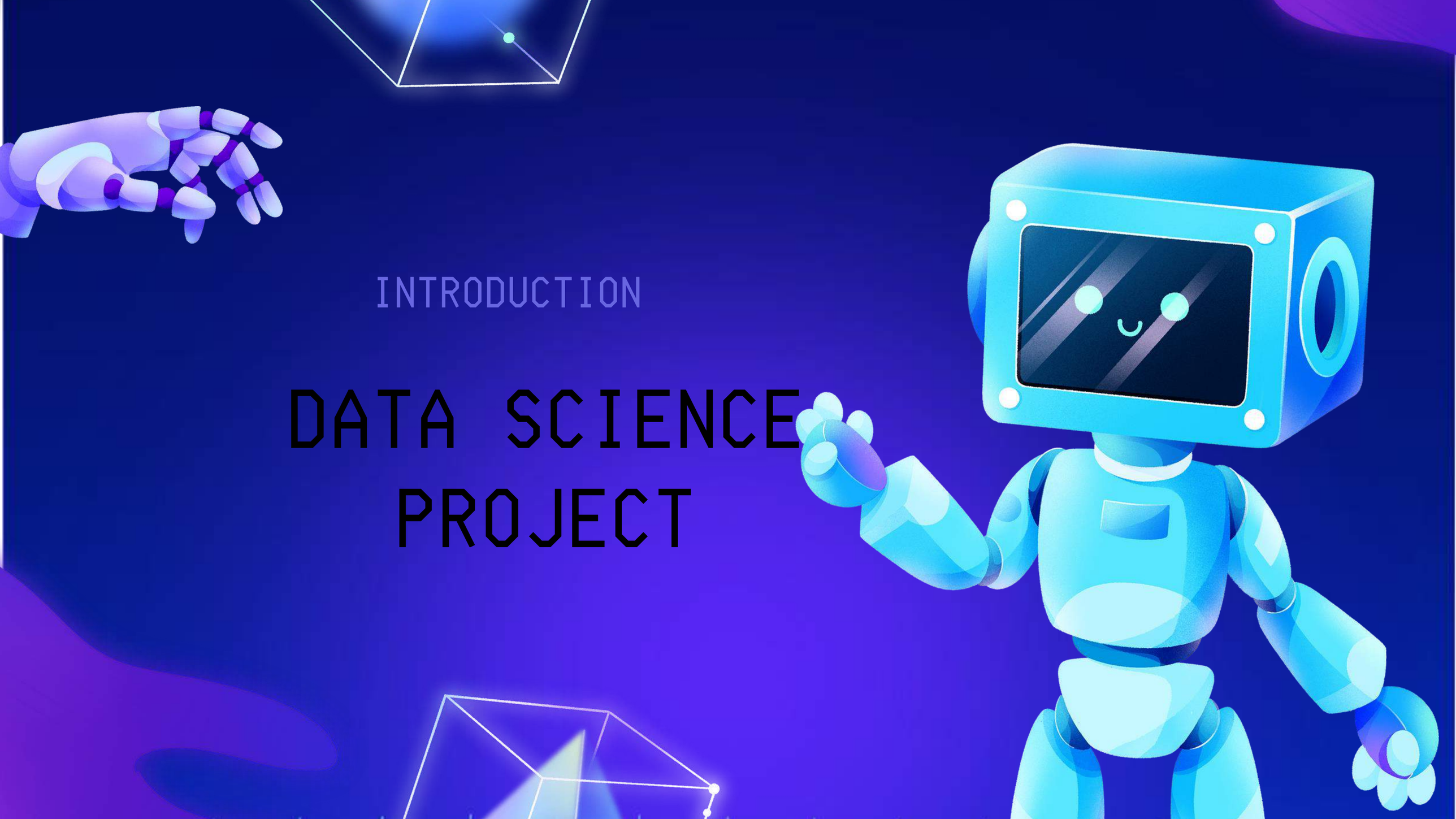
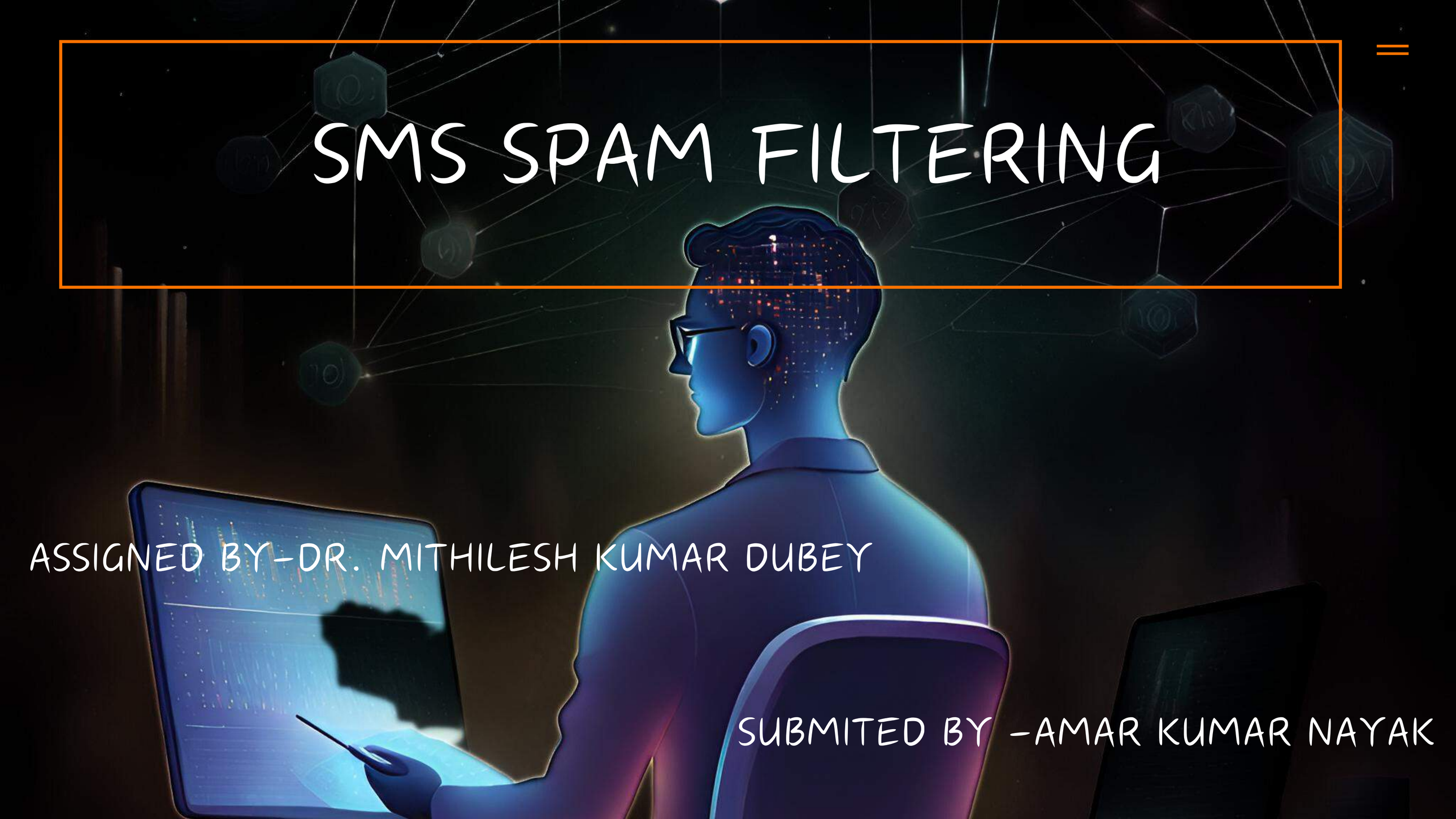
INTRODUCTION

DATA SCIENCE PROJECT



SMS SPAM FILTERING



ASSIGNED BY-DR. MITHILESH KUMAR DUBEY

SUBMITED BY -AMAR KUMAR NAYAK

Introduction

The proliferation of short message service (SMS) communication has become an integral part of daily life, with users relying on this medium for personal and professional communication. However, the prevalence of SMS spam poses a significant challenge, leading to potential inconveniences and security risks for users. In this project, we aim to address this issue through the implementation of an SMS spam filtering system using RapidMiner.

HARDWARE REQUIREMENTS

- Processor: A multi-core processor with sufficient processing power for data preprocessing and machine learning tasks.
- Memory (RAM): Minimum of 8GB RAM for handling moderate-sized datasets and model training. Larger datasets may require more RAM.
- Storage: Adequate storage space for storing datasets, RapidMiner project files, and the trained machine learning model. SSDs are recommended for faster data access.

SOFTWARE REQUIREMENTS

Software Requirements:

• Operating System: Windows, macOS, or Linux. Ensure that the chosen operating system is compatible with the RapidMiner version you plan to use.

• RapidMiner: Install the latest version of RapidMiner Studio, the data science platform that will be used for data preprocessing, feature engineering, model development, and evaluation. Download from RapidMiner's official website.

