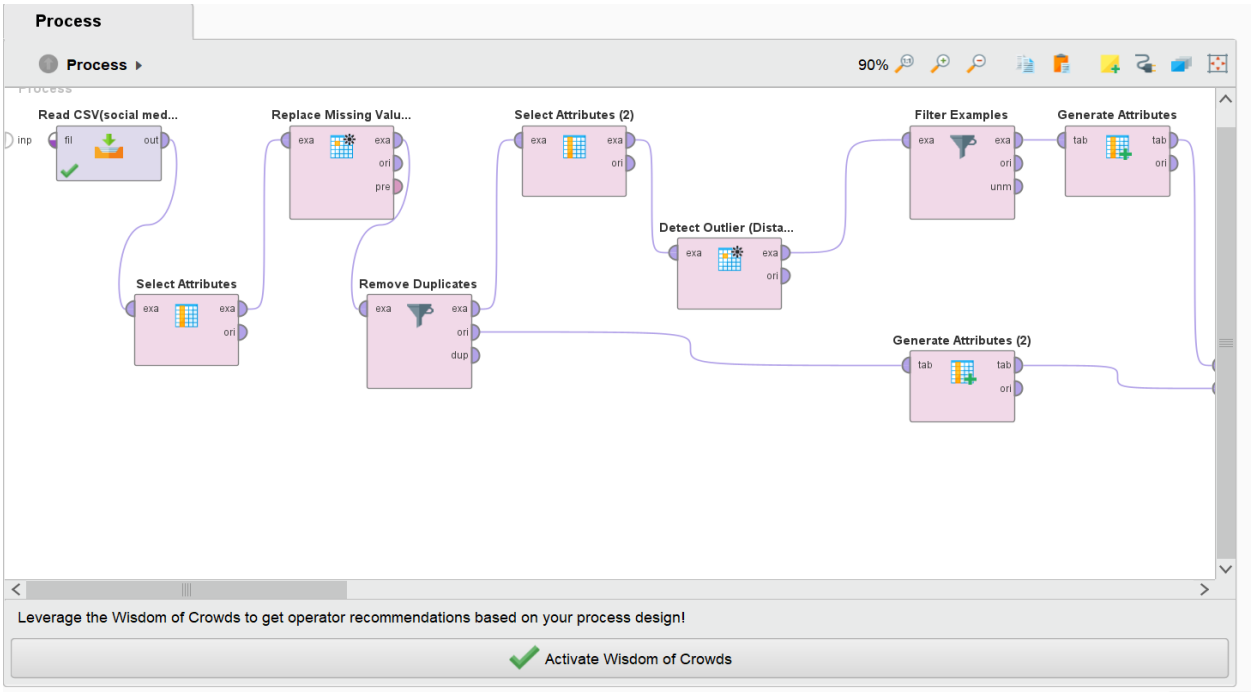
COMPLETE PROCESS SNAPSHOT:

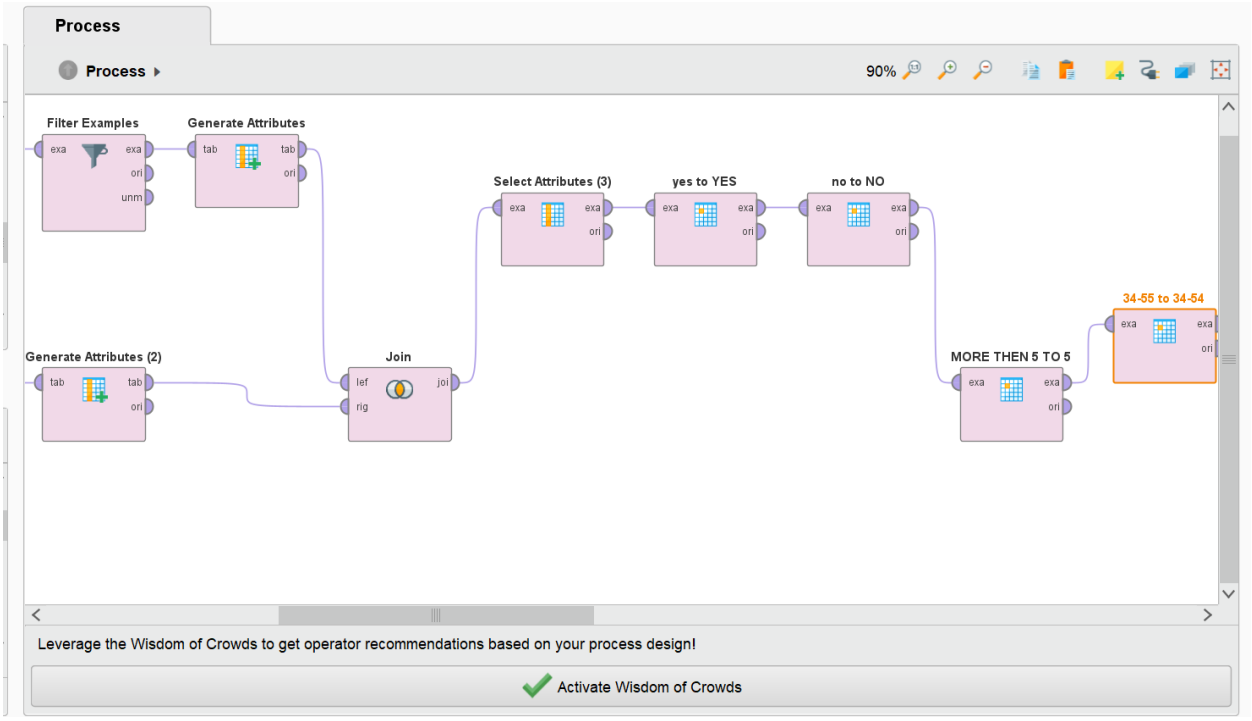
- Decision tree is suitable for my dataset because it has the capability to handle both numeric and categorical data .
- The model's decisions can be visualized as a tree, making it easier to understand which factors (platform type or hours spent etc) lead to a "YES" or "NO" emotional response
- Decision Trees can show feature importance, helping identify which factors have the strongest influence on emotional response.
- Here decision tree is applied to the preprocessed data which I did before .

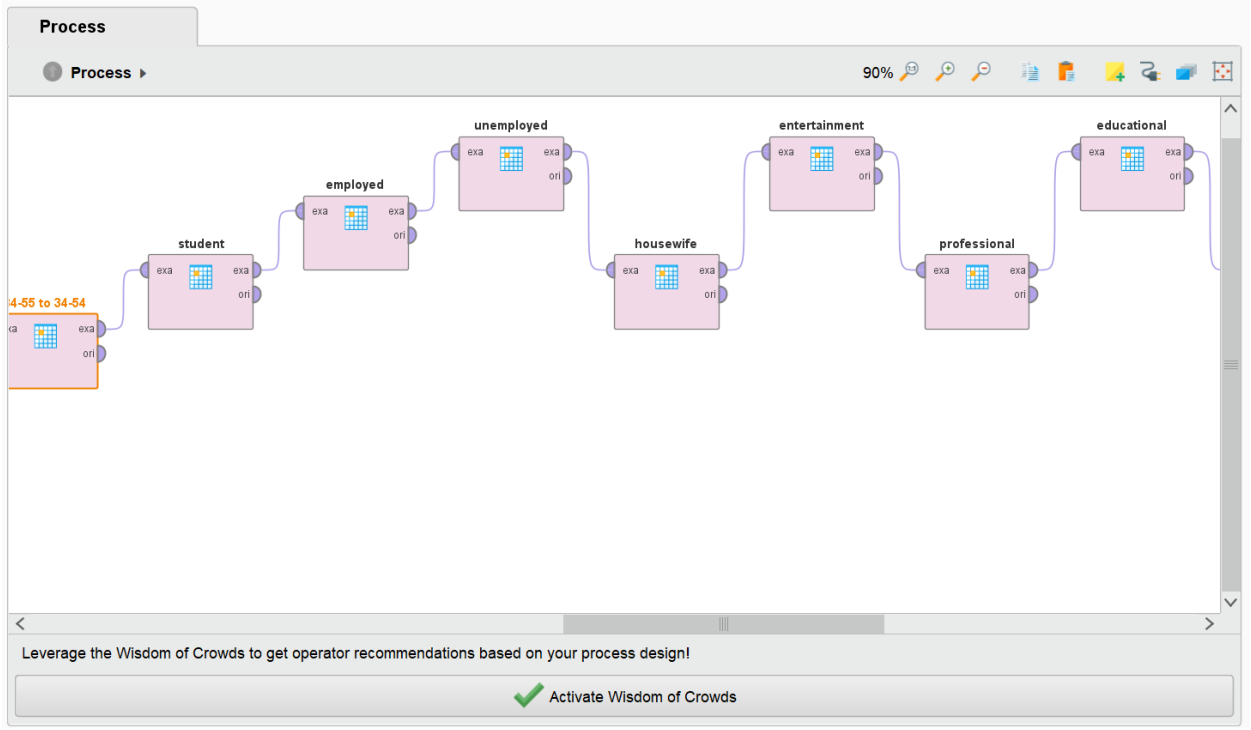
"Emotional response" is a binary categorical feature (e.g., 'YES' or 'NO'), making it a classification problem.

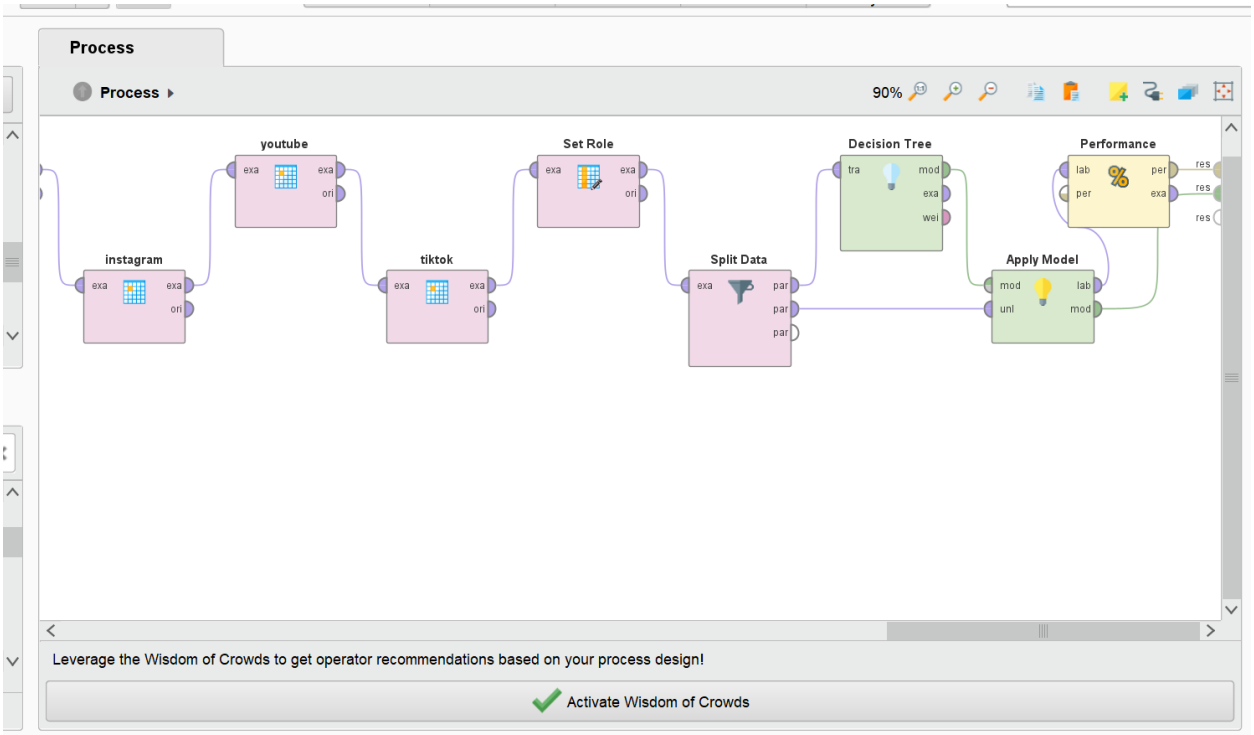
- Here decision tree is used to explain why a person has emotional response based on several factors (e.g hours of phone usage, type of content, etc) and also based on demographics information.

