جامعة جدة
University of Jeddah

DATA MINING | CCDS 223

FINAL PROJEC REPORT

CLIMATE CHANGE

M3B| 18520

November 9, 2022

Prepared by:

Marwh AL-hadi	2007881
Enas khan	2007145
Samah Saad	2006293
Dania Alshehri	2006276

Supervised by:

Dr .Ines Boufateh

2022 – 2023

TABLE OF CONTENTS

phase 1

phase 2

INTRODUCTION

On an everyday basis, many people use weather forecasts to determine (e.g what to wear on a given day). forecasts can be used to plan activities ahead.

Knowing the expected weather is a matter that affects farmers, business owners, and various groups of society in different aspects, especially in Indonesia, because it is a tourist country with a volatile climate and sensitive infrastructure.

So we decided to achieve our goal of predicting the expected weather on a specific date or a specific region in Indonesia and for that, we used the climate data in Indonesia to predict whether the day is hot or cold

It is also important to study and start the nature of the weather in Indonesia so that we can know when winter or summer often begins

ABOUT DATA

The data in our project is about daily climate data covering almost all regions of Indonesia from 2010 to 2020, This dataset belong to (BMKG) Indonesia (Meteorology, Climatology, and Geophysical Agency in English) and contains climatic details for each region and province as it covered 34 provinces according to The date, a statement of the maximum and minimum temperature, including wind direction and speed, humidity, latitude and longitude, the percentage of the duration of the sun's brightness, and also the stations that recorded all of that.

The data distributed in three files:

PROVINCE DETAIL:

province_id:id of the province
province_name: name of the province

STATION DETAIL:

station_id: station id which record the data
station_name:name of the station
region_name:city and region (kabupaten) in Indonesia
latitude: lat
longitude: long
region_id:city and region (kabupaten) name

CLIMATE DETAILE:

date

Tn: Min temperature (°C)

Tx:Max temperature (°C)

Tavg:Avg temperature (°C)

RH_avg:Avg humidity (%)

RR: rainfall (mm), how much water falls as rain in a certain period of time

ss: sunshine duration (hour), duration of sunshine in a day

ff_x :max wind speed (m/s)

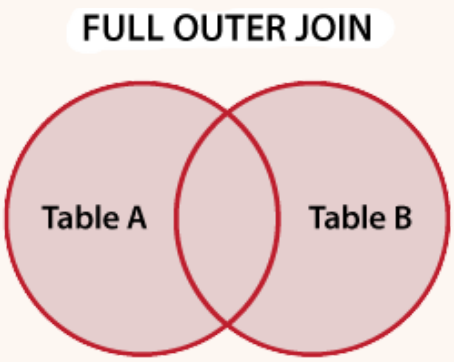
ddd_x: wind direction at maximum speed (°)

ff_avg: avg wind speed (m/s)

ddd_car:most wind direction (°)

IMPORTING DATA

combine all datasets



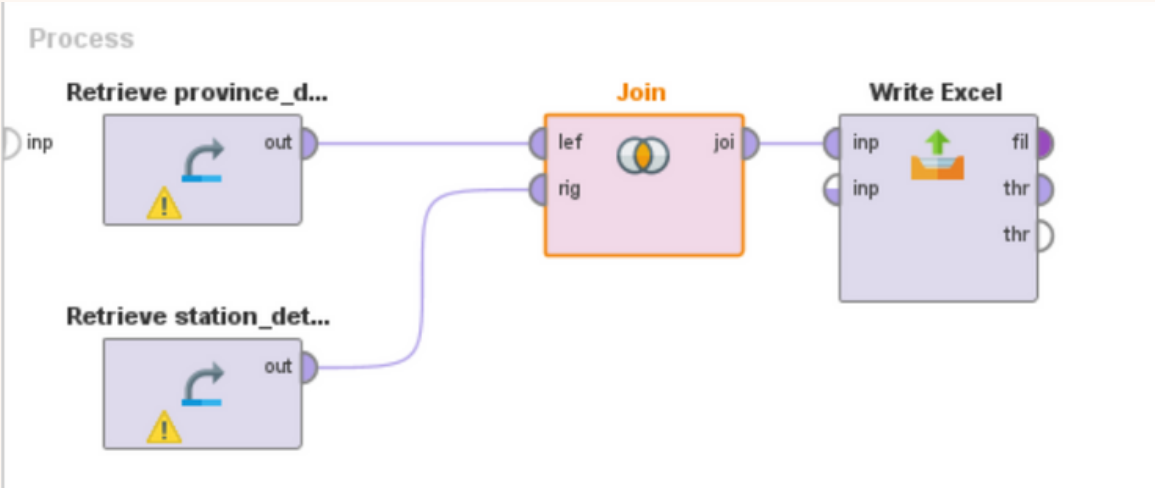
- collected into one file using :

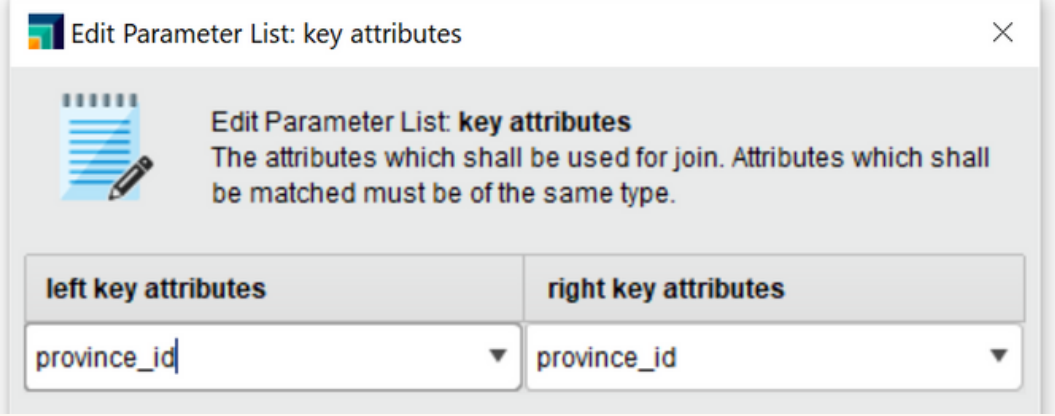
join operator

- type of join:

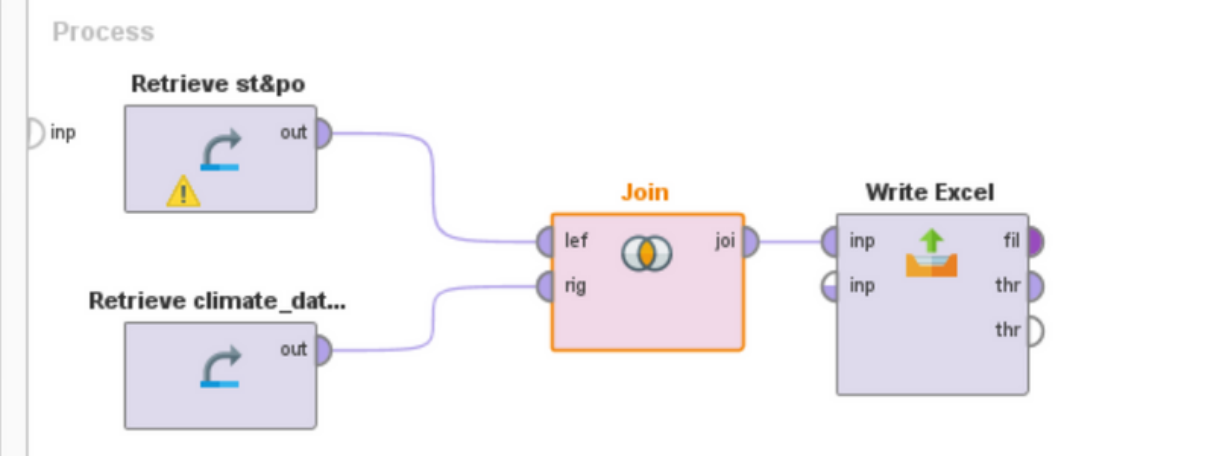
Outer join: return matched values and unmatched values from either or both tables.

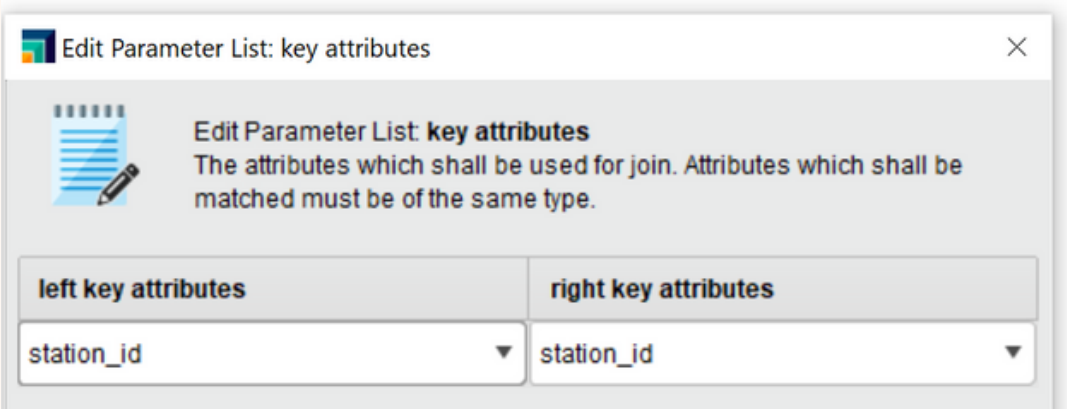
1. first combine the province detail and station detail by Shared key





2. secondly using the data that exported by the last process and join it with climate data





- The data set after combining has 589285 rows and 19 features including a label column.
- according to RapidMiner license , we decide to get sample with 5000 rows using sample operator
- The label column specifies the status of the Out of the 19 features in the data, we have 6 categorical variables, 12 numerical variables, and 1 date variable

MAJOR CHALLENGES:


Missing values


Redundant attributes


Inconsisetant data


Outliers

we will discuss the **cleaning process** but before that let us see how these problems arise in the dataset in next slides.

Attributes Properties				
name:	type	mean	median	variance
province_id	nominal	n/a	n/a	n/a
province_name	nominal	n/a	n/a	n/a
station_id	nominal	n/a	n/a	n/a
station_name	nominal	n/a	n/a	n/a
ddd_car	nominal	n/a	n/a	n/a
region_id	nominal	n/a	n/a	n/a
date	date	n/a	n/a	n/a
latitude	numeric	-2.997	-3.045	4.071
longitude	numeric	114.539	112.783	11.338
Tn	numeric	23.312	24	2.281
Tx	numeric	31.529	31.8	2.312
Tavg	numeric	26.855	27.2	1.940
RH_avg	numeric	82.489	83	14.338
RR	numeric	8.681	1	17.929

Attributes Properties				
name:	type	mean	median	variance
ff_x	numeric	4.710	4	2.612
ddd_x	numeric	188.488	180	107.657
ff_avg	numeric	1.957	2	1.803
ss	numeric	5.083	5.3	3.262

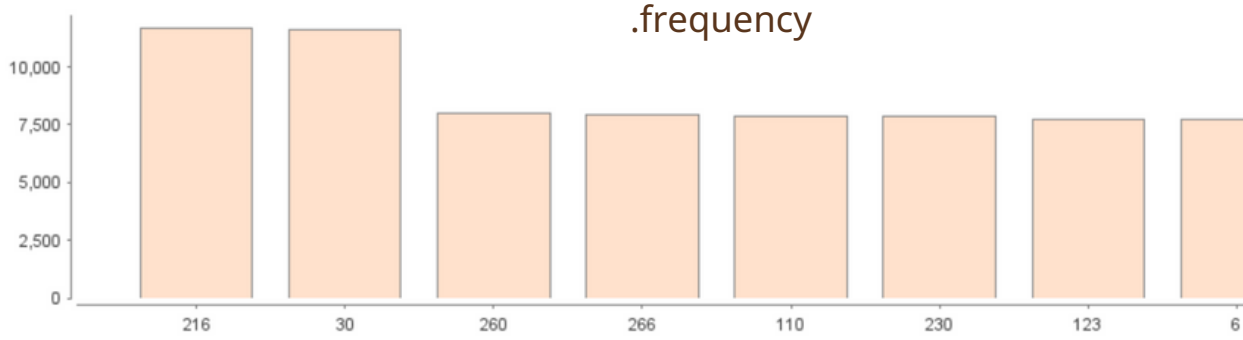
region_id

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 1.95%
Valid: 98.02%

Top Values



histogram shows us the ragion frequency in our used data. We note that most of them are of course close in .frequency

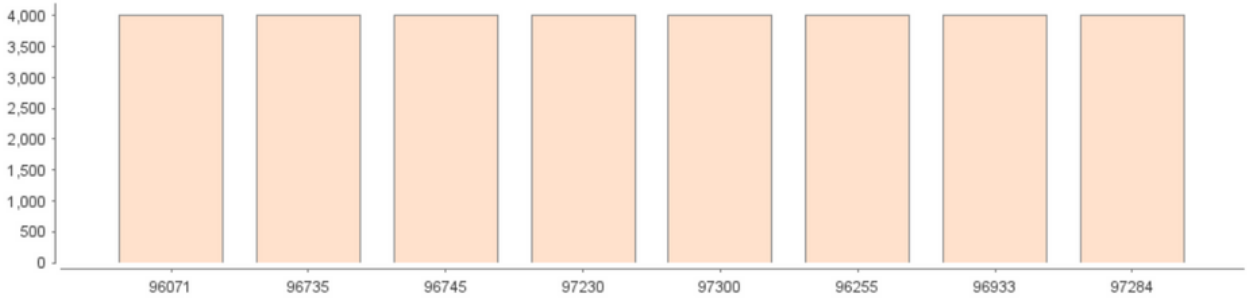
station_id

Summary

Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 0.86%
Valid: 99.11%

Top Values



histogram shows us the station frequency in our used data. We note that most of them are of course equals in frequency.

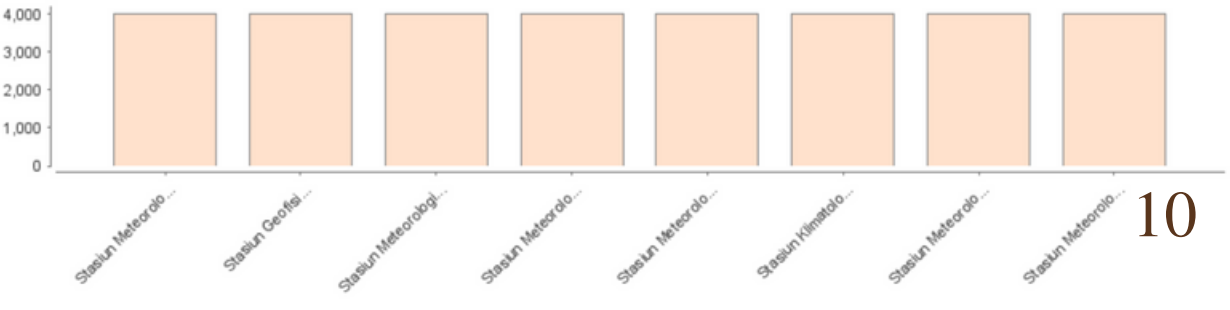
station_name

Summary

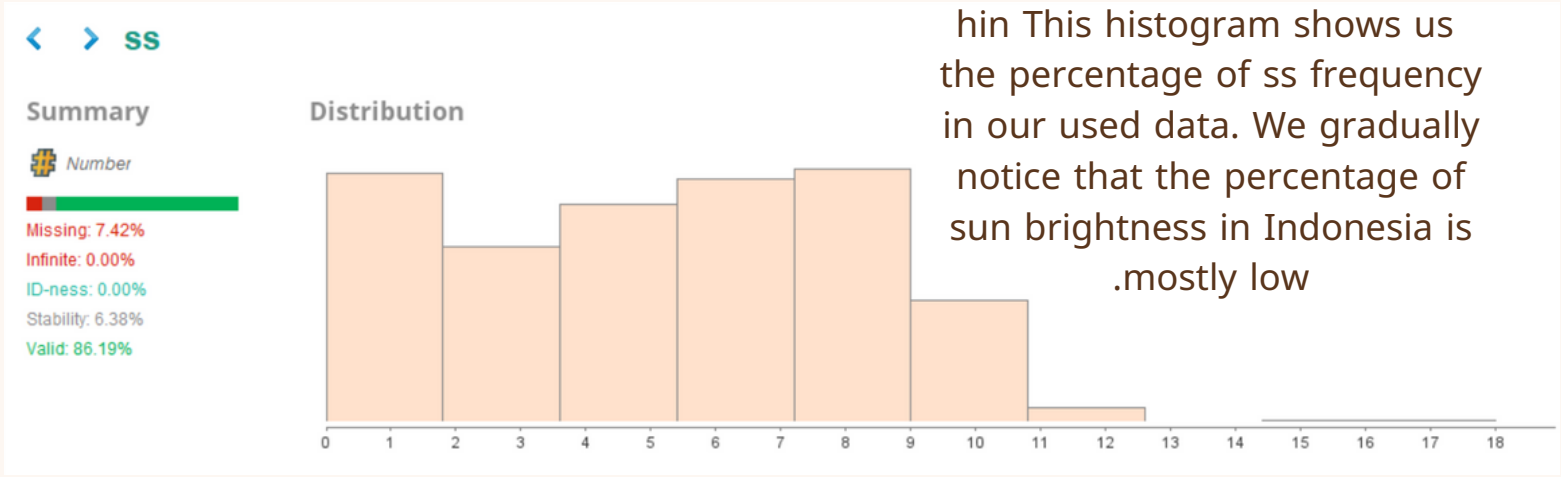
Category

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 0.86%
Valid: 99.11%

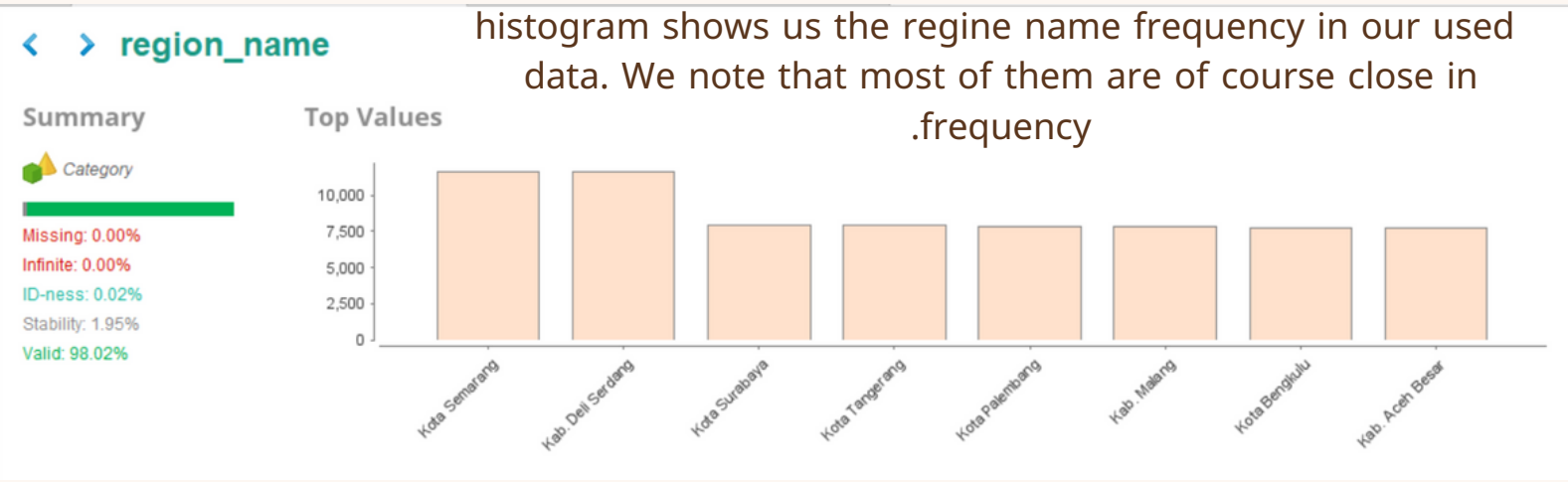
Top Values



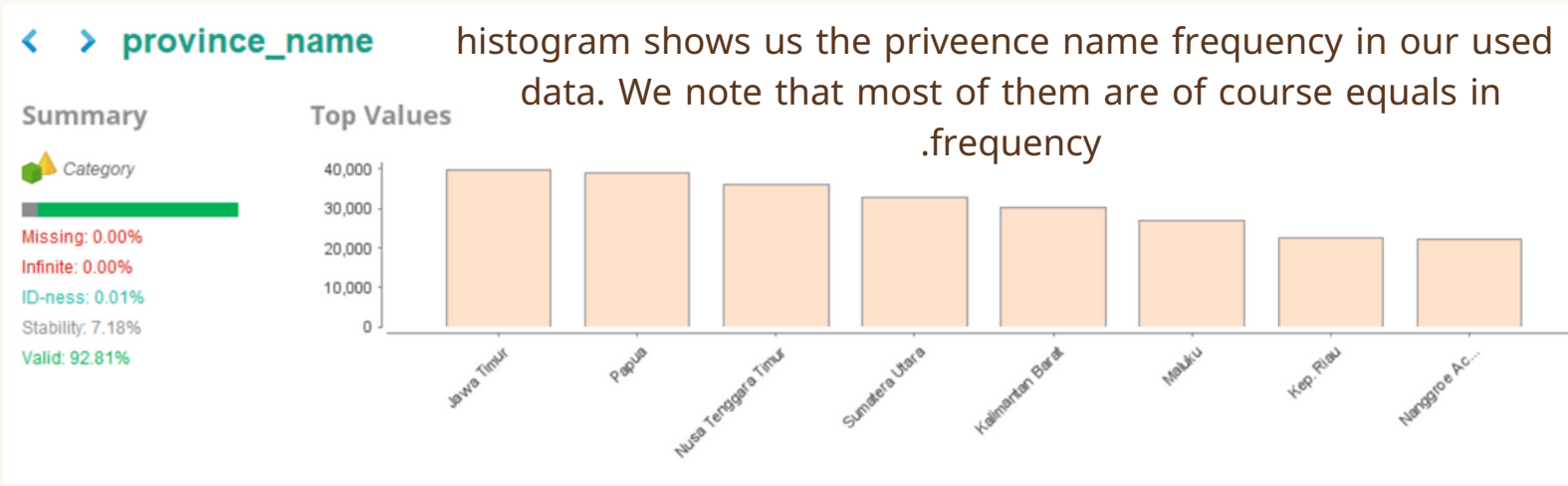
histogram shows us the staton name frequency in our used data. We note that most of them are of course equals in .frequency