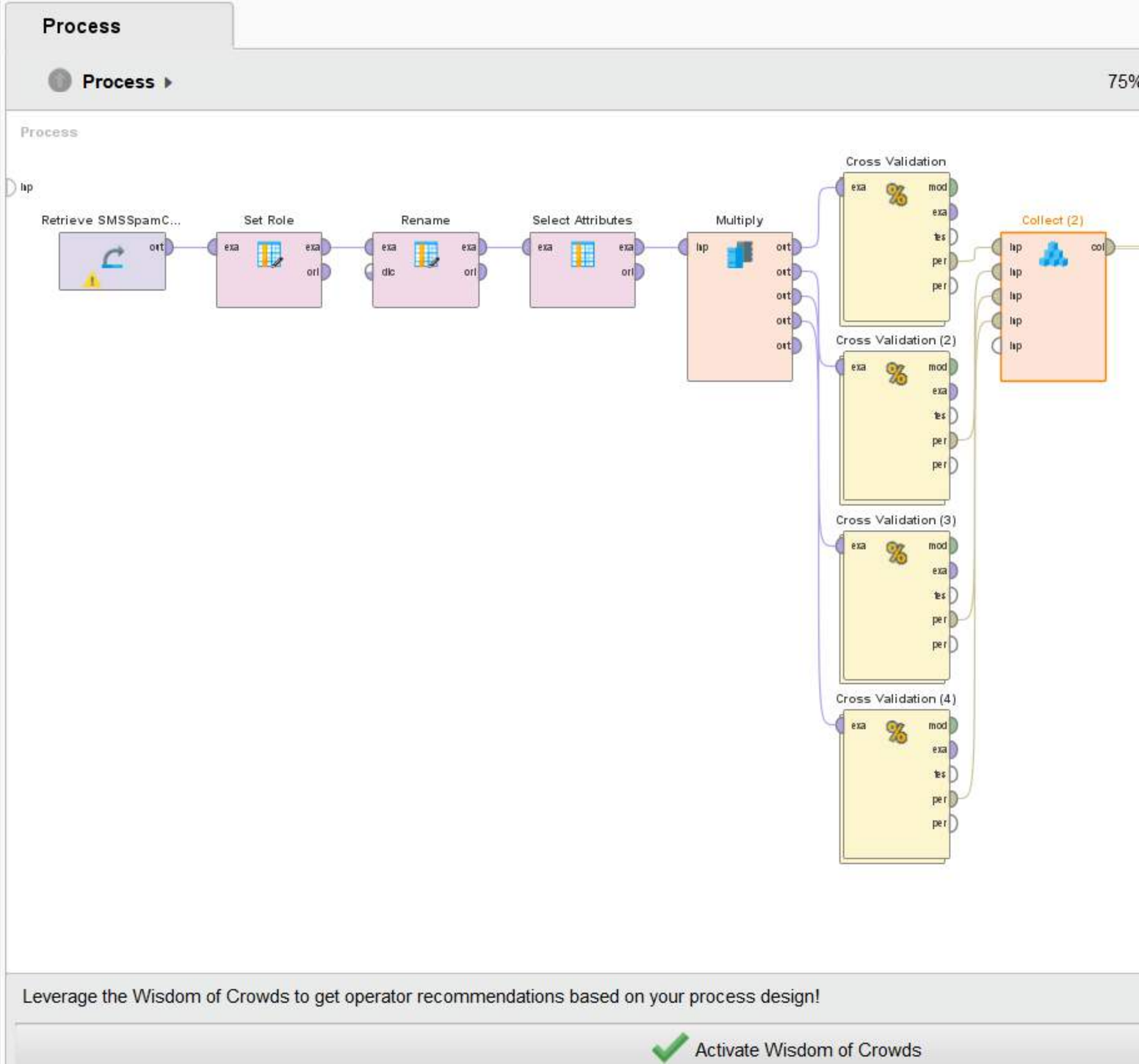
• Programming Language: While not strictly necessary, a basic understanding of Python may be beneficial for additional customization and scripting within RapidMiner.

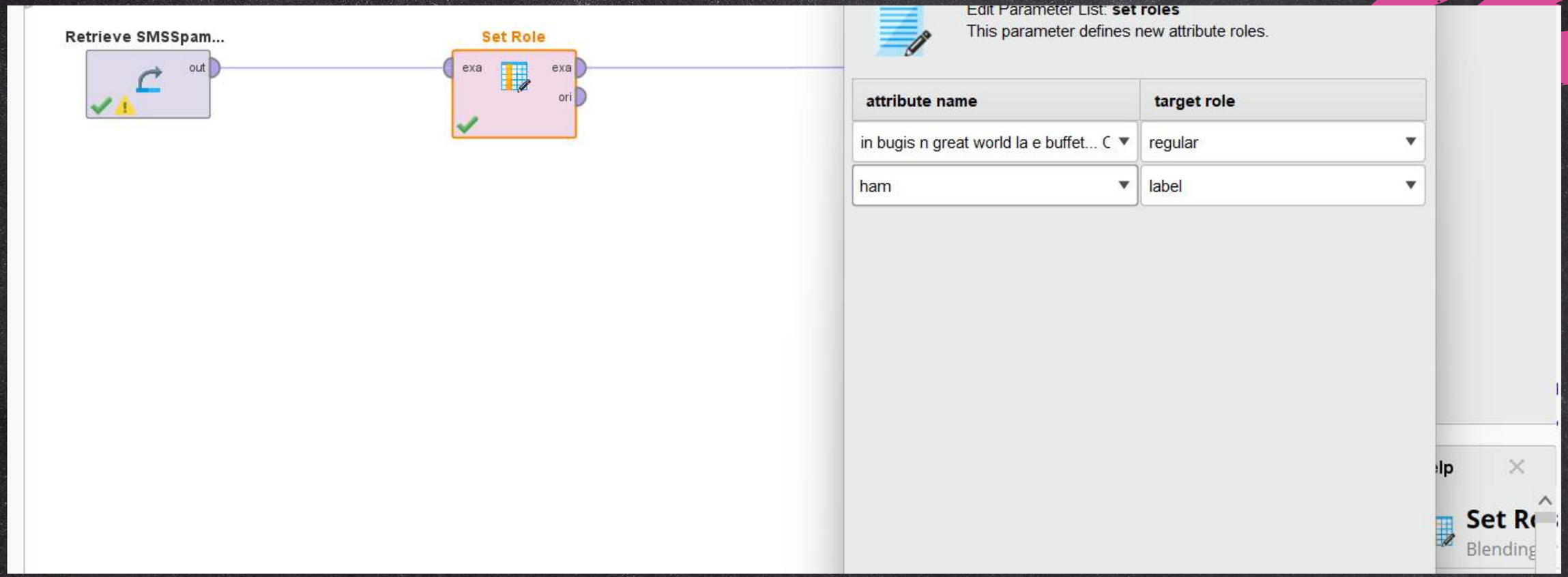
• Python Libraries (Optional): Depending on the specific needs of your project, you might need Python libraries for additional data preprocessing or analysis. Common libraries include NumPy, Pandas, and scikit-learn.

• Text Editor or IDE: A text editor or integrated development environment (IDE) for any additional scripting or code editing. Examples include Visual Studio Code, Atom, or PyCharm.