Complete snapshot of process:





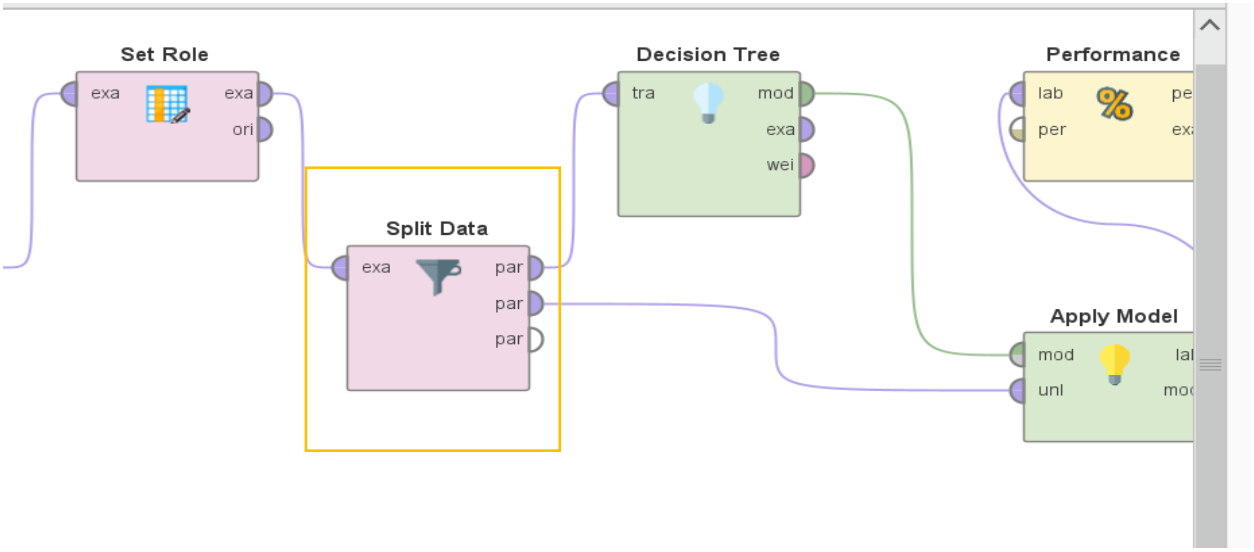


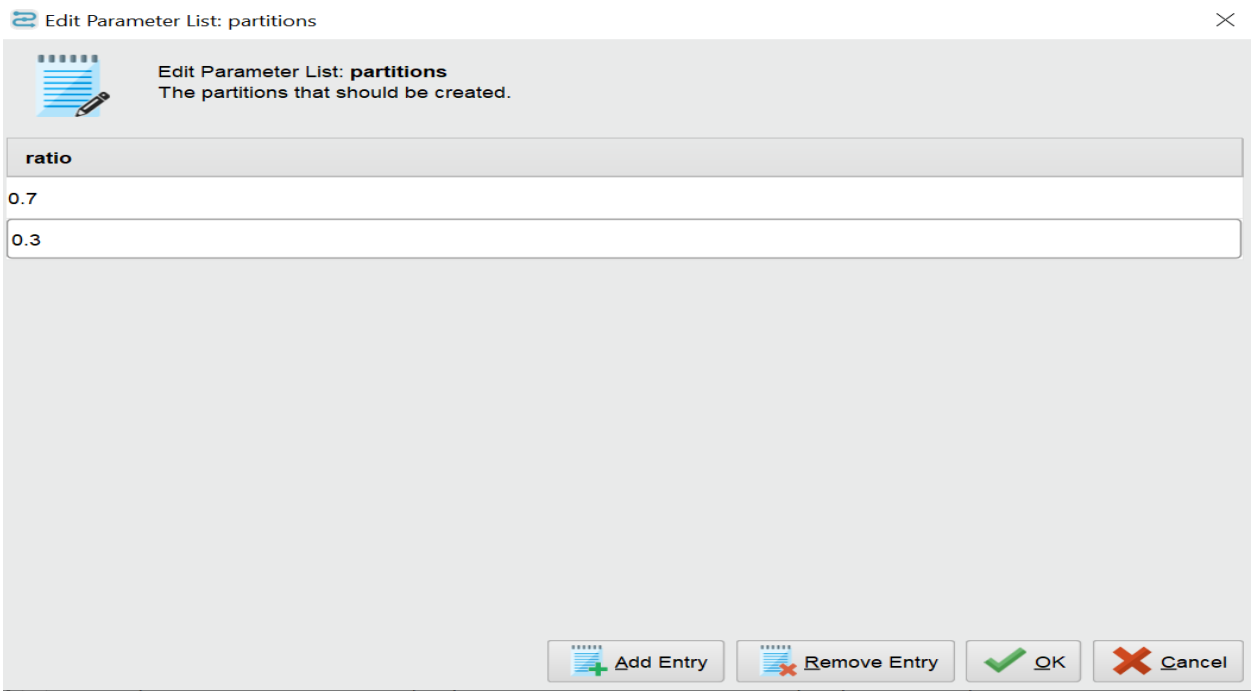


Model training and evaluation:

- Data is split into 70% training data and 30% testing data.

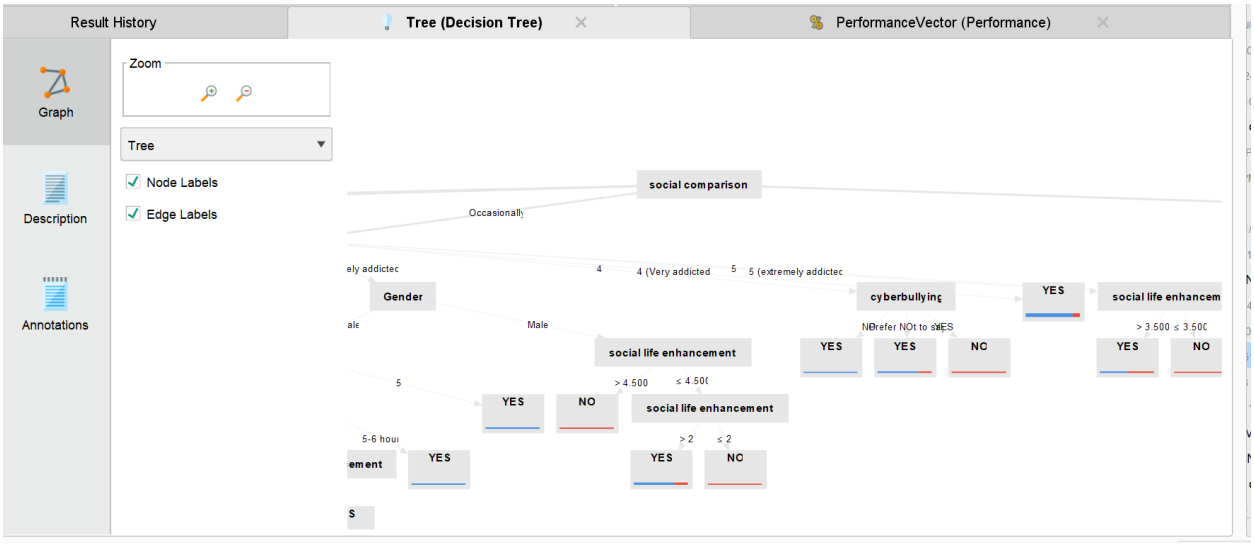
- 70% data is used to train the model (decision tree) based on which 30% data is checked.
- 30% data is attached to apply model operator and performance is checked based on how accurately the model is working.

