Sunny

32°

Let's discuss

Climate Data Daily IDN

Swipe to know more

TABLE OF CONTENT

intro

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

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

Project Goal

the main purpose of the project is predicting the expected weather on a specific date or a specific region in Indonesia and for that, we used the climate data in IDN to predict whether the day is hot or cold

about data

The data is about daily climate data covering almost all regions of Indonesia from 2010 to 2020, This dataset belongs to (BMKG) Indonesia .

contains climatic details for each

- region and province as it covered 34 provinces ,8 region according to The date.
- max and min temperature
-wind direction and speed
- humidity ,latitude and longitude
- how much water falls as rain in a certain period of time
-the percentage of the duration of the sun's brightness.
-finally the stations that recorded all of that

MAJOR CHALLENGES:

Missing values

replacing missing values by avg.

ddd_car
E
E
E
SW
?
?
?
E

ddd_car
W
S
W
S
SW
NE
W
S

Inconsisetant data

using excel features select "**data text to column**"

	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

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

MAJOR CHALLENGES:

Outliers

by using the "**detect outlier**" operator in RapidMiner

Redundant attributes

discovering the correlation between them and then deleting repeated groups, normalization

Duplication Data:

by using the "Remove Duplicates" operator in RapidMiner

Classification and prediction

measure accuracy

the pre-processing phase had already been done and we improved it by using different pre-processing techniques.

- decision tree --> 10 methods
- Naïve Bayes --> 3 methods
- natural network --> 3 methods
- Random Forest --> 6 methods
- K-NN -->3 methods

total combination: 25

table of accuracy

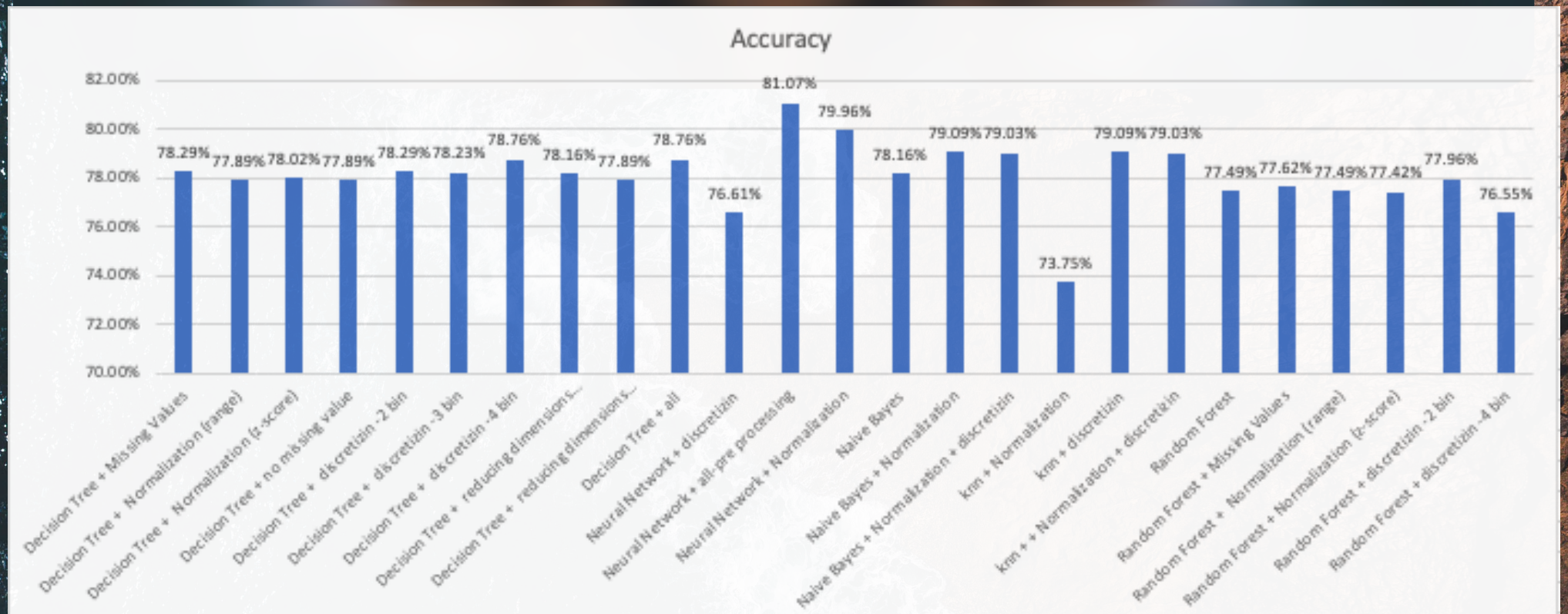
Experiment	Accuracy
Decision Tree + Missing Values	78.29%
Decision Tree + Normalization (range)	77.89%
Decision Tree + Normalization (z-score)	78.02%
Decision Tree + no missing value	77.89%
Decision Tree + discretizin -2 bin	78.29%
Decision Tree + discretizin -3 bin	78.23%
Decision Tree + discretizin -4 bin	78.76%
Decision Tree + reducing dimensions (corraltions)	78.16%
Decision Tree + reducing dimensions (information gain)	77.89%
Decision Tree + all	78.76%
Neural Network + discretizin	76.61%
Neural Network + all-pre processing	81.07%