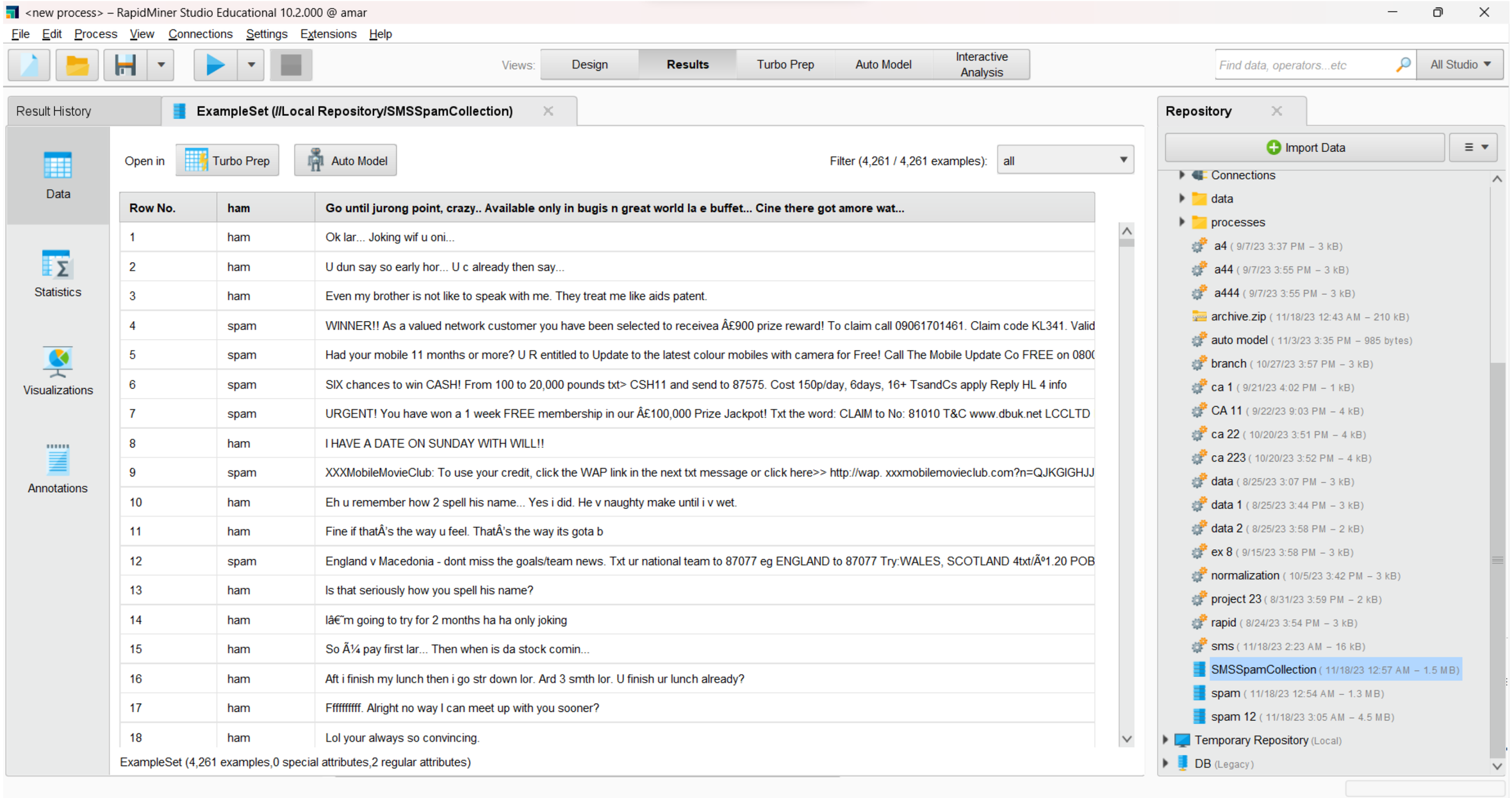
OPERATORS

- SET ROLE
- RENAME
- SELECT ATTRIBUTES
- MULTIPLY
- CROSS VALIDATION(4)
- COLLECT



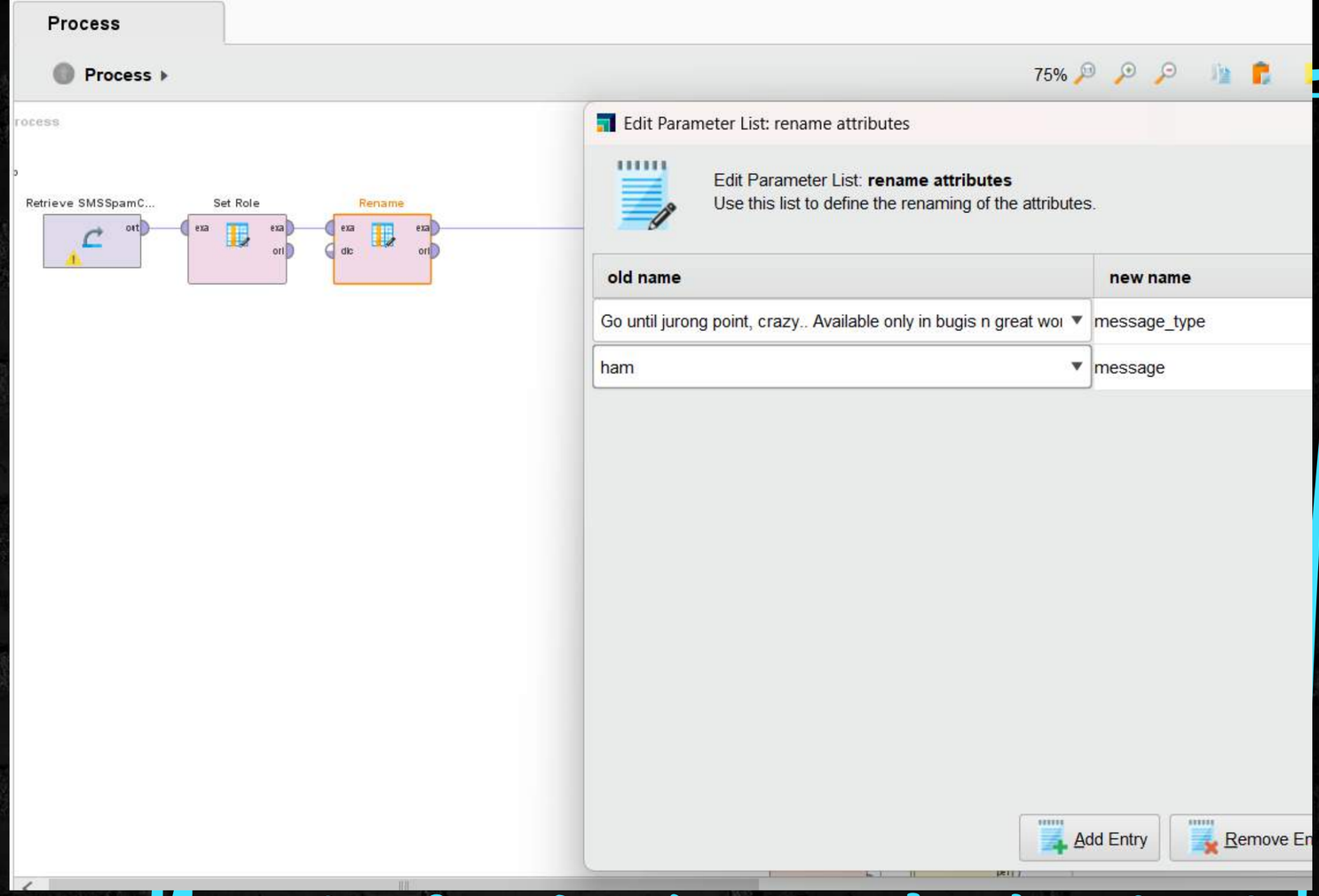


I ALREADY DOWNLOAED SMS SPAM COLLECTION DATA SET AND I WILL JUST CLEARLY DRAG IT OVER TO THIS END HERE AND TO MAKE SURE THAT IT'S WORKING WELL . THEN TO USE THE SET ROLE OPERATOR WHICH IS THE RULE OVER ATTRIBUTE DESCRIBED HOW OTHER OPERATORS HANDLE THIS ATTRIBUTE. THIS IS WHY I AM USING A SIGNAL OPERATOR BECAUSE I WANT TO MAKE IT UNIQUE HOW ARE THE ATTRIBUTES OR OTHER PARAMETERS HANDLE THIS ATTRIBUTES. SO DRAG IT AND CHANGES ATTRIBUTES NAME AND TARGET ALSO AS I MENTIONED IN PICTURE.


<new process> – RapidMiner Studio Educational 10.2.000 @ amar
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
ExampleSet (//Local Repository/SMSSpamCollection)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (4,261 / 4,261 examples): all
Row No. ham Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
1 ham Ok lar... Joking wif u oni...
2 ham U dun say so early hor... U c already then say...
3 ham Even my brother is not like to speak with me. They treat me like aids patent.
4 spam WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid
5 spam Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 080(
6 spam SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
7 spam URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD
8 ham I HAVE A DATE ON SUNDAY WITH WILL!!
9 spam XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJ
10 ham Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
11 ham Fine if thatÂ's the way u feel. ThatÂ's the way its gota b
12 spam England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ã¹1.20 POB
13 ham Is that seriously how you spell his name?
14 ham Iâ€˜m going to try for 2 months ha ha only joking
15 ham So Ã¼ pay first lar... Then when is da stock comin...
16 ham Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
17 ham Ffffffffff. Alright no way I can meet up with you sooner?
18 ham Lol your always so convincing.
ExampleSet (4,261 examples,0 special attributes,2 regular attributes)
Repository
Import Data
Connections
data
processes
a4 (9/7/23 3:37 PM – 3 kB)
a44 (9/7/23 3:55 PM – 3 kB)
a444 (9/7/23 3:55 PM – 3 kB)
archive.zip (11/18/23 12:43 AM – 210 kB)
auto model (11/3/23 3:35 PM – 985 bytes)
branch (10/27/23 3:57 PM – 3 kB)
ca 1 (9/21/23 4:02 PM – 1 kB)
CA 11 (9/22/23 9:03 PM – 4 kB)
ca 22 (10/20/23 3:51 PM – 4 kB)
ca 223 (10/20/23 3:52 PM – 4 kB)
data (8/25/23 3:07 PM – 3 kB)
data 1 (8/25/23 3:44 PM – 3 kB)
data 2 (8/25/23 3:58 PM – 2 kB)
ex 8 (9/15/23 3:58 PM – 3 kB)
normalization (10/5/23 3:42 PM – 3 kB)
project 23 (8/31/23 3:59 PM – 2 kB)
rapid (8/24/23 3:54 PM – 3 kB)
sms (11/18/23 2:23 AM – 16 kB)
SMSSpamCollection (11/18/23 12:57 AM – 1.5 MB)
spam (11/18/23 12:54 AM – 1.3 MB)
spam 12 (11/18/23 3:05 AM – 4.5 MB)
Temporary Repository (Local)
DB (Legacy)

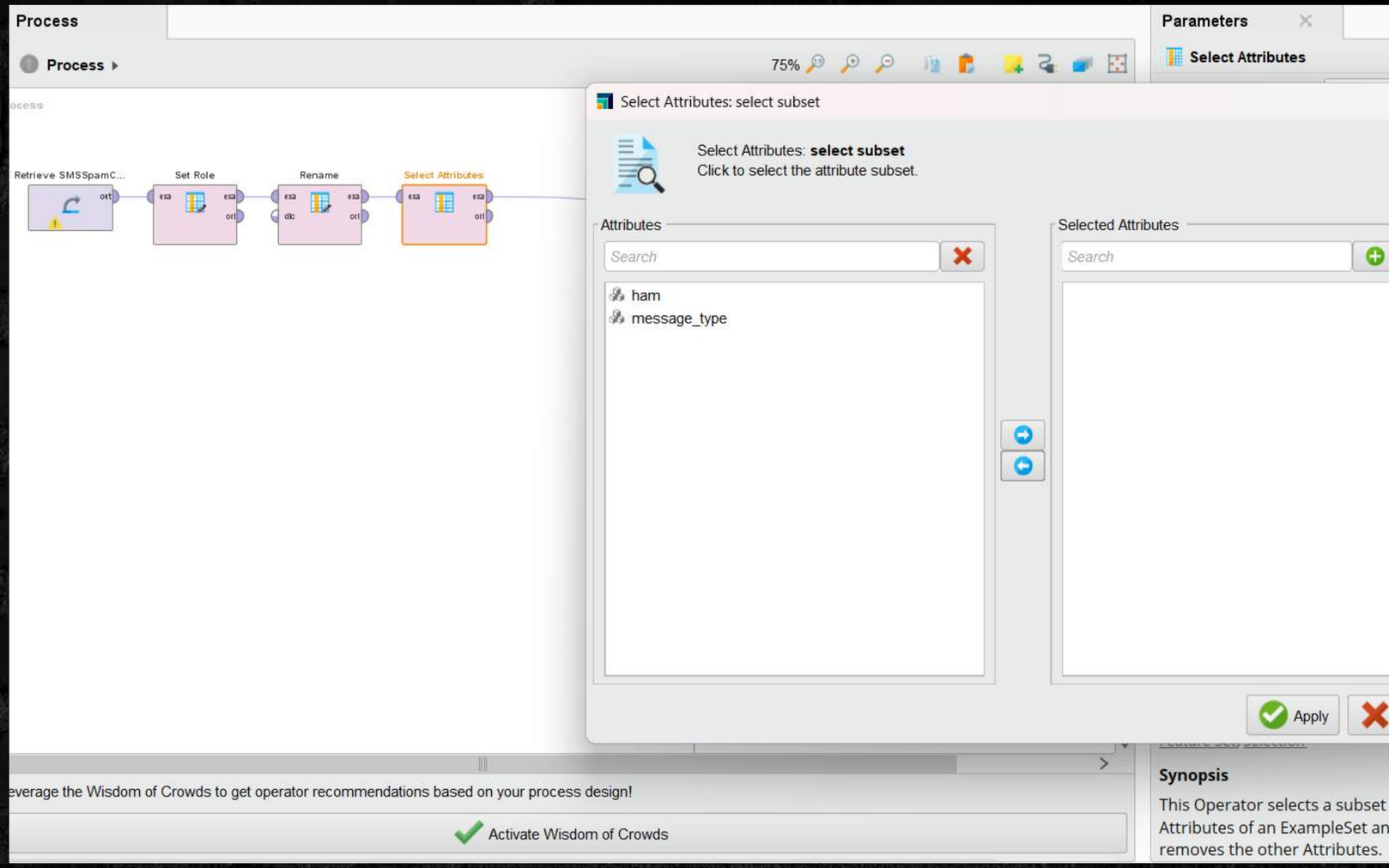
RENAME

Next operator is the rename operator so drag and drop here and then some changes do in parameters. click rename attributes and changes old name to new name. I have do this step below picture