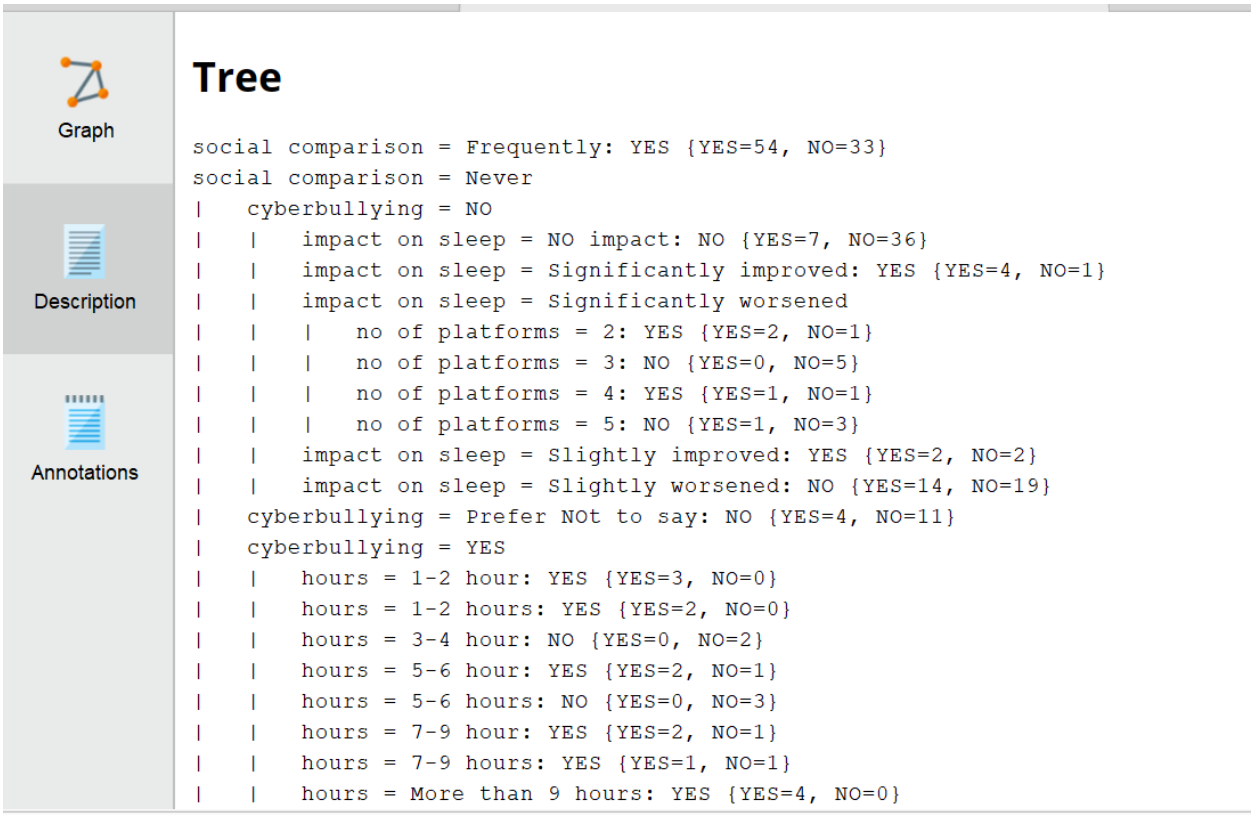


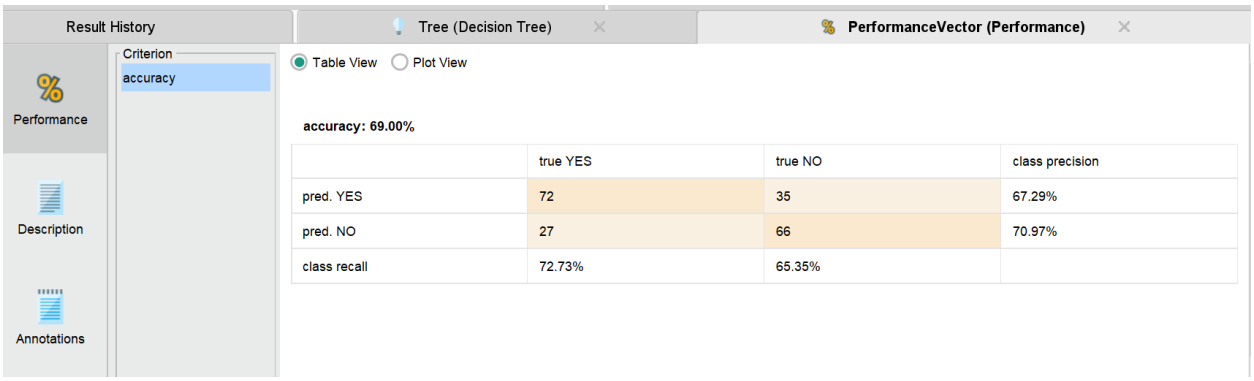


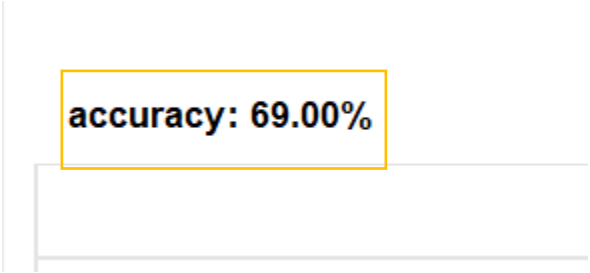
Model Evaluation:

- At each decision node, the model selects the feature that best splits the data based on a criterion


Result History
Tree (Decision Tree)
PerformanceVector (Performance)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
social comparison
Occasionally
Gender
Male
social life enhancement
YES
NO
cyberbullying
Prefer NOt to say
social life enhancem


Graph
Description
Annotations
Tree
social comparison = Frequently: YES {YES=54, NO=33}
social comparison = Never
| cyberbullying = NO
| | impact on sleep = NO impact: NO {YES=7, NO=36}
| | impact on sleep = Significantly improved: YES {YES=4, NO=1}
| | impact on sleep = Significantly worsened
| | | no of platforms = 2: YES {YES=2, NO=1}
| | | no of platforms = 3: NO {YES=0, NO=5}
| | | no of platforms = 4: YES {YES=1, NO=1}
| | | no of platforms = 5: NO {YES=1, NO=3}
| | impact on sleep = Slightly improved: YES {YES=2, NO=2}
| | impact on sleep = Slightly worsened: NO {YES=14, NO=19}
| cyberbullying = Prefer NOt to say: NO {YES=4, NO=11}
| cyberbullying = YES
| | hours = 1-2 hour: YES {YES=3, NO=0}
| | hours = 1-2 hours: YES {YES=2, NO=0}
| | hours = 3-4 hour: NO {YES=0, NO=2}
| | hours = 5-6 hour: YES {YES=2, NO=1}
| | hours = 5-6 hours: NO {YES=0, NO=3}
| | hours = 7-9 hour: YES {YES=2, NO=1}
| | hours = 7-9 hours: YES {YES=1, NO=1}
| | hours = More than 9 hours: YES {YES=4, NO=0}





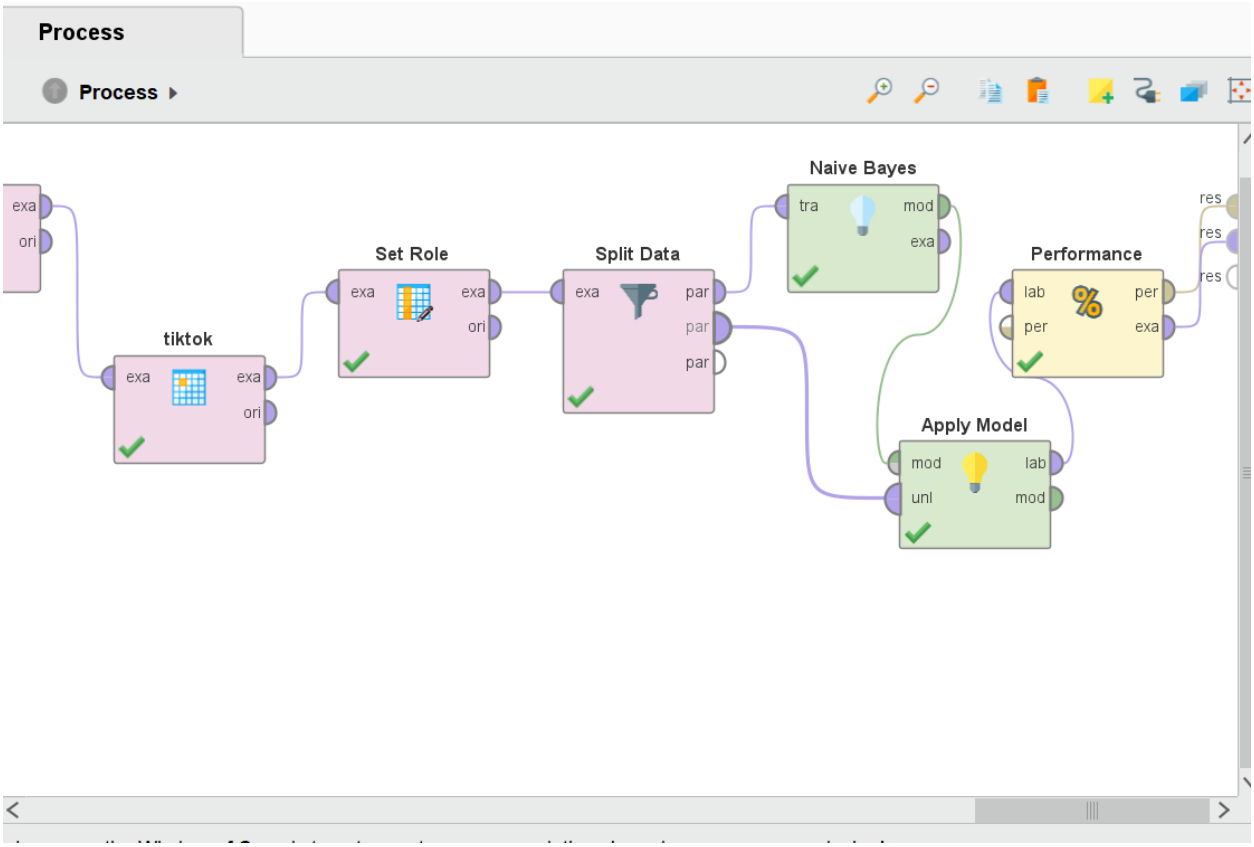
- **Total samples:** 667
- **Correct predictions:** 460
- **Incorrect predictions:** 667-460=207
- **Accuracy:** 69%

MODEL 2: NAIVE BAYES:

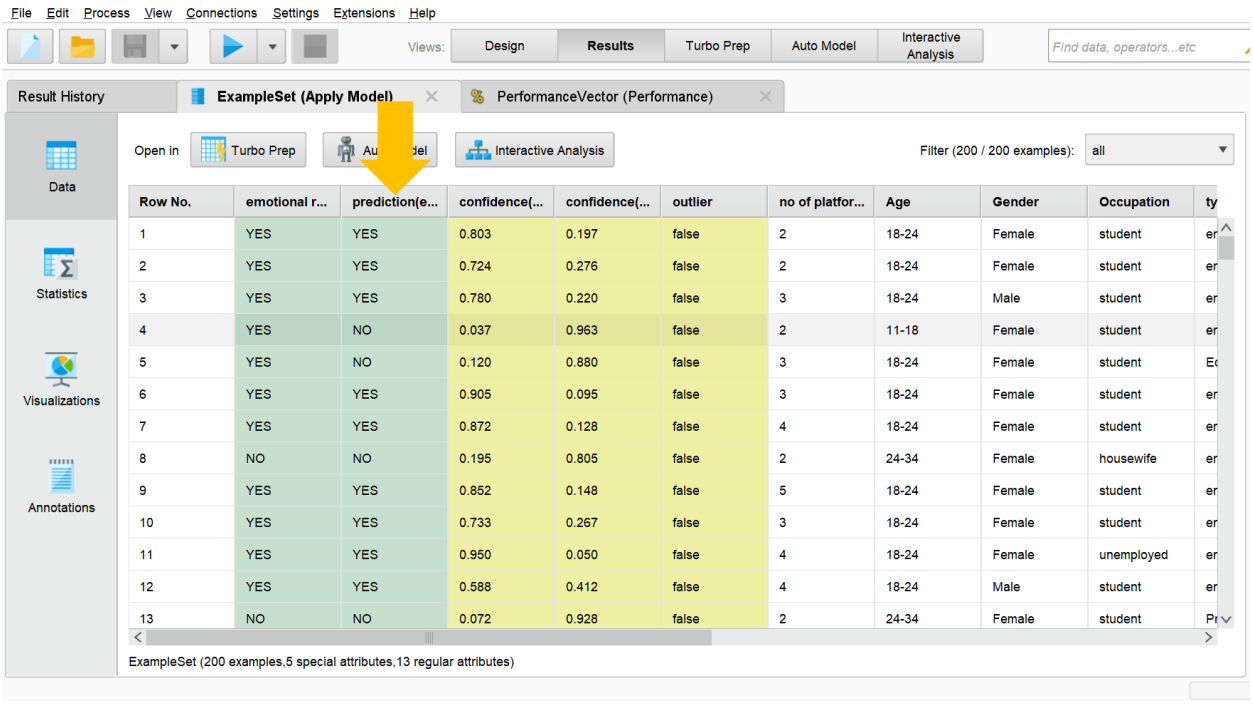
- This model is used in my dataset bcz it also works well on categorical as well as numerical data like mine where decision is made whether a person has emotional response based on the probabilities of other factors (like impact on sleep, type of content, addiction rate).
- It provides probabilities for each class, offering insights into the confidence of predictions (e.g chances of a "YES" emotional response given the type of platform or frequency of social comparison or other input features).
- I've applied this model on same preprocessing steps which are done earlier.

