

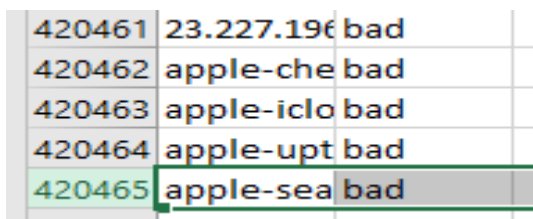
URL classification:

Executive Overview:

1. The provided data collection contains data about website links that are indicated by good and bad URLs. Bad URLs are viewed as malicious internet links that attempt to offer counterfeits.
2. To make use of the data, we must analyze it, perform various operations, such as cleansing and applying a naïve bayes model, to it.
3. The cross-validation approach and testing by qualified operators are utilized to validate the model.
4. Based on the results, predictions on the dataset can be made, which can be used to predict harmful websites in the future.

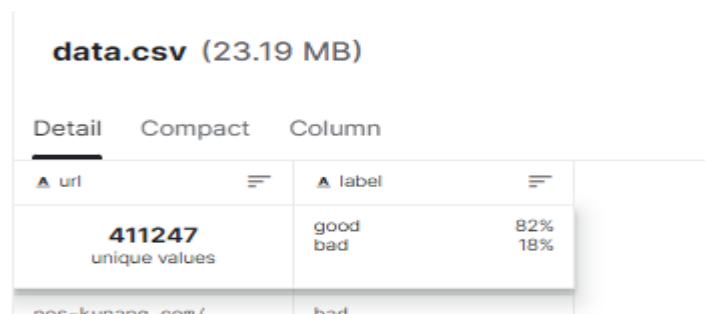
Data Analysis and Transformation:

1. By looking at below images, dataset has 411247 unique values and 420465 values in it which include duplicates and missing data in it. With this information dataset can be considered as huge data.



420461	23.227.196	bad
420462	apple-che	bad
420463	apple-iclo	bad
420464	apple-upt	bad
420465	apple-sea	bad

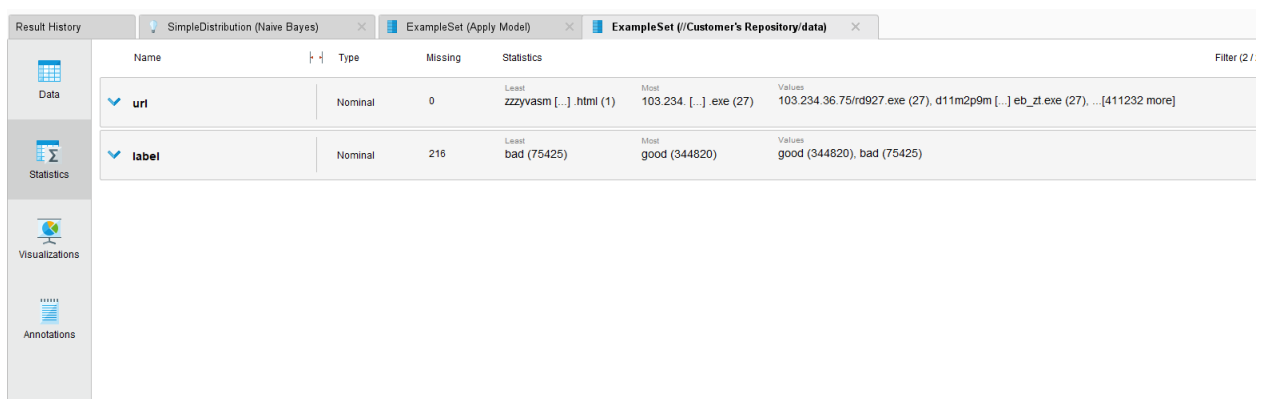
Screenshots of data size



data.csv (23.19 MB)		
Detail	Compact	Column
▲ url		▲ label
411247 unique values	good bad	82% 18%

2. When the dataset is imported on rapid miner, the statistics says it has 216 missing values and considering each URL as unique since it has different Ip addresses in it.