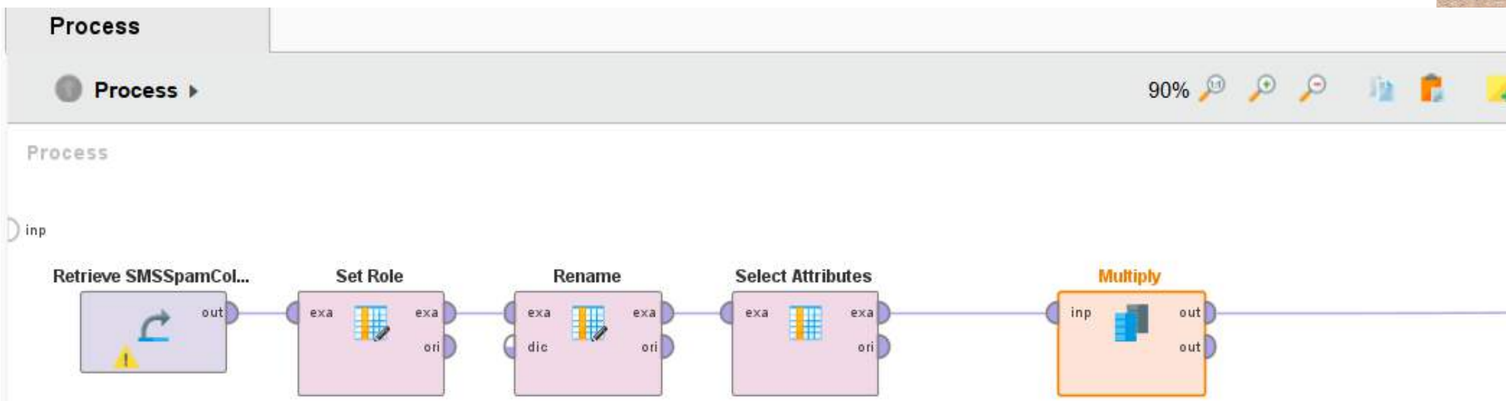
SELECT ATTRIBUTE

So next operator is select attribute operator.Click attribute filter type and select a subset then click select subset -select attribute . A pop window open . Then select all attributes and take it right side i.e select attribute side then click apply.



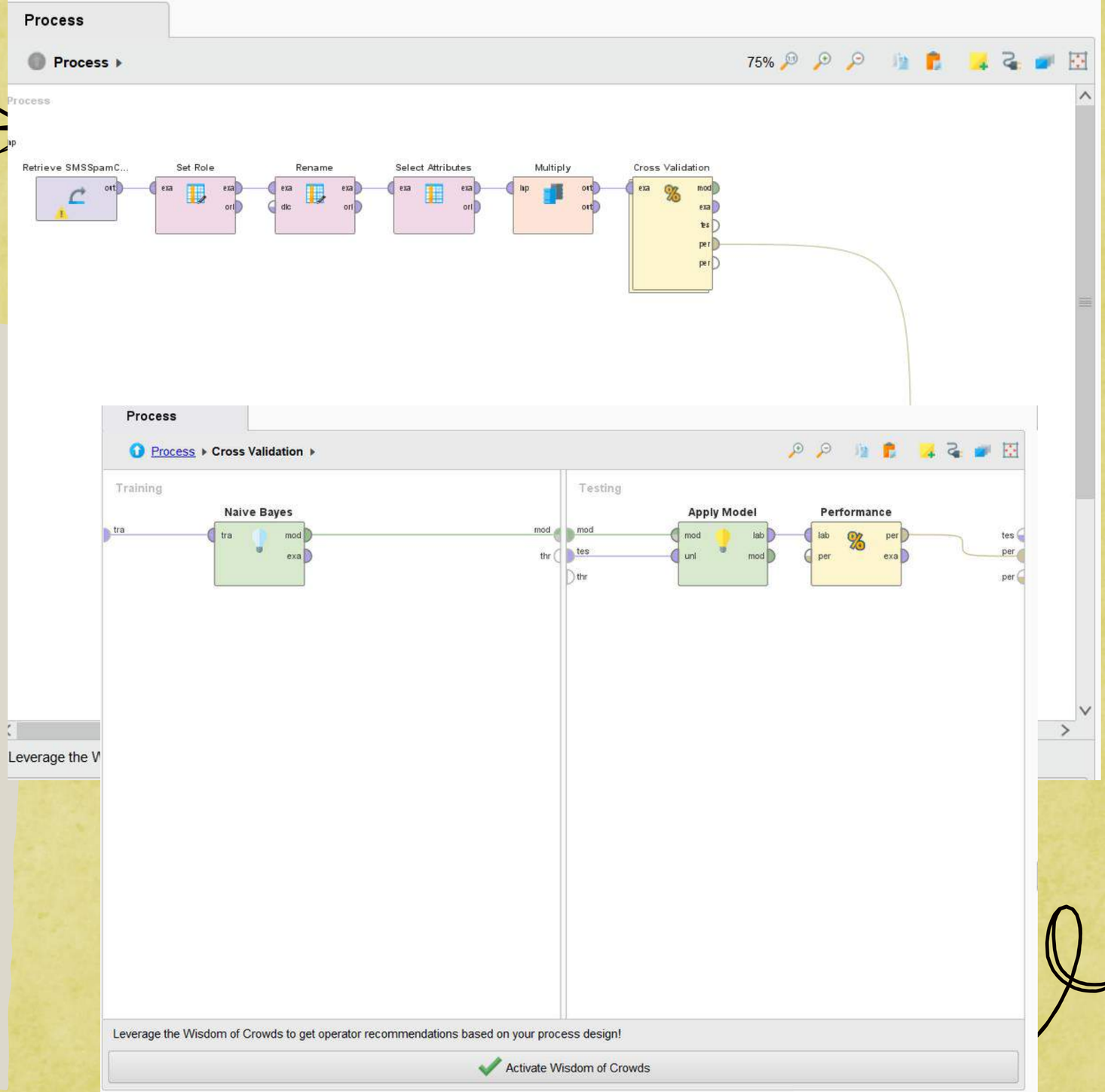
MULTIPLY

Next operator is multiply operator. So drag this operator and drop. Multiply operator creates copies of a operator takes the rapid miner object from the import board copies of it to the output port so since we are using different classifiers .This is going to be quite and important parts of our project so just like that right in there so the example ports will be connected to input ports and now it has output ports .



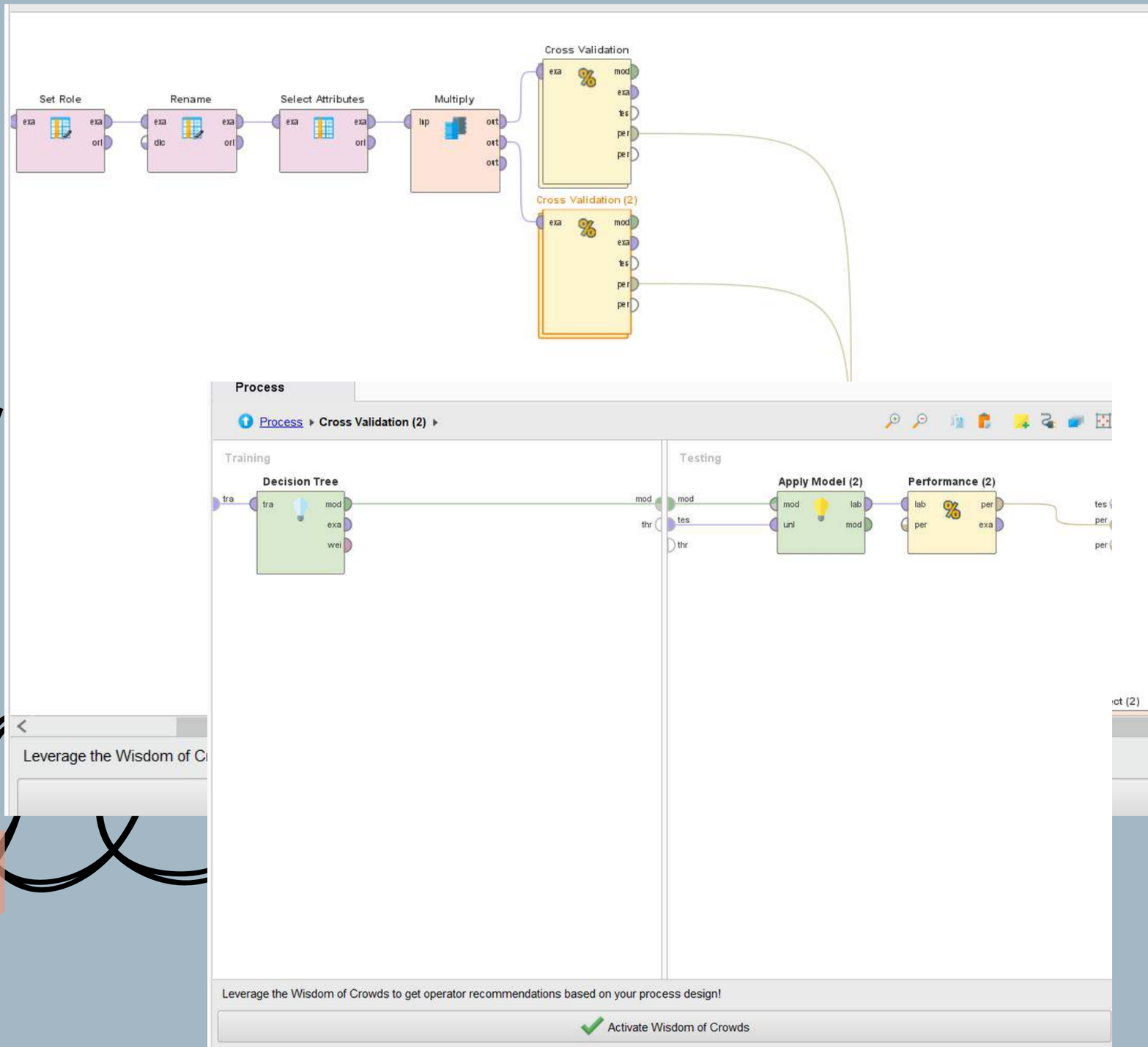
CROSS VALIDATION

Next I take training and testing datasets i.e I use cross validation operator. This operator creates a training and testing database I double click on it and have sub process in this process that's where I apply classifier which is the nine things classifier. I put the naïve bayes classifier the training porter then I use apply model operator which applies a model on the example sets . Then I use performance operator and connect performance input to performance porter. Then click go back process