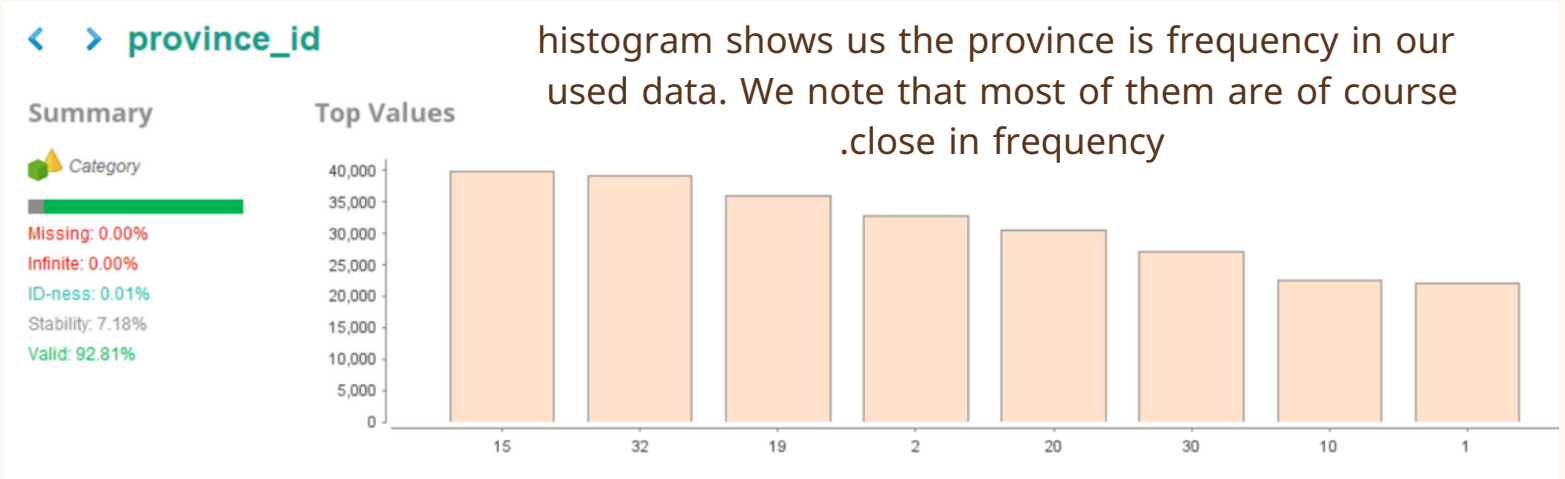
hin This histogram shows us the percentage of ss frequency in our used data. We gradually notice that the percentage of sun brightness in Indonesia is .mostly low



histogram shows us the regine name frequency in our used data. We note that most of them are of course close in .frequency



histogram shows us the priveence name frequency in our used data. We note that most of them are of course equals in .frequency



histogram shows us the province is frequency in our used data. We note that most of them are of course .close in frequency

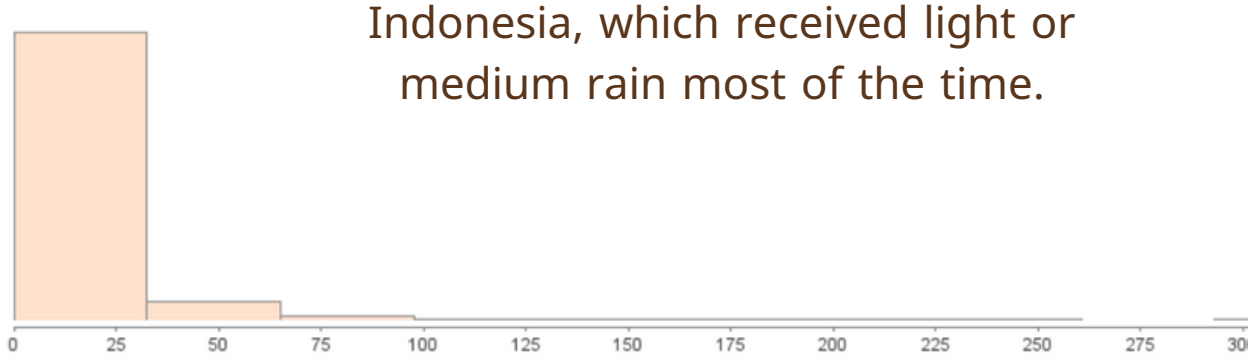
RR

Summary

Number

Missing: 21.28%
Infinite: 0.00%
ID-ness: 0.02%
Stability: 40.64%
Valid: 38.06%

Distribution



We see here that the percentage of rainfall in our selected sample was Indonesia, which received light or medium rain most of the time.

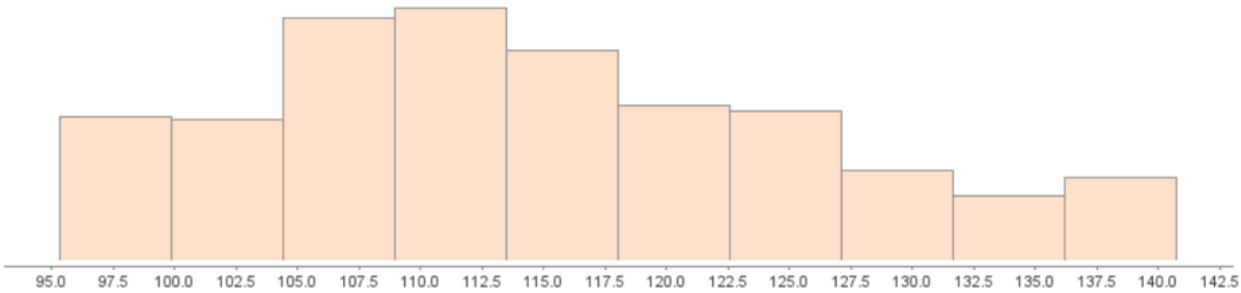
longitude

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 1.32%
Valid: 98.68%

Distribution



We note the histogram of longitude due to its geographical location

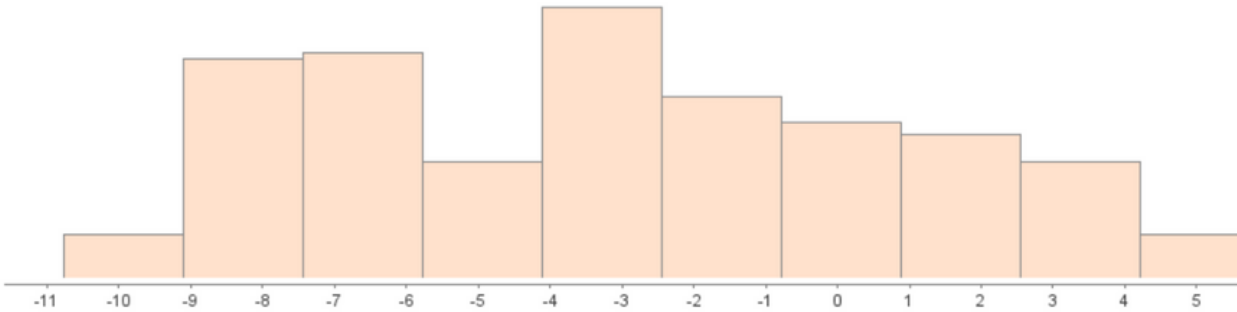
latitude

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 1.30%
Valid: 98.70%

Distribution



We note the histogram of Latitude due to its geographical location

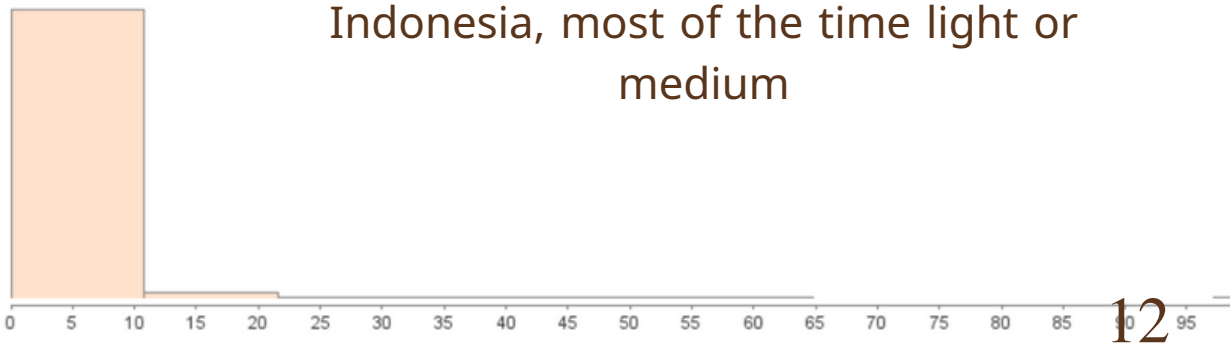
ff_x

Summary

Number

Missing: 1.74%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 23.91%
Valid: 74.35%

Distribution



We see here that the percentage of wind gusts in our selected sample was Indonesia, most of the time light or medium

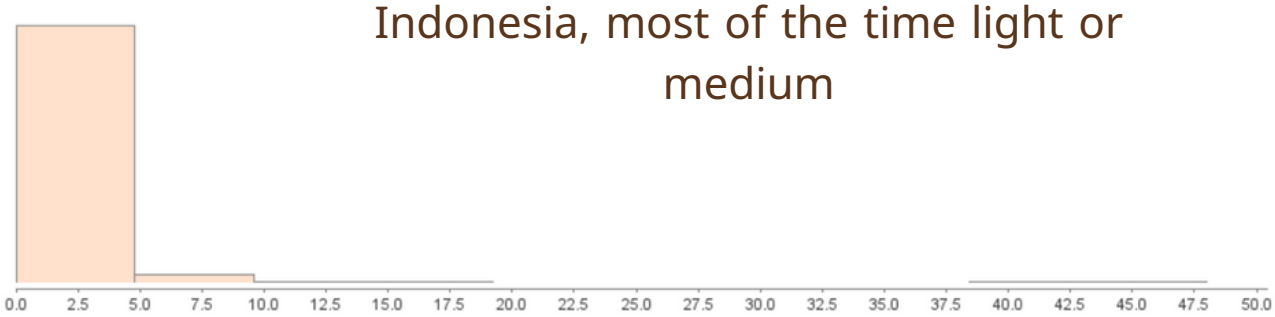
ff_avg

Summary

Number

Missing: 1.72%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 35.60%
Valid: 62.68%

Distribution



We see here that the percentage of avg wind gusts in our selected sample was Indonesia, most of the time light or medium

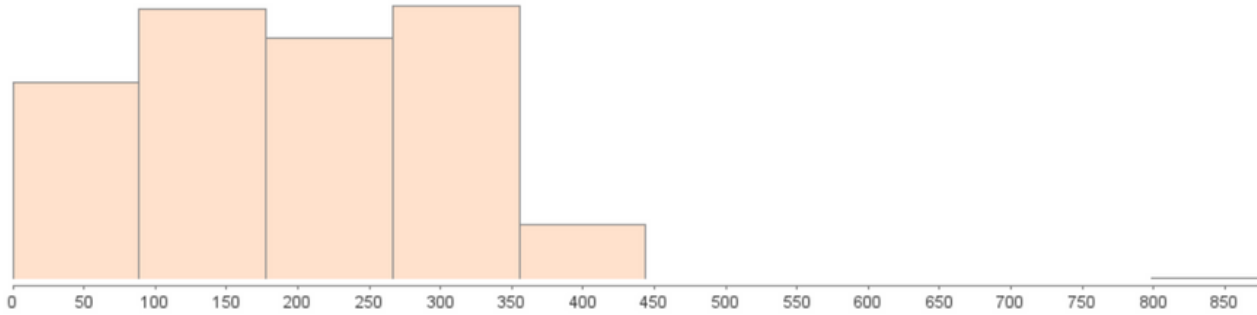
ddd_x

Summary

Number

Missing: 2.23%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 6.76%
Valid: 90.98%

Distribution



We see here that the percentage of speed of wind gusts in our selected sample was Indonesia, most of the time light or medium

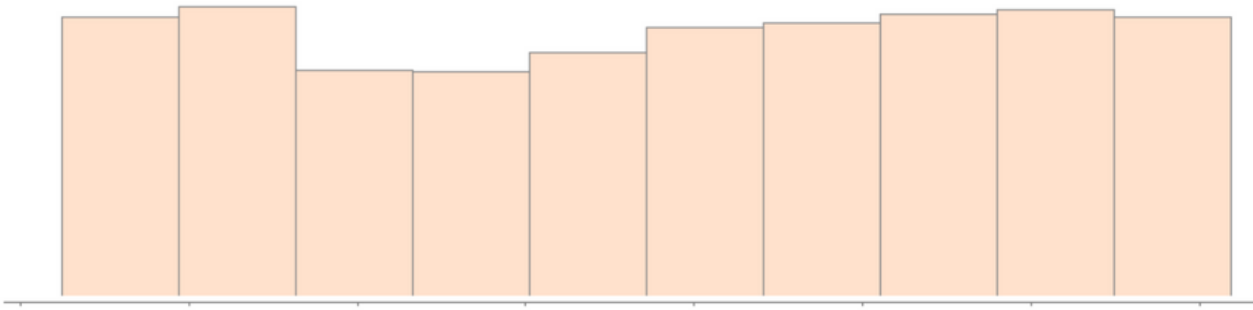
date

Summary

Date / Time

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 0.11%
Valid: 99.89%

Distribution



In this data, the data is recorded during the day repeatedly, so we notice a high frequency on all dates

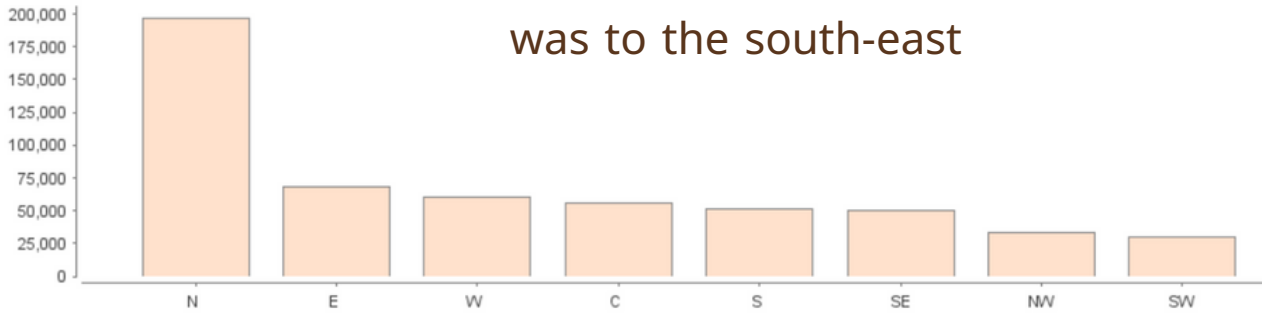
ddd_car

Summary

Category

Missing: 2.33%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 33.96%
Valid: 63.71%

Top Values



In this sample, most of the wind direction was to the north, and the least was to the south-east

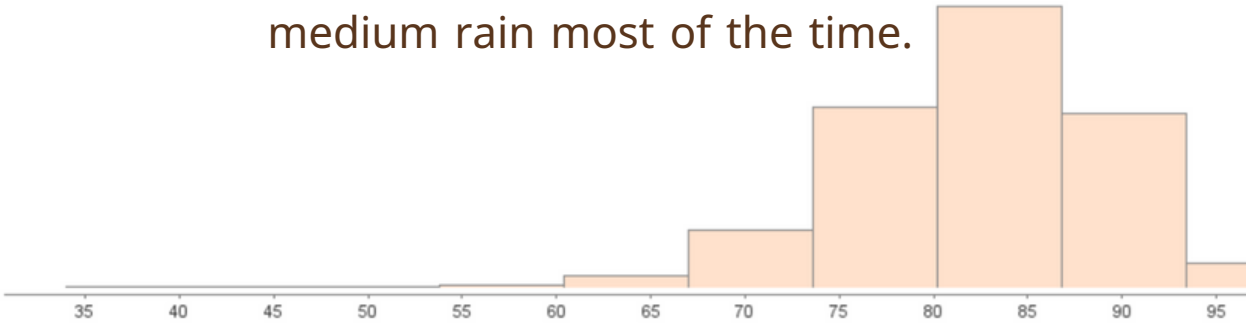
RH_avg

Summary

Number

Missing: 8.18%
Infinite: 0.00%
ID-ness: 0.01%
Stability: 8.05%
Valid: 83.76%

Distribution



We see here that the percentage of avg rainfall in our selected sample was Indonesia, which received light or medium rain most of the time.

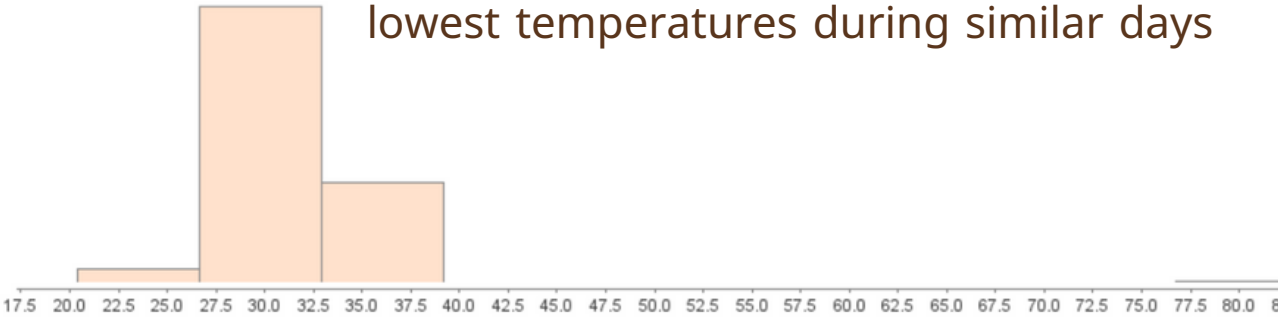
Tx

Summary

Number

Missing: 6.41%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 4.56%
Valid: 89.03%

Distribution



We note here the recurrence of the lowest temperatures during similar days

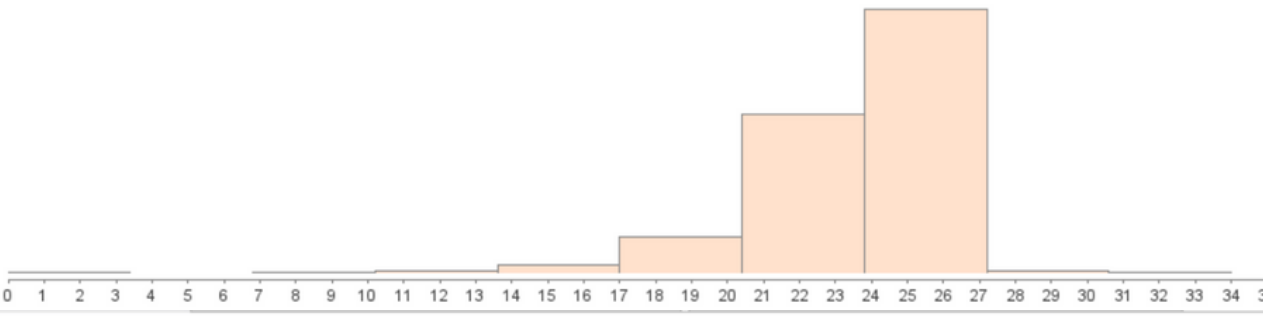
Tn

Summary

Number

Missing: 3.97%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 22.64%
Valid: 73.39%

Distribution



We note here the recurrence of the maximum temperatures during similar days

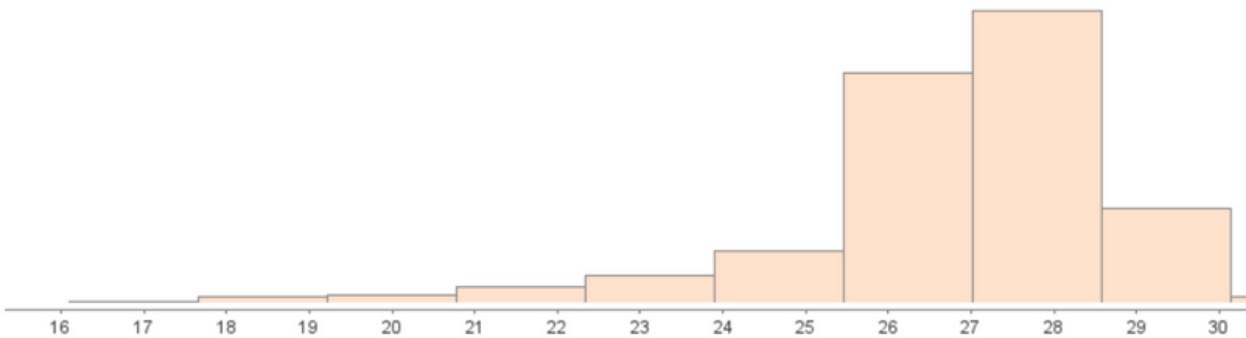
Tavg

Summary

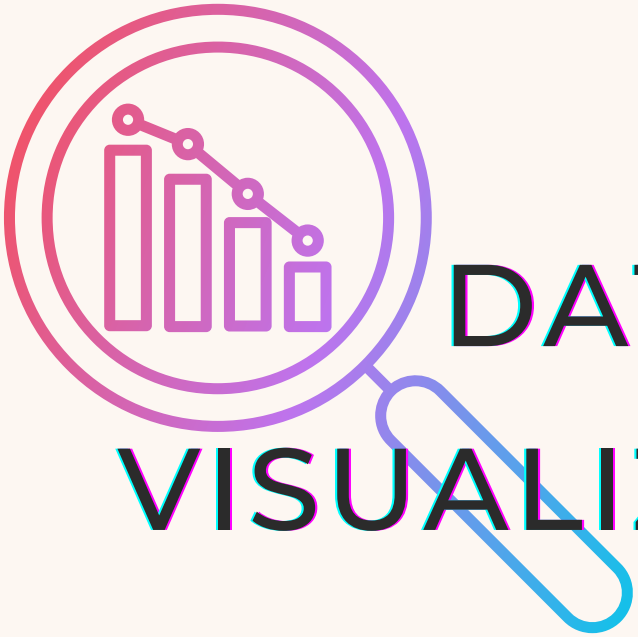
Number

Missing: 7.66%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 3.29%
Valid: 89.05%