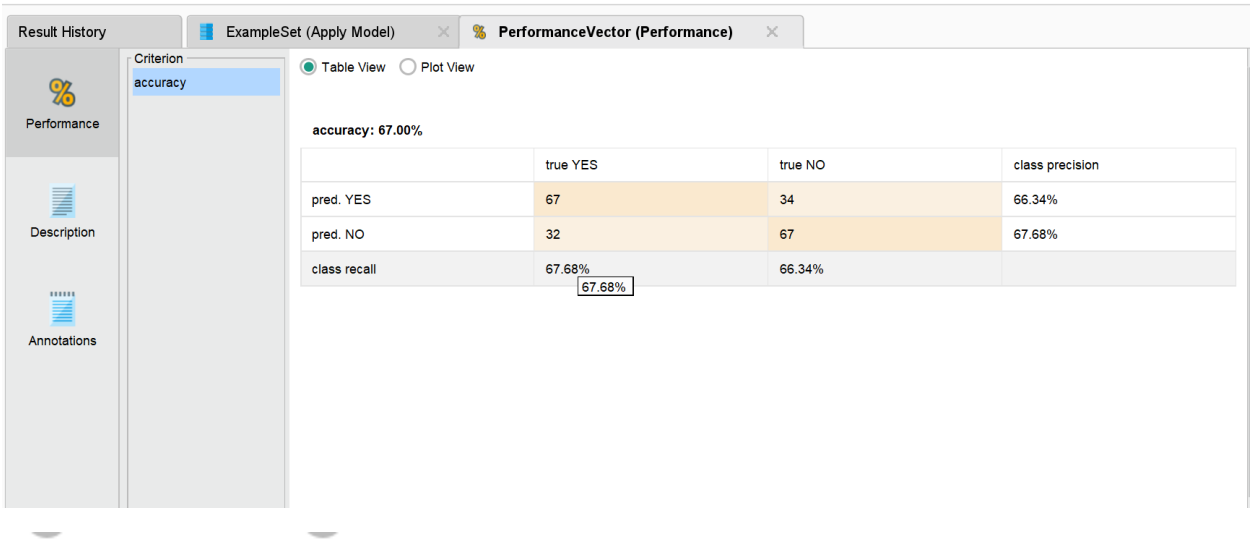
Model training and evaluation:

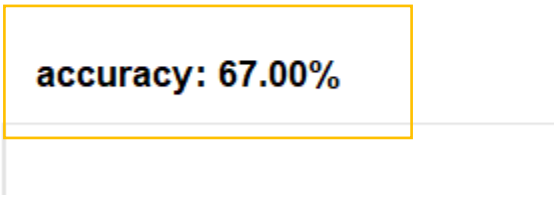
- Data is split into 70% training data and 30% testing data.
- 70% data is used to train the model (naïve bayes) based on which 30% data is checked.
- 30% data is attached to apply model operator and performance is checked based on how accurately the model is working.



- **Confidence Scores**: Just like the Decision Tree, Naive Bayes provides confidence scores (probabilities) for each prediction. For instance, if the confidence score for 'YES' is 0.9 (90%), it indicates a high chances of emotional response.







- **Total samples:** 667
- **Correct predictions:** 448
- **Incorrect predictions:** 219

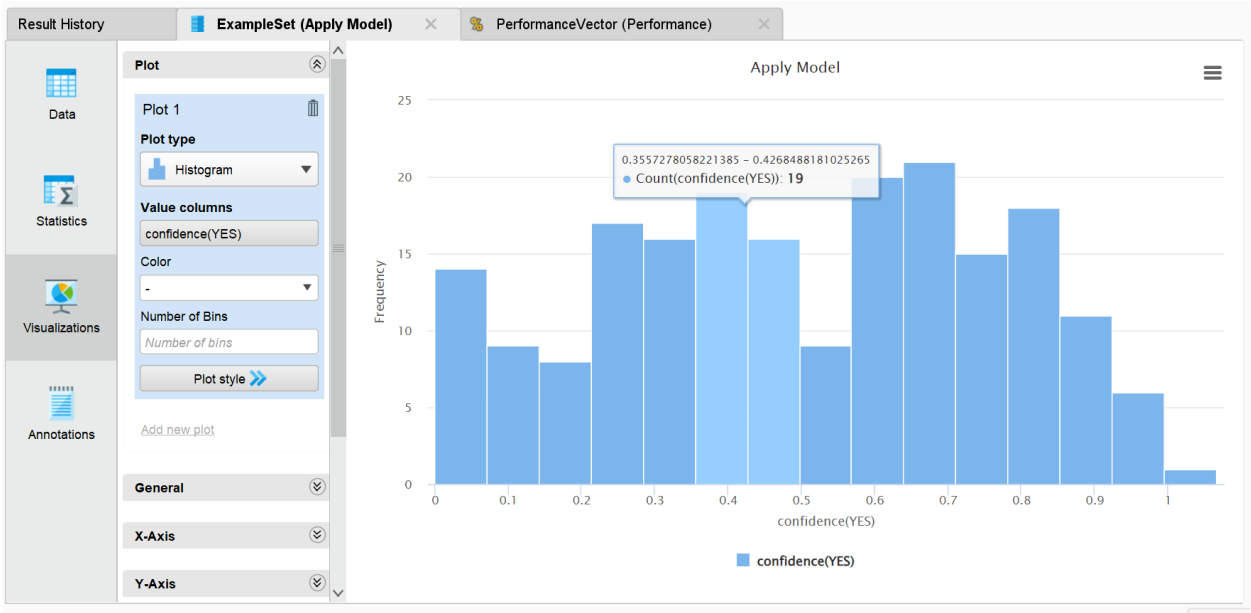
CHAPTER 4:

VISUALIZATION AND MODEL COMPARISON

FOR NAIVE BAYES MODEL:

Visualization 1: Histogram

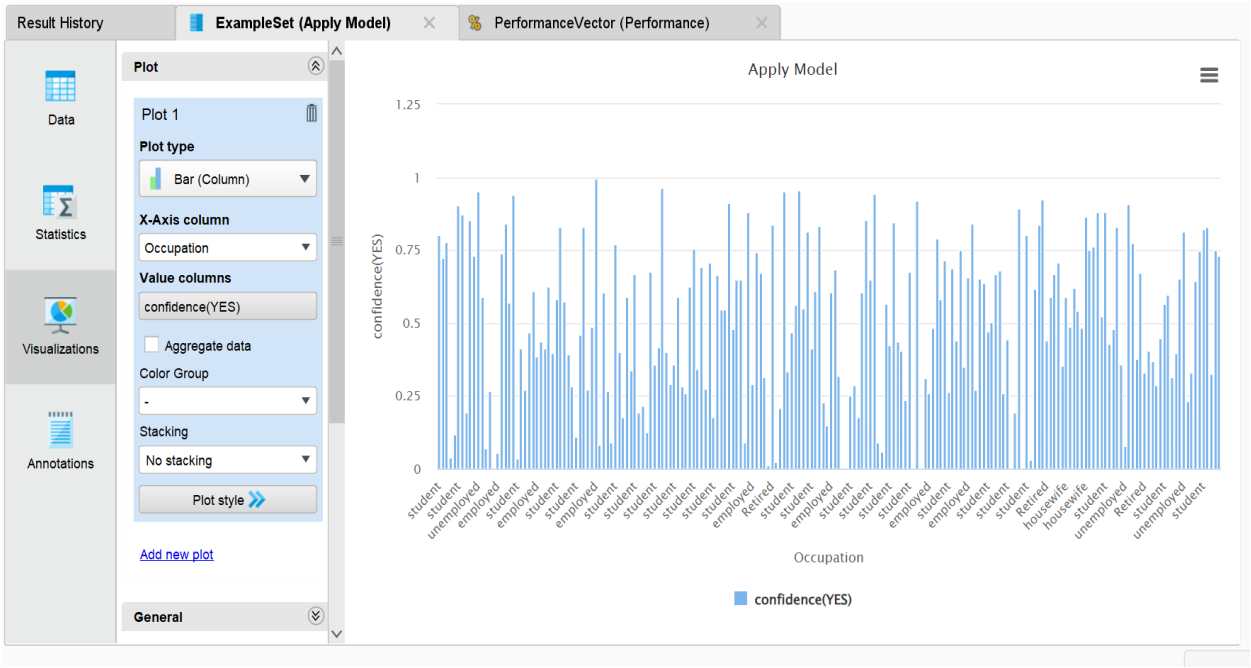
- The histogram shows how often the model predicts ("YES" outcome). The taller the bars, the more often the model predicts "YES" with a certain level of confidence.
- For example, the tallest bar is around 0.65, which means the model often predicts "YES" with a confidence level between 0.6 and 0.7.