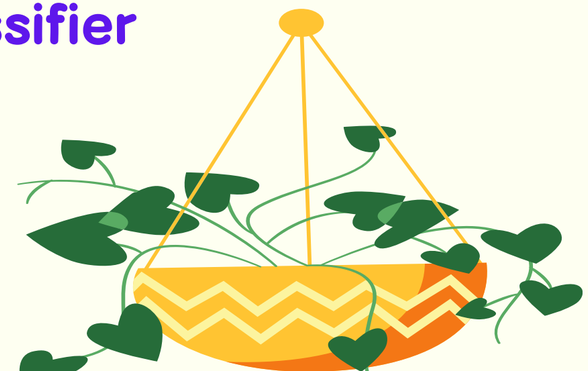
CROSS VALIDATION 2

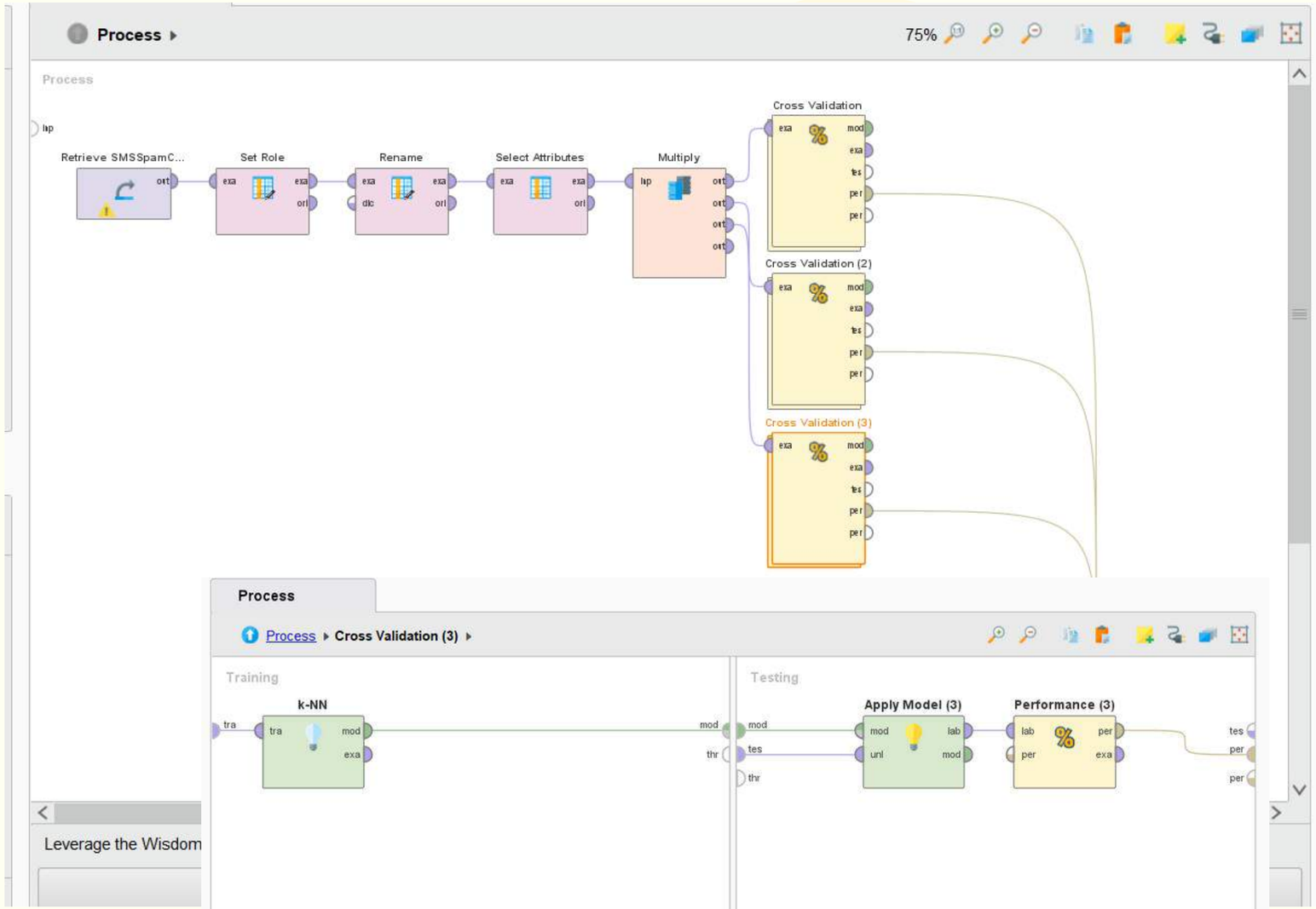
Next I'm copy and paste cross validation operator because I am going to pretty much use the same method only different classifiers. I connect the example to the output and the performance here and then change to the classifier. I use decision tree operator.



cross validation (3)

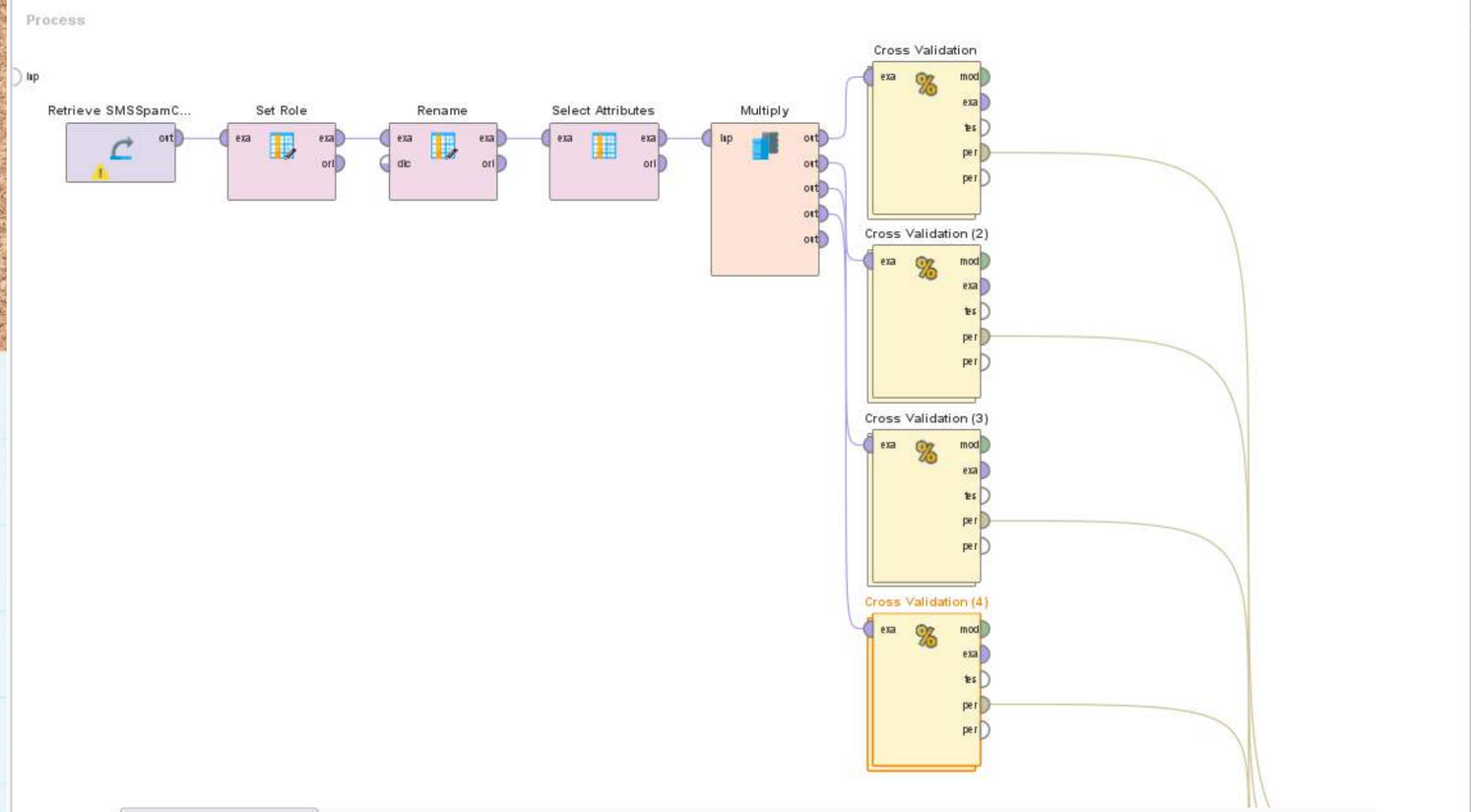
Next I'm again copy and paste cross validation operator because I am going to again use the same method only different classifiers. I connect the example to the output and the performance here and then change to the classifier. Then cliclk double click cross validation operator then change decision tree classifier to k-NN classifier

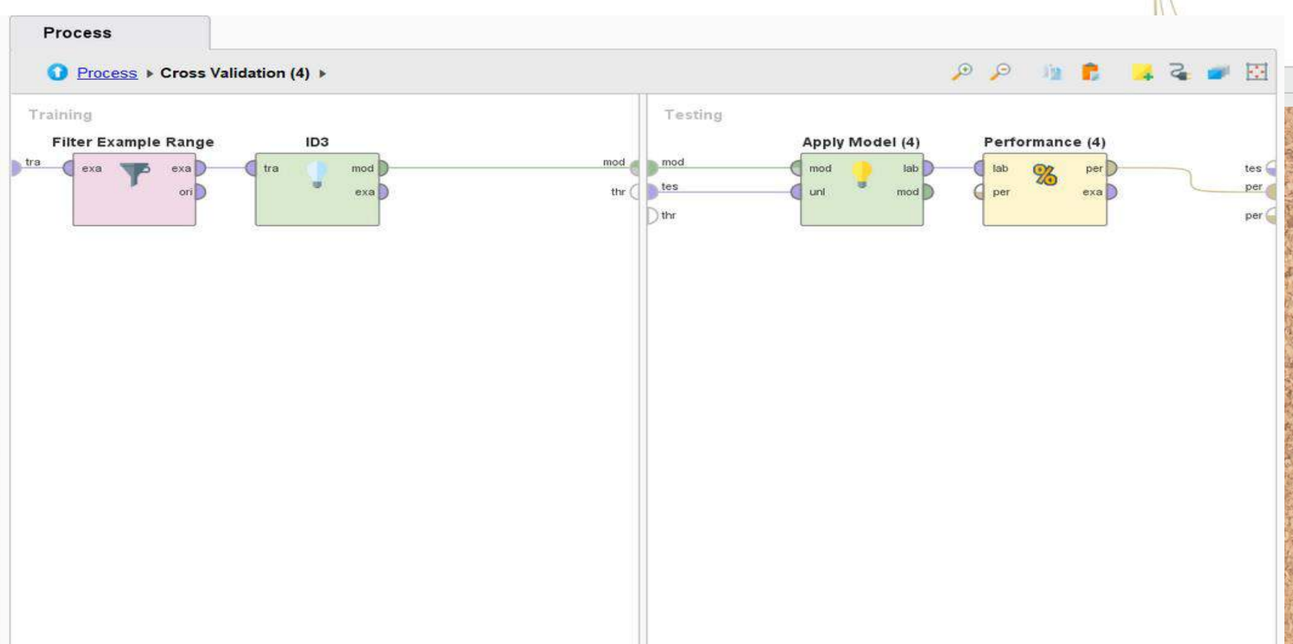




cross validation (4)