Distribution

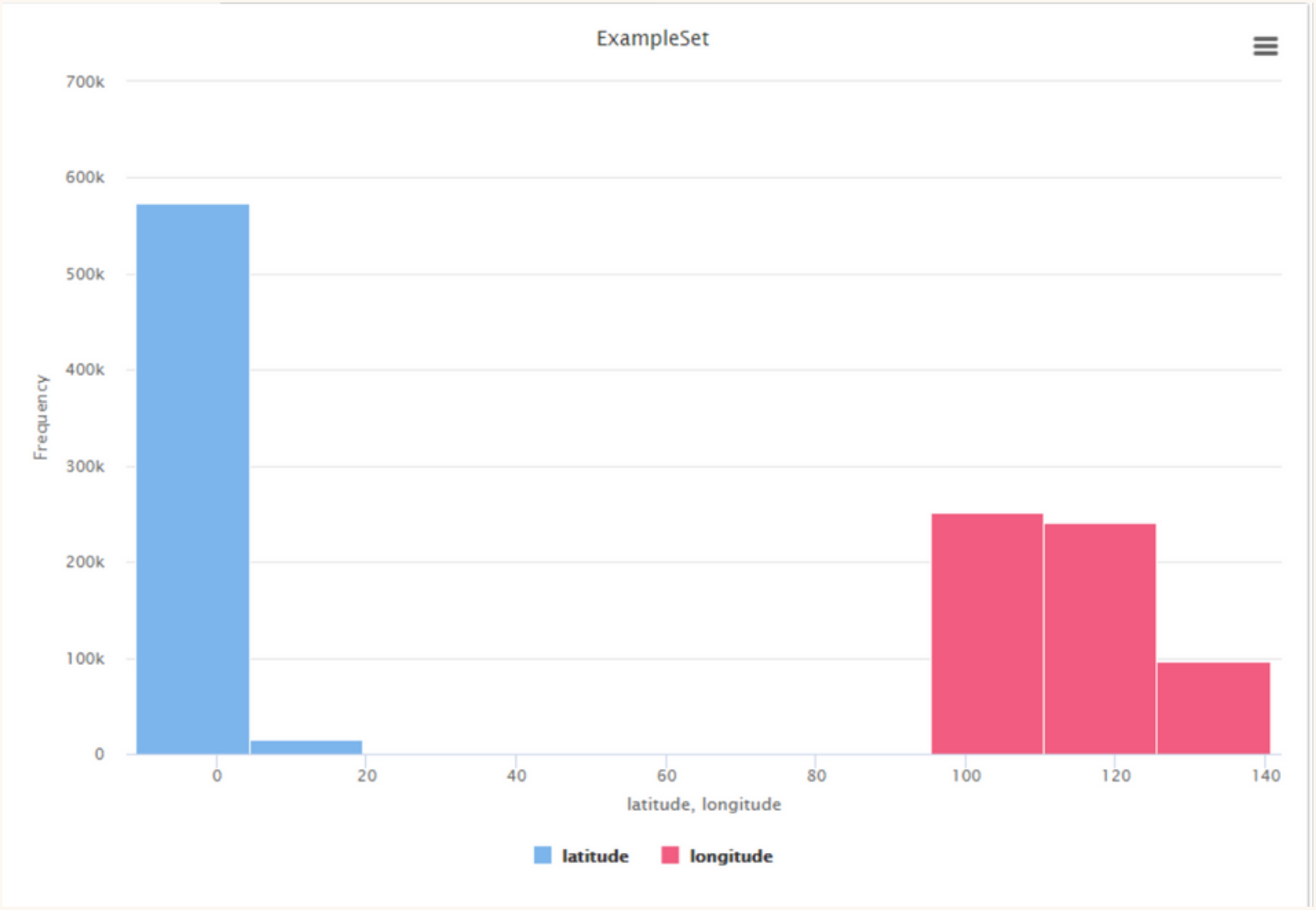


We note here the recurrence of the avg temperatures during similar days



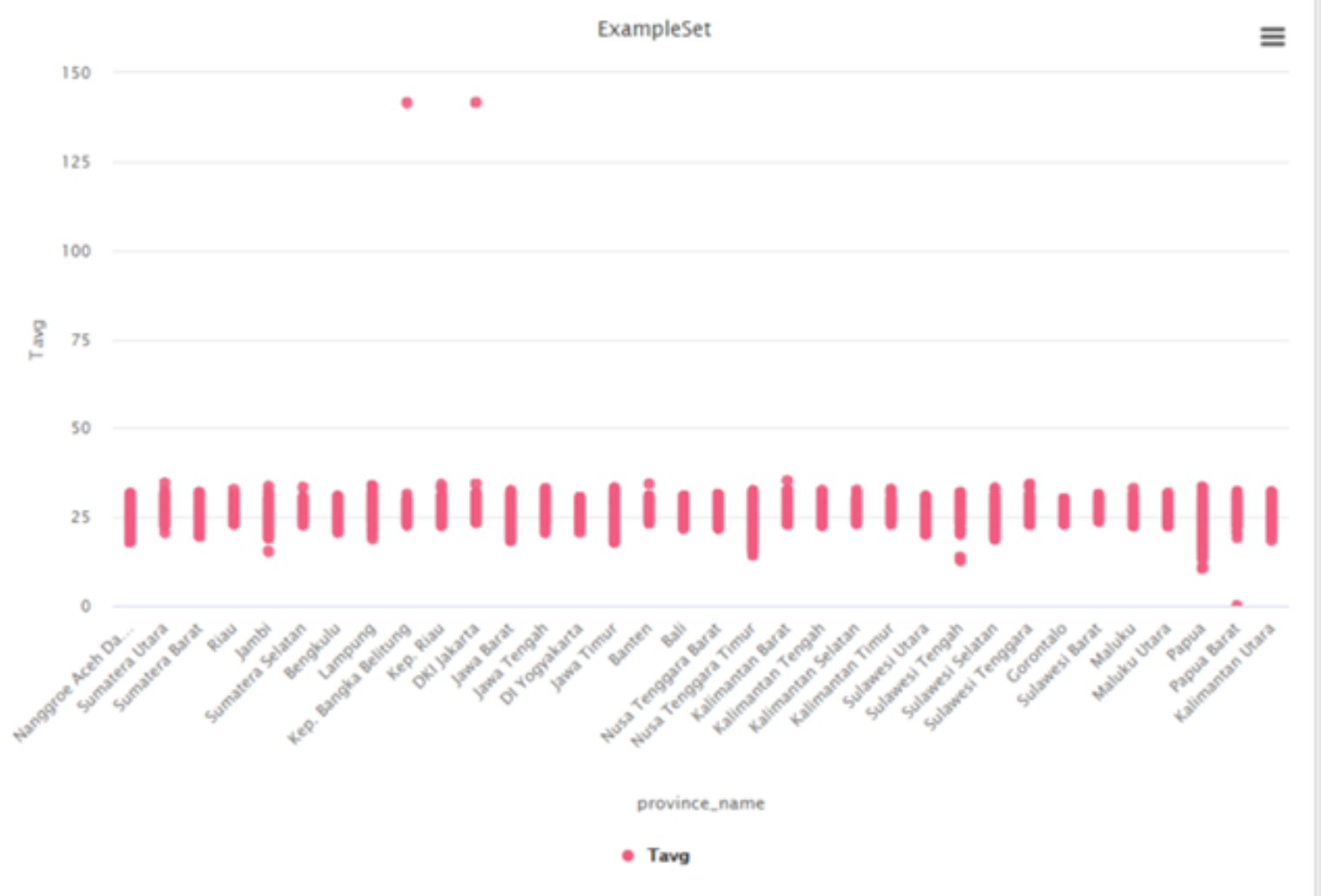
DATA VISUALIZATION

Histogram



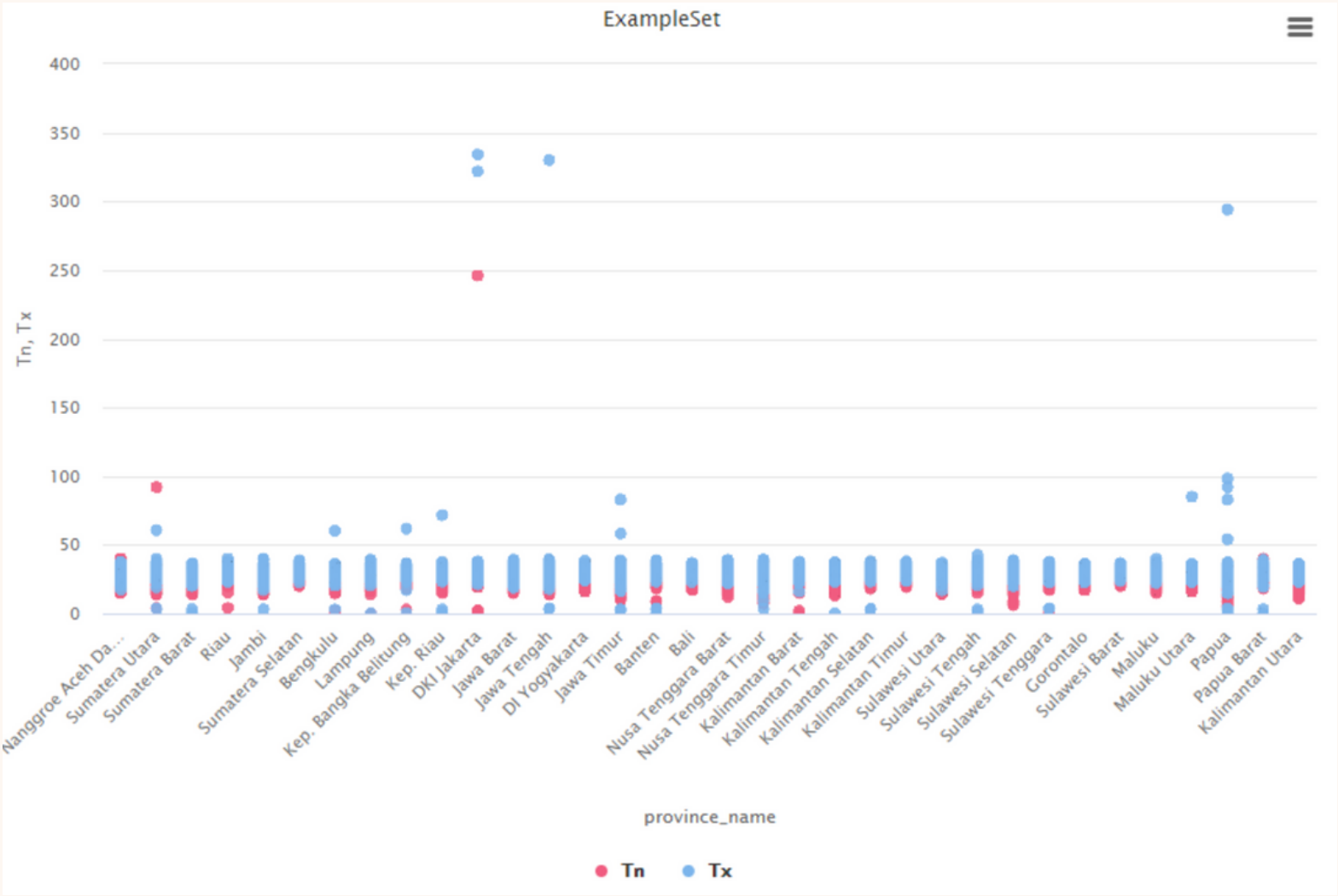
The figure shows latitude and longitude, and latitude lines are the most frequent in this data

Scatter plot 1



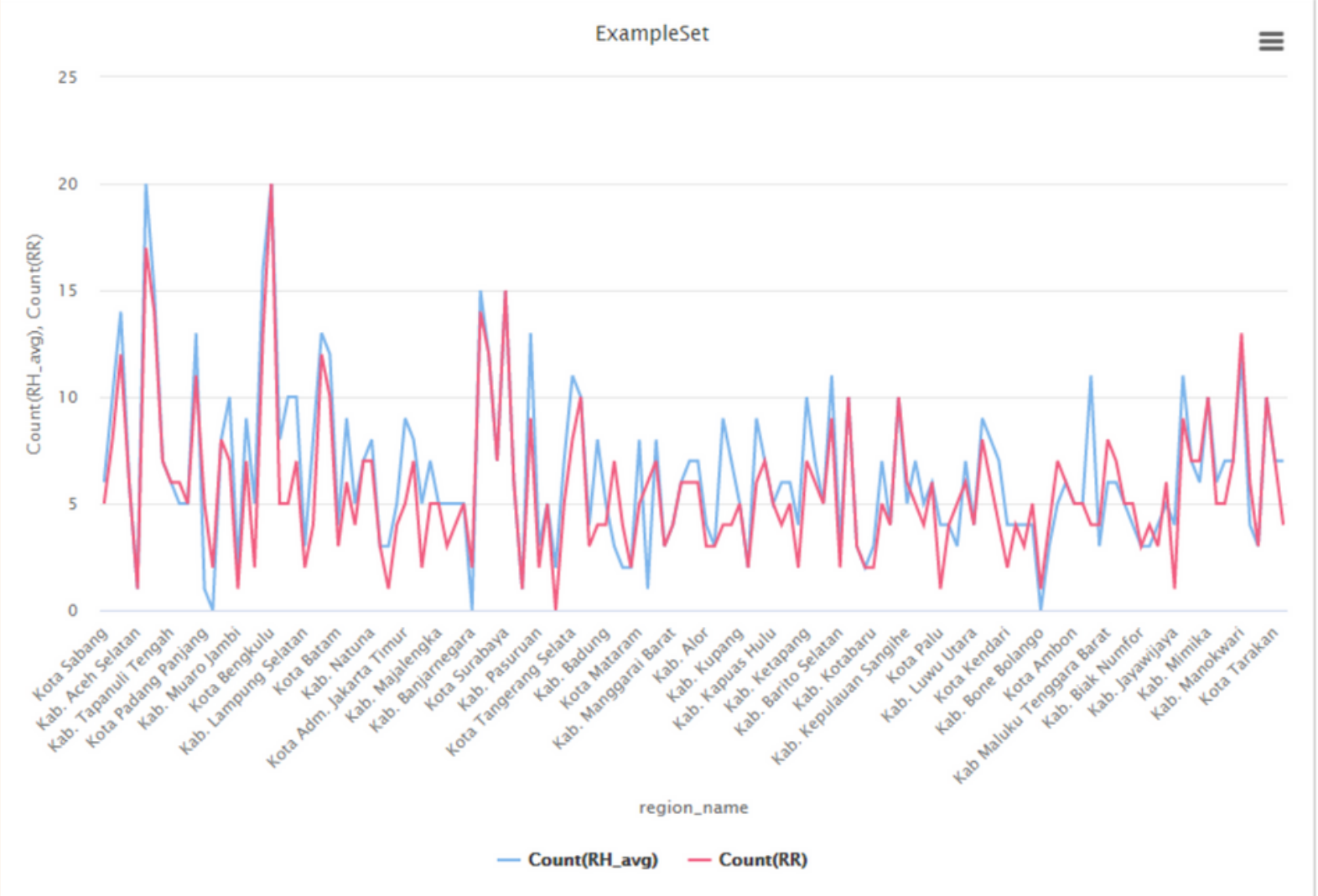
The figure shows the average temperature for all provinces, and it appears that there are outliers in some provinces

Scatter plot 2



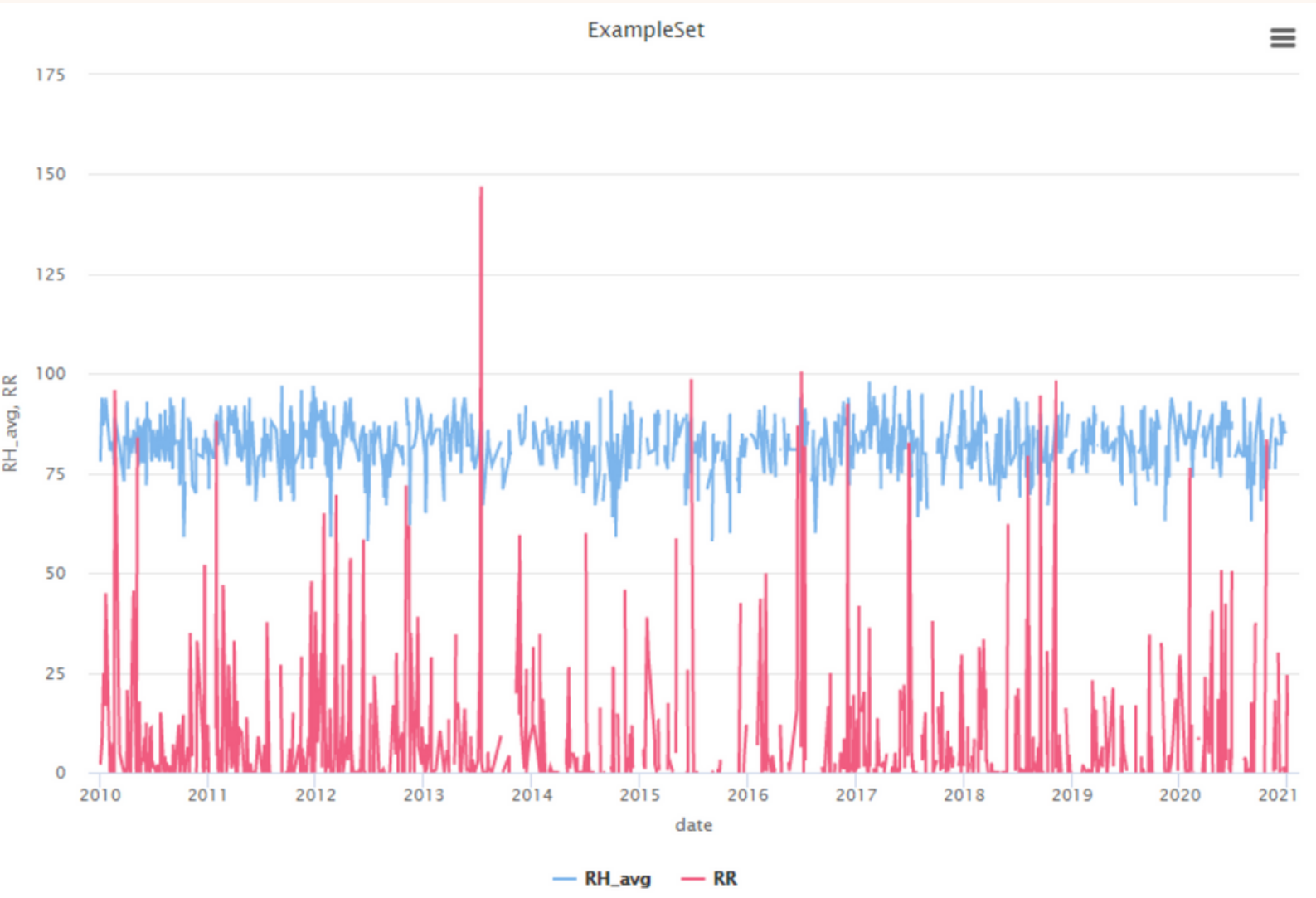
Scatterplot-(T)min-max: shows the maximum and minimum temperature degree on each province

line graph 1



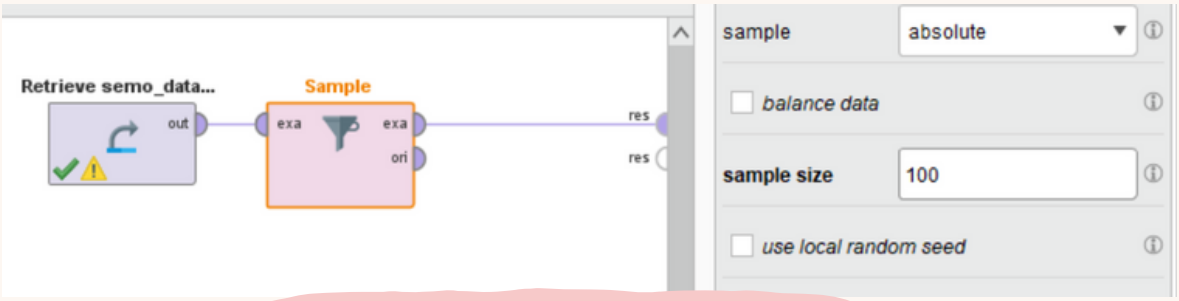
The figure shows both the degree of humidity and the amount of rain for each region, and it is clear that there is a direct relationship between humidity and rain

line graph 2

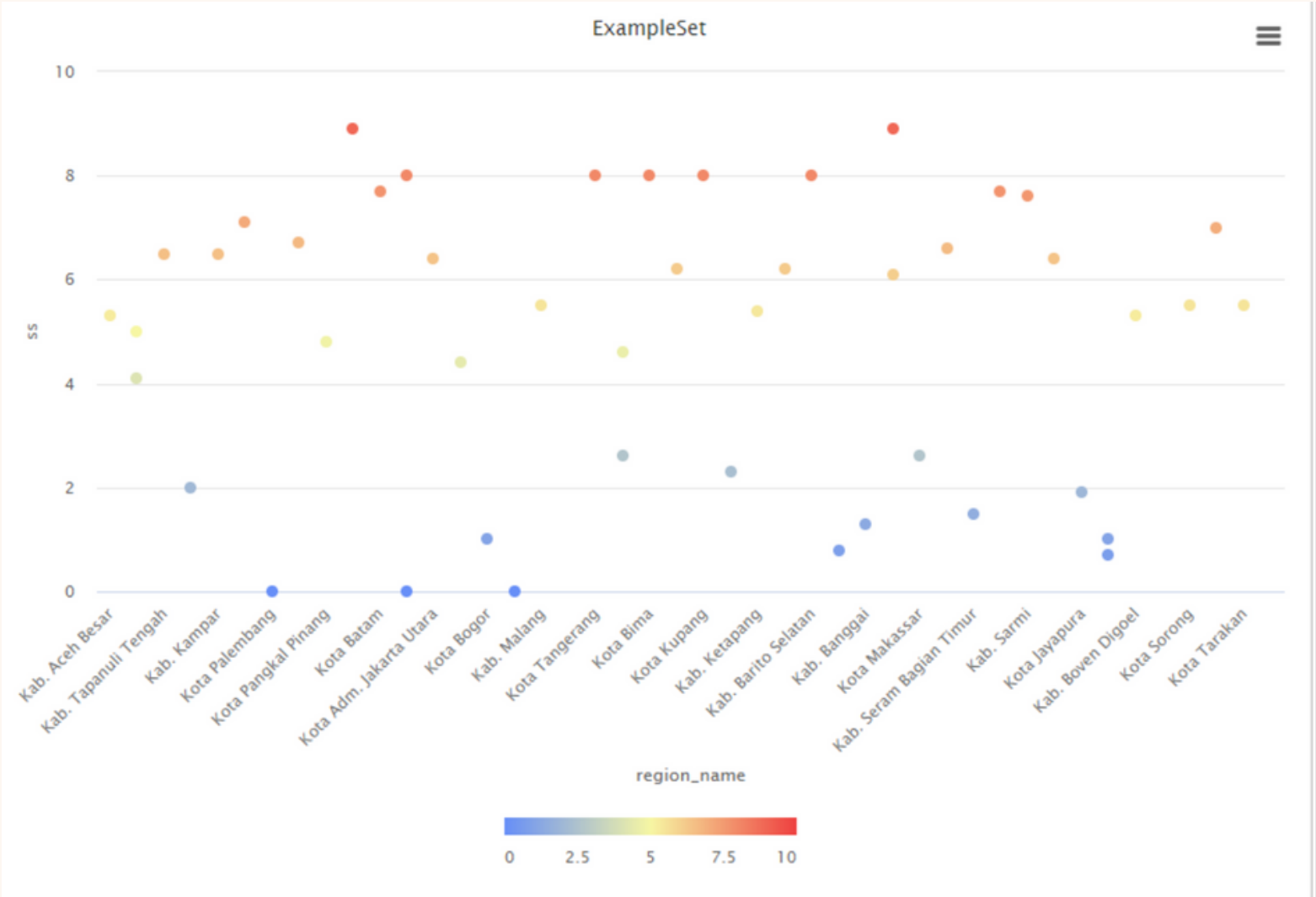


The figure shows both the degree of humidity and the amount of rain for each year, and it is clear that There is a discrepancy in the amount of rain per year, while the humidity is constant