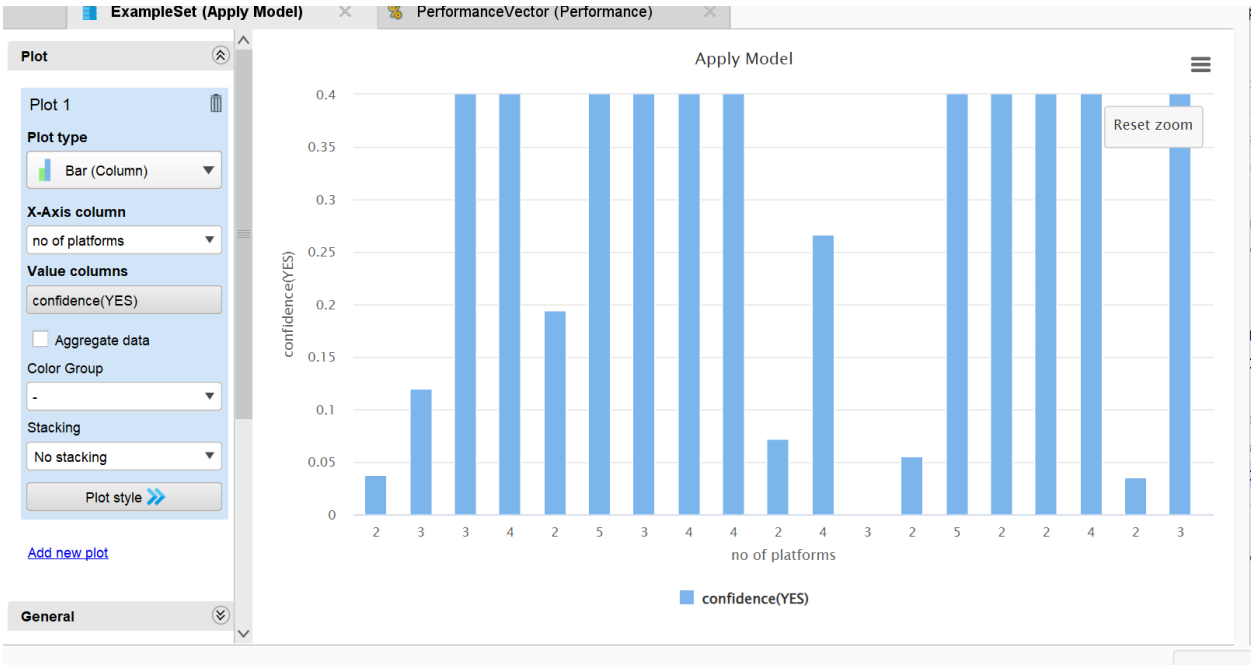
Visualization 2: bar column

- The bar chart illustrates the confidence levels of a model's predictions for different occupations.
- The height of each bar represents the model's confidence in predicting a "YES" outcome for that specific occupation.

- The chart reveals varying levels of confidence across different occupations, indicating potential performance differences.



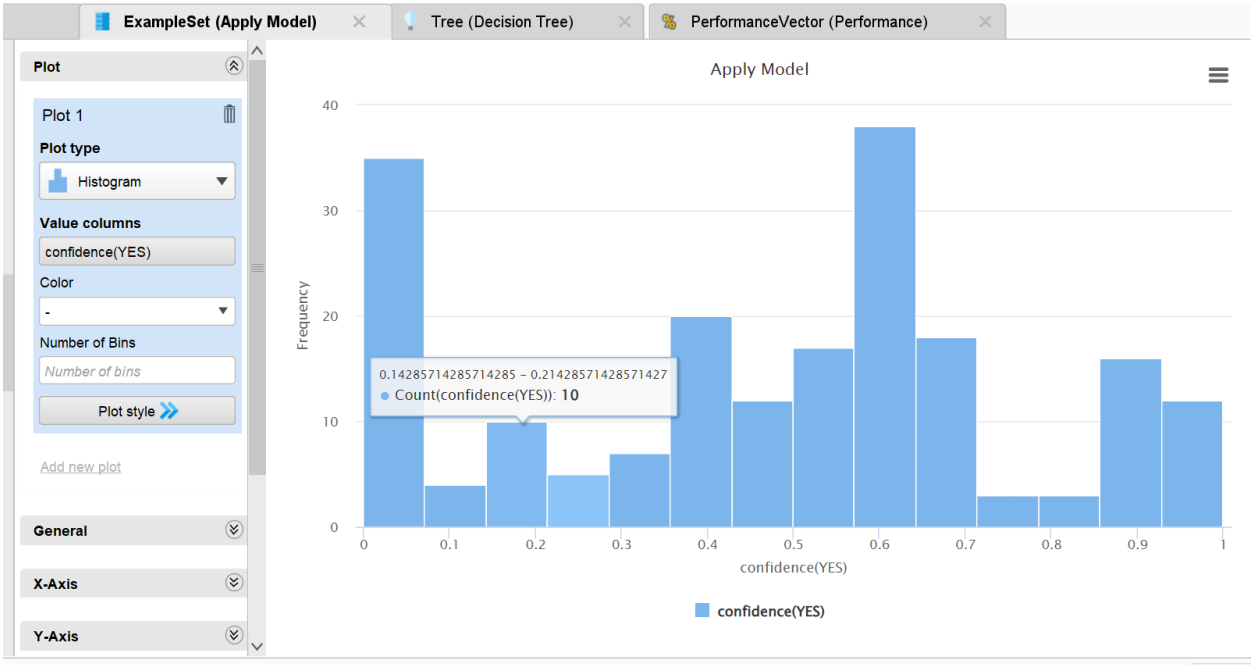
- Similarly the bar chart below shows how number of platform effect emotional response.



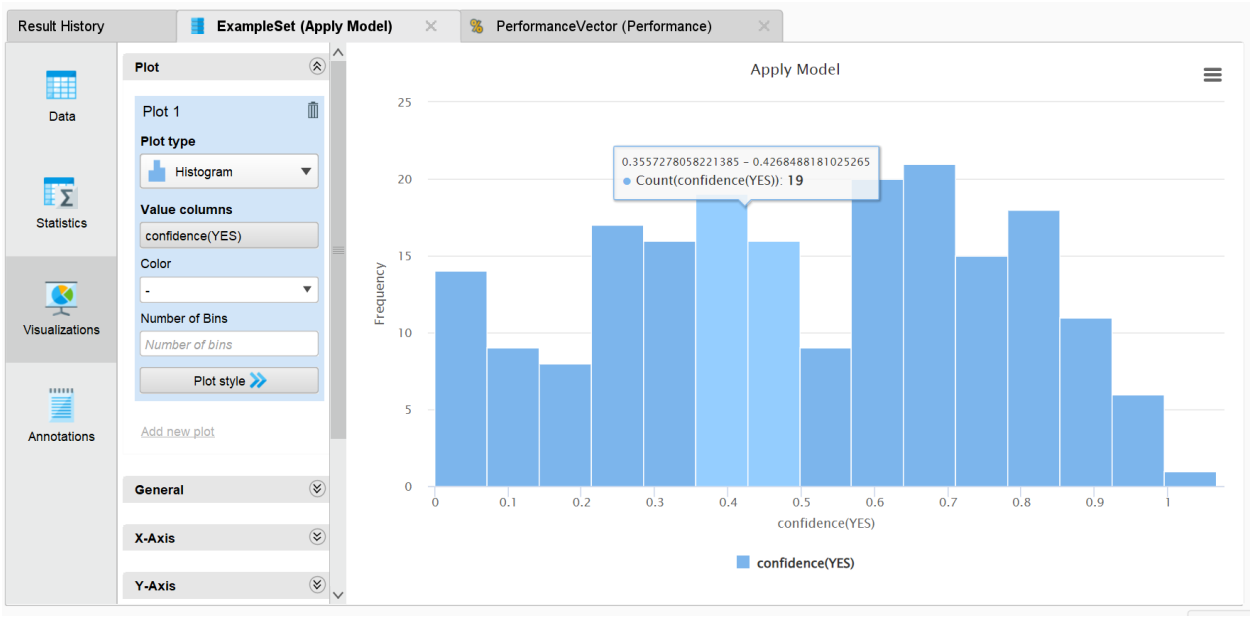
Visualization 2: model comparison

Histogram :

Decision tree Histogram:



Naive bayes Histogram:



Decision Tree:

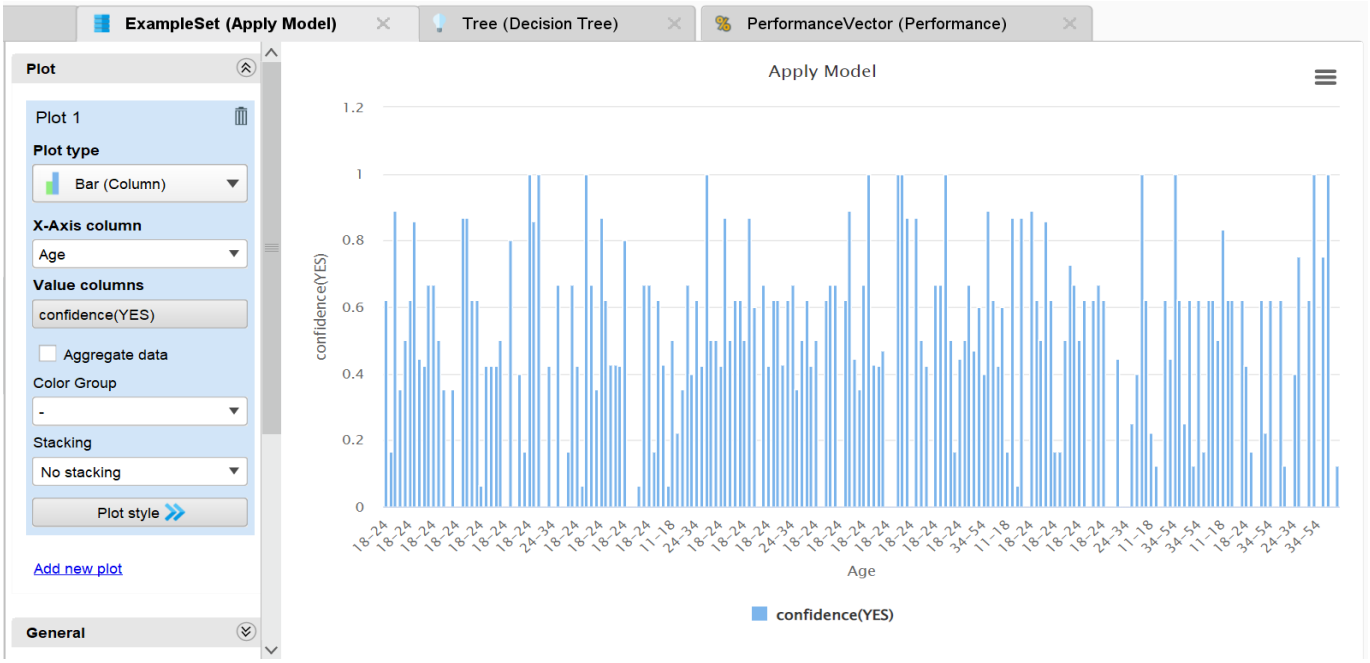
- The histogram shows a wider range of confidence scores, with some predictions having very high confidence (close to 1) and others having lower confidence.
- This suggests that the decision tree model is more confident in some predictions than others. It might be better at capturing complex relationships in the data but might also be more prone to overfitting.