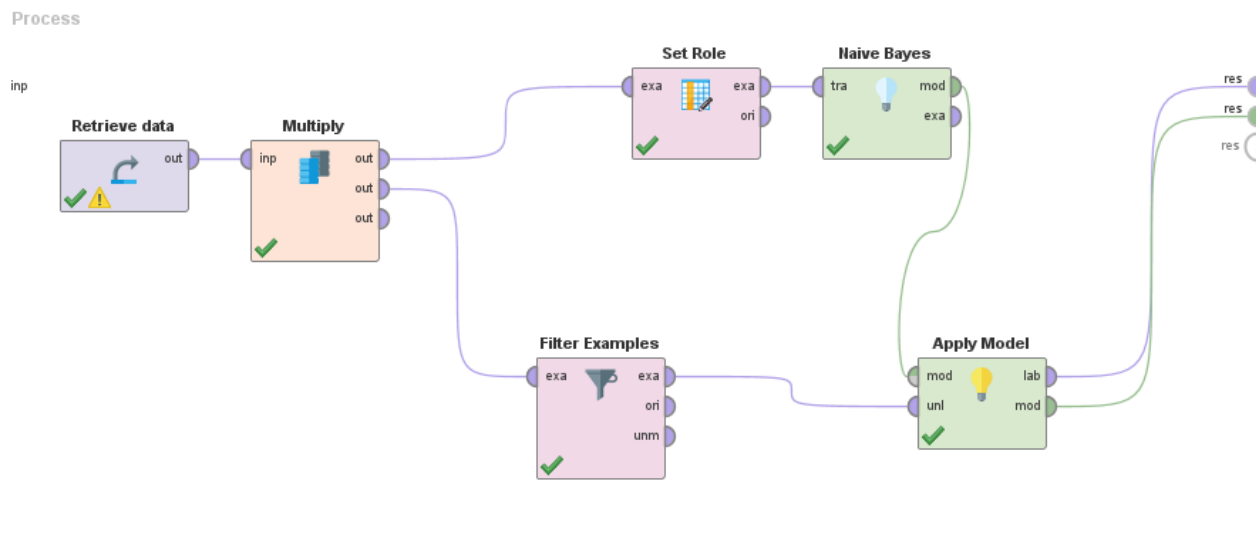
Screenshot of data statistics



Name	Type	Missing	Statistics
url	Nominal	0	Least: zzyvasm [...] .html (1) Most: 103.234. [...] .exe (27) Values: 103.234.36.75/rd927.exe (27), d11m2p9m [...] eb_2t.exe (27), ... [411232 more]
label	Nominal	216	Least: bad (75425) Most: good (344820) Values: good (344820), bad (75425)

- Now applying the model as shown on the given data set. Here Navie Bayes model is used which even predicts on missing data on label.

Screenshot of applying Navie Bayes model



- From the above image, all the label i.e., good, and bad was set to special attributes as label by using set role operator which is connected to naïve Bayes for predicting the values. Also, filter examples are used to filter out missing data. Then apply model is used which gives the below results as shown.

Screenshots of prediction values and distribution values of naïve bayes on dataset

Result History

SimpleDistribution (Naive Bayes) | ExampleSet (Apply Model) | ExampleSet (Customer's Repository/data)

Filter (216 / 216 examples): All

Row No.	prediction(label)	confidence...	confidence...	url	label
1	good	0.500	0.500	japansesman...	?
2	good	0.500	0.500	shopovate...	?
3	good	0.500	0.500	dat.info.pat...	?
4	good	0.500	0.500	52.55.48.83...	?
5	good	0.500	0.500	cafecompet...	?
6	good	0.500	0.500	findufind.info...	?
7	good	0.500	0.500	yak.bytewap...	?
8	good	0.500	0.500	pipersoperah...	?
9	good	0.500	0.500	hitonmotor.c...	?
10	good	0.500	0.500	ec2-52-48-48...	?
11	good	0.500	0.500	pipersoperah...	?
12	good	0.500	0.500	carshoping.co...	?
13	good	0.500	0.500	pipersoperah...	?
14	good	0.500	0.500	shopper.scri...	?
15	good	0.500	0.500	shopper.scri...	?
16	good	0.500	0.500	guardbancard...	?
17	good	0.500	0.500	carshoping.co...	?
18	good	0.500	0.500	altomismail...	?
19	good	0.500	0.500	altomismail...	?
20	good	0.500	0.500	classcunnet...	?
21	good	0.500	0.500	classcunnet...	?
22	good	0.500	0.500	oportunidade...	?
23	good	0.500	0.500	oportunidade...	?
24	good	0.500	0.500	fb.noticestat...	?
25	good	0.500	0.500	wp.cash.co...	?

ExampleSet (216 examples, 3 special attributes, 2 regular attributes)

Calculating result: Distribution Table

54:1 PM 8/1/2022

SimpleDistribution

Distribution model for label attribute label

Class bad (0.179)
1 distributions

Class good (0.820)
1 distributions

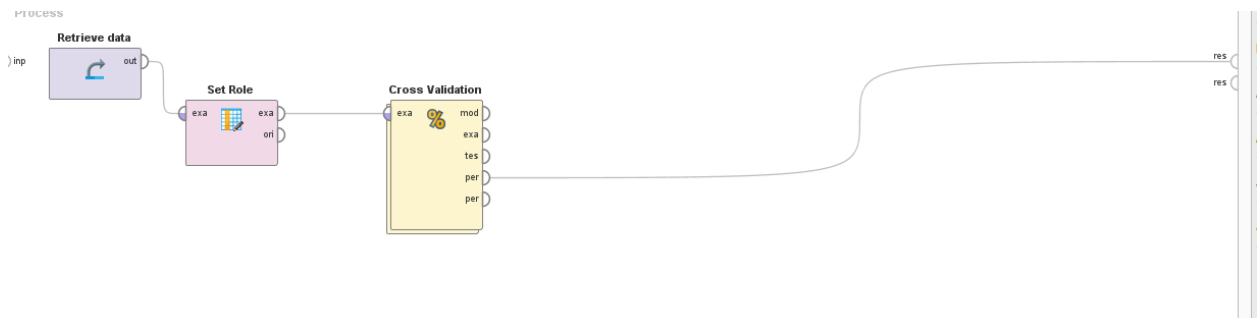
Why Navie Bayes?