After the sample to illustrate some attributes that Rapid Miner is unable to represent in large quantities

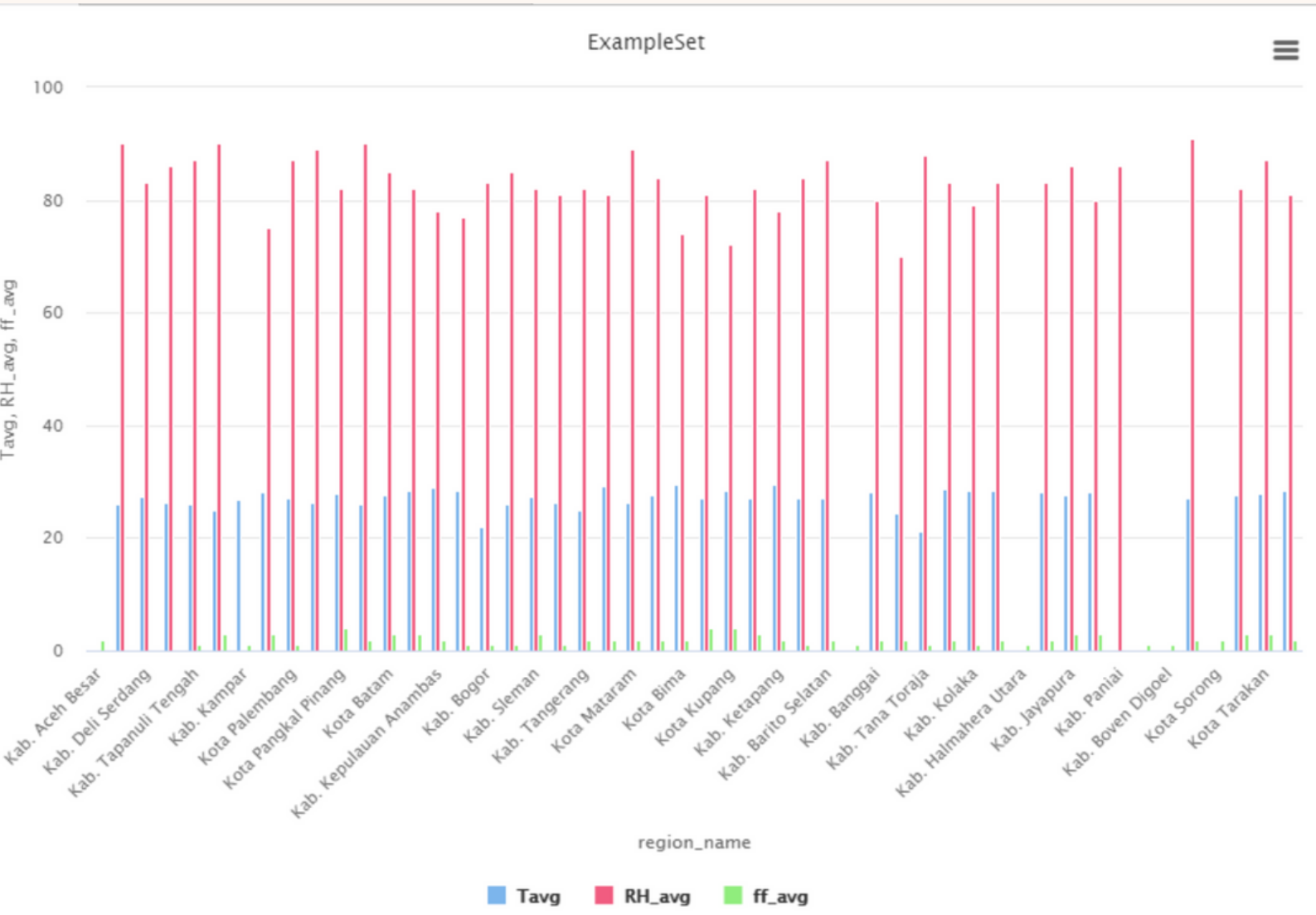


Scatter plot 3



The figure shows the percentage of the sun's brightness for each region, and it appears that there are three regions with the lowest brightness, which are Kota Palembang, Kab Malang and Jakarta.

bar(vertical)

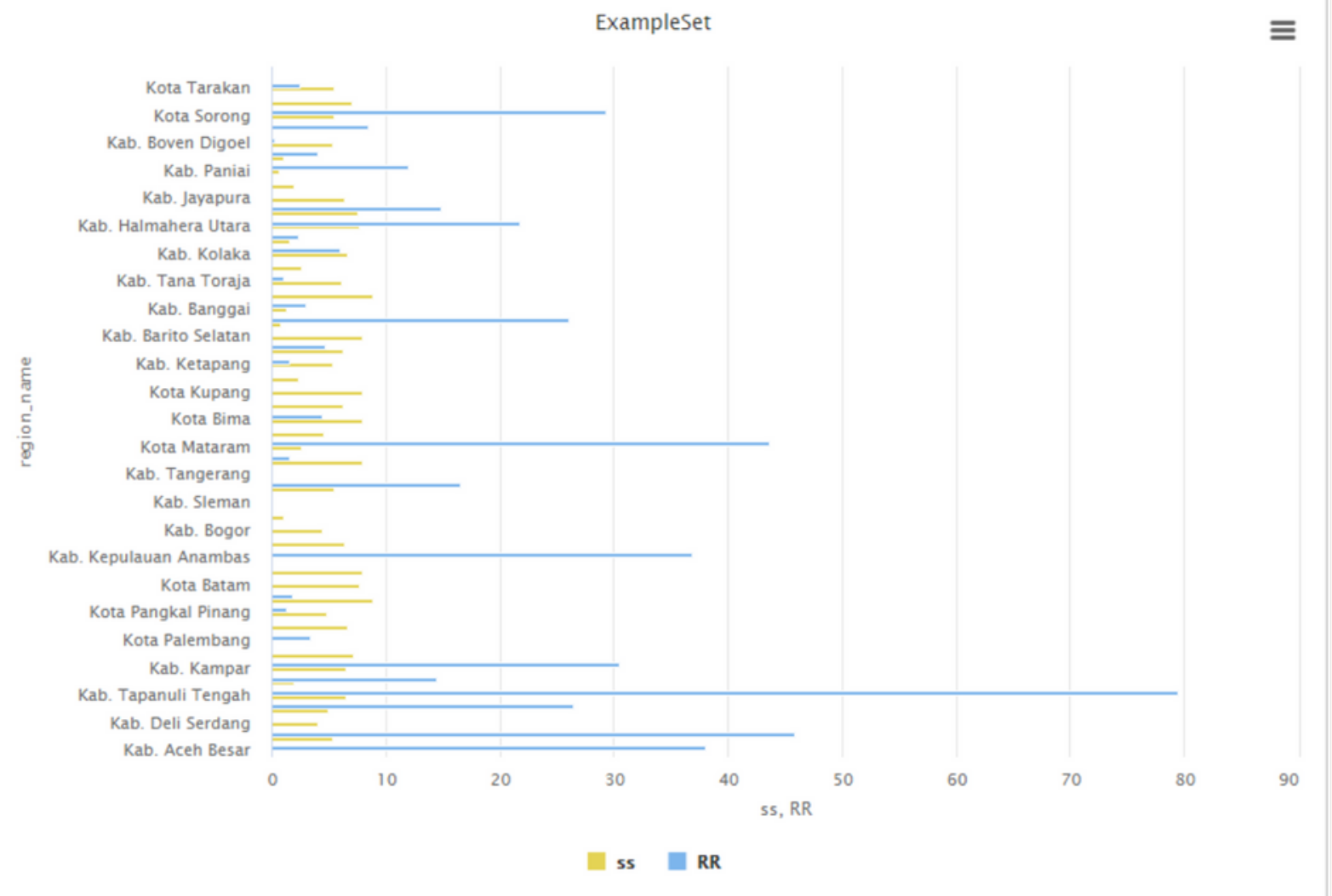


bar(vertical)-region: The figure shows the climate factors, average temperature, humidity, and average maximum wind speed for each region. We can note that all regions are similar for both of the three factors.

bar(horizontal)

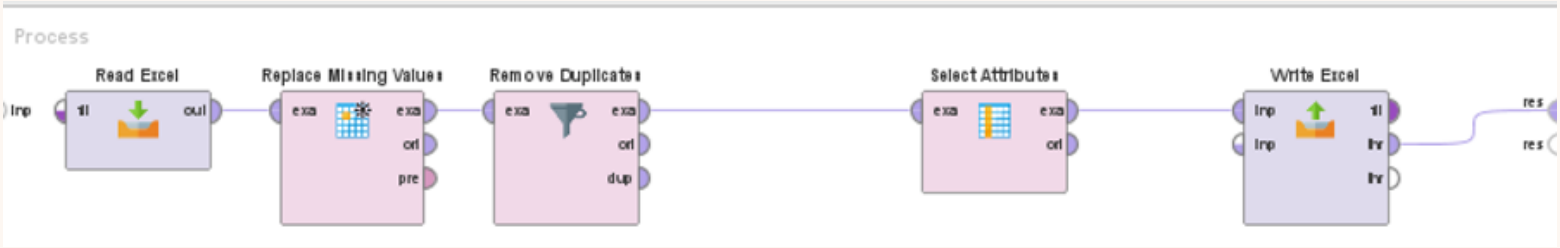
bar(horizontal)-Station: The figure shows the climate factors, average temperature, humidity, and average maximum wind speed for each station that recorded them. We can note that all stations have a similar amount of data.

bar(horizontal)



bar(horizontal)-The figure shows the percentage of the sun's brightness and the amount of rain for each area, and it seems that there is an inverse relationship between the sun's brightness and rain

DATA CLEANING



replace Missing values:

the data with missing values:

Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
27.100	82	9	0.500	7	90	5	E
25.700	95	24	0.200	6	90	4	E
24.500	98	63	0	5	90	4	E
25.800	90	0	0.100	4	225	3	SW
26.700	90	2	0.400	?	?	?	?
26.100	93	11	0.300	?	?	?	?
25.400	96	2	0.100	?	?	?	?
26.800	91	3	0.600	5	90	4	E

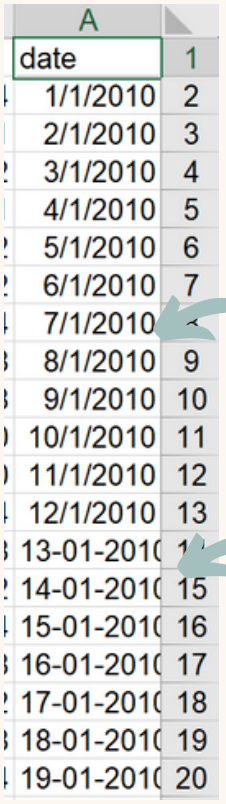
	Missing Values	% Value
RR	125384	21
RH_avg	48182	8
Tavg	45105	7
ss	43721	7
Tx	37736	6
Tn	23383	3
ddd_car	13739	2
ddd_x	13128	2
ff_x	10214	1
ff_avg	10127	1
date	0	0
station_id	0	0

the data after replacing missing values by avg :

ss	ff_x	ddd_x	ff_avg	ddd_car
11.800	7	280	3	W
11.800	6	210	4	S
11.600	9	300	4	W
11.500	5	150	2	S
11.500	7	250	4	SW
11.500	7	60	2	NE
11.500	6	270	3	W
11.500	4	190	2	S

Click to sort

inconsisetant data:



A	
date	1
1/1/2010	2
2/1/2010	3
3/1/2010	4
4/1/2010	5
5/1/2010	6
6/1/2010	7
7/1/2010	8
8/1/2010	9
9/1/2010	10
10/1/2010	11
11/1/2010	12
12/1/2010	13
13-01-2010	14
14-01-2010	15
15-01-2010	16
16-01-2010	17
17-01-2010	18
18-01-2010	19
19-01-2010	20

with different format / and -

using excel features select

"**data text to column**"

date
1/1/2010
2/1/2010
3/1/2010
4/1/2010
5/1/2010
6/1/2010
7/1/2010
8/1/2010
9/1/2010
10/1/2010
11/1/2010
12/1/2010

Redundant attributes:

the dataset contains many attributes called **redundant** meaning that an attribute it can be derived from any other attribute or set of attributes.

Inconsistencies in attribute or dimension naming can also lead to the redundancies in data set.

Row No.	province_id	province_na...	station_id	station_name	region_name	latitude	longitude	regio
1	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
2	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
3	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
4	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
5	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
6	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
7	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
8	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
9	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20
10	1	Nanggroe Ac...	96001	Stasiun Mete...	Kota Sabang	5.877	95.338	20

We dealt with this problem by discovering the correlation between them and then deleting repeated groups, normalization..etc,

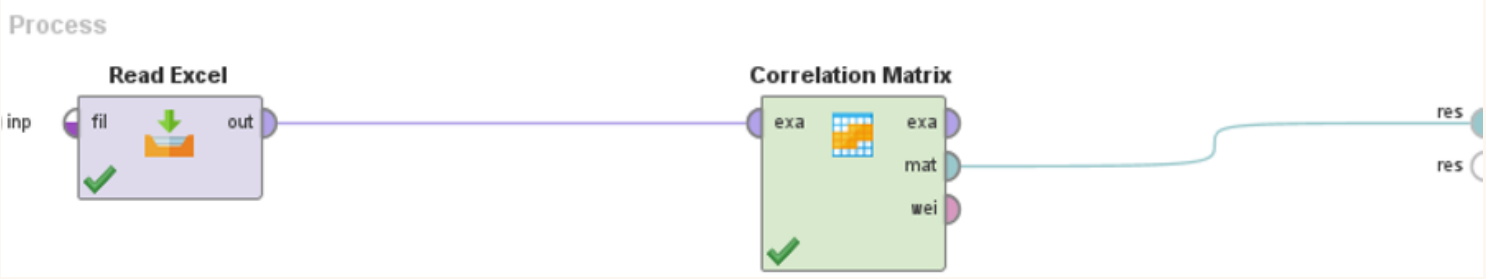
Duplication Data:

Data duplication happens when you store information about the same entity multiple times, instead of updating a single record

We dealt with this problem by discovering the correlation between them and then deleting repeated groups, normalization..etc, using removes duplicat operator

correlated attributes

Correlation is a statistical technique that can show whether and how strongly pairs of Attributes are related.



Correlation matrix operator determines correlation between all Attributes and it can produce a weights vector based on these correlations.

Attribut...	latitude	longitude	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg
latitude	1	-0.269	0.006	0.038	0.020	0.216	0.025	-0.208	-0.131	0.019	-0.129
longitude	-0.269	1	0.030	-0.084	-0.011	0.010	0.017	0.087	0.061	-0.014	0.053
Tn	0.006	0.030	1	0.485	0.785	-0.073	-0.078	0.072	0.112	0.008	0.079
Tx	0.038	-0.084	0.485	1	0.770	-0.432	-0.195	0.277	0.081	-0.046	0.031
Tavg	0.020	-0.011	0.785	0.770	1	-0.391	-0.167	0.216	0.119	-0.045	0.096
RH_avg	0.216	0.010	-0.073	-0.432	-0.391	1	0.298	-0.385	-0.197	0.103	-0.298
RR	0.025	0.017	-0.078	-0.195	-0.167	0.298	1	-0.196	-0.061	0.056	-0.085
ss	-0.208	0.087	0.072	0.277	0.216	-0.385	-0.196	1	0.093	-0.084	0.086
ff_x	-0.131	0.061	0.112	0.081	0.119	-0.197	-0.061	0.093	1	0.051	0.520
ddd_x	0.019	-0.014	0.008	-0.046	-0.045	0.103	0.056	-0.084	0.051	1	0.011
ff_avg	-0.129	0.053	0.079	0.031	0.096	-0.298	-0.085	0.086	0.520	0.011	1

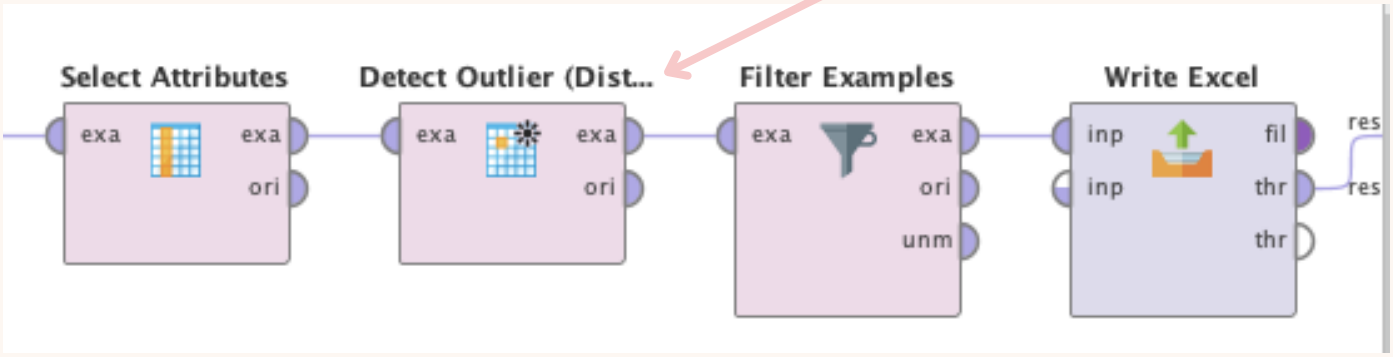
Tn and Tavg strongly correlated

Tx and Tavg strongly correlated

Outliers

We discovered in **visualization** that the dataset contains outliers a values that "lies outside" (is much smaller or larger than) most of the other values in a set of data.

So we dealt with this problem by using "**detect outlier**" operator in RapidMiner



As you see there is a column generated (name"**outlier**") with **false** which means that there are just rows that haven't an outlier

ddd_car	outlier
SW	false
S	false
N	false
N	false
N	false
E	false

In last of **cleaning process**, we generate a new sample of data without outliers , missing values , inconsistent or even redundant attributes

phase 2

Sittings of process

Generate Attributes operator

This operator constructs new user defined attributes using mathematical expressions.

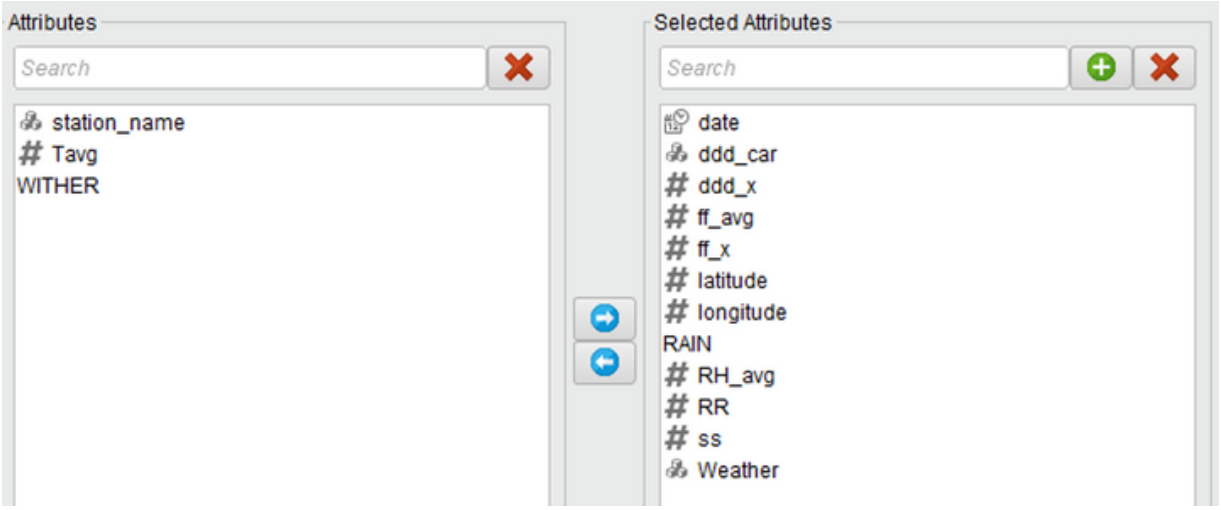
we create new column for new attribute "weather" and fills this column with corresponding values of this attribute.