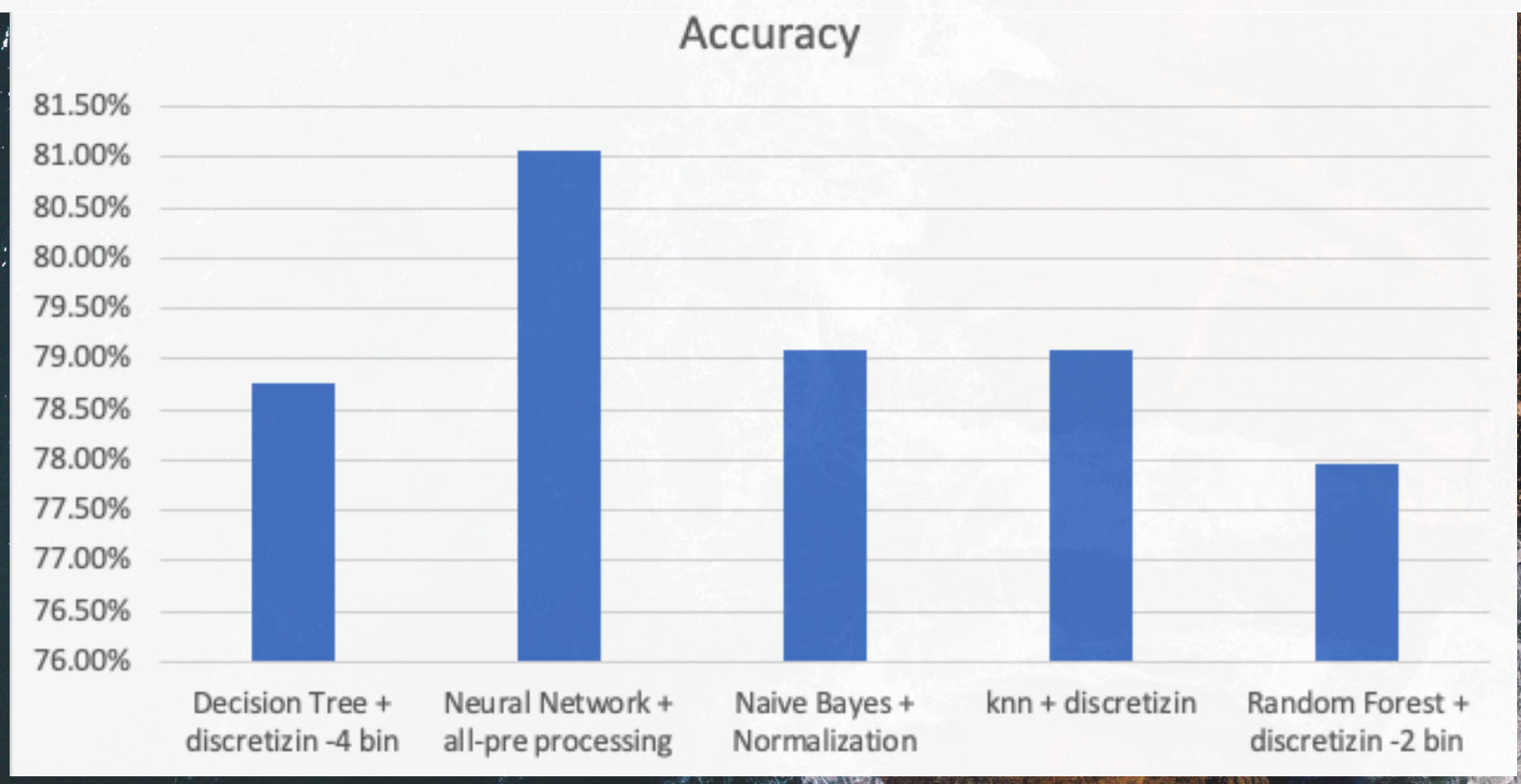
table of accuracy

Experiment	Accuracy
Naive Bayes	78.16%
Naive Bayes + Normalization	79.09%
Naive Bayes + Normalization + discretizin	79.03%
knn + Normalization	73.75%
knn + discretizin	79.09%
knn + + Normalization + discretizin	79.03%
Random Forest	77.49%
Random Forest + Missing Values	77.62%
Random Forest + Normalization (range)	77.49%
Random Forest + Normalization (z-score)	77.42%
Random Forest + discretizin -2 bin	77.96%
Random Forest + discretizin -4 bin	76.55%

histogram

as we see the table here shows the best accuracy for each models, then we decide to choose the best accuracy for all of models and the result is : Neural Network+ all pre-processing