Naive Bayes:

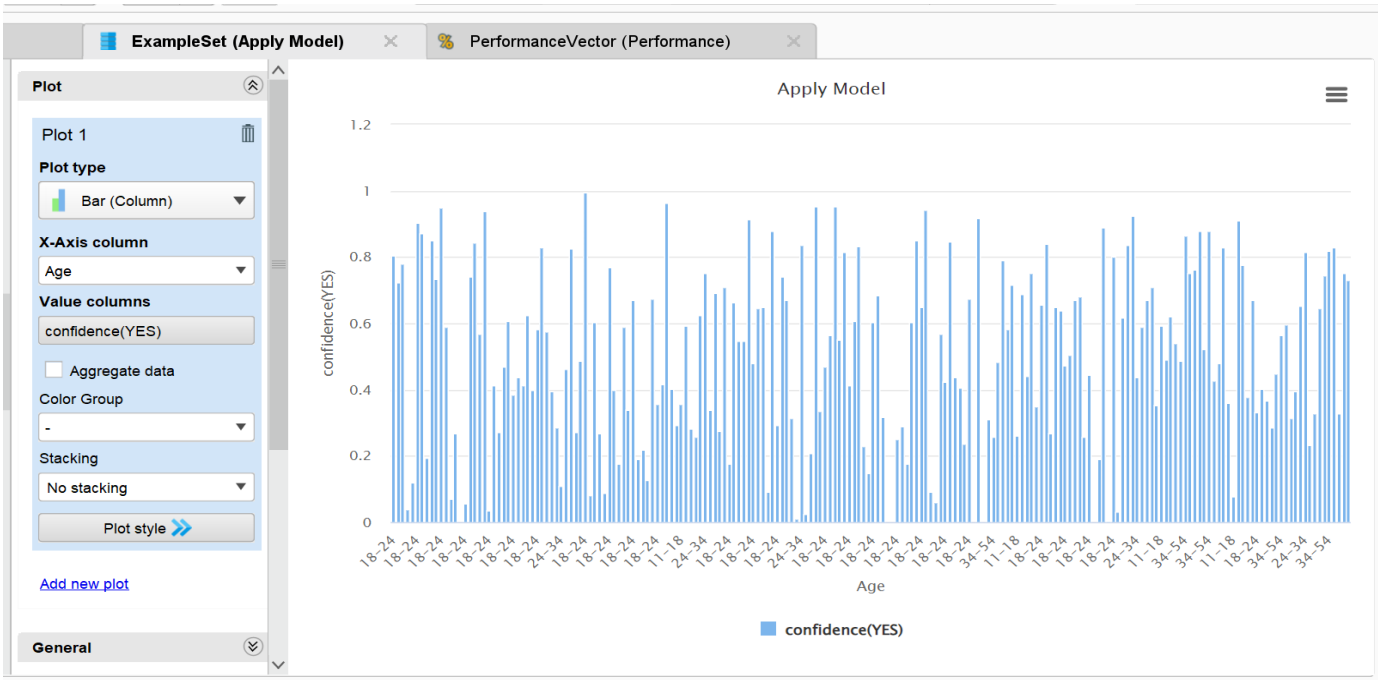
- The histogram shows a narrower range of confidence scores, with most predictions falling within a moderate confidence range.
- This suggests that the naive Bayes model is generally less confident in its predictions.

Bar chart:

Decision tree bar chart (age based analysis):



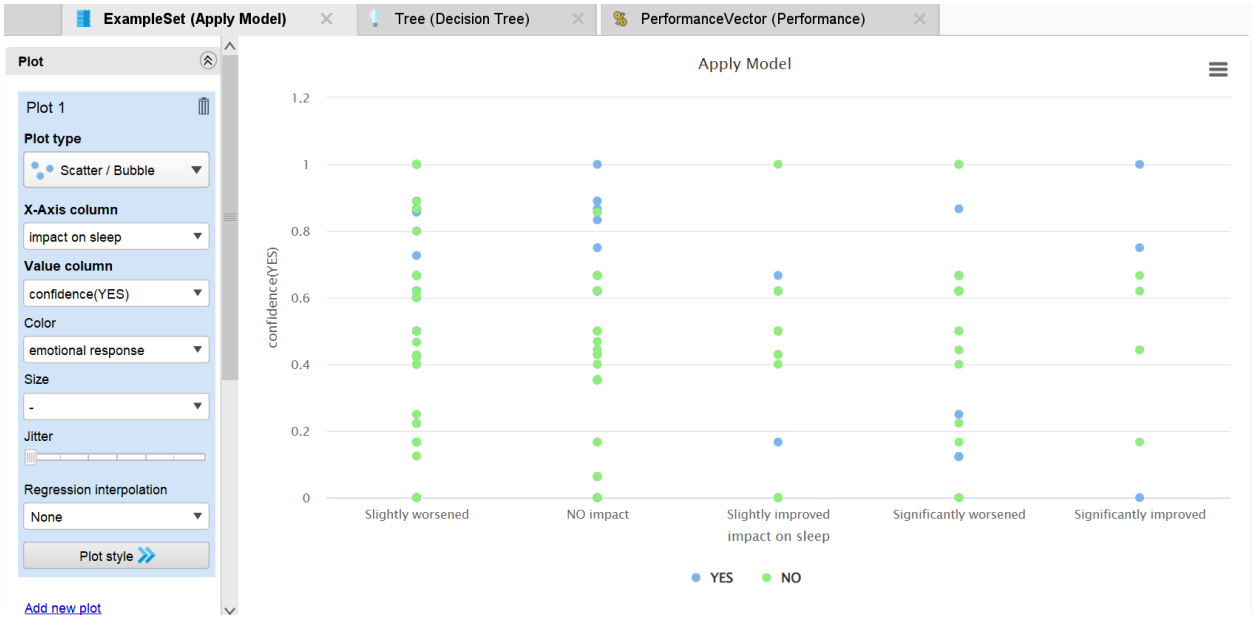
Naïve bayes bar chart (age based analysis):



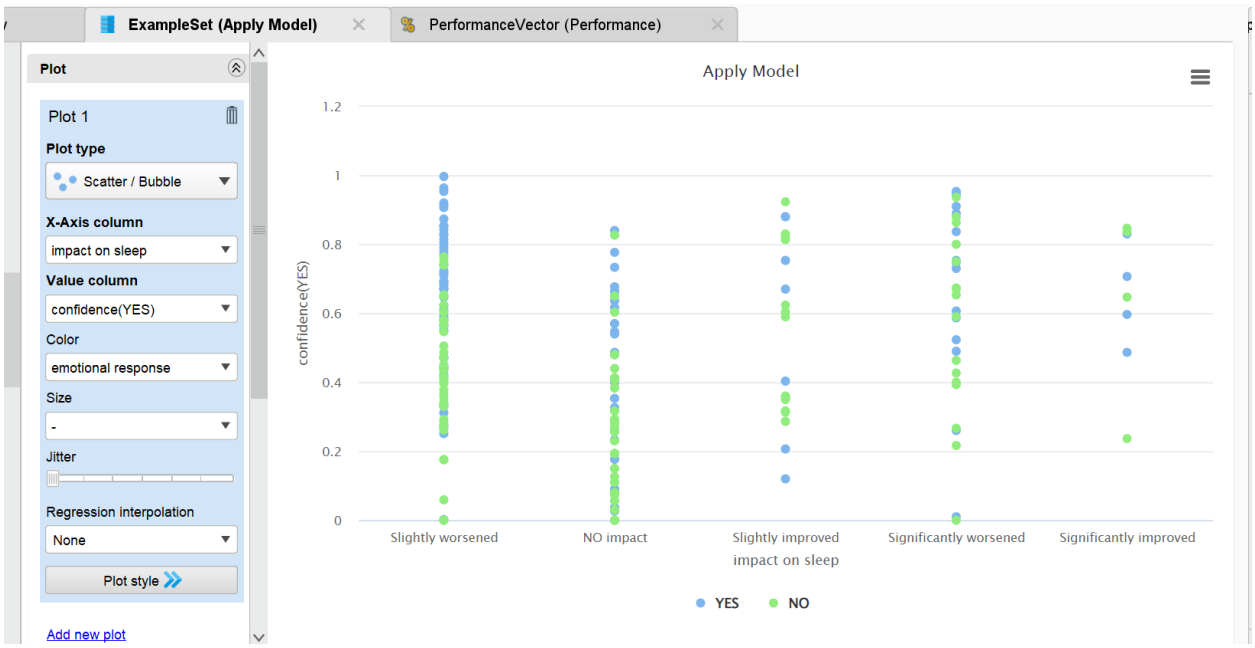
- **Decision Tree:** The chart shows confidence values that fluctuate significantly, with many predictions near 0 or 1. This indicates that the decision tree often makes "all-or-nothing" predictions, displaying high certainty when it categorizes an age group as likely to have a "YES" response.
- **Naive Bayes:** The confidence values are more evenly distributed and tend to be closer to mid-range. This suggests that Naive Bayes makes more balanced predictions, as it calculates probabilities based on the assumption that attributes are independent.

Scatter Bubble :

For decision tree:



For naïve bayes


ExampleSet (Apply Model)
PerformanceVector (Performance)
Plot
Plot 1
Plot type
Scatter / Bubble
X-Axis column
impact on sleep
Value column
confidence(YES)
Color
emotional response
Size
-
Jitter
Regression interpolation
None
Plot style
Add new plot
Apply Model
confidence(YES)
1.2
1
0.8
0.6
0.4
0.2
0
Slightly worsened
NO impact
Slightly improved
Significantly worsened
Significantly improved
impact on sleep
YES
NO

CHAPTER 5:

CONCLUSION

Key Steps:

- data collected through survey forms underwent preprocessing in rapid miner which involve handling missing values, removing duplicates, outliers and handling inconsistent data .
- the next step was model selection .this step was crucial because we have to select the model which best fit our requeirement and perform better
- two machine learning models were selected , decision tree and naïve bayes
- both models were trained based on the factor of emotional response to social media.
- Both models were evaluated using metrics like accuracy, precision, and recall.
- Confusion matrices were used to analyze the correct and incorrect predictions

Results:

- For **decision tree**, achieved an accuracy of 69%, with balanced precision and recall across the classes (YES and NO).
- The Decision Tree model was able to identify key features such as time spent on social media and engagement with entertainment content as major decision nodes.
- For **naive bayes,** achieved an accuracy of 67%, slightly lower than the Decision Tree model.
- Provided probability-based insight for emotional responses based on user features.