If the value in Tavg > 28 then "weather" is **HOT**

If the value in Tavg <= 28 then "weather" is **COLD**

Select Attributes operator

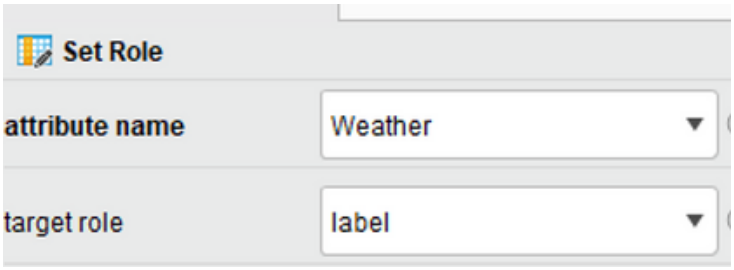
This Operator selects a subset of Attributes and removes the other Attributes.



We use it to selects attributes that we need and removes the others.

Set Role operator

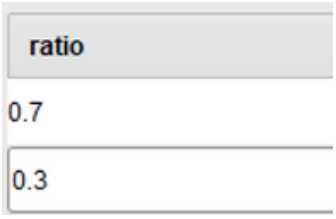
This Operator is used to change the role of one or more Attributes.



we use "WEATHER" as label

Split Data operator

This Operator takes input and delivers the subsets through its output ports



we split the data into two parts.
one part (70%) for training and other (30%) for testing the model.

our pre-processing(in ph1(replace missing values , etc...)) was applied in a new file excel and we continue with that file to improve pre-processing

old file (file with all pre-processing **except** replace missing values)

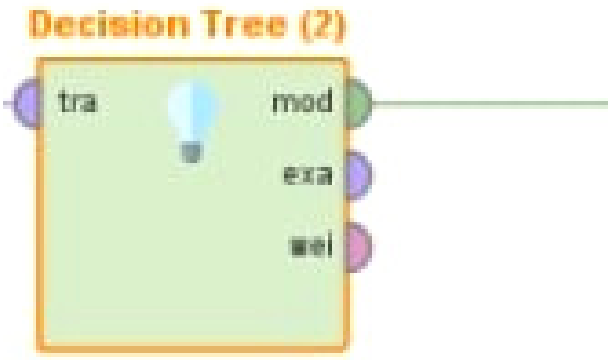
new file (file with all pre-processing + replace missing values)

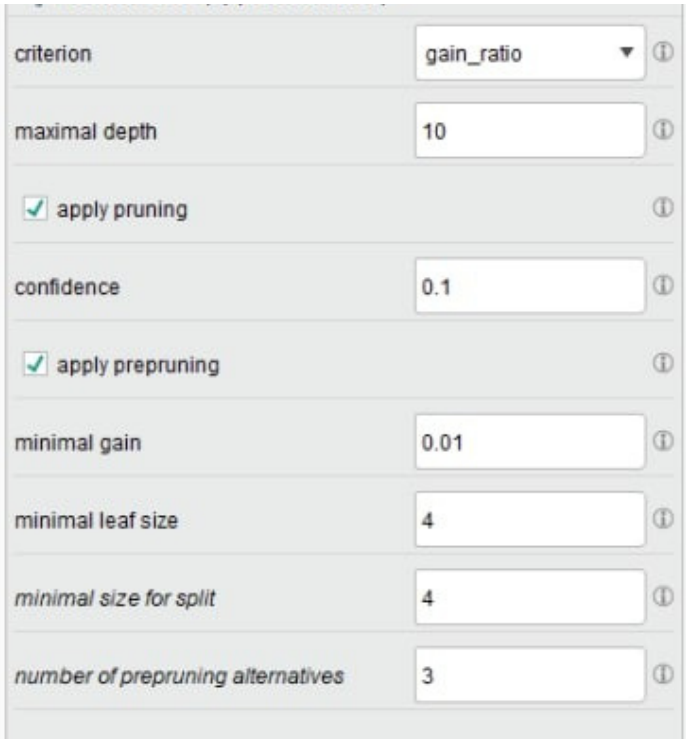
Decision Tree

A decision tree is a tree like collection of nodes intended to create a decision on values affiliation to a class or an estimate of a numerical target value. Each node represents a splitting rule for one specific Attribute.

the first model is the decision tree with different pre-processing techniques one by one, we aim to achieve the best accuracy.
to observe which pre-processing technique has improved the accuracy under the same classifier or even due to diff classifier models on the same dataset would the accuracy be the same for all or not?
in order to observe this let's begin

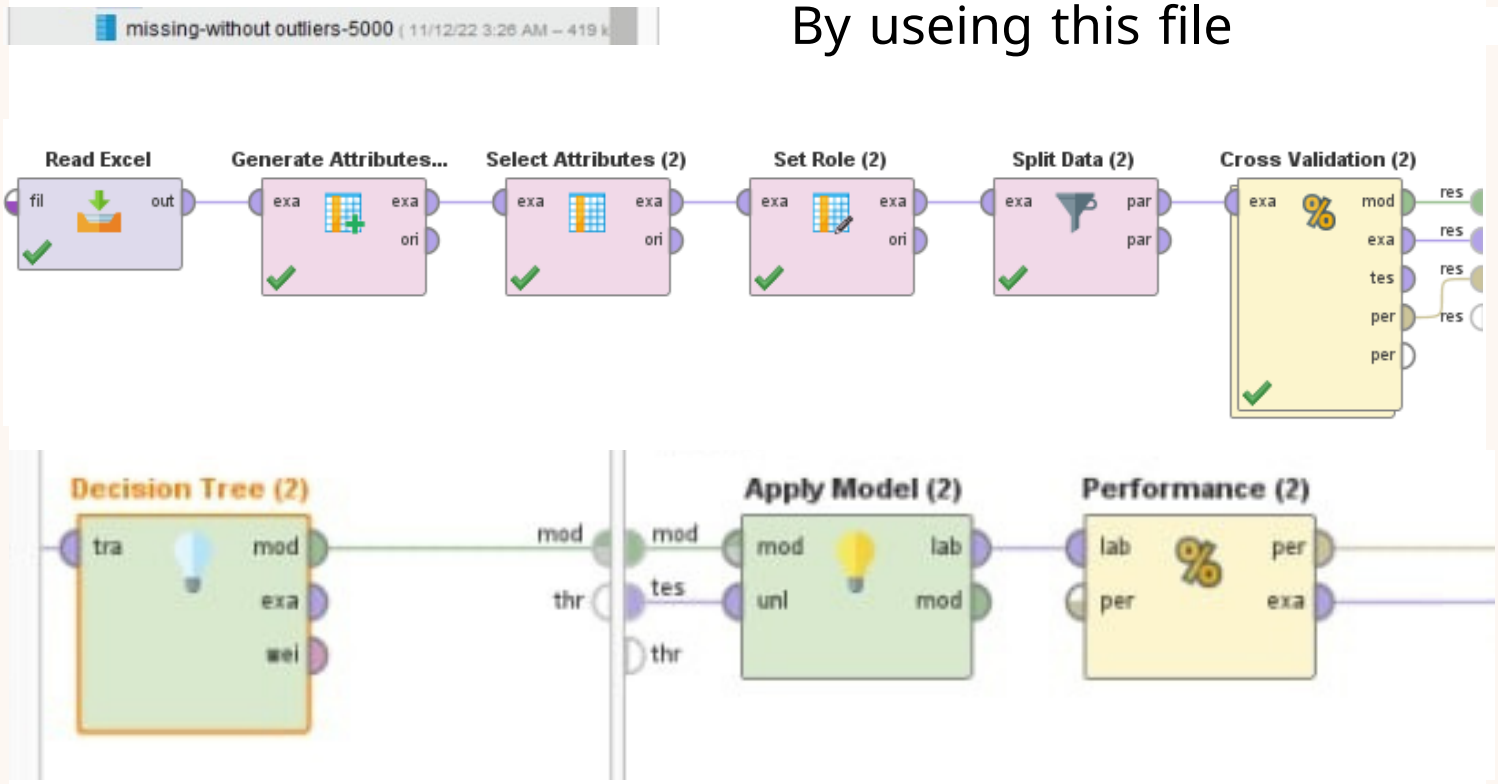
Decision Tree settings





with missing value(old file)

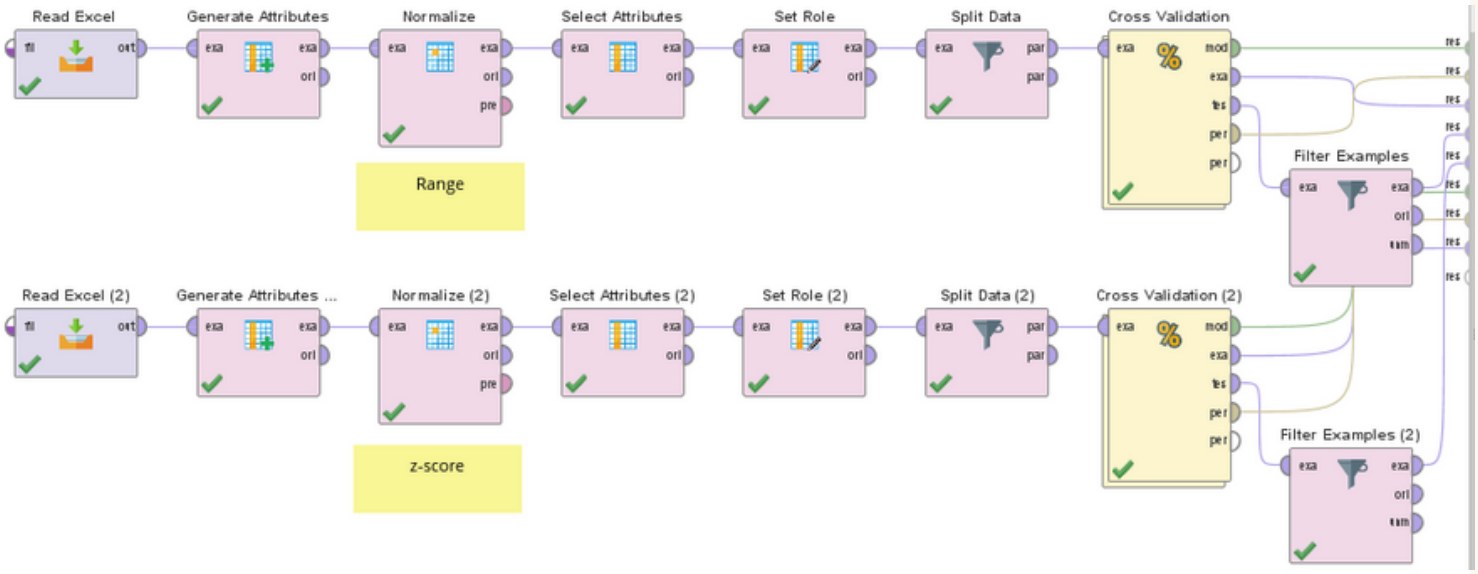
By useing this file



accuracy: 78.29% +/- 2.24% (micro average: 78.29%)

	true cold	true hot	class precision
pred. cold	1089	263	80.55%
pred. hot	62	83	57.24%
class recall	94.61%	23.99%	

a- normalizing the data(new file)



type1-Range

accuracy: 77.89% +/- 2.25% (micro average: 77.89%)

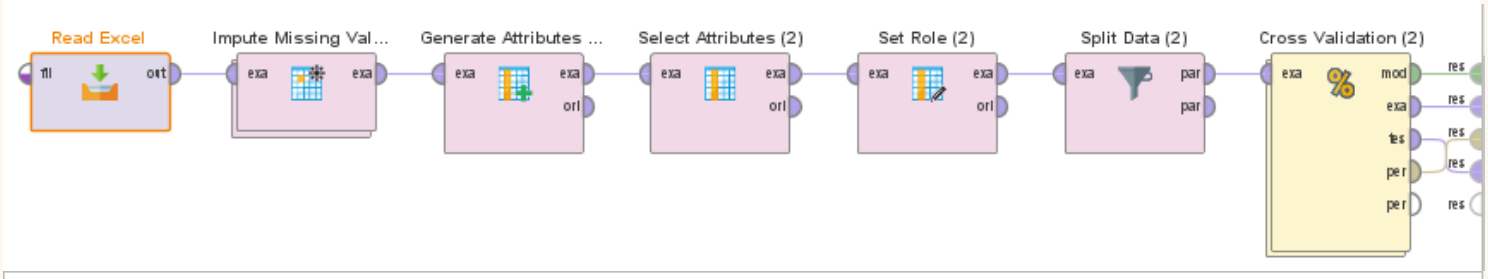
	true cold	true hot	class precision
pred. cold	1086	266	80.33%
pred. hot	65	80	55.17%
class recall	94.35%	23.12%	

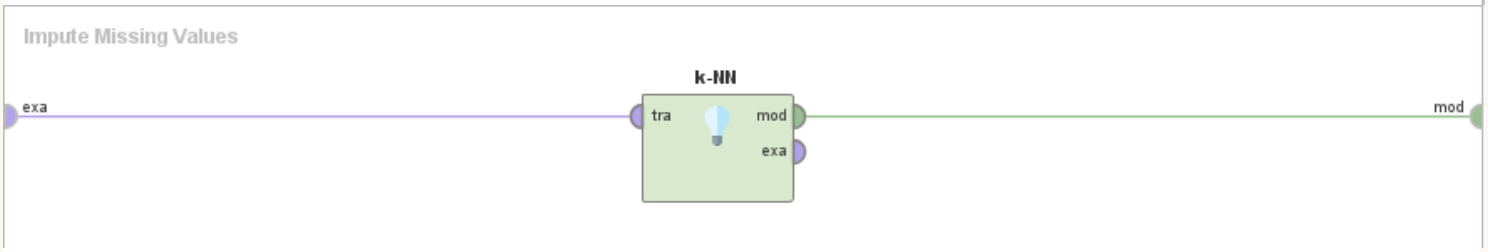
type2-z-score

accuracy: 78.02% +/- 2.12% (micro average: 78.02%)

	true cold	true hot	class precision
pred. cold	1097	275	79.96%
pred. hot	54	71	56.80%
class recall	95.31%	20.52%	

b- imputing missing values(old file)

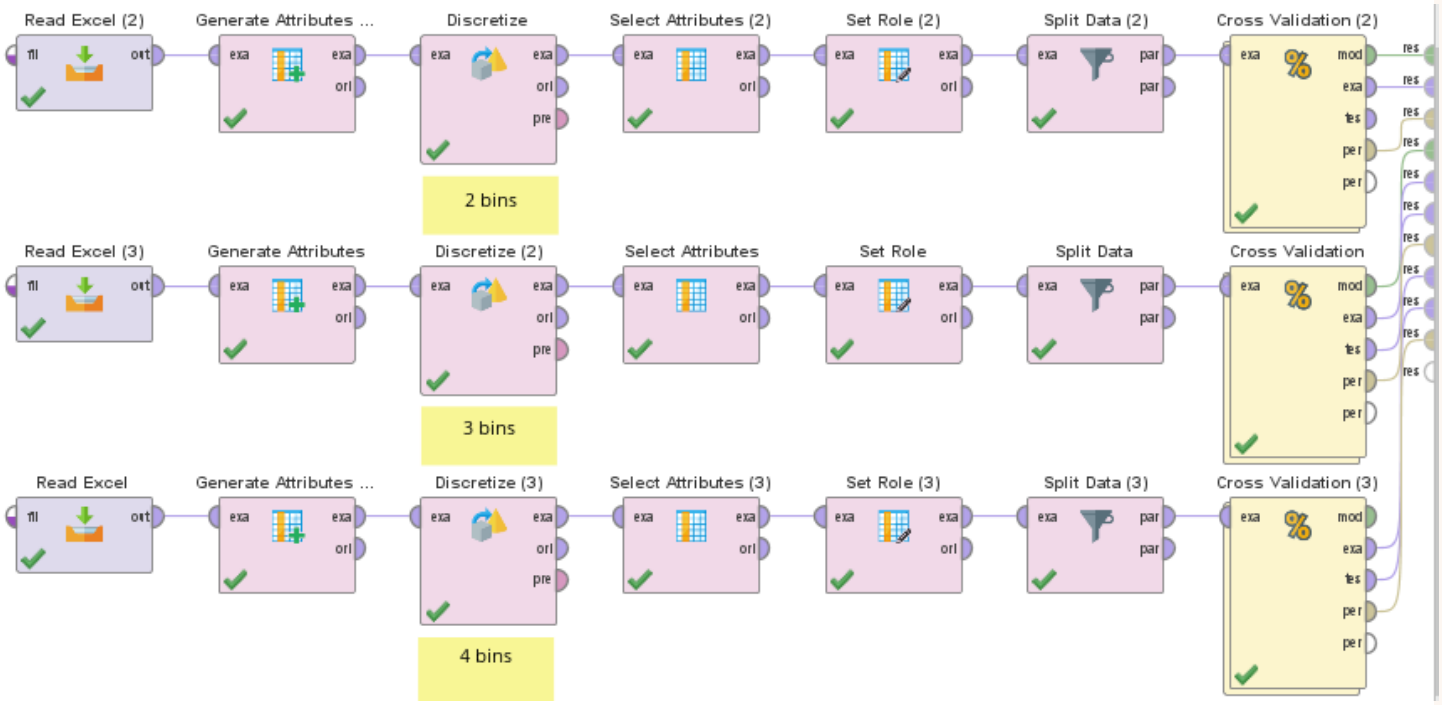




accuracy: 77.89% +/- 2.25% (micro average: 77.89%)

	true cold	true hot	class precision
pred. cold	1086	266	80.33%
pred. hot	65	80	55.17%
class recall	94.35%	23.12%	

c- discretizing 2, 3, and 4 bins



2 bins

accuracy: 78.29% +/- 1.58% (micro average: 78.29%)

	true cold	true hot	class precision
pred. cold	1088	262	80.59%
pred. hot	63	84	57.14%
class recall	94.53%	24.28%	

3 bins

accuracy: 78.23% +/- 2.07% (micro average: 78.22%)

	true cold	true hot	class precision
pred. cold	1097	272	80.13%
pred. hot	54	74	57.81%
class recall	95.31%	21.39%	

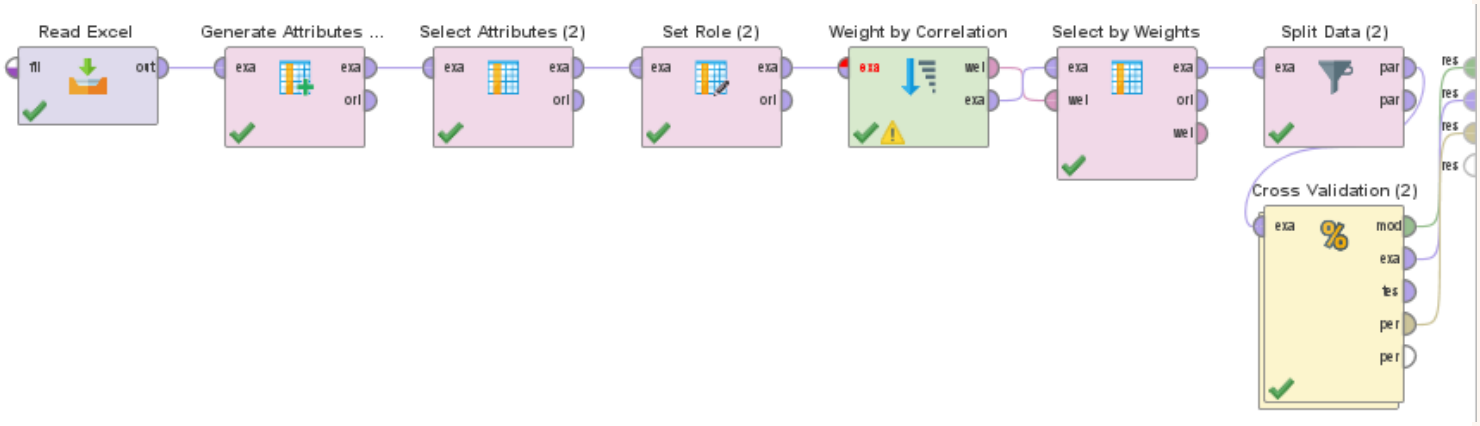
4 bins

accuracy: 78.76% +/- 2.09% (micro average: 78.76%)

	true cold	true hot	class precision
pred. cold	1096	263	80.65%
pred. hot	55	83	60.14%
class recall	95.22%	23.99%	

d- reducing dimensions(new file)

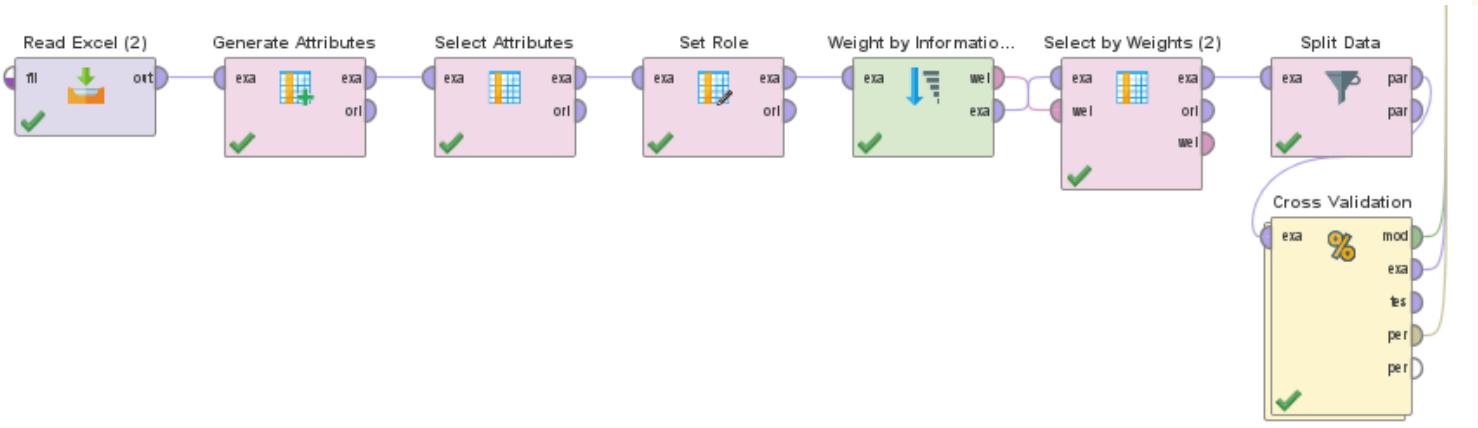
type1-weight by coralationns



accuracy: 78.16% +/- 2.11% (micro average: 78.16%)

	true cold	true hot	class precision
pred. cold	1098	274	80.03%
pred. hot	53	72	57.60%
class recall	95.40%	20.81%	