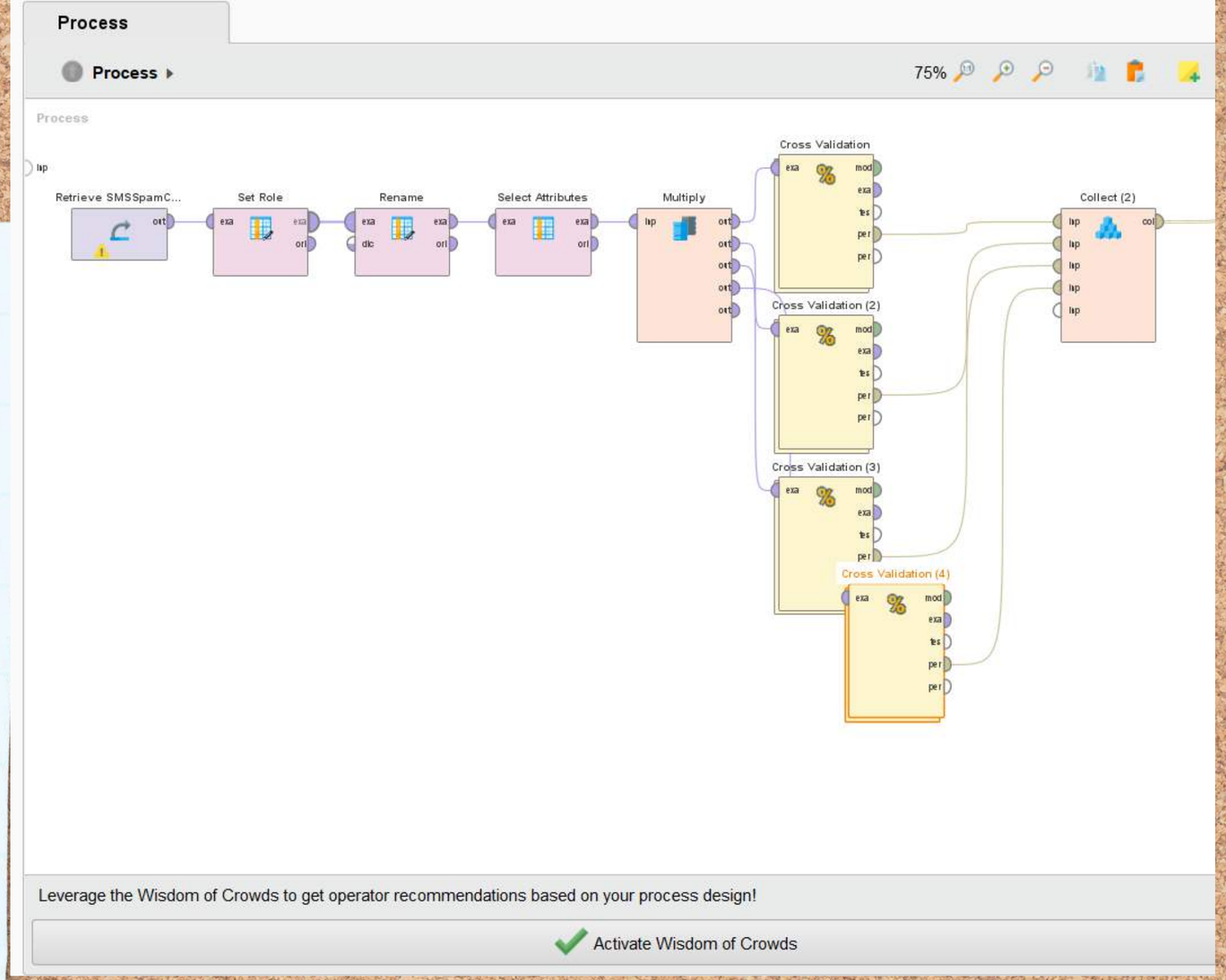
Next I'm again copy and paste cross validation operator because I am going to again use the same method only different classifiers. I connect the example to the output and the performance here and then change to the classifier. Then double click cross validation operator then change K-NN operator to filter example operator and ID3 operator connect as I have designed in below picture.