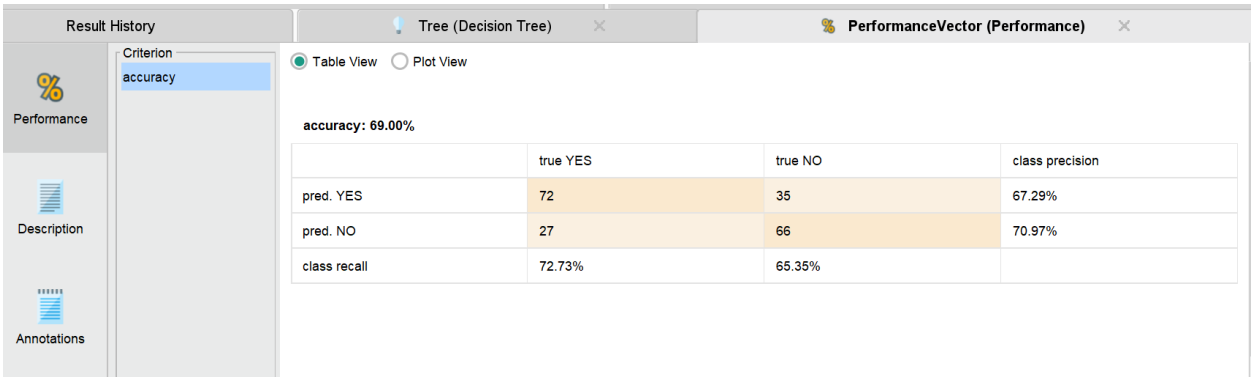
Findings:

- Young respondents (18-24 years old) and students showed higher engagement with social media and were more likely to report negative emotional responses.
- A significant portion of users experienced worsened sleep quality and increased anxiety or FOMO due to prolonged social media usage.
- Frequent use of platforms like Instagram and TikTok was associated with higher levels of social comparison and emotional response.

Model Comparison:

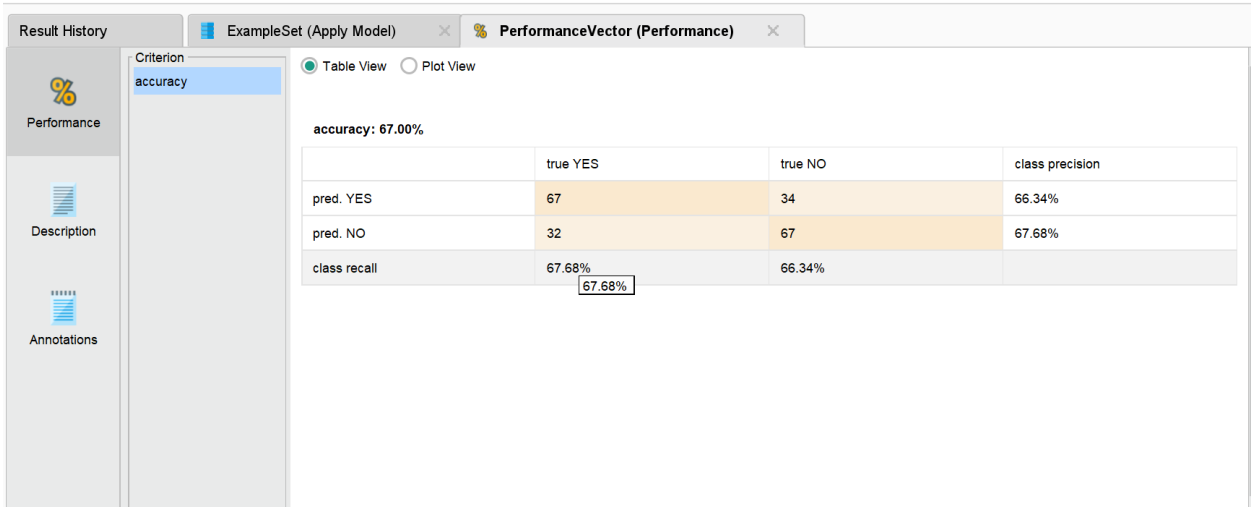
- The Decision Tree model performed better with a higher accuracy (69%) compared to Naive Bayes (67%).
- The Decision Tree offered better interpretability due to its clear decision rules and visual representation of feature importance.
- It highlights key features (e.g., time spent on social media, type of content engaged) that play a major role in predicting emotional response.
- Based on confusion matrix,

Decision tree confusion matrix:



	true YES	true NO	class precision
pred. YES	72	35	67.29%
pred. NO	27	66	70.97%
class recall	72.73%	65.35%	

Naïve Bayes confusion matrix :



	true YES	true NO	class precision
pred. YES	67	34	66.34%
pred. NO	32	67	67.68%
class recall	67.68%	66.34%	

The comparison between the Decision Tree and Naive Bayes models reveals that the Decision Tree performed better in terms of accuracy and balanced metrics. Specifically, the Decision Tree achieved an accuracy of 69%, whereas Naive Bayes reached 67%.

In terms of precision, the Decision Tree scored 67.29% for the YES class (indicating emotional response) and 70.97% for the NO class, showing slightly better handling of false positives compared to Naive Bayes, which had precisions of 66.34% for YES and 67.68% for NO. This implies that the Decision Tree made fewer errors when predicting each class, particularly for the NO category.

In recall, the Decision Tree also performed better then Naive Bayes. For the YES class, it achieved a recall of 72.73%, meaning it successfully identified more cases with an emotional response, while Naive Bayes captured only 67.68%. The recall for the NO class was 65.35% for the Decision Tree, which was slightly less than its YES recall but remained effective overall, whereas Naive Bayes had a slightly higher recall for NO at 66.34%.

The Decision Tree model performed better overall because it had a good balance of precision and recall, especially when predicting YES (emotional response). Its higher accuracy (69% compared to Naive Bayes' 67%) shows that it handled the relationships between features better. Naive Bayes may have struggled because it assumes all features are independent, which wasn't true in this dataset. Because of its better accuracy and consistent results, the Decision Tree is the preferred model. However, Naive Bayes is still useful when you need quick predictions or want to see the probability of each prediction, even though it is slightly less accurate.

Issues:

PREPROCESSING 2:

Label encoding:

Another way I tried of preprocessing is **label encoding** which means I assigned numeric values to each label/category to ensure that data remains consistent. As my dataset had issues in spelling as well as I mentioned earlier. Multiple replace operators were used for this purpose

Original table:

Open in Turbo Prep Auto Model Interactive Analysis Filter (677 / 677 examples): all

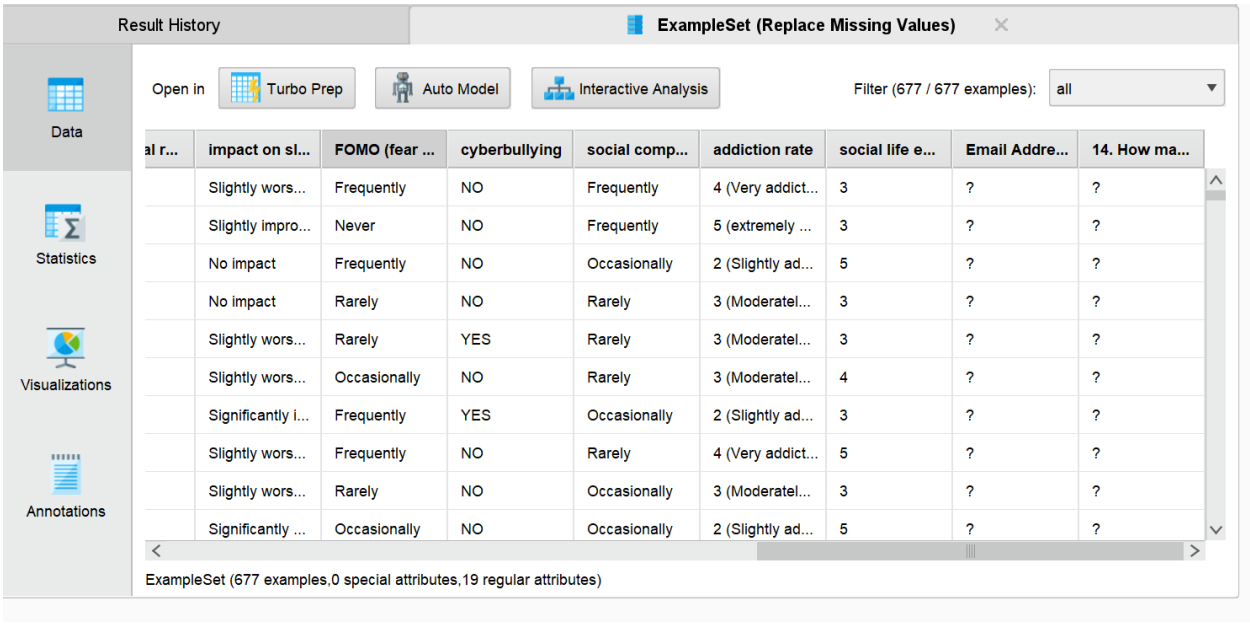
Row No.	Timestamp	Age	Gender	province	Occupation	type of cont...	hours	platfo
1	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Instag
2	?	18-24	Female	Rawalpindi	Student	Entertainmen...	7-9 hour	TikTo
3	?	18-24	Male	Taxila	Student	Entertainmen...	1-2 hour	Instag
4	?	18-24	Female	Islamabad	Student	News and cu...	1-2 hour	Instag
5	?	18-24	Female	Taxila	Student	Entertainmen...	3-4 hour	Youtu
6	?	18-24	Male	Rawalpindi	Student	Educational ...	5-6 hour	Instag
7	?	18-24	Male	punjab	Student	Entertainmen...	1-2 hour	Instag
8	?	18-24	Male	Rawalpindi	Student	Entertainmen...	7-9 hour	Instag
9	?	18-24	Female	Wah	Student	Entertainmen...	3-4 hour	TikTo
10	?	18-24	Female	Wah	Student	Entertainmen...	1-2 hour	Youtu

ExampleSet (677 examples,0 special attributes,19 regular attributes)

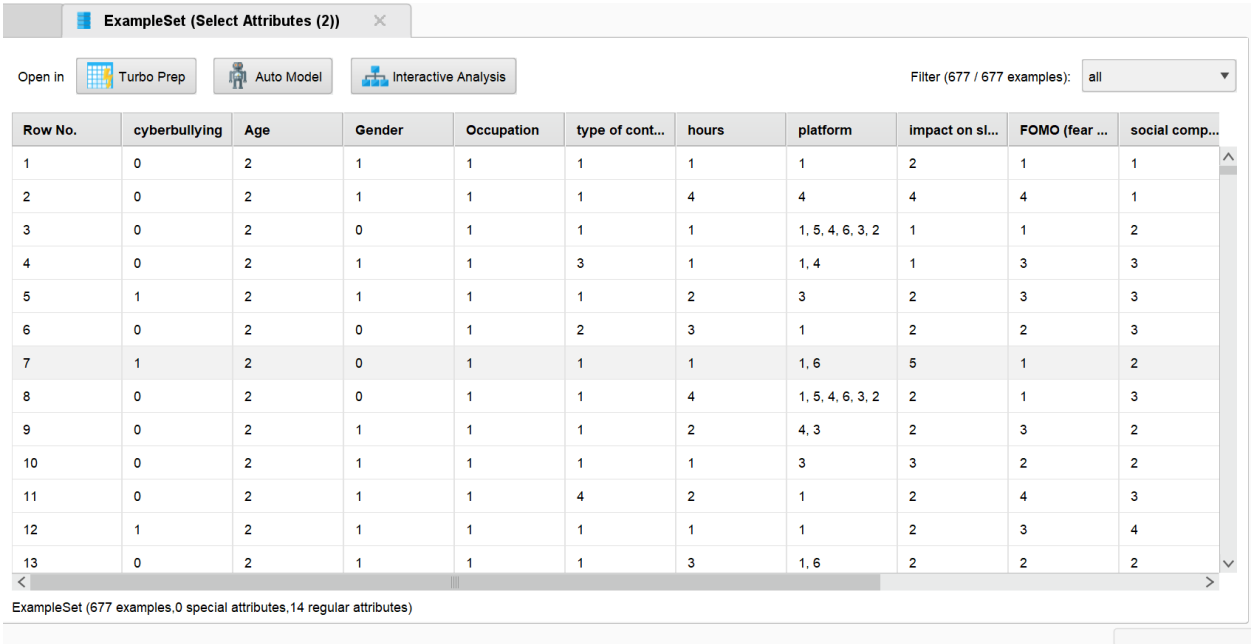
Open in Turbo Prep Auto Model Interactive Analysis Filter (677 / 677 examples): all