Navie Bayes is one of the classification models which works on fundamental assumption with given value and independent value of other resulting in Navie Bayes predictive model.

Proposed Solutions and Testing:

1. For validation, Cross validation method is used to validate the applied model is the right one.

Screenshot of applying Cross validation



Screenshot of Subset of Cross validation

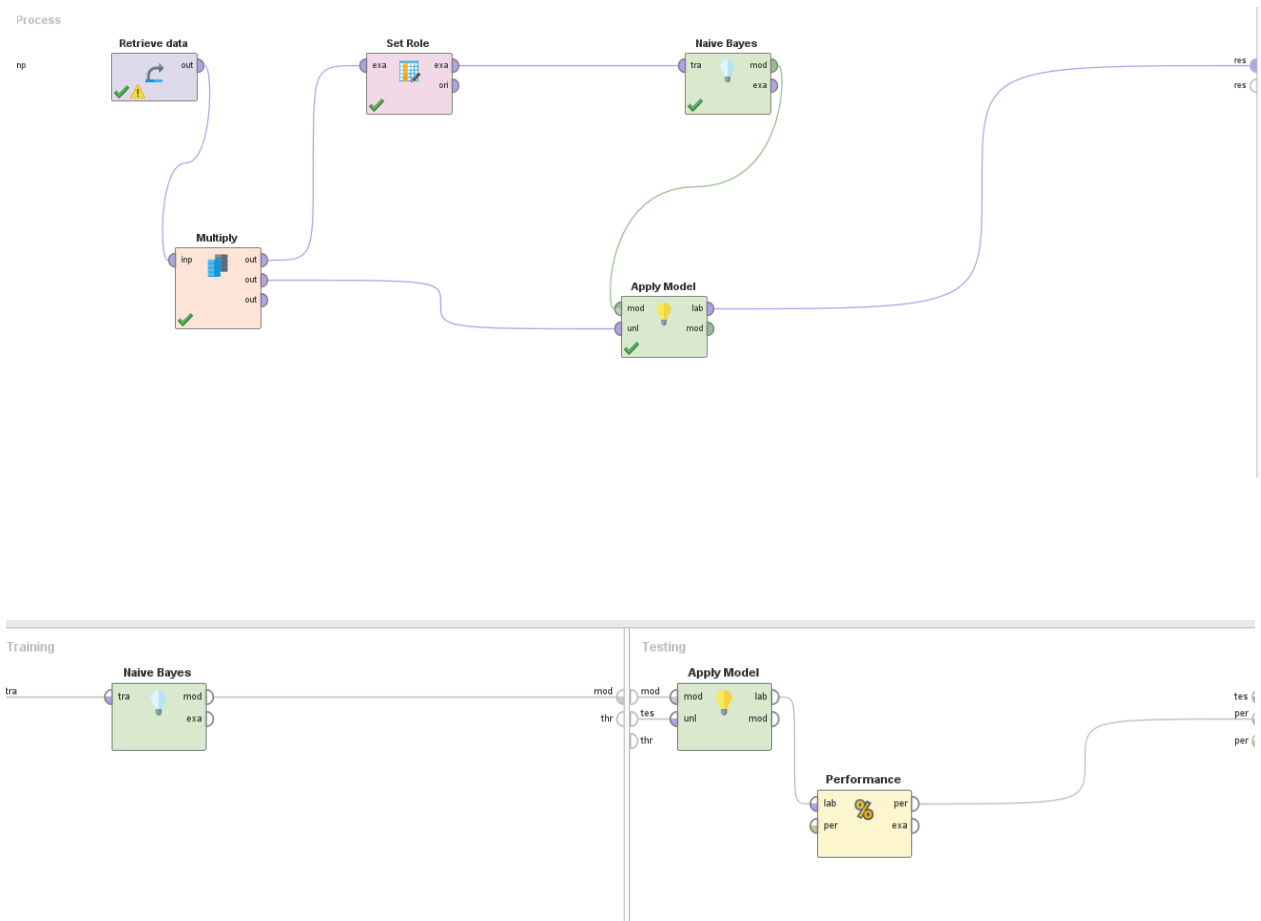
accuracy: 85.79% +/- 0.09% (micro average: 85.79%)

	true bad	true good	class precision
pred. bad	15725	0	100.00%
pred. good	59700	344820	85.24%
class recall	20.85%	100.00%	

Screenshot of results of Cross Validation

- The above images show how the sub process of Cross Validation is to apply and also results of it.
- For testing, below operators are used on rapid miner i.e., taking the naïve bayes predictions and the data set attributes from the given dataset are applied as input on apply model. Now by comparison, accurate predictions can be confirmed as shown below.