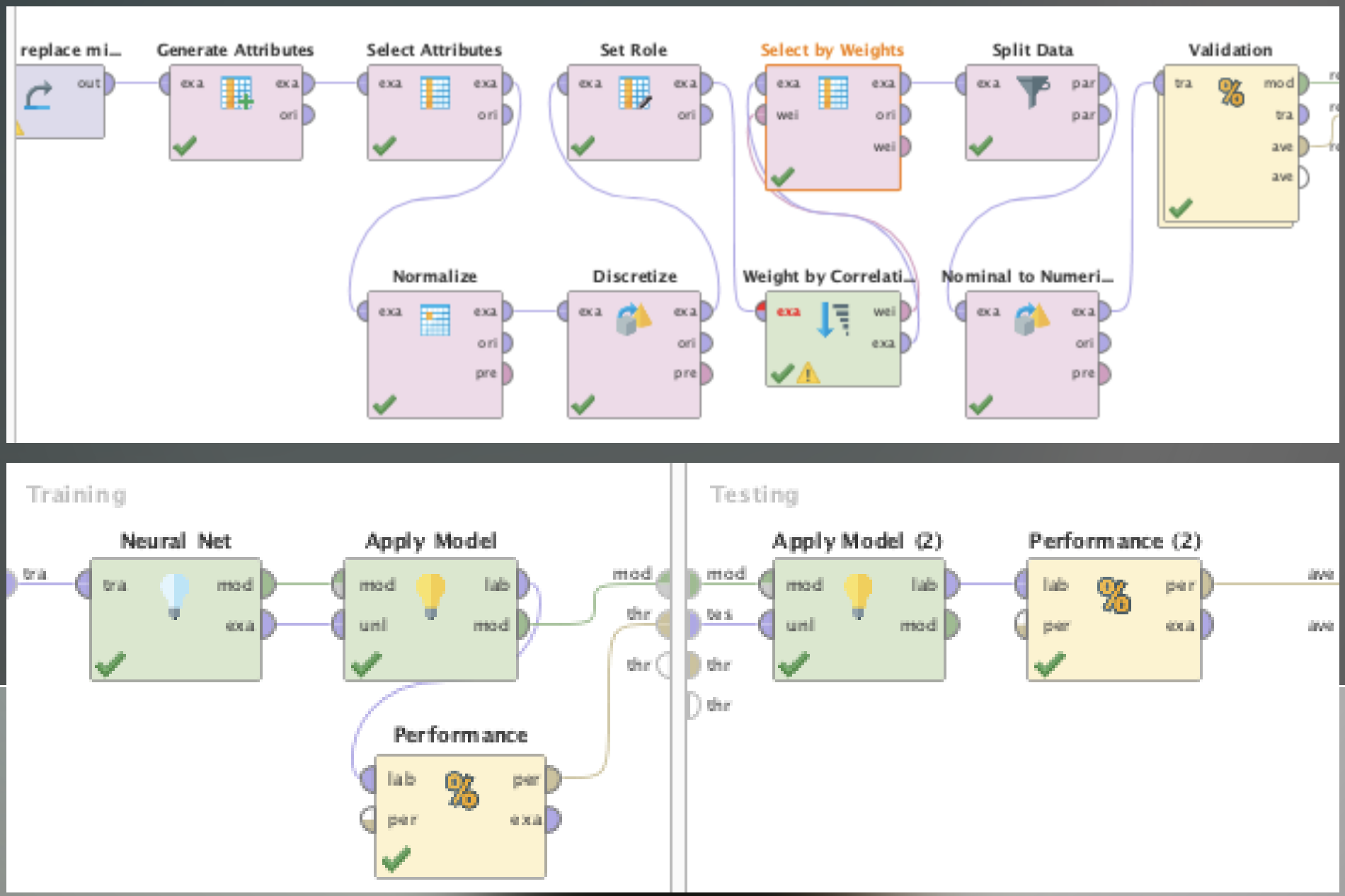




Analyze the best model

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:

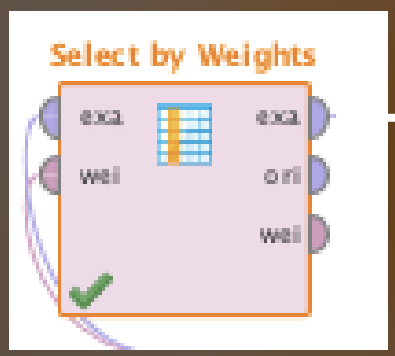


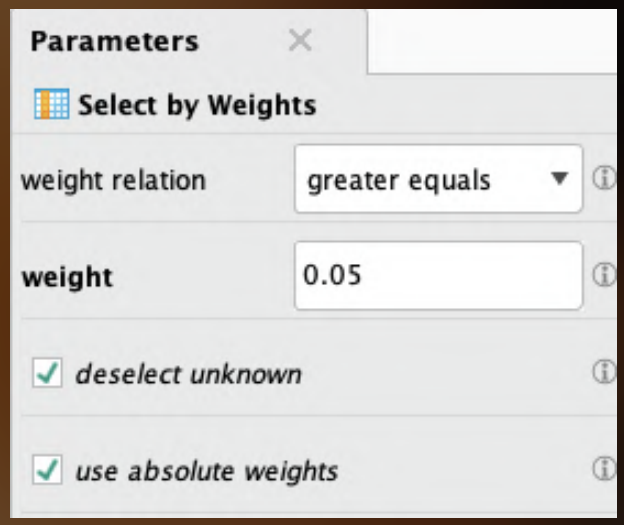
Neural network sittings

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

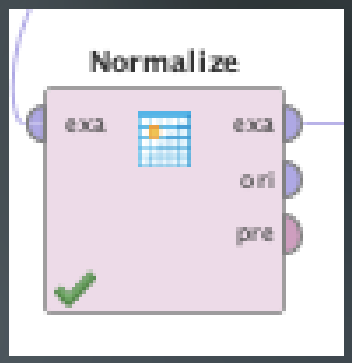
it will remove all this attribute

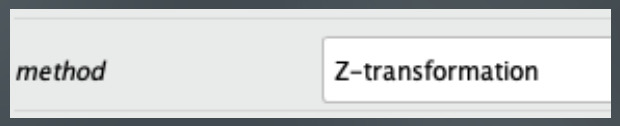
by selecting a weight to delete all attributes that are less than 0.05



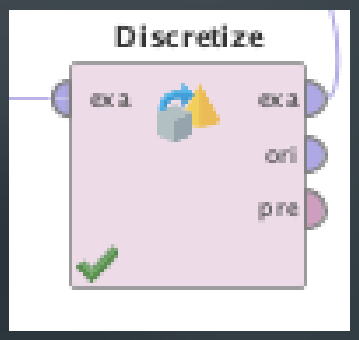


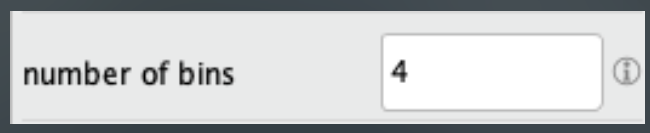
normalization is an essential process for professionals that deal with large amounts of data. useful for data consistency & reduces redundancy