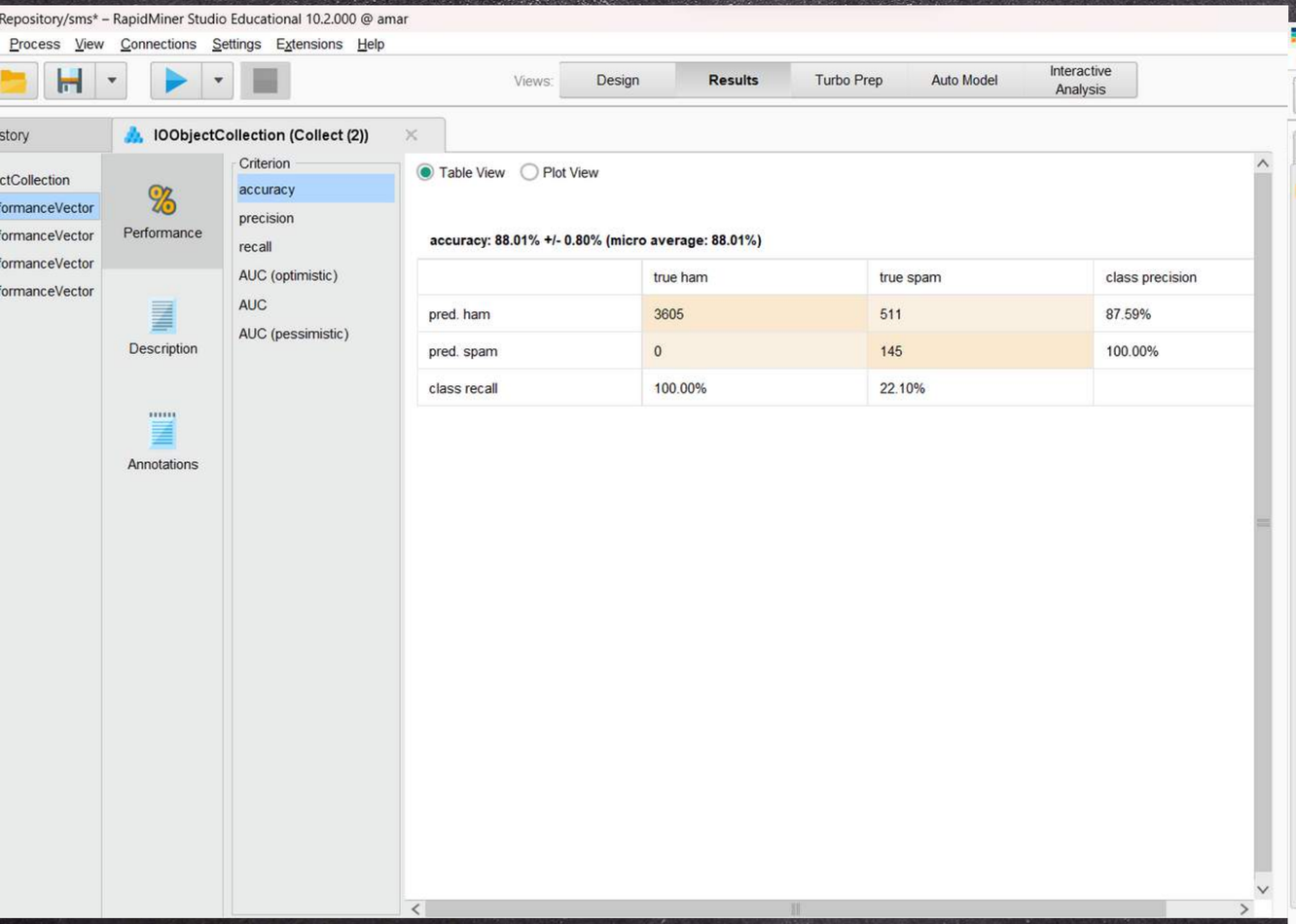


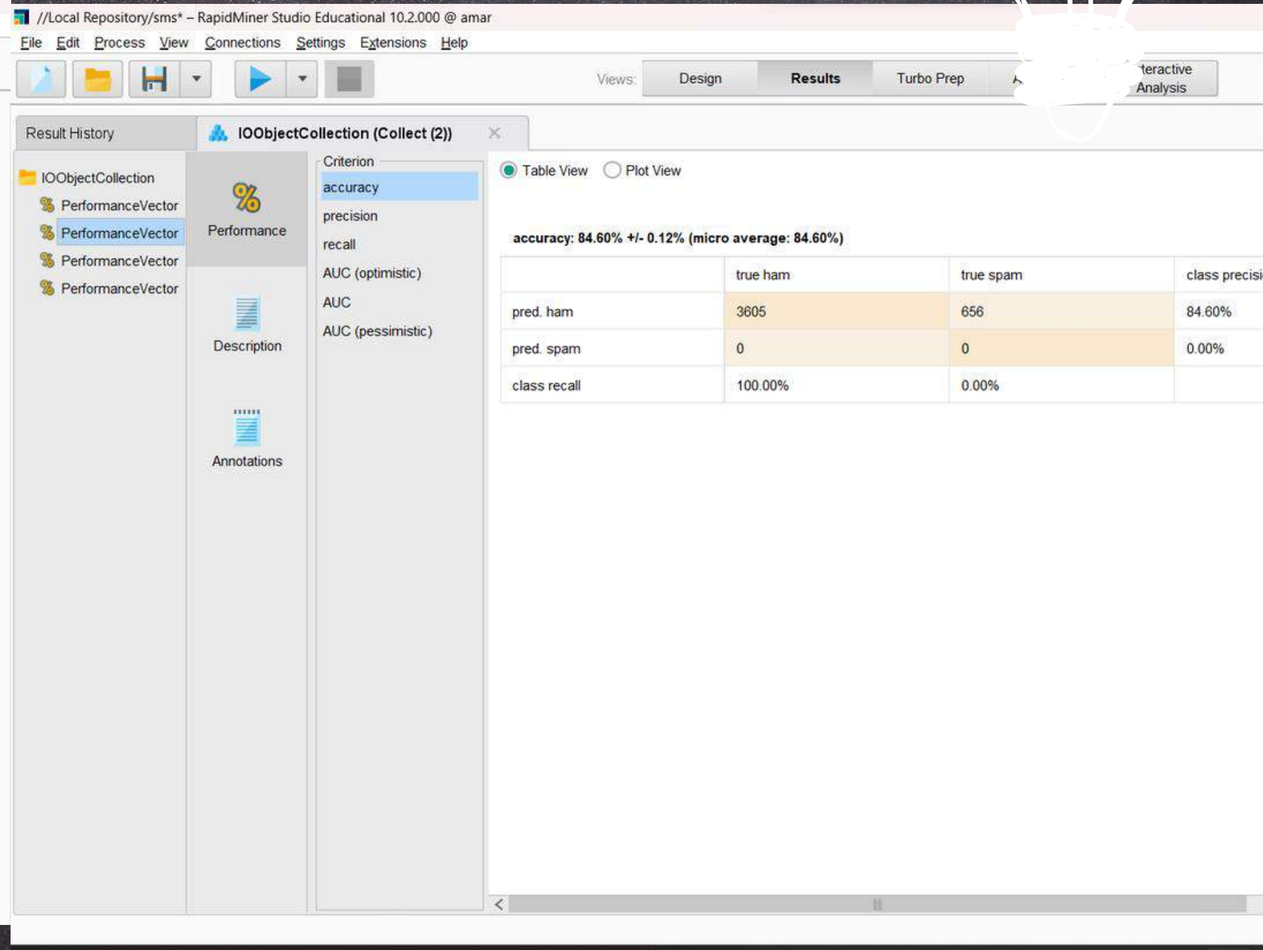
collect

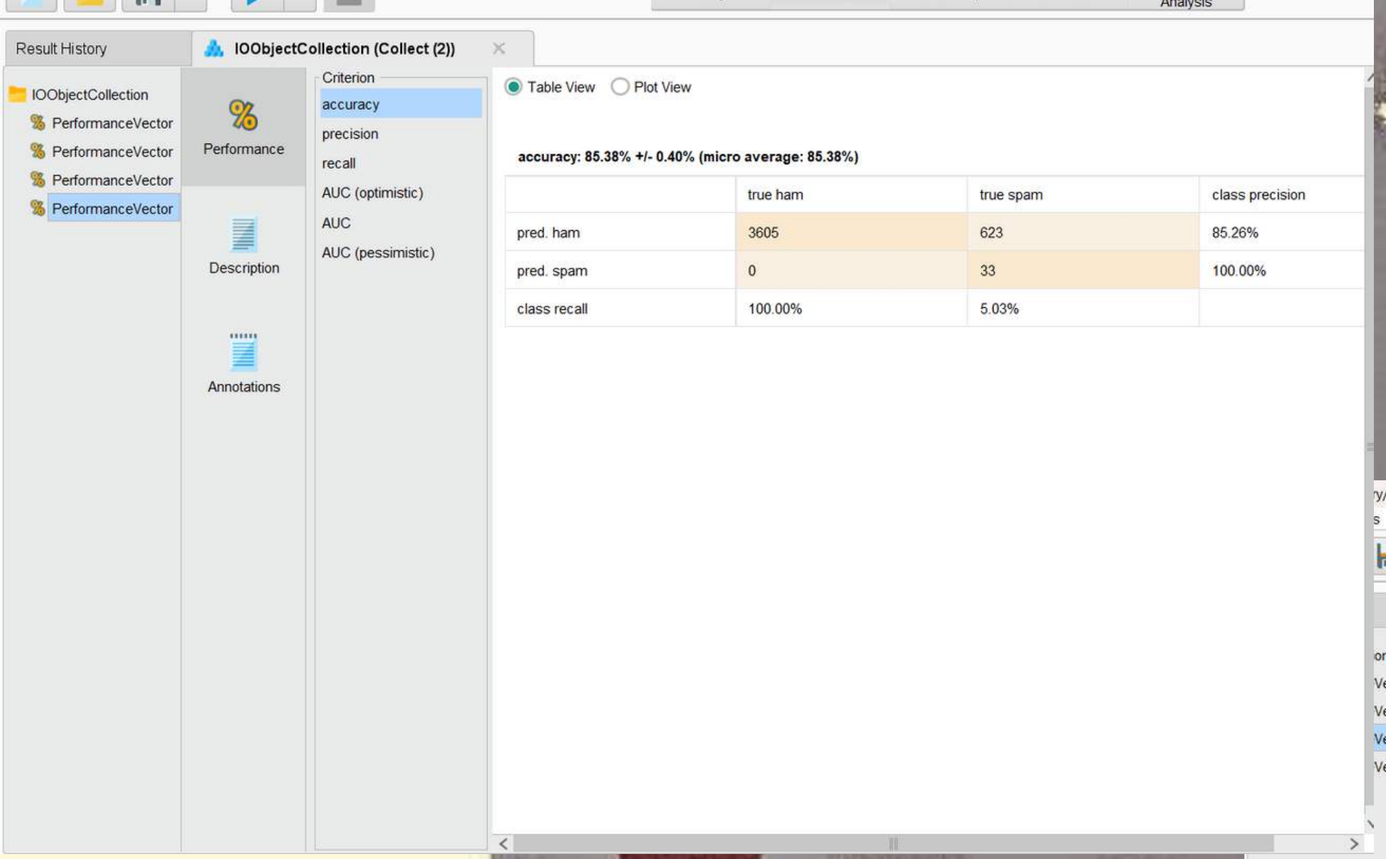
Another collect operator. The collect operator will be able to collect both methodologies so what they collect overrated does is it collects or combines multiple inputs into a single collection that way things don't get missing. So connect all 4 cross validation operator performance ports to inputs of collect operator. So everything can be down in one column so everything has been collected in one that things don't get too noisy and then when play it we can easily compare . Finally all dataset has been arranged and run the process