emotional r...	impact on sl...	FOMO (fear ...	cyberbullying	social comp...	addiction rate	no of platfor...	social life e...	sleep
YES	Slightly wors...	Frequently	NO	Frequently	4 (Very addict...	2	3	1.5 ho
YES	Slightly impro...	Never	NO	Frequently	5 (extremely ...	4	3	6
NO	No impact	Frequently	NO	Occasionally	2 (Slightly ad...	4	5	24 ho
YES	No impact	Rarely	NO	Rarely	3 (Moderatel...	2	3	0
YES	Slightly wors...	Rarely	YES	Rarely	3 (Moderatel...	2	3	21
NO	Slightly wors...	Occasionally	NO	Rarely	3 (Moderatel...	more than 5	4	5
NO	Significantly i...	Frequently	YES	Occasionally	2 (Slightly ad...	2	3	5
YES	Slightly wors...	Frequently	NO	Rarely	4 (Very addict...	3	5	26
NO	Slightly wors...	Rarely	NO	Occasionally	3 (Moderatel...	3	3	2
NO	Significantly ...	Occasionally	NO	Occasionally	2 (Slightly ad...	2	5	8

ExampleSet (677 examples,0 special attributes,19 regular attributes)



Preprocessed table:

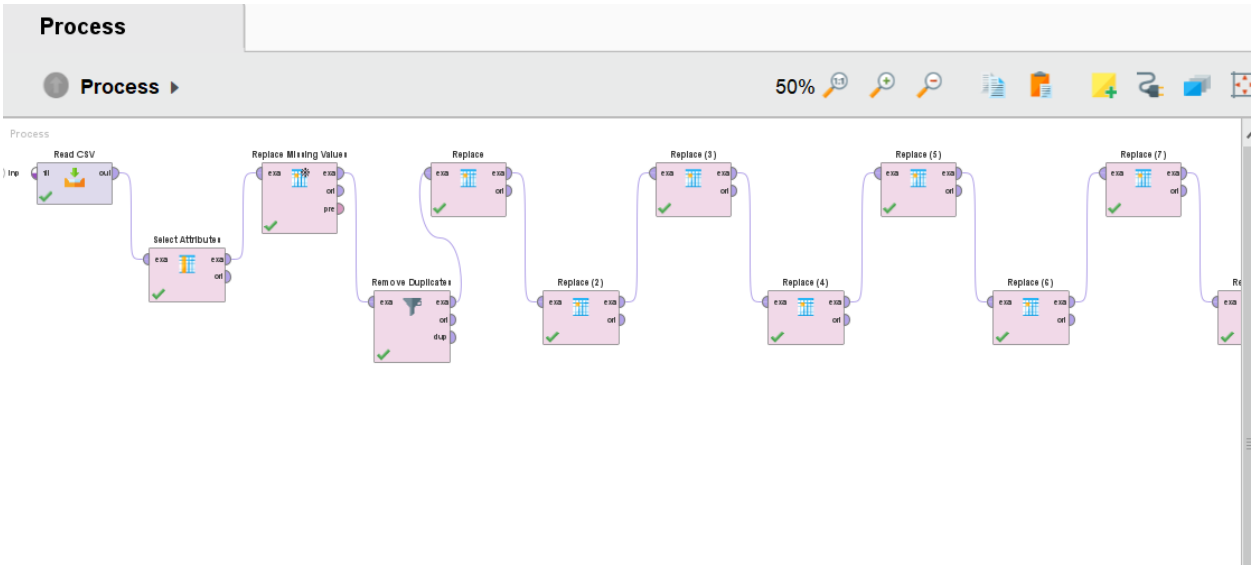


ExampleSet (Select Attributes (2))

Open in Turbo Prep Auto Model Interactive Analysis Filter (677 / 677 examples): all

Occupation	type of cont...	hours	platform	impact on sl...	FOMO (fear ...	social comp...	addiction rate	no of platfor...	emotional r...	social life e...
1	1	1	1	2	1	1	4	2	YES	3
1	1	4	4	4	4	1	5	4	YES	3
1	1	1	1, 5, 4, 6, 3, 2	1	1	2	2	4	NO	5
1	3	1	1, 4	1	3	3	3	2	YES	3
1	1	2	3	2	3	3	3	2	YES	3
1	2	3	1	2	2	3	3	5	NO	4
1	1	1	1, 6	5	1	2	2	2	NO	3
1	1	4	1, 5, 4, 6, 3, 2	2	1	3	4	3	YES	5
1	1	2	4, 3	2	3	2	3	3	NO	3
1	1	1	3	3	2	2	2	2	NO	5
1	4	2	1	2	4	3	3	2	NO	3
1	1	1	1	2	3	4	3	5	YES	3
1	1	3	1, 6	2	2	2	3	3	YES	3

ExampleSet (677 examples,0 special attributes,14 regular attributes)



Age:

11-18: 1

18-24: 2

24-34: 3

34-54: 4

Gender:

Yes : 1

No : 0

Occupation:

Student: 1

Employed : 2

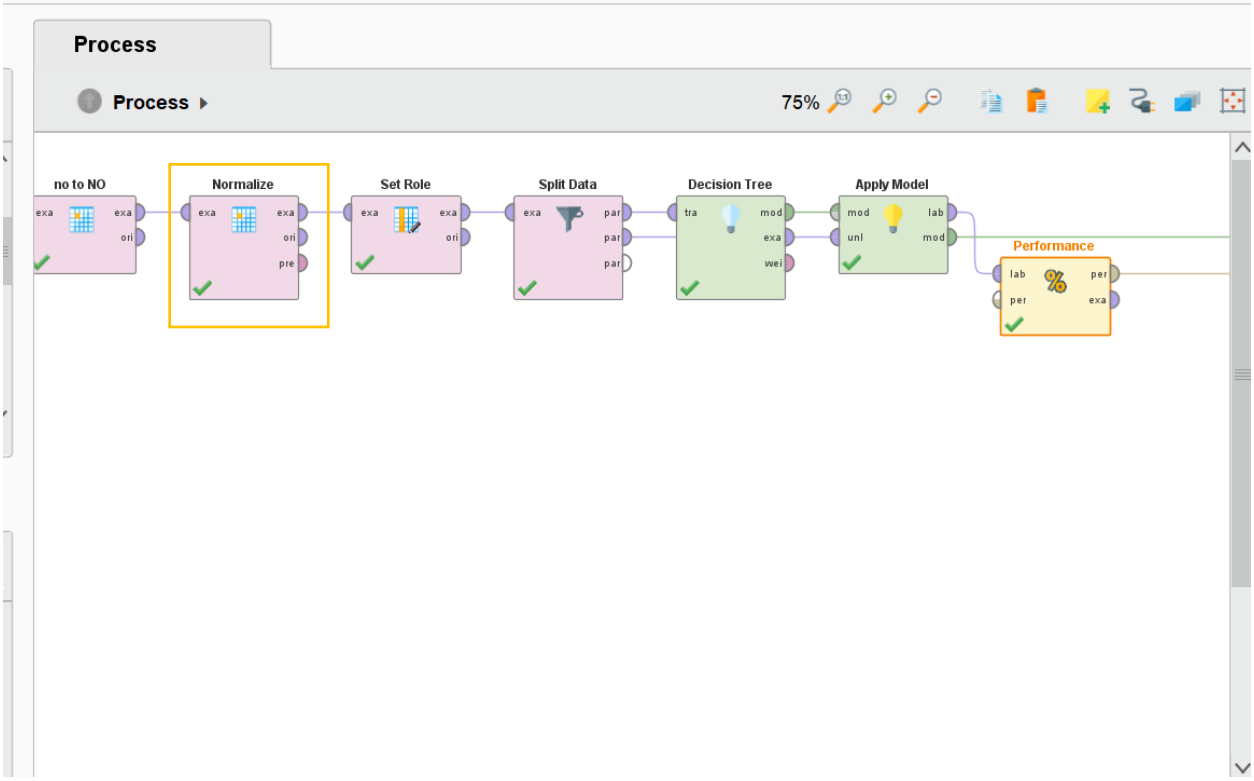
Unemployed: 3

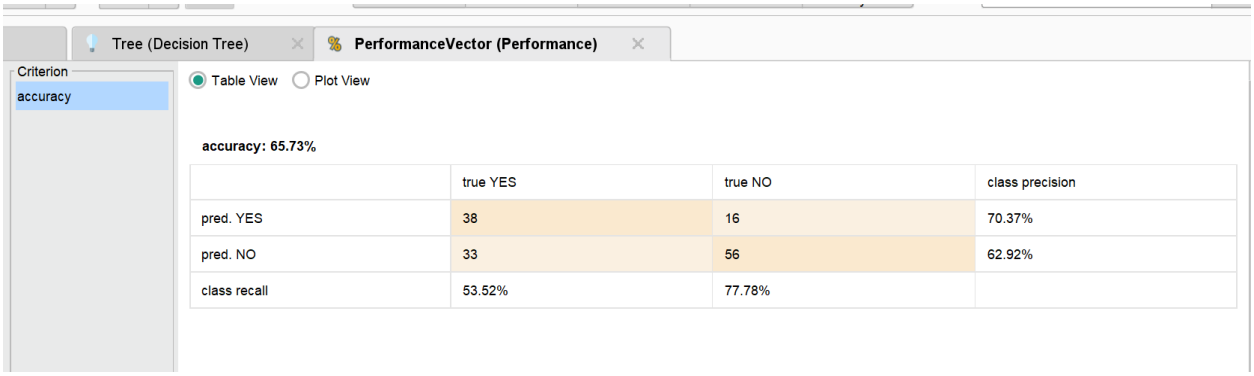
Housewife: 4

Retired: 5

Similarly all others are done this way .

After that I normalized numerical columns and applied decision tree.




Tree (Decision Tree)
PerformanceVector (Performance)
Criterion
accuracy
Table View
Plot View
accuracy: 65.73%
true YES
true NO
class precision
pred. YES
38
16
70.37%
pred. NO
33
56
62.92%
class recall
53.52%
77.78%