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





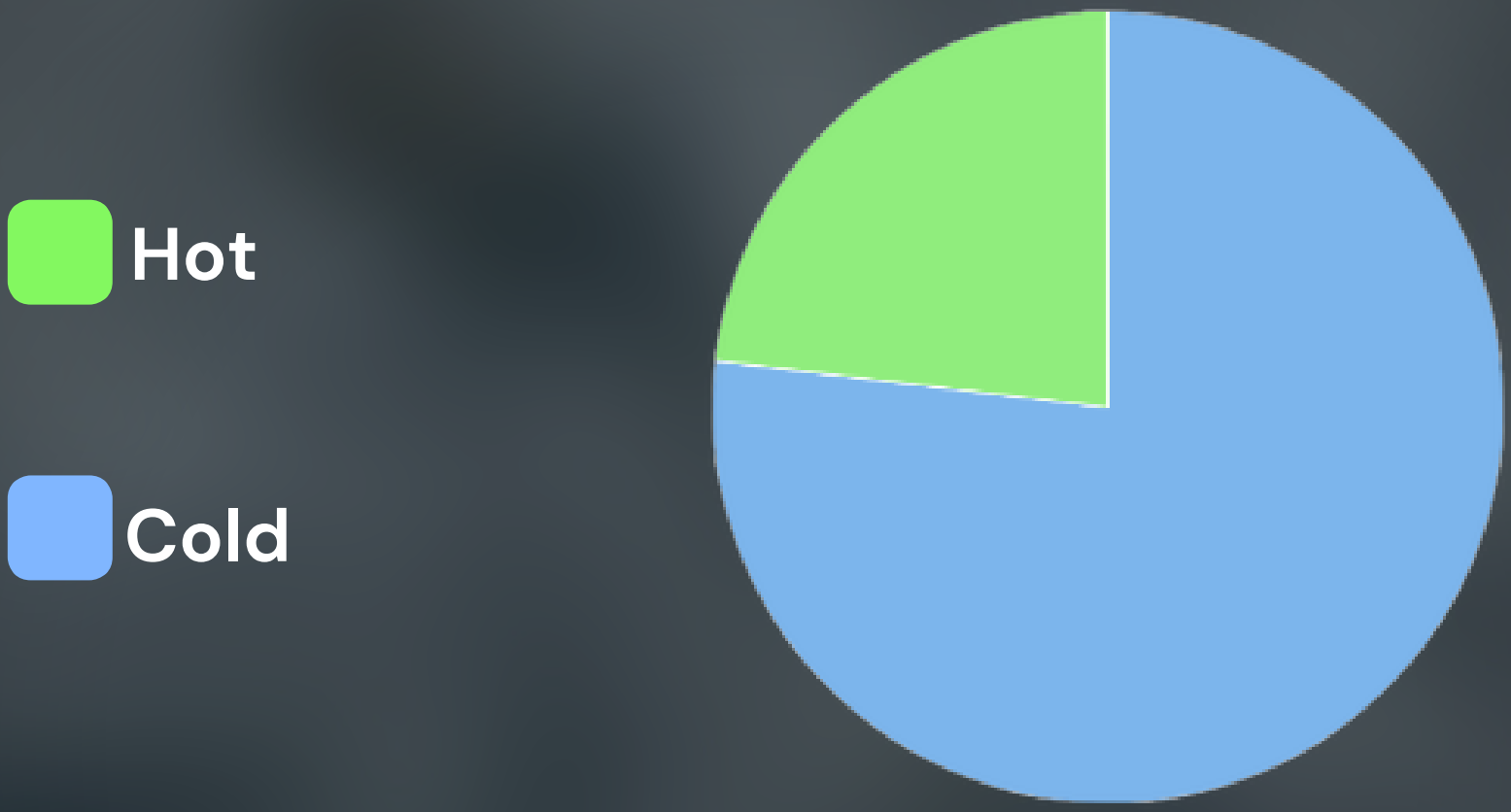
Accuracy of model

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%	

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.

conclusion

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

Group Members ?

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