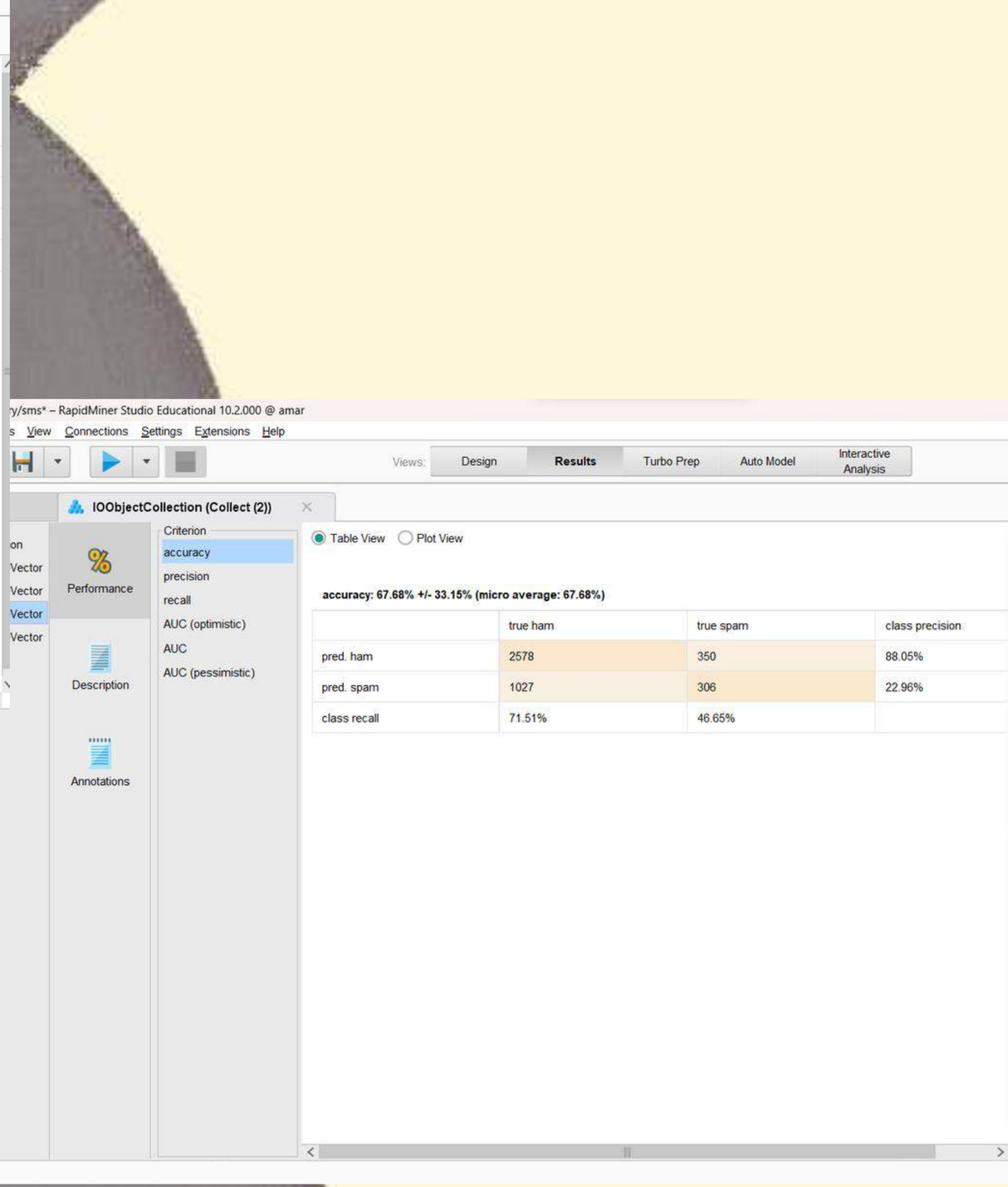


YOU CAN SEE ACCURACY OF PREDICTION OF HAM AND PREDICTION OF SPAM ARE SHOW DIFFERENT PERFORMANCE VECTOR



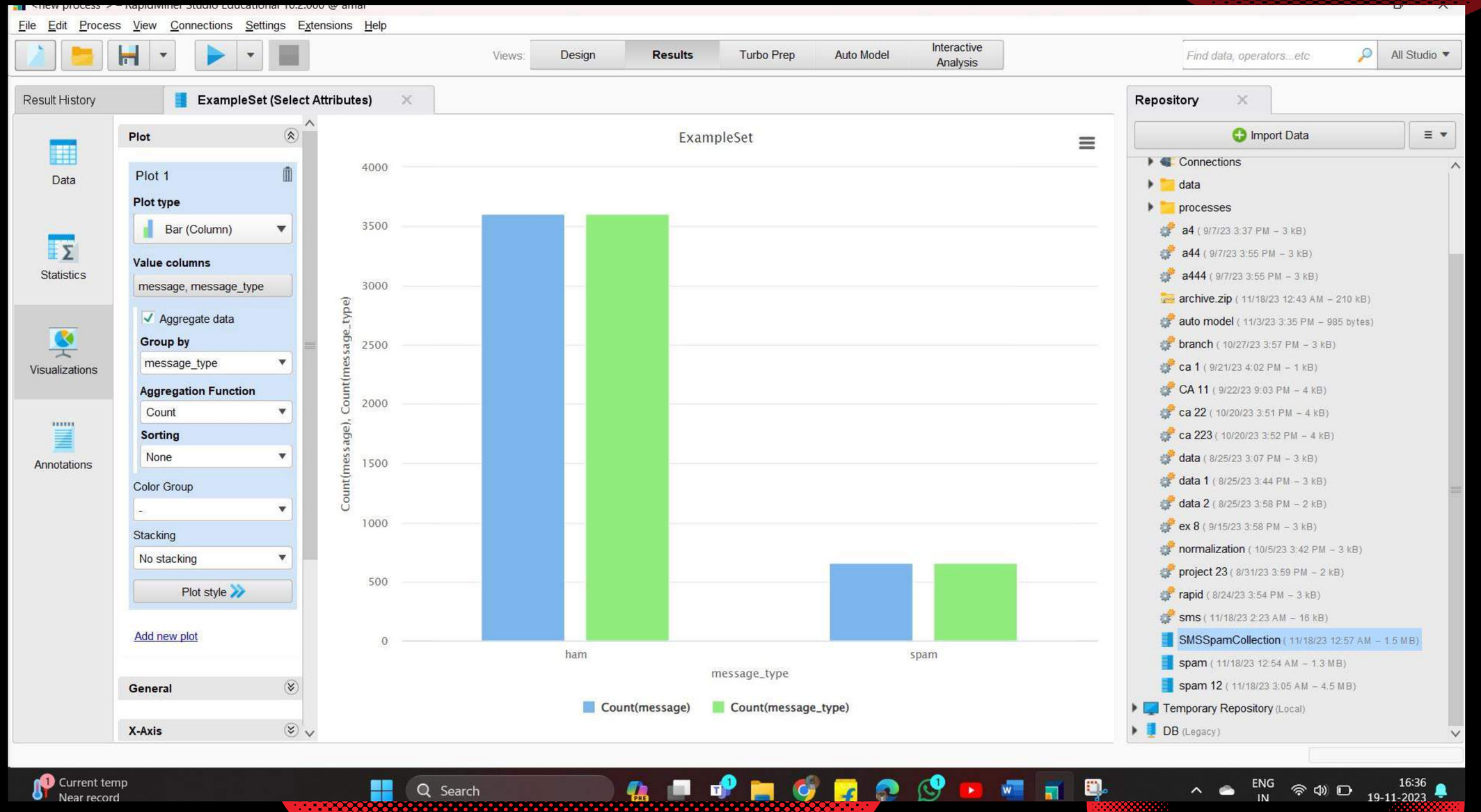



Result History
IOObjectCollection (Collect (2))
IOObjectCollection
PerformanceVector
PerformanceVector
PerformanceVector
PerformanceVector
Performance
Description
Annotations
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
Table View
Plot View
accuracy: 85.38% +/- 0.40% (micro average: 85.38%)
true ham
true spam
class precision
pred. ham
3605
623
85.26%
pred. spam
0
33
100.00%
class recall
100.00%
5.03%

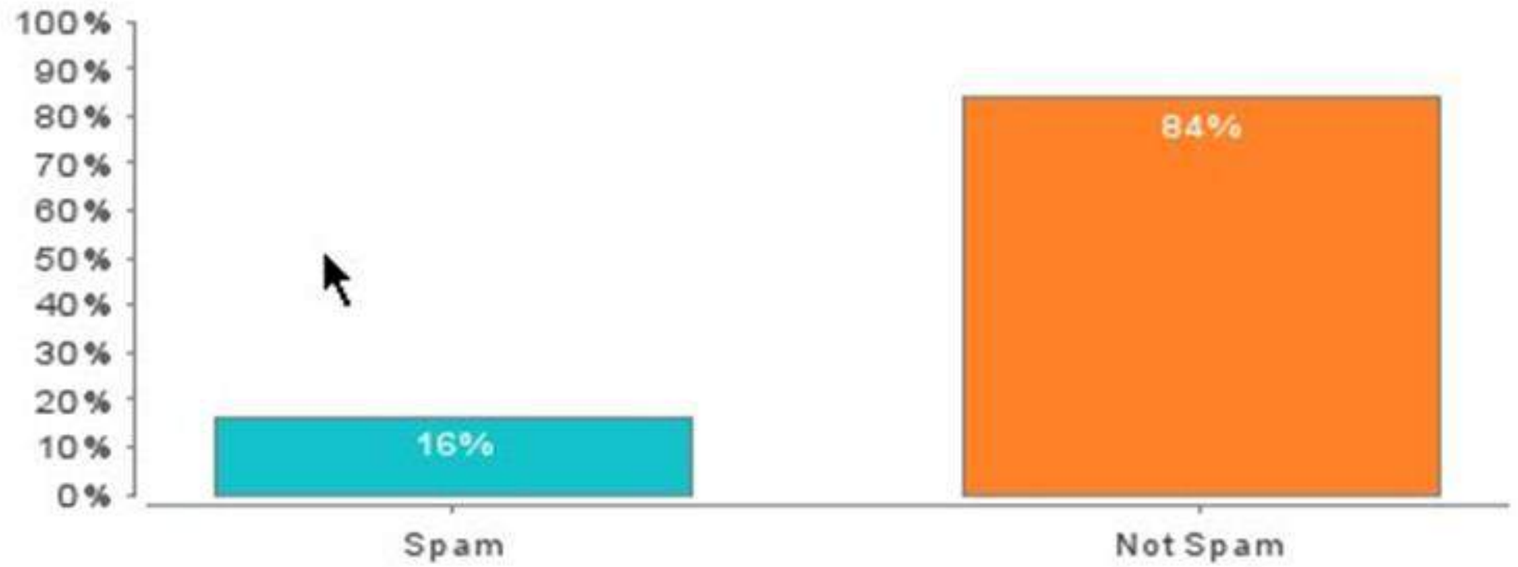

y/sms* – RapidMiner Studio Educational 10.2.000 @ amar
View
Connections
Settings
Extensions
Help
Views:
Design
Results
Turbo Prep
Auto Model
Interactive Analysis
IOObjectCollection (Collect (2))
Performance
Description
Annotations
Criterion
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)
Table View
Plot View
accuracy: 67.68% +/- 33.15% (micro average: 67.68%)
true ham
true spam
class precision
pred. ham
2578
350
88.05%
pred. spam
1027
306
22.96%
class recall
71.51%
46.65%

Plot visualization:

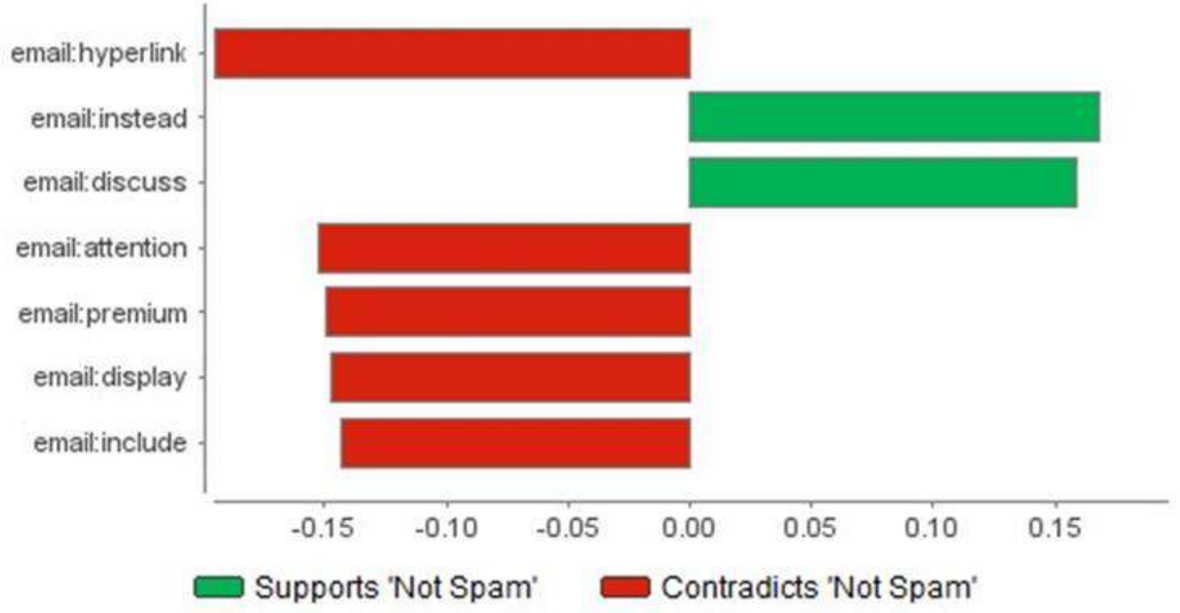
You can see in bar column how many ham message and spam messages are predicted.



Most Likely: Not Spam



Important Factors for Not Spam



CONCLUSION

In conclusion, the SMS spam filtering project conducted using RapidMiner has yielded valuable insights and outcomes in the realm of automated text message classification. The primary objectives of this project were to develop an efficient and accurate spam filtering system, leveraging the capabilities of the RapidMiner platform.

REFERENCES

1. Cormack, G.V., "Email spam filtering: A systematic review", Foundations and Trends® in Information Retrieval, Vol. 1, No. 4, (2008), 335-455. 2. Almeida, T.A., Hidalgo, J.M.G. and Yamakami, A., "Contributions to the study of sms spam filtering: New collection and results", in Proceedings of the 11th ACM symposium on Document engineering., (2011), 259-262. 3. Parandeh Motlagh, F. and Khatibi Bardsiri, A., "Detecting fake websites using swarm intelligence mechanism in human learning", International Journal of Engineering, Transactions A: Basics, Vol. 31, No. 10, (2018), 1642-1650. 4. Mohammadi, A. and Hamidi, H., "Analysis and evaluation of privacy protection behavior and information disclosure concerns in online social networks", International Journal of Engineering, Transactions B: Applications, Vol. 31, No. 8, (2018), 1234-1239. 5. Jain, A.K. and Gupta, B.B., "Phishing detection: Analysis of visual similarity based approaches", Security and Communication Networks, Vol. 2017, No., (2017).

THANK YOU!