Screenshot of design of testing



Screenshot of checking the prediction values with label values

ExampleSet (420,461 / 420,461 examples)

Row No.	prediction	confidence	confidence	url	label
471	bad	1.000	0.000	directre.com...	bad
472	bad	1.000	0.000	youtube.com	bad
473	bad	1.000	0.000	youtube.com	bad
474	bad	1.000	0.000	ukonline.hc0...	bad
475	bad	1.000	0.000	youtube.com	bad
476	bad	1.000	0.000	ukonline.hc0...	bad
477	bad	1.000	0.000	tdms.saglik.g...	bad
478	bad	1.000	0.000	sunny99.chol...	bad
479	bad	1.000	0.000	rooversadvoc...	bad
480	bad	1.000	0.000	malest.com	bad
481	bad	1.000	0.000	d01.fadmr.c...	bad
482	bad	1.000	0.000	d01.fadmr.c...	bad
483	bad	1.000	0.000	rebeccacella...	bad
484	bad	1.000	0.000	lowes-plano...	bad
485	bad	1.000	0.000	lcbcad.co.uk...	bad
486	bad	1.000	0.000	v.inigiplan.ru...	bad
487	bad	1.000	0.000	mailfoto.com...	bad
488	bad	1.000	0.000	guycards.co...	bad
489	bad	1.000	0.000	guycards.co...	bad
490	bad	1.000	0.000	teamedata.co...	bad
491	bad	1.000	0.000	wcbi3ghk.h...	bad
492	bad	1.000	0.000	instuminahu...	bad
493	bad	1.000	0.000	server1.extra...	bad
494	bad	1.000	0.000	nutret.irdih...	bad
495	bad	1.000	0.000	etlehabib.co...	bad

ExampleSet (420,461 examples, 3 special attributes, 2 regular attributes)

4. From the above image we can see the prediction value and the label value which are mostly accurate.

Results:

1. From all images above, results can be seen how Navie Bayes model predicted even on the missing data.
2. Navie Bayes is one which can be used for cases which have scenarios like for the given data.
3. This model has accurate rate resulting in utilized data with its prediction values.

2. Malicious Server Hack:

Executive Overview:

1. The given data set has the attributes of anonymized variables which can be used to decide when the hack is going to happen.
2. For this data set, k-means model can be used for prediction.
3. Cross validation and required operators are used for validation and testing the model.
4. With the clusters of the predicted values, data can be developed for detecting anomalies.

Data Analysis and Transformation:

1. Since the first step is importing data on rapid miner, the below statistics can be observed.

Name	Type	Missing	Statistics
X_1	Integer	0	Min 0 Max 5 Average 2.456
X_2	Integer	0	Min 1 Max 19 Average 6.154
X_3	Integer	0	Min 0 Max 18 Average 4.877
X_4	Integer	0	Min 0 Max 99 Average 0.972
X_5	Integer	0	Min 0 Max 6 Average 4.924
X_6	Integer	0	Min 1 Max 90 Average 1.245
X_7	Integer	0	Min 0 Max 332 Average 206.955
X_8	Real	182	Min 0 Max 90 Average 0.974
X_9	Integer	0	Min 0 Max 116 Average 85.237
X_10	Integer	0	Min 0 Max 142 Average 72.674
X_11	Integer	0	Min 0 Max 50 Average 33.465
MALICIOUS_OFFENSE	Integer	0	Min 0 Max 1 Average 0.955

- ### Screenshot of replacing values on turbo prep

- Once it is done, auto model option can be used which directs as below and selecting the clusters as shown below.

Predict

Want to predict the values of a column?

Clusters

Want to identify groups in your data?

Outliers

Want to detect outliers in your data?

4.Now, selecting the inputs by unmarking the malicious offense and incident_id and taking the other attributes as shown below.

Screenshot choosing input values on auto model

Load Data Select Task Prepare Target Select Inputs Model Types Results

⏮️ RESTART ⏪ BACK NEXT

Selected: 15 / Total: 18

● Deselect Red ✔ Select All ✖ Deselect All

Selected	Status ↑	Quality	Name	Correlation	ID-ness	Stability	Missing	Text-ness
<input type="checkbox"/>	●	<div><div></div></div>	INCIDENT_ID	?	100.00%	0.00%	0.00%	37.09%
<input type="checkbox"/>	●	<div><div></div></div>	MALICIOUS_OFFENSE	?	?	95.52%	0.00%	0.00%
<input type="checkbox"/>	●	<div><div></div></div>	DATE	?	38.23%	0.09%	0.00%	16.74%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_1	?	?	79.80%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_2	?	?	16.89%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_3	?	?	16.89%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_4	?	?	23.04%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_5	?	?	30.89%	0.00%	0.00%
<input checked="" type="checkbox"/>	●	<div><div></div></div>	X_6	?	?	14.51%	0.00%	0.00%

4. Since this on auto model which gives the required model as shown below.

Screenshot of auto model-K-Means Clustering

Load Data Select Task Prepare Target Select Inputs Model Types Results

⏮️ RESTART ⏪ BACK ▶️ RUN

Models

☒ **k-Means Clustering**
 Number of Clusters:

☒ **x-Means Clustering**
 Maximal Number of Clusters:

Data Preparation

☐ **Remove Columns with Too Many Values**
 Maximum Number of Values:

☐ **Extract Date Information**

☐ **Extract Text Information**
 Select Text Columns (0)...