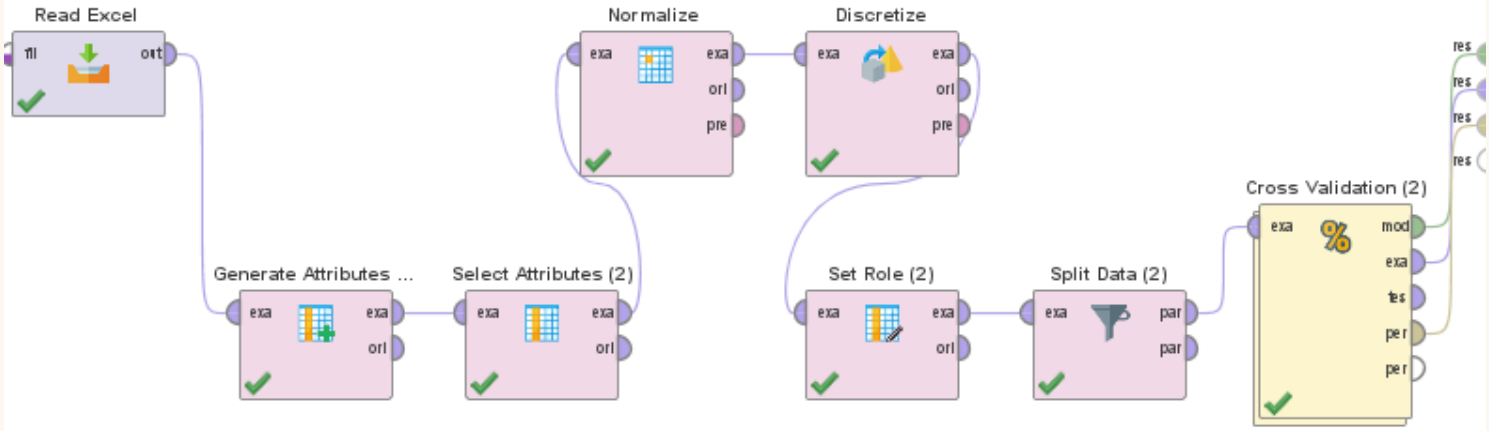
type2- weight by information gain



accuracy: 77.89% +/- 2.25% (micro average: 77.89%)

	true cold	true hot	class precision
pred. cold	1086	266	80.33%
pred. hot	65	80	55.17%
class recall	94.35%	23.12%	

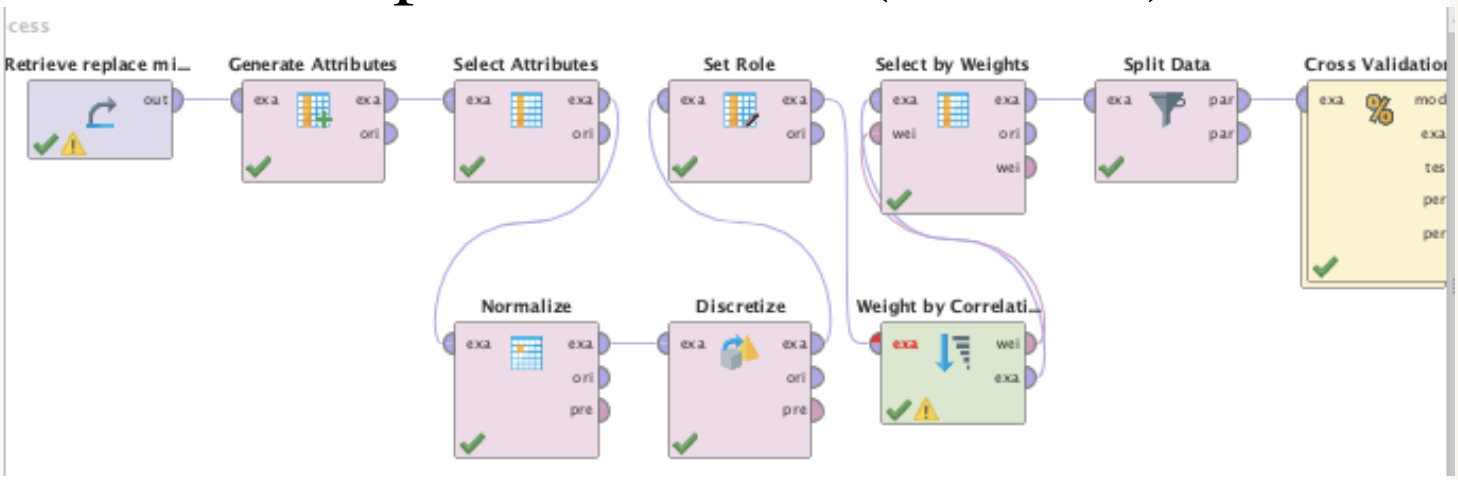
e- all previous tasks exclude reduce dimension (new file)



accuracy: 78.76% +/- 2.09% (micro average: 78.76%)

	true cold	true hot	class precision
pred. cold	1096	263	80.65%
pred. hot	55	83	60.14%
class recall	95.22%	23.99%	

f- all previous tasks (new file)



accuracy: 78.83% +/- 2.40% (micro average: 78.82%)

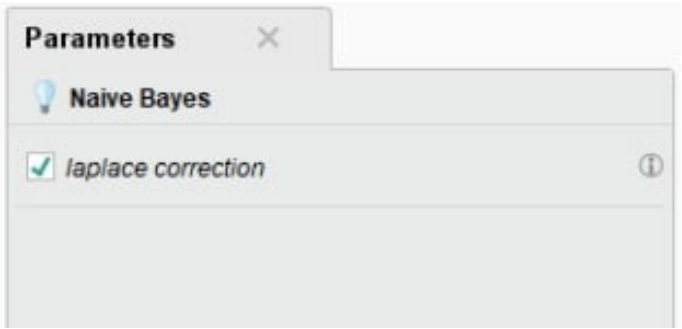
	true cold	true hot	class precision
pred. cold	1073	239	81.78%
pred. hot	78	107	57.84%
class recall	93.22%	30.92%	

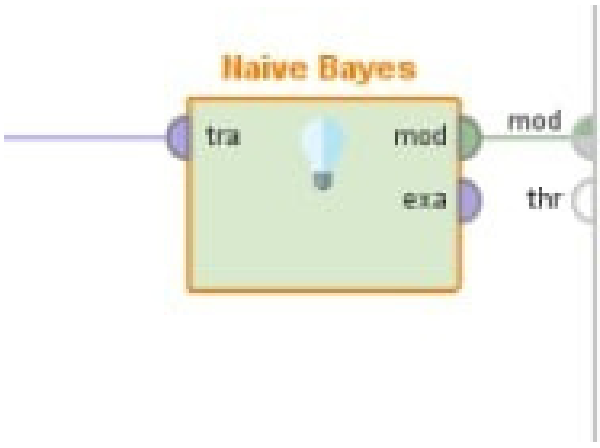
Naive Bayes

Naive Bayes is a high-bias, low-variance classifier, Typical use cases involve text categorization, including spam detection, sentiment analysis, and recommender systems.

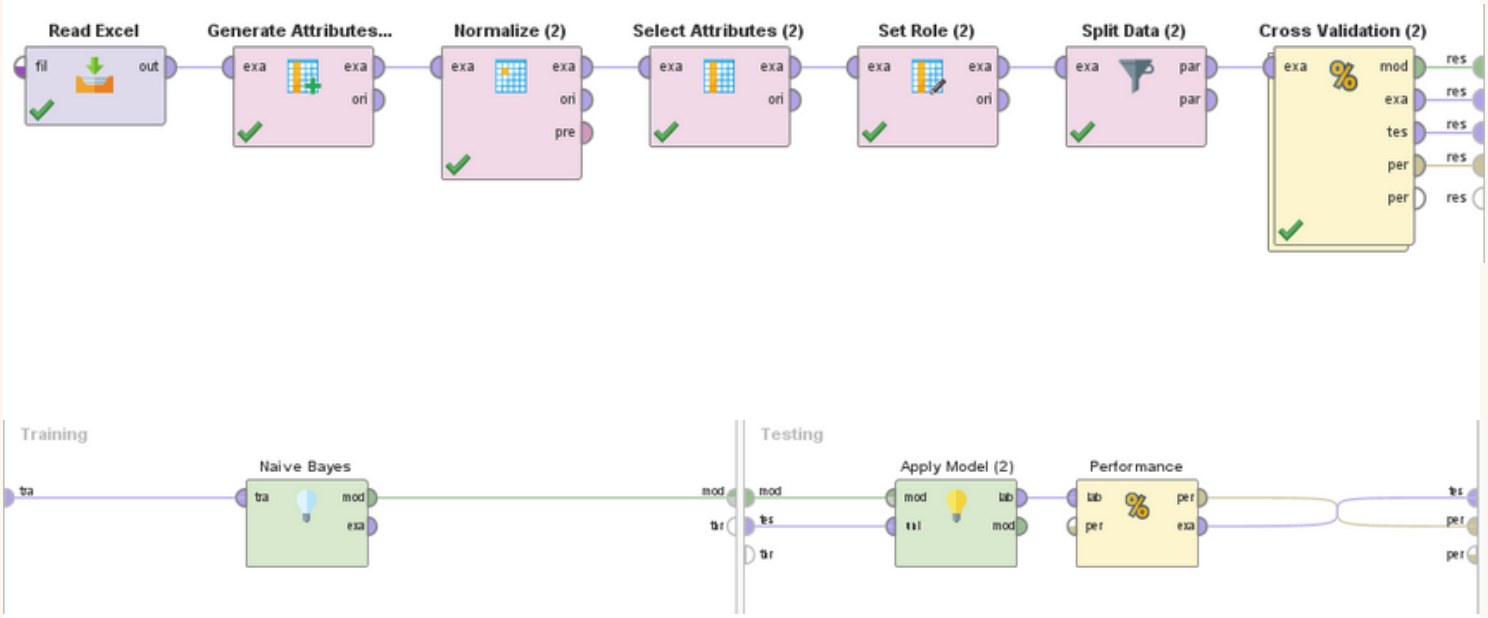
applied Naïve Bayes classifier in order to gain accuracy and then do comparison. Lets have a look on model using classifier Naïve Bayes.

Naive Bayes settings





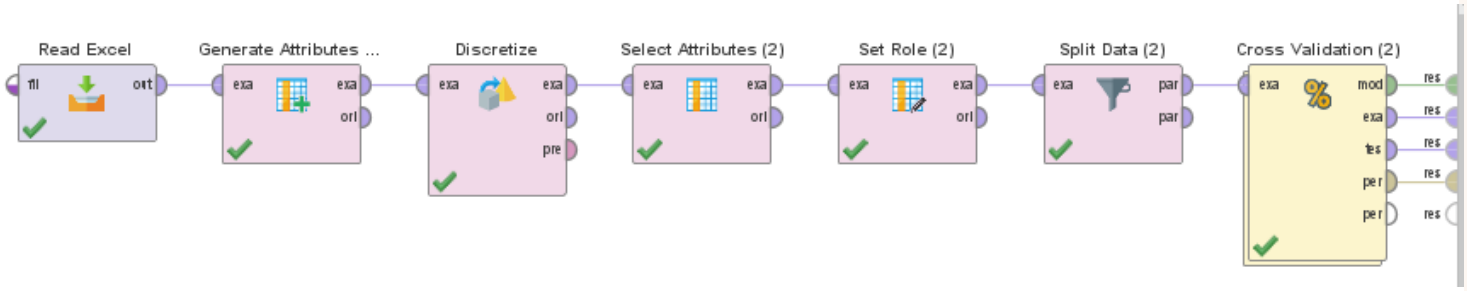
1-normalizing the data(new file)



accuracy: 78.16% +/- 2.94% (micro average: 78.16%)

	true cold	true hot	class precision
pred. cold	970	146	86.92%
pred. hot	181	200	52.49%
class recall	84.27%	57.80%	

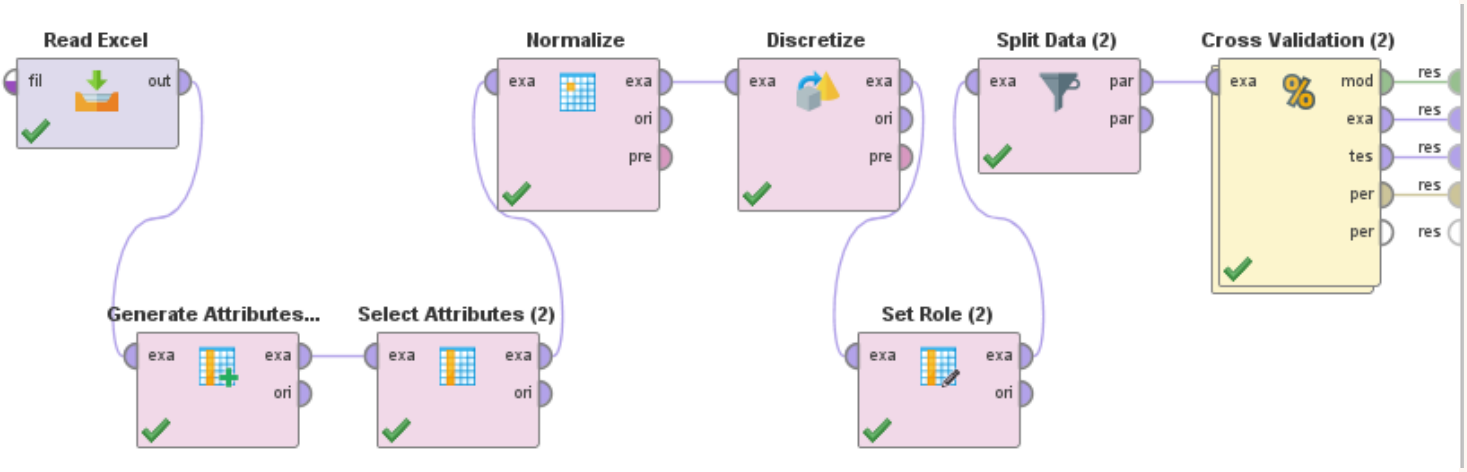
2- discretizing (new file)



accuracy: 79.09% +/- 1.95% (micro average: 79.09%)

	true cold	true hot	class precision
pred. cold	1033	195	84.12%
pred. hot	118	151	56.13%
class recall	89.75%	43.64%	

3- all previous tasks exclude reduce dimension(new file)



accuracy: 79.03% +/- 2.03% (micro average: 79.02%)

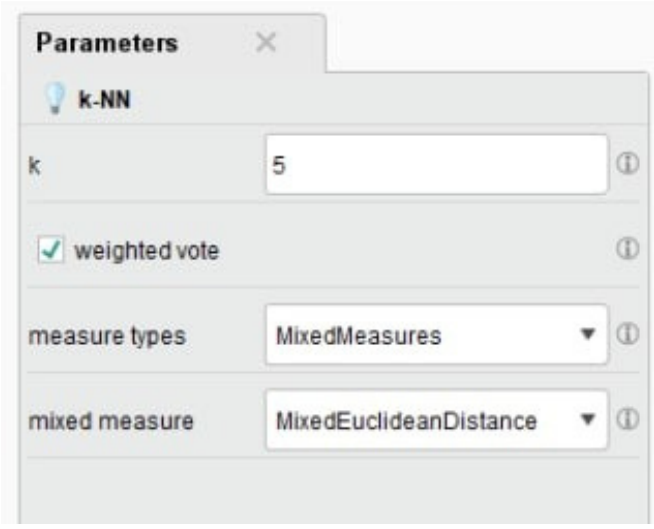
	true cold	true hot	class precision
pred. cold	1033	196	84.05%
pred. hot	118	150	55.97%
class recall	89.75%	43.35%	

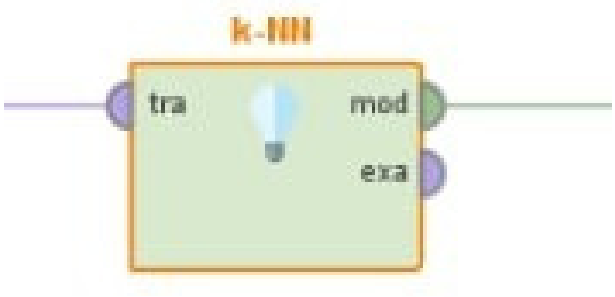
KNN

The k-Nearest Neighbor algorithm is based on comparing an unknown Example with the k training Examples which are the nearest neighbors of the unknown Example.

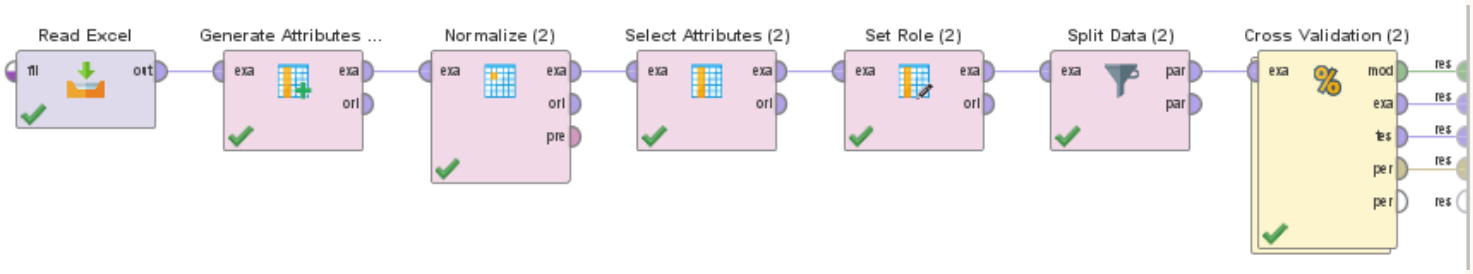
We have applied KNN classifier in order to gain accuracy and then do comparison. Lets have a look on model using classifier KNN.

KNN settings





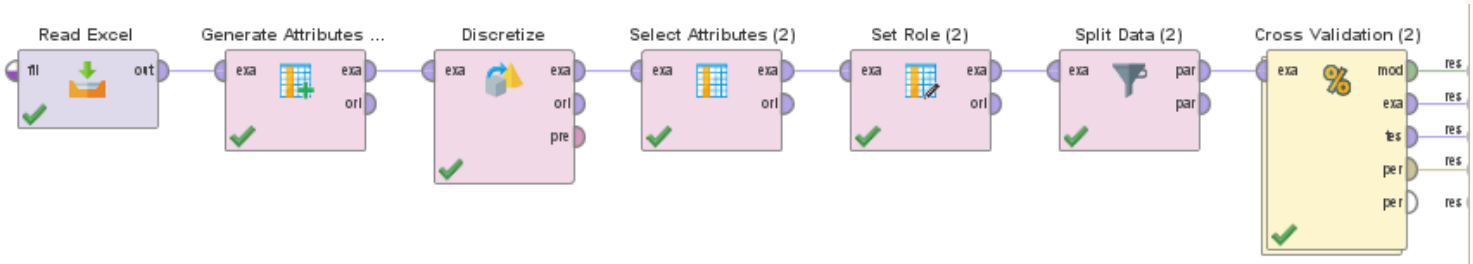
1-normalizing the data(new file)



accuracy: 73.75% +/- 3.09% (micro average: 73.75%)

	true cold	true hot	class precision
pred. cold	1038	280	78.76%
pred. hot	113	66	36.87%
class recall	90.18%	19.08%	

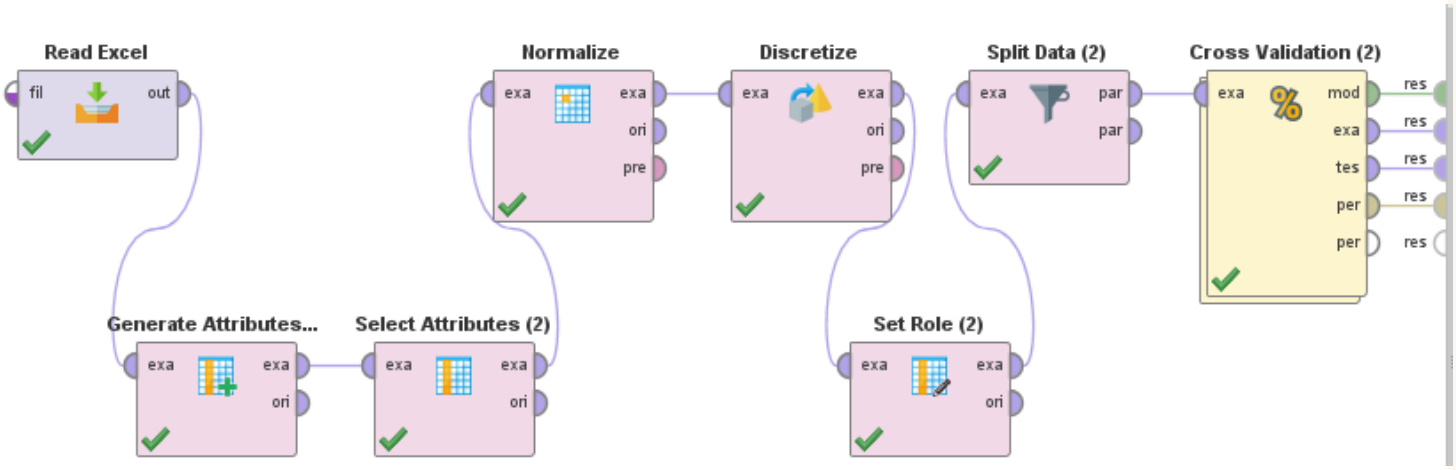
2- discretizing (new file)



accuracy: 79.09% +/- 1.95% (micro average: 79.09%)

	true cold	true hot	class precision
pred. cold	1033	195	84.12%
pred. hot	118	151	56.13%
class recall	89.75%	43.64%	

3- all previous tasks (new file)



accuracy: 79.03% +/- 2.03% (micro average: 79.02%)