Number of Extracted Features:

☐ **Automatic Feature Selection**
 Additional Time (in Minutes):

Final Feature Set should be

Column Analysis

☒ **Correlations between Columns**

- When the run button is selected which directly shows the results as below.

Screenshot of results

k-Means - Summary

Number of Clusters: 2

Cluster 0

16,810

X_5 is on average 33.30% smaller, X_2 is on average 33.24% larger, X_3 is on average 33.22% larger

Cluster 1

7,046

X_5 is on average 79.43% larger, X_2 is on average 79.31% smaller, X_3 is on average 79.25% smaller

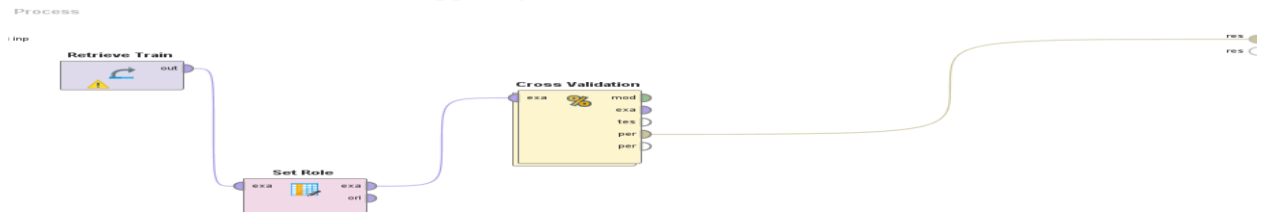
Why K-Means model?

K-Nearest Neighbor model is one of the classification or regression models which is based on comparing the most unknown examples which are nearly neighbors.

Proposed Solution and Testing:

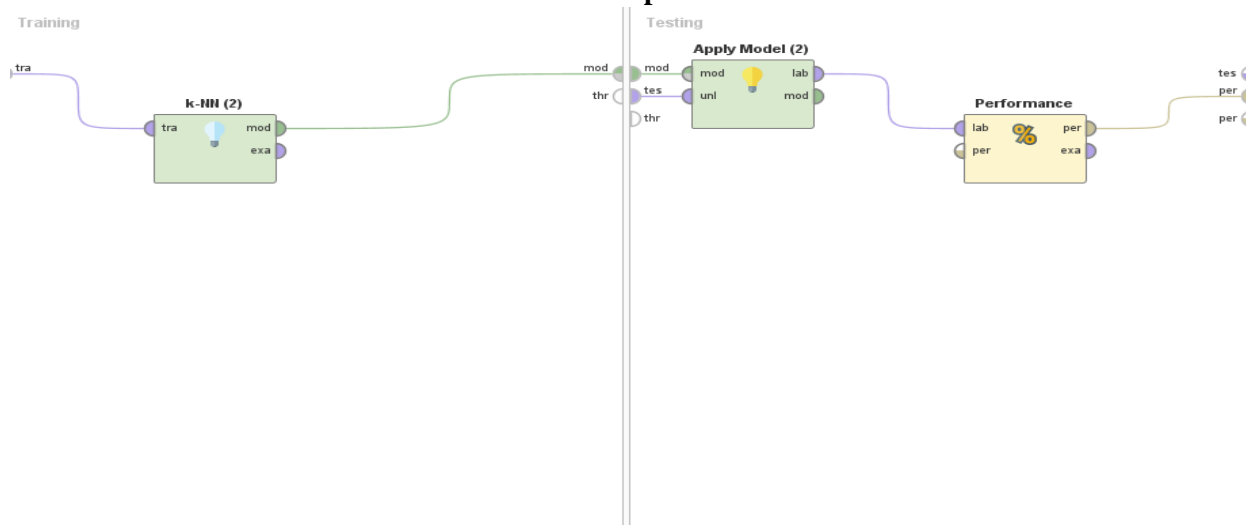
- For validation, Cross validation method is used on rapid miner as shown below.

Screenshot Applying Cross Validation Method



- Cross validation is a sub-process in which it has connected as shown below.

Screenshot of sub process of Cross validation



- From the above image, K-NN is connected to apply model and then to performance which results as shown below.