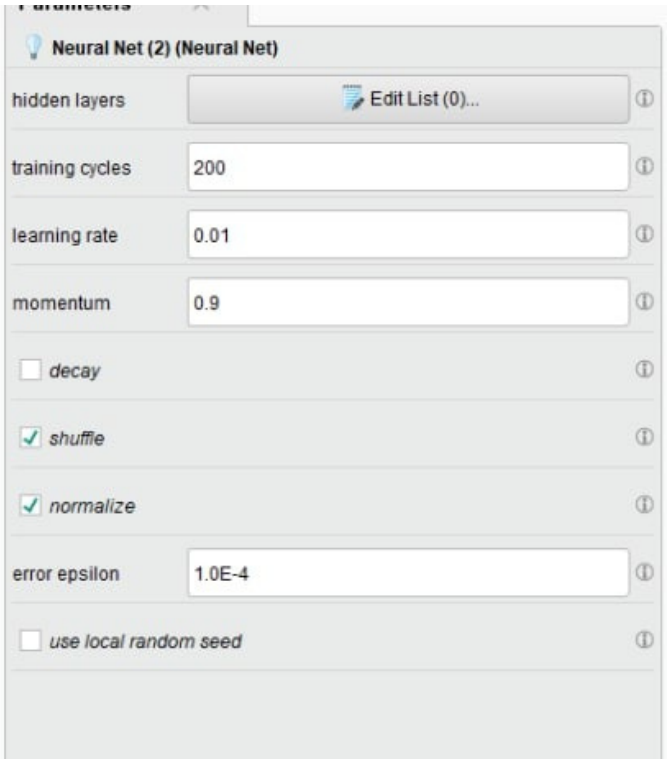
	true cold	true hot	class precision
pred. cold	1033	196	84.05%
pred. hot	118	150	55.97%
class recall	89.75%	43.35%	

Neural Network

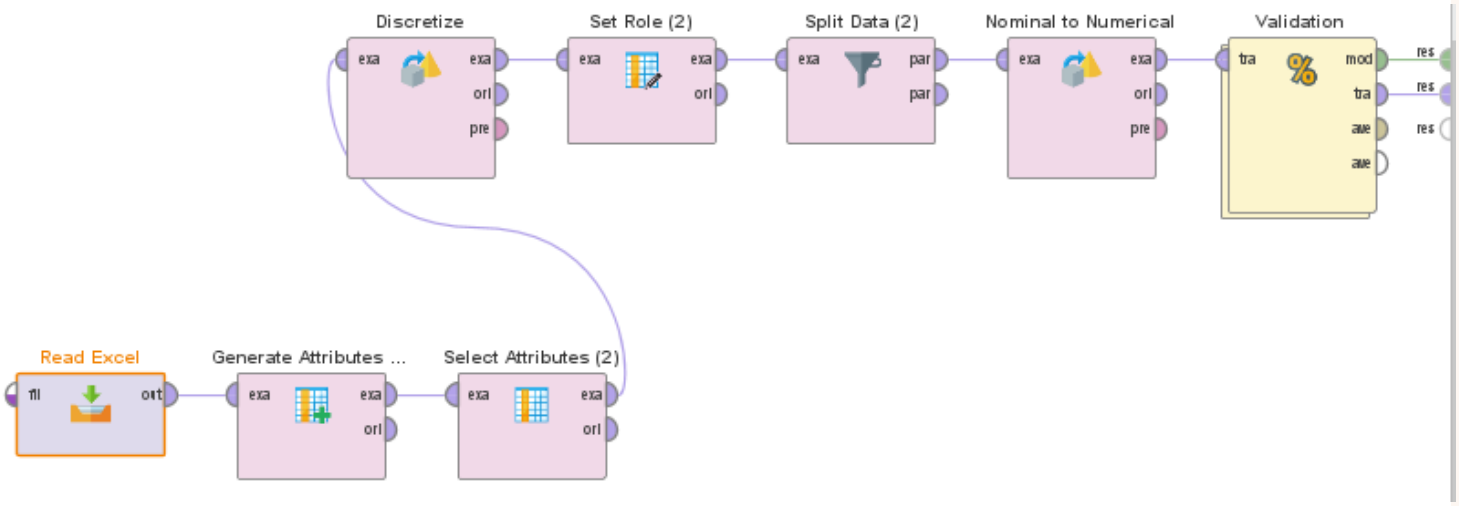
This operator learns a model by means of a feed-forward neural network trained by a back propagation algorithm (multi-layer perceptron). This operator cannot handle polynominal attributes.

we convert all the polynomial attributes to numeric to be able to use the neural network operator because it only works with numeric attributes.

Neural Network settings



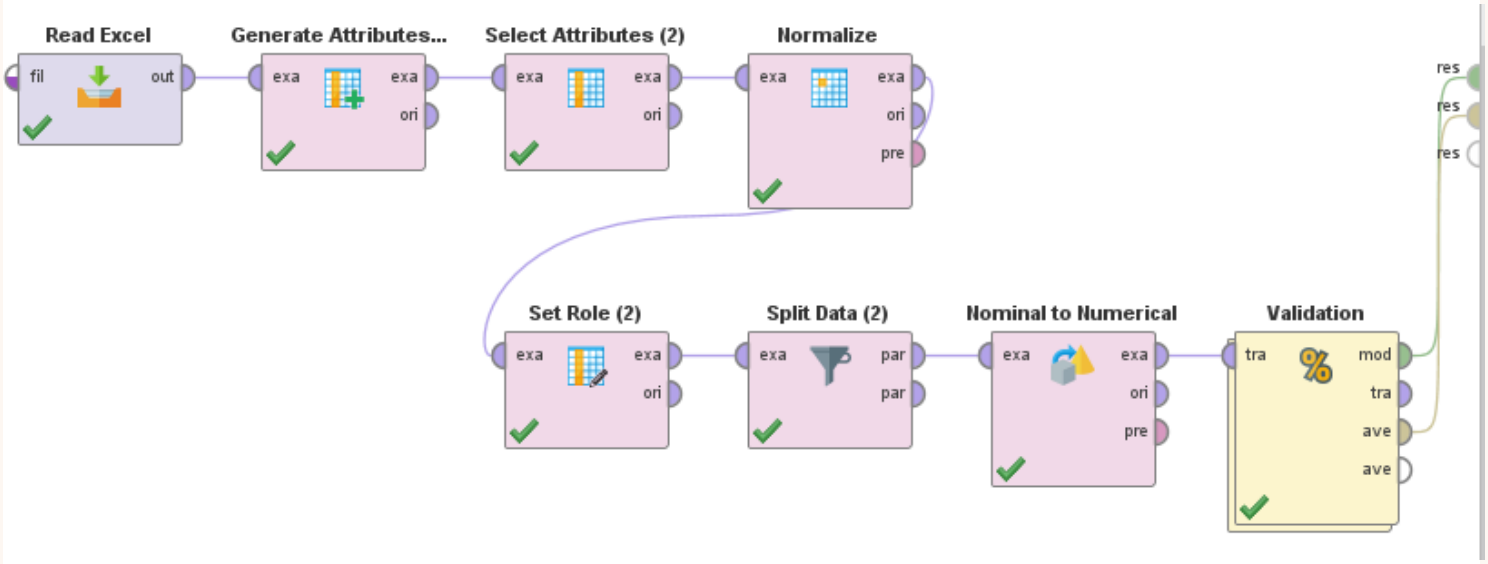
1- discretizing (new file)



accuracy: 76.61%

	true cold	true hot	class precision
pred. cold	302	62	82.97%
pred. hot	43	42	49.41%
class recall	87.54%	40.38%	

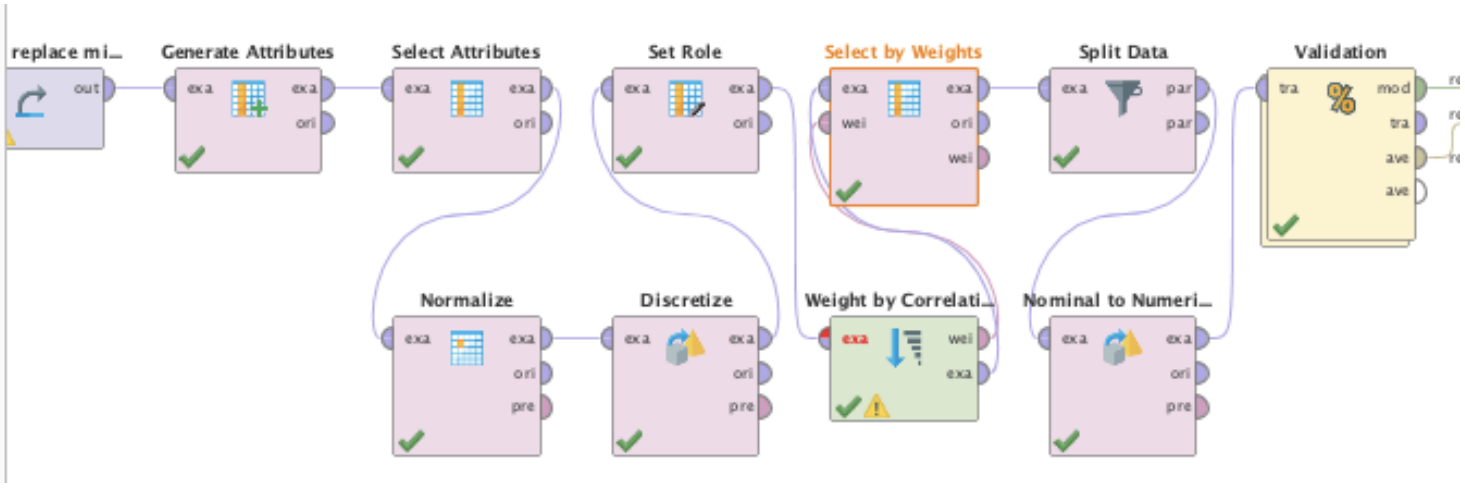
2-normalizing the data(new file)



accuracy: 79.96%

	true cold	true hot	class precision
pred. cold	305	50	85.92%
pred. hot	40	54	57.45%
class recall	88.41%	51.92%	

3- all previous tasks (new file)



accuracy: 81.07%

	true cold	true hot	class precision
pred. cold	329	69	82.66%
pred. hot	16	35	68.63%
class recall	95.36%	33.65%	

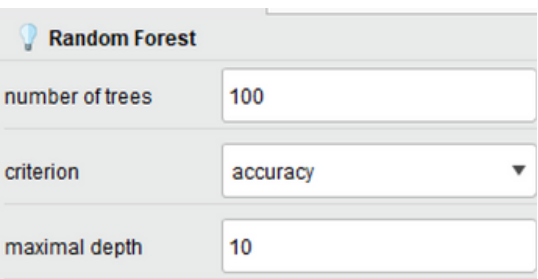
Random Forest

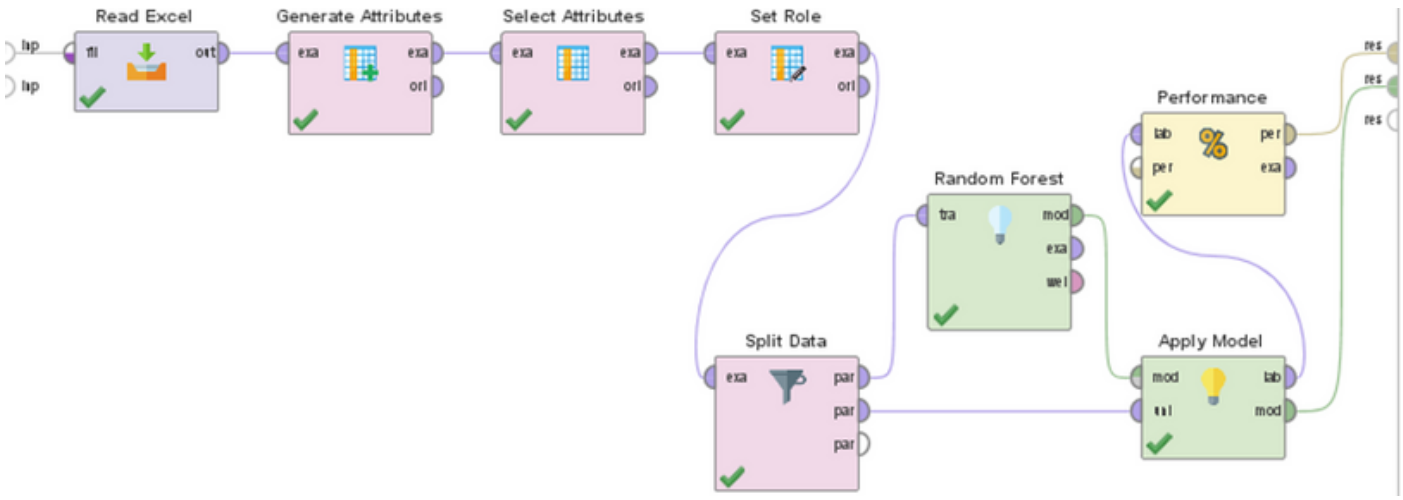
Random forest generate great outcome even outwardly hyper- parameter optimization. Because of its diversity and simplicity, it is most used algorithm. It can be used for regression and classification.

Random Forest sittings

Random Forest operator

This Operator generates a random forest model, which can be used for classification and regression.





Accuracy before cleaning

accuracy: 77.62%

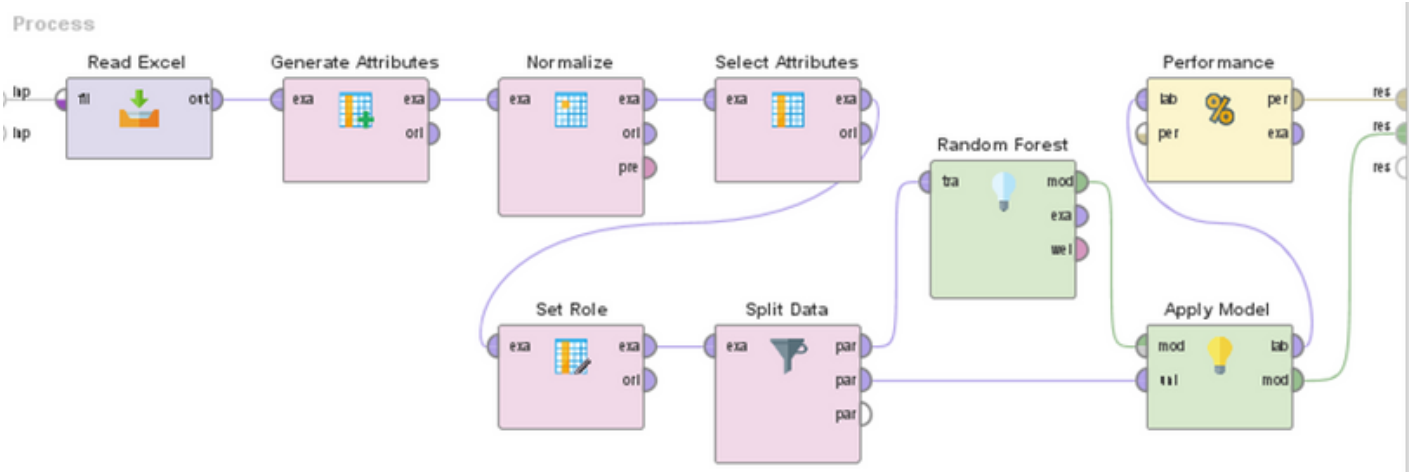
	true cold	true hot	class precision
pred. cold	1101	285	79.44%
pred. hot	50	61	54.95%
class recall	95.66%	17.63%	

1) Accuracy after cleaning

accuracy: 77.49%

	true cold	true hot	class precision
pred. cold	1102	288	79.28%
pred. hot	49	58	54.21%
class recall	95.74%	16.76%	

2-normalizing the data(new file)



method	range transformation
min	0.0
max	1.0

Accuracy

accuracy: 77.49%

	true cold	true hot	class precision
pred. cold	1102	288	79.28%
pred. hot	49	58	54.21%
class recall	95.74%	16.76%	

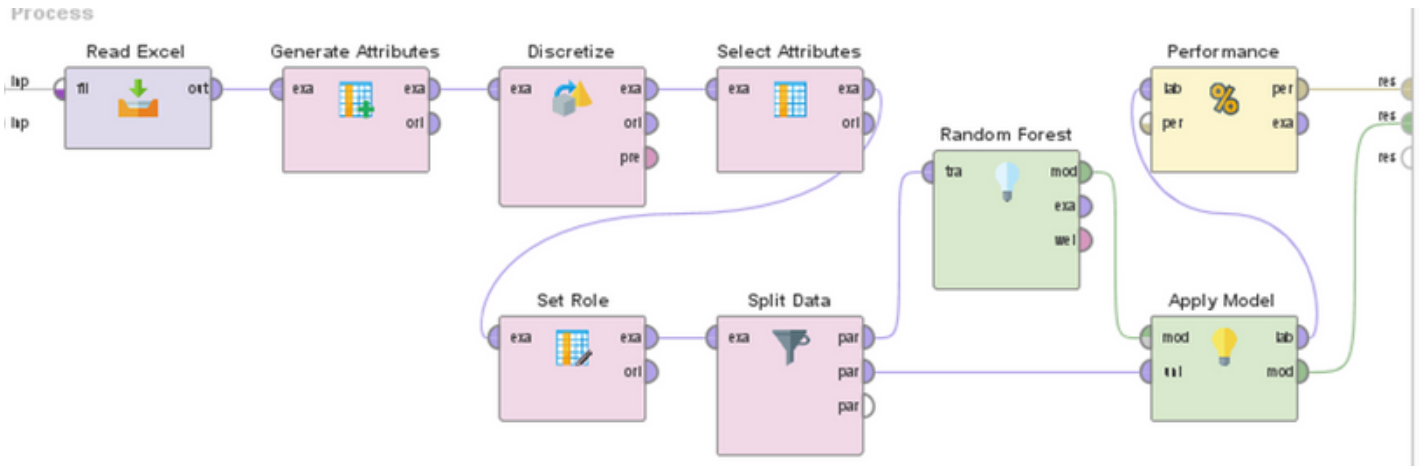
method	Z-transformation

Accuracy

accuracy: 77.42%

	true cold	true hot	class precision
pred. cold	1102	289	79.22%
pred. hot	49	57	53.77%
class recall	95.74%	16.47%	

3- discretizing (new file)



bins = 2

Accuracy

accuracy: 77.96%

	true cold	true hot	class precision
pred. cold	1089	268	80.25%
pred. hot	62	78	55.71%
class recall	94.61%	22.54%	

bins = 4

Accuracy

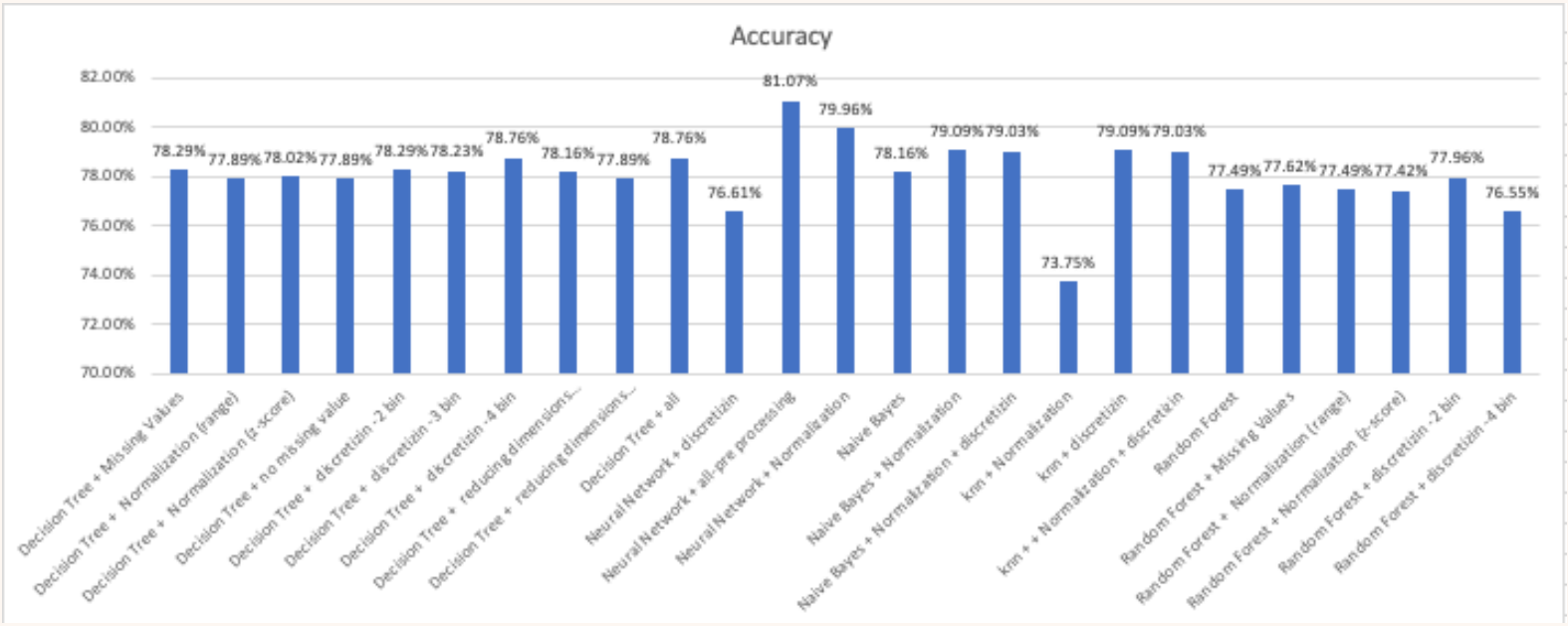
accuracy: 76.55%

	true cold	true hot	class precision
pred. cold	1052	252	80.67%
pred. hot	99	94	48.70%
class recall	91.40%	27.17%	

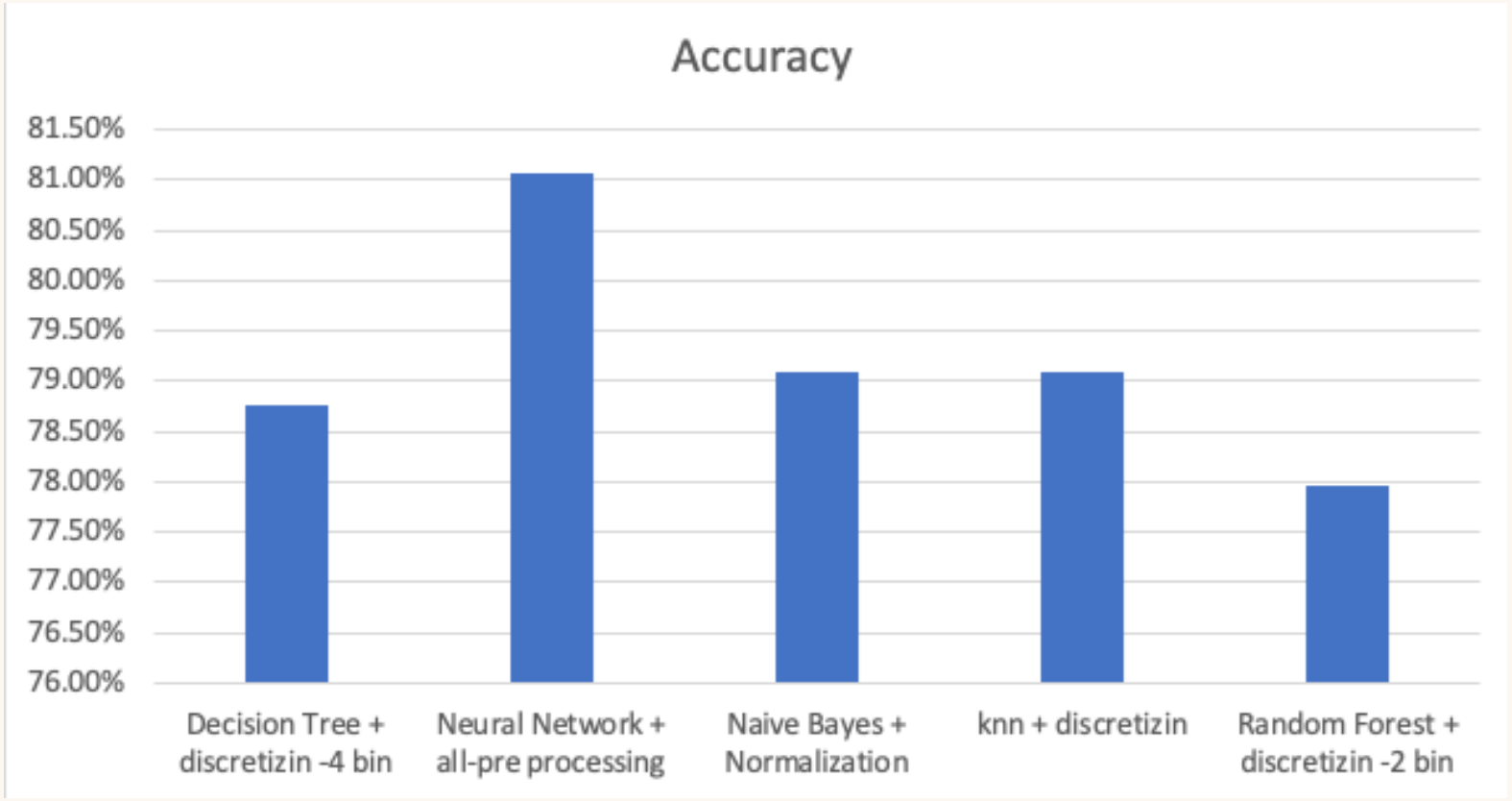
MODELS SUMMARY

Experiment	Accuracy	Experiment	Accuracy
Decision Tree + Missing Values	78.29%	Naive Bayes + Normalization+ discretizin	79.03%
Decision Tree + Normalization (range)	77.89%	knn + Normalization	73.75%
Decision Tree + Normalization (z-score)	78.02%	knn + discretizin	79.09%
Decision Tree + no missing value	77.89%	knn + Normalization + discretizin	79.03%
Decision Tree + discretizin -2 bin	78.29%	Random Forest	77.49%
Decision Tree + discretizin -3 bin	78.23%	Random Forest + Missing Values	77.62%
Decision Tree + discretizin -4 bin	78.76%	Random Forest + Normalization (range)	77.49%
Decision Tree + reducing dimensions (corraltions)	78.16%	Random Forest + Normalization (z-score)	77.42%
Decision Tree + reducing dimensions (info gain)	77.89%	Random Forest + Discretizin -2 bin	77.96%
Decision Tree	78.76%	Random Forest + Discretizin -4 bin	76.55%
Neural Network + discretizin	76.61%		
Neural Network + Normalization	79.96%		
Neural Network + all-pre processing	**81.07%**	←	
Naive Bayes	78.16%		
Naive Bayes + Normalization	79.09%		

histogram with accuracy for each models



histogram with best accuracy for each models

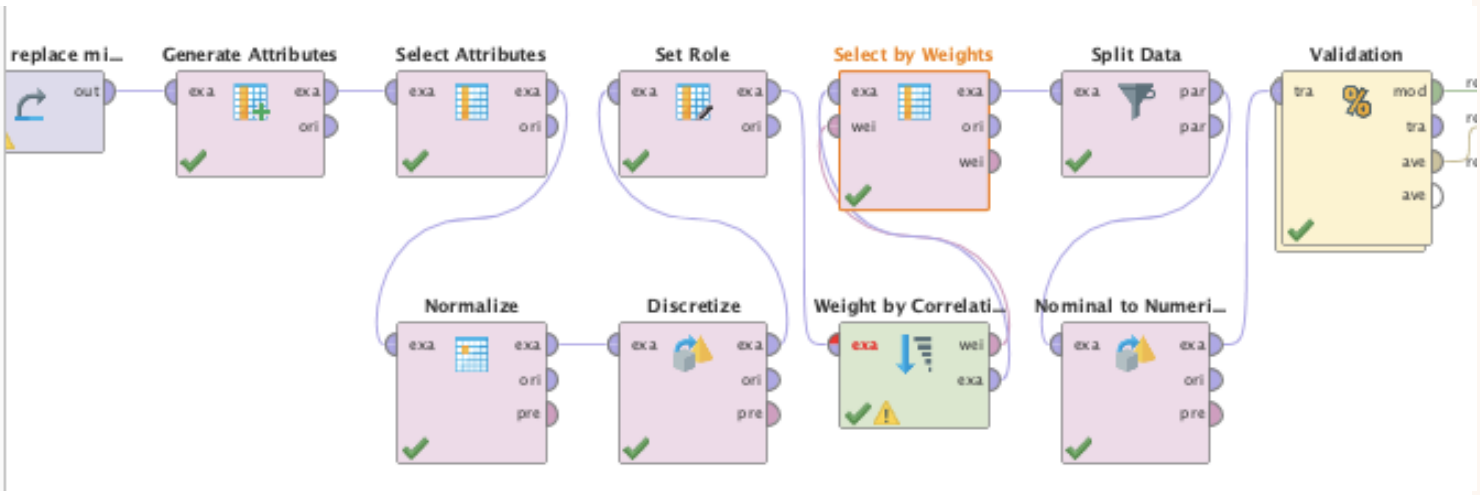


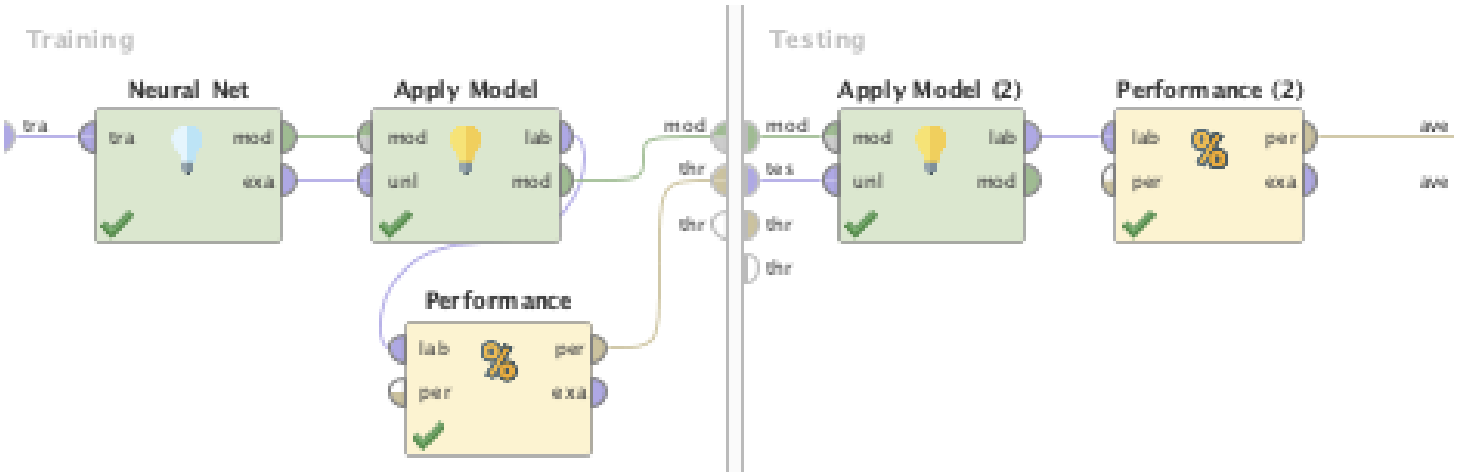
Analyze the best model

After all of these attempts, we reached to the best model by **neural network** technique.

Neural Network + with app pre-processing give us accuracy equal to 81.07%

so now we going to look closer to the model and describe some parameters that help us to reach to that ratio:



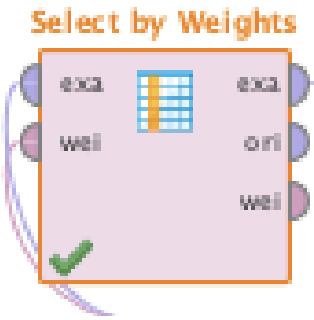


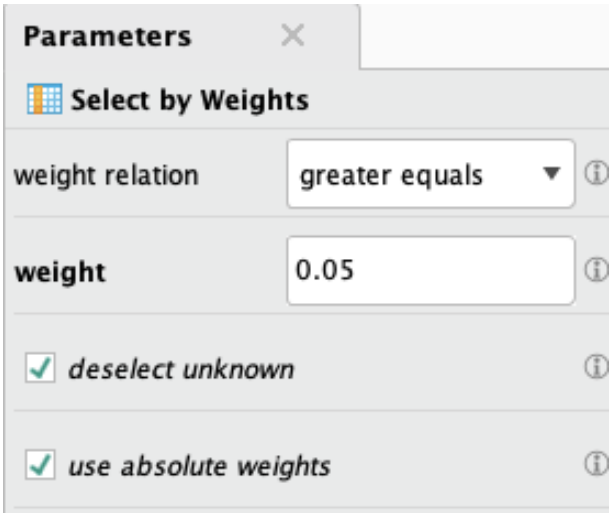
neural network Sittings

attribute	weight
ddd_car	0.001
ff_x	0.013
ff_avg	0.016
station_...	0.026
longitude	0.042
date	0.047
ddd_x	0.070
RR	0.071
latitude	0.087
ss	0.188
RH_avg	0.362

it will remove all this attribute

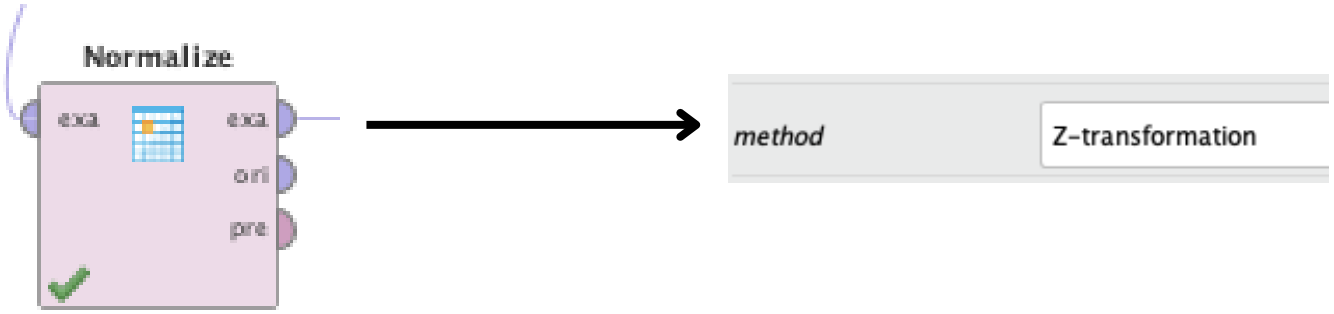
by selecting a weight to delete all attributes that are less than 0.05



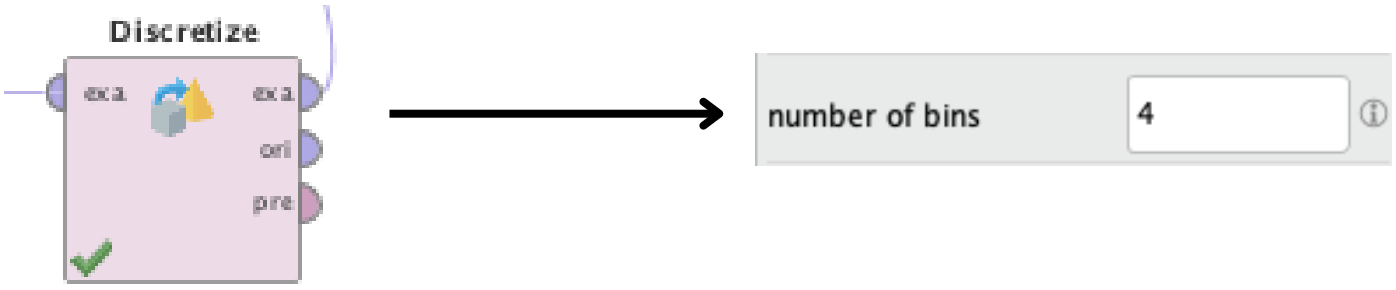


neural network Sittings

normalization is an essential process for professionals that deal with large amounts of data. useful for data consistency & reduces redundancy



use discretize by Binning for data smoothing that helps to group a huge number of continuous values into smaller values.



the result

after running the model, the result show that we get the **best** accuracy

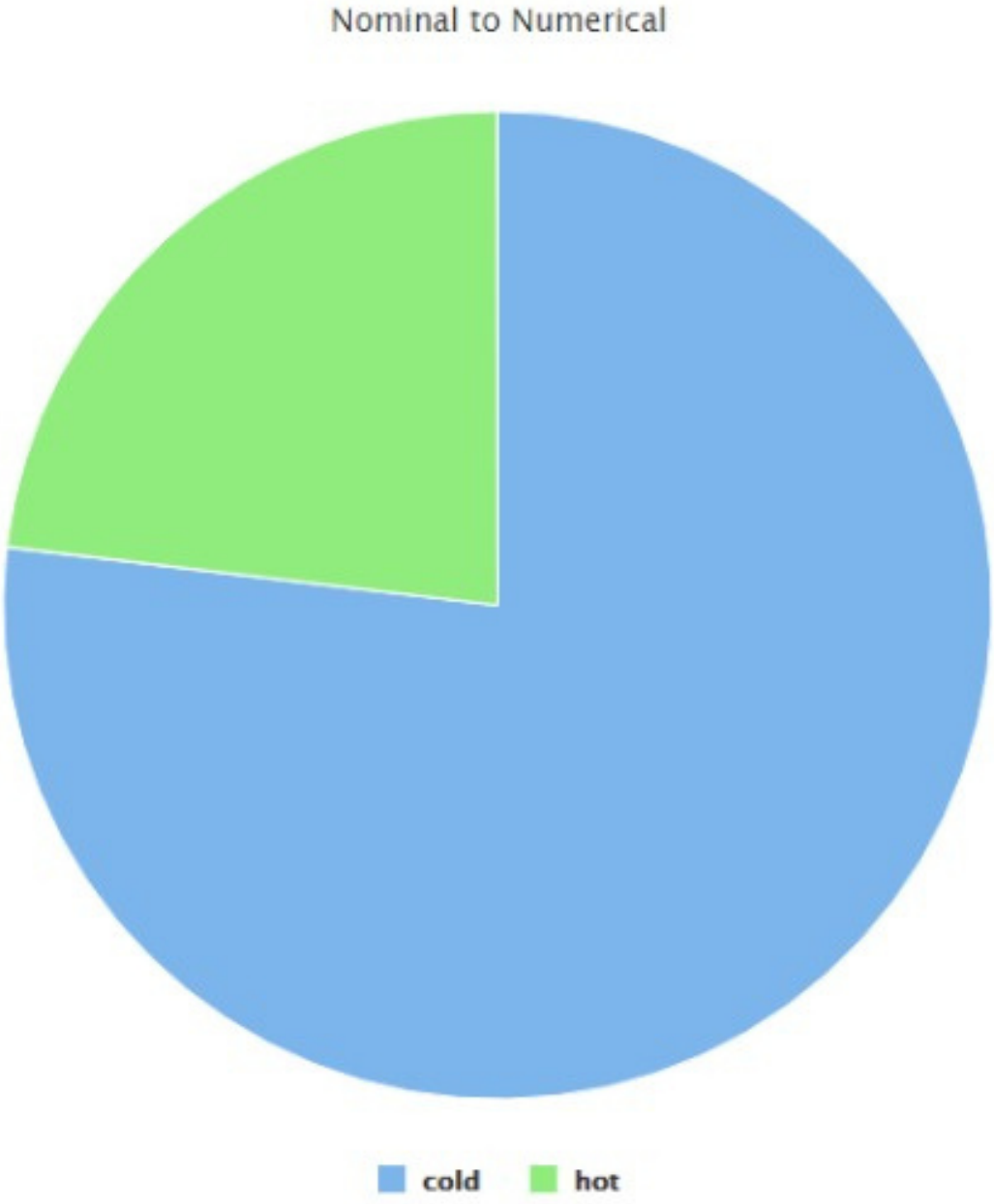
accuracy: 81.07%

	true cold	true hot	class precis
pred. cold	329	69	82.66%
pred. hot	16	35	68.63%
class recall	95.36%	33.65%	

Open in Turbo Prep Auto Model Filter (1,497 / 1,497 examples): all

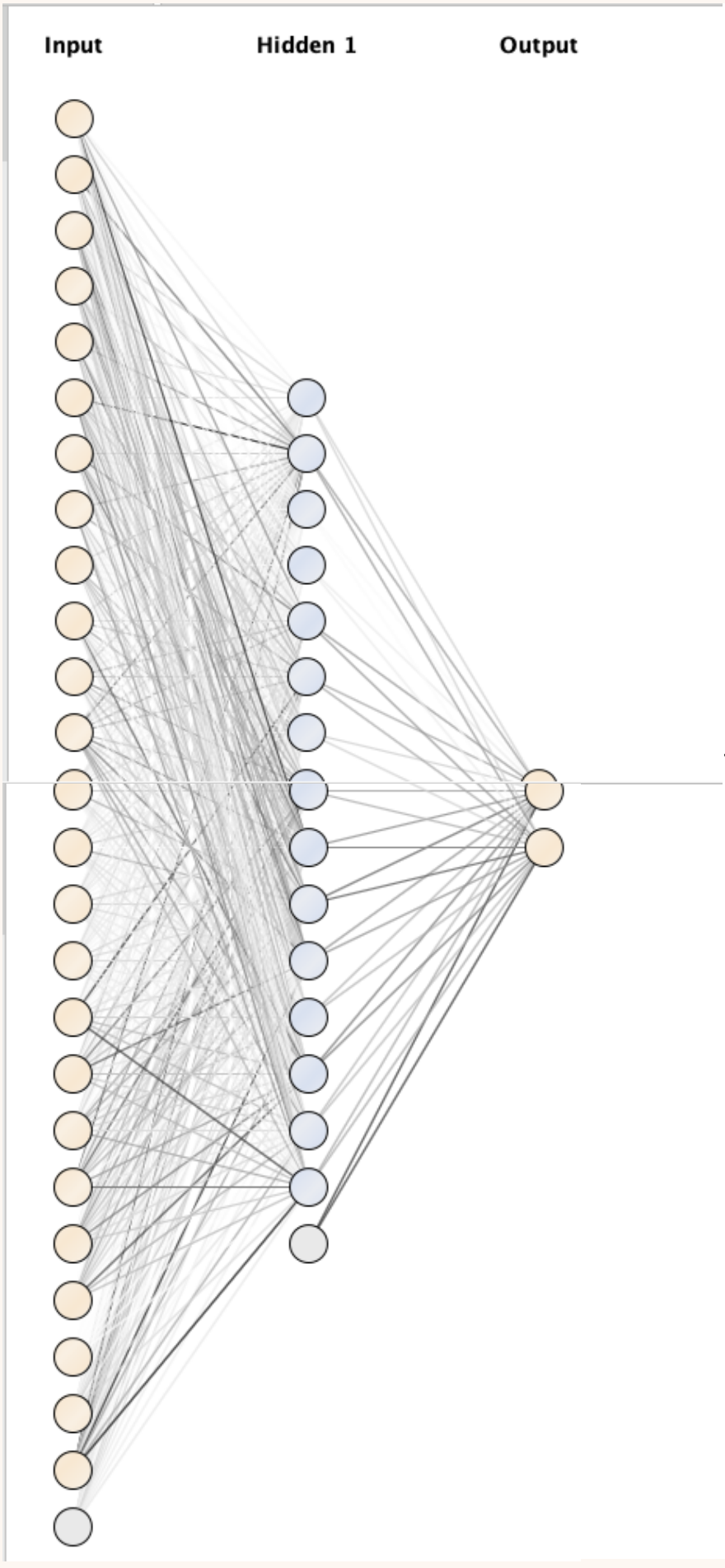
Row No.	the withear	latitude = r...	latitude = r...	latitude = r...	latitude = r...	RH_avg = r...	RH_avg = r...	RH_
1	hot	0	0	0	1	0	1	0
2	hot	0	0	0	1	0	0	1
3	cold	0	0	0	1	0	0	1
4	cold	0	0	0	1	0	0	1
5	cold	0	0	0	1	0	0	1
6	cold	0	0	0	1	0	0	1
7	cold	0	0	0	1	0	0	0
8	hot	0	0	0	1	0	0	1
9	cold	0	0	0	1	0	0	1
10	cold	0	0	0	1	0	0	1

Pie char of model



Here we see the percentage of cold and hot days expected in Indonesia at a station(Meteorologi Maimun Saleh), so we expect from this that cold weather will prevail in this station,
the state will be able to anticipate decisions

neural network Digram



We chose A neural network that consists of three layers - because it is good for our prediction because it shows us the input and output

conclusion

Weather forecasting contributes greatly to facilitating the life of the individual and society, and weather forecasting is of particular importance to the owners of marketing companies and affected areas.

The weather forecast has helped the governments of Indonesia to educate people when there is a possibility of severe cold or high temperatures, and this will greatly contribute to saving their lives, given that it is a developing country, so the effect of weather will be. stronger than people
Indonesia is characterized by many agricultural regions. Knowing the temperature is of great importance to the farmers in order to reduce the type of seed, harvest time, etc.

We have tried many techniques and methods to get the best accuracy such as Naïve Bayes, Random Forest, and K-NN to predict whether the temperature will be hot or cold. The best algorithm that gives us the best accuracy is Neural Grid which gives an accuracy of 81.07%. There is no difference between remembering the values for almost every category. This means that there is no bias