Screenshot of performance of the K-NN model

Result History

PerformanceVector (Performance)

PerformanceVector

PerformanceVector:

root_mean_squared_error: 0.157 +/- 0.007 (micro average: 0.157 +/- 0.000)

squared_error: 0.025 +/- 0.002 (micro average: 0.025 +/- 0.129)

Performance

Description

Annotations

- As auto-model is applied on this data set assuming the testing the data would give the same accuracy in prediction.

Results:

- From the results of above, we can observe how clusters are divided the attributes of anonymized variables which can be used for the prediction.

2. However, by root_mean_square_error which has low value can be considered the model is a good fit which shows the outcomes are accurate one.

SMS spam prediction:

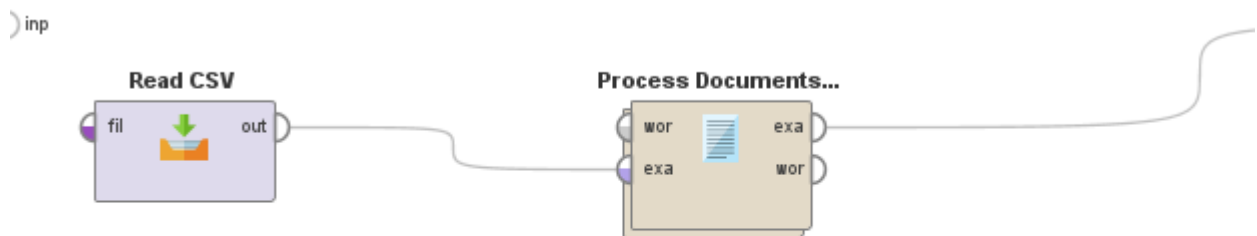
Executive Overview:

1. The given data set is text data which must verified whether it is spam or ham.
2. Text mining with Natural Language Processing (NPL) is a challenging task since this is done based on the prediction on word count and character count.
3. Navie Bayes model is applied for the given data set with few other operators to predict the word count.
4. Cross validation method is used to validate the model.

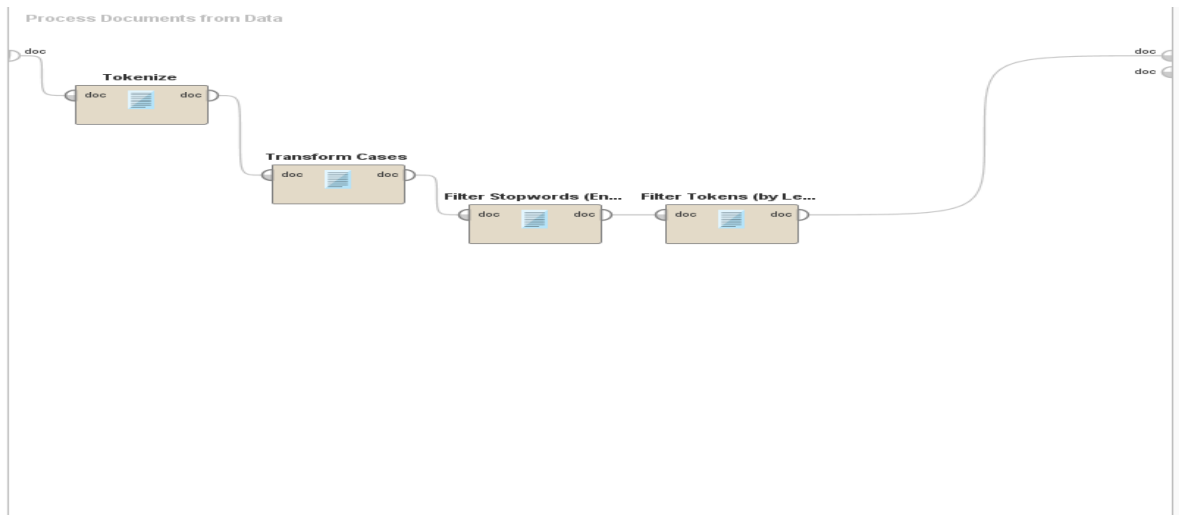
Data Analysis and Transformation:

1. Given data set has two attributes which one has text and other as spam or ham. Since the second attribute is considered as Natural Language processing.
2. Since text words need to be counted based on the number of occurrences in whole dataset rapid miner needs to load few extensions such as text processing and web mining from the marketplace.
3. Now the below operators are used to counting the words in the text as shown below.

Screenshot of operators used on data set



Screenshot of sub process of process document to data



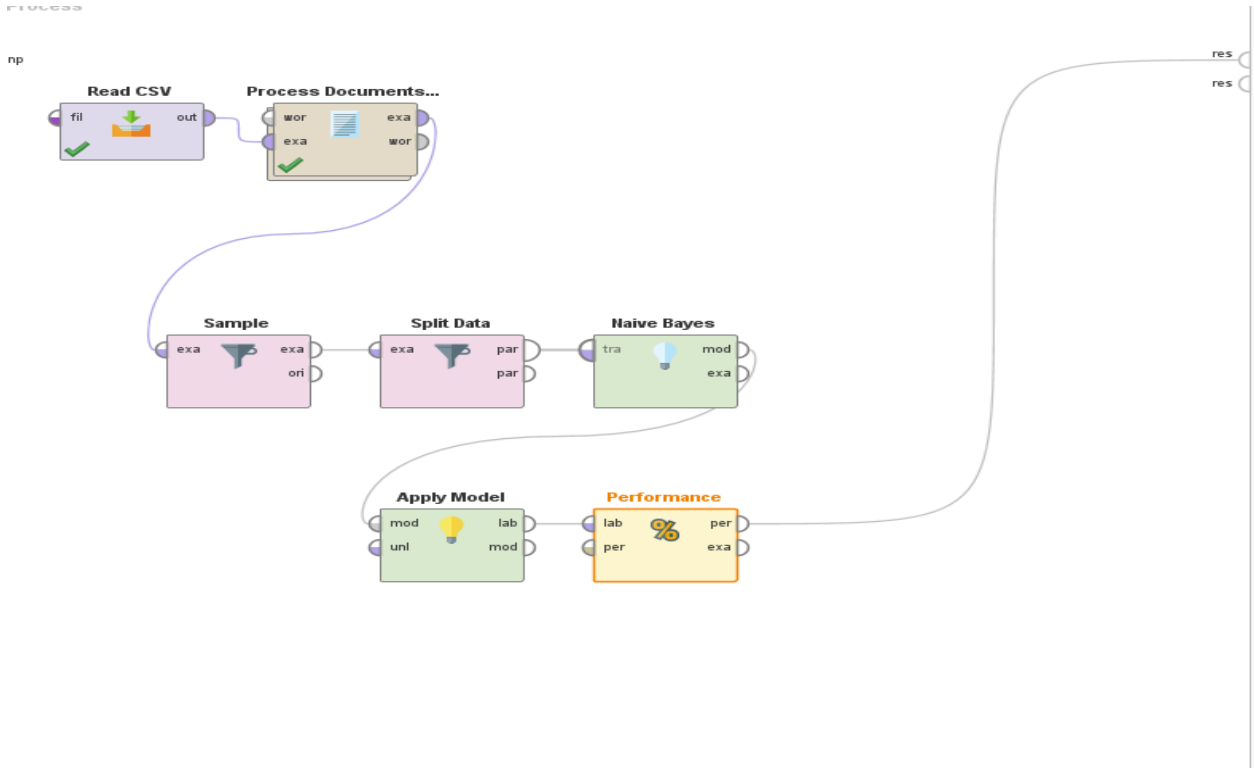
4. Below are the results of the design to count the number of occurrences in the data.

Screenshot of word's count in given data set

Word	Attribute Name	Total Occurences	Document Occurences	spam ↑	ham	ham*****
aathi	aathi	6	6	0	6	0
ability	ability	2	2	0	2	0
abiola	abiola	11	11	0	11	0
able	able	24	24	0	24	0
absolutly	absolutly	2	2	0	2	0
aburo	aburo	2	2	0	2	0
accept	accept	9	6	0	9	0
accidentally	accidentally	4	4	0	4	0
accounts	accounts	2	2	0	2	0
ache	ache	4	4	0	4	0
acted	acted	2	2	0	2	0
acting	acting	2	2	0	2	0
activities	activities	4	4	0	4	0
actor	actor	2	2	0	2	0
actual	actual	2	2	0	2	0
actually	actually	32	32	0	32	0
addicted	addicted	4	2	0	4	0
addie	addie	3	3	0	3	0
admit	admit	2	2	0	2	0
adore	adore	3	3	0	3	0
adoring	adoring	2	2	0	2	0
advance	advance	7	7	0	7	0
adventure	adventure	2	2	0	2	0
advice	advice	5	5	0	5	0
aeronautics	aeronautics	2	2	0	2	0

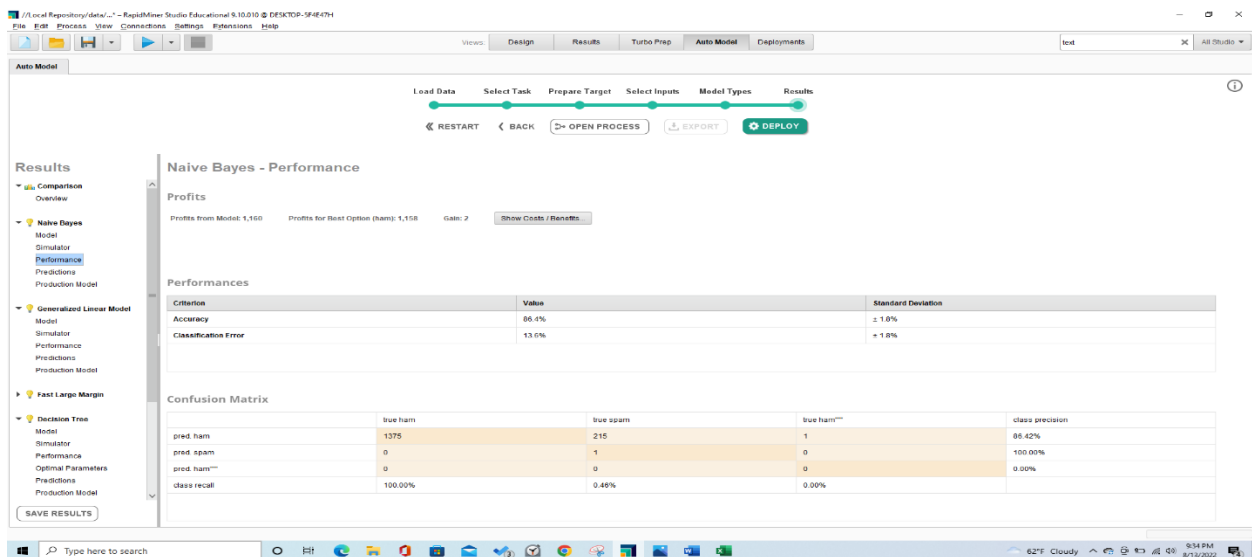
3. Now applying Navie bayes model we can see the below results as shown below.

Screenshot of Applying Naïve Bayes model on the given dataset



4. However, this can be done through auto model which gives the below results.

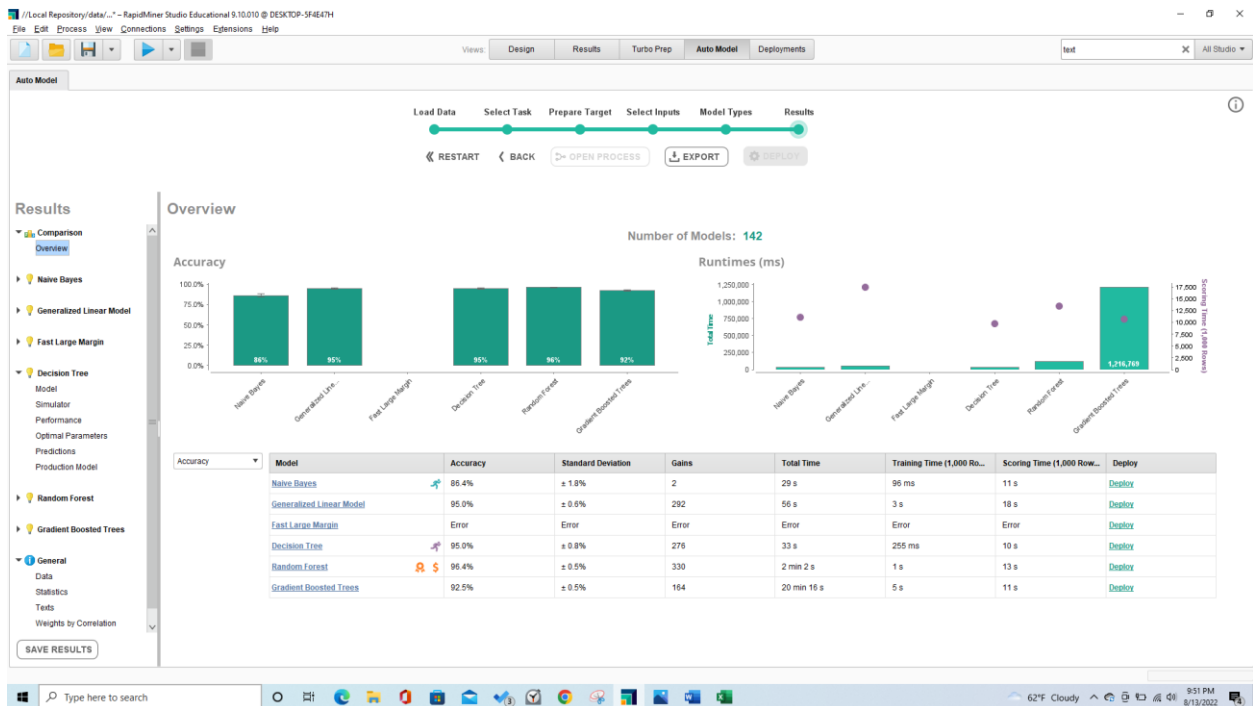
Screenshot of results of given dataset



Proposed Solution and Testing:

1. To validate the model, cross validation is used to verify if the model is stable or not.
2. After applying this gives accurate results on the predictions.
3. Assuming the model and testing are the accurate as auto model is used here to verify.
4. Below is the image of overview of auto model

Screenshot of Overview



Results:

1. Even the text has NPL, the data is retrieved by transforming data using tokenization, transforming into lower cases, and cutting the words by length by using token by length.
2. From the above results we can see that many more models can be used on this dataset.

Why Cross validation?

In all the above datasets, cross validation is used. Cross validation estimates the accuracy of model. It is a nested Operator which has a training subprocess and testing subprocess in it.

