EKSPLORASI DAN EKSPERIMEN CLASSIFICATION TASK DENGAN GAUSSIAN NB DAN RANDOM FOREST

Muhammad Shiba Kabul
1301183457
IF-42-12

R. ArditYo Cahyo Putro Hutomo
1301183507
IF-42-12

Pendahuluan

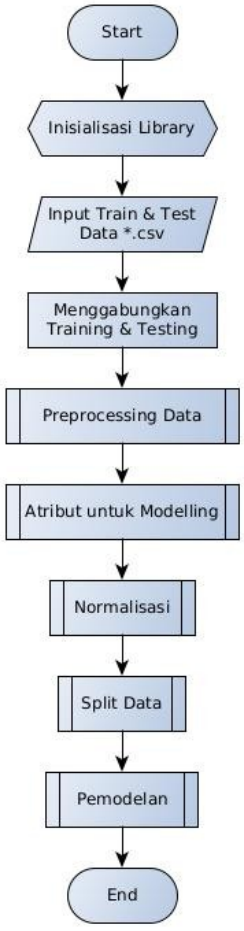
Diberikannya suatu data-set berformat *.csv yakni Training-set dan Testing-set mengenai pendataan curah hujan. Dalam file tersebut dijelaskan tiap tanggal yang tertulis terdapat banyak kolom data seperti Kode Lokasi, Suhu Min, Suhu Max, Hujan, Penguapan, Sinar Matahari, Arah Angin Terkencang, Kecepatan Angin Terkencang, Arah Angin 9am, Arah Angin 3pm, Kecepatan Angin 9am, Kecepatan Angin 3pm, Kelembaban 9am, Kelembaban 3pm, Tekanan 9am, Tekanan 3pm, Awan 9am, Awan 3pm, Suhu 9am, Suhu 3pm, Bersalju Hari Ini, dan Bersalju Besok. Dari Semua data tersebut kami diminta untuk mengelola data tersebut agar dapat mengetahui prediksi Besok akan turun salju atau tidak. Dalam kesempatan kali ini, Kami menggunakan bahasa Python dan menggunakan beberapa library pembantu agar data mudah untuk diolah.

Eksperimen

Dalam eksplorasi yang dilakukan oleh kami, sebelum melakukan Classification, diharuskannya terlebih dahulu untuk melakukan Preprocessing. Didalam Preprocessing sebenarnya terdapat beberapa metode yang dilakukan (sesuai kondisi). Beberapa diantaranya yakni,

a) Mengidentifikasi dan mengatasi Missing Values,
b) Data Formatting,
c) Data Normalization,
d) Data Binning,
e) Mengubah Categorical ke Numerical.

Namun dalam pengerjaan program ini tidak semua metode diatas dilakukan dikarenakan didalam data-set yang diberikan beberapa metode telah diterapkan. Untuk model program dapat dilihat pada Gambar 1.

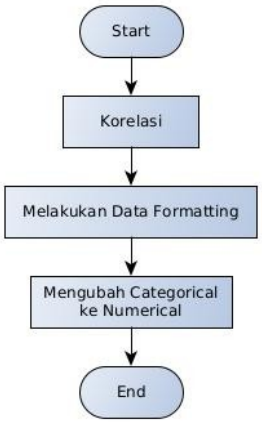


Gambar 1: Model Progam

Analisis Hasil Eksperimen

1. Preprocessing

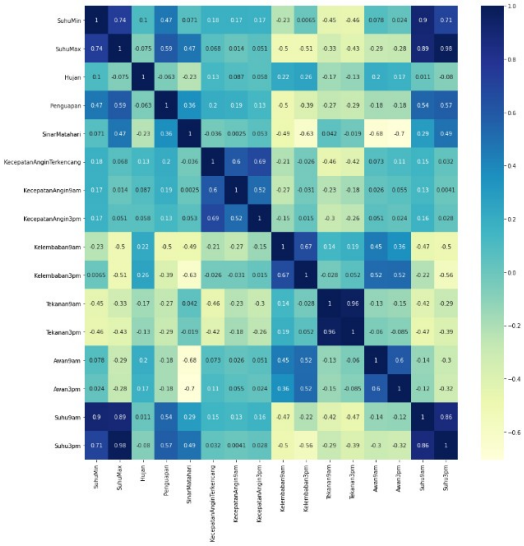
Dalam penerapan model diatas, *preprocessing* dilakukan dengan Korelasi, *Data Formatting*, dan Mengubah *Categorical* ke *Numerical* sesuai pada Gambar 2.



Gambar 2: Model Preprocessing

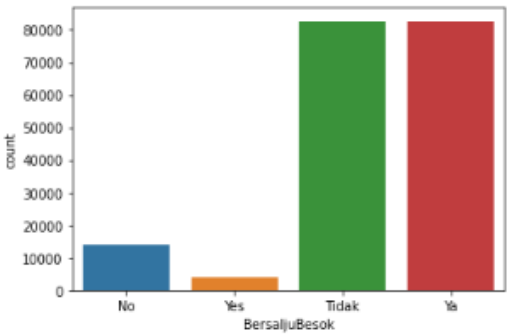
1.a. Korelasi

Tahap ini dilakukan oleh kami untuk menentukan pengambilan atribut. Berikut korelasi beserta visualisasinya.

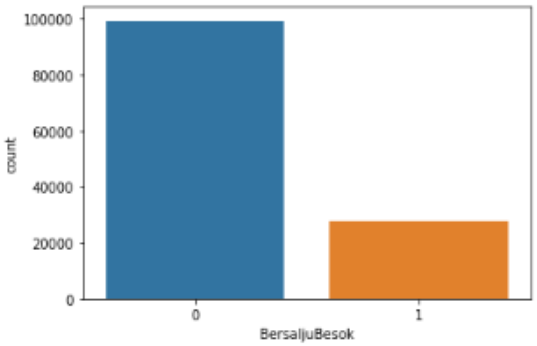


1.b. Data Formatting

Dalam Data Formatting dilakukannya pengubahan data yang berpotensi dapat merusak hasil akhir, pada file training dan testing yang diberikan ditemukannya data yang tidak konsisten, pada training tertulis "Ya" dan "Tidak" (dalam Bahasa Indonesia) sedangkan untuk testing tertulis "Yes" dan "No" (dalam Bahasa Inggris). Maka kami ganti dengan 0 dan 1.



Gambar 3: Sebelum Data Formatting



Gambar 4: Sesudah Data Formatting

1.c. Mengubah *Categorical* ke *Numerical*

Pada kolom ArahAngin9am, ArahAngin3pm, ArahAnginTerkencang, dan Kode lokasi dilakukan encode, ini dilakukan untuk mengubah kata menjadi angka, disinggung pada Gambar 5 dan Gambar 6.

	ArahAnginTerkencang	ArahAngin9am	ArahAngin3pm	KodeLokasi
0	WSW	W	W	C39
1	WNW	W	NW	C35
2	SSW	NE	N	C18
3	SW	E	SSE	C31
4	NW	W	WNW	C14
...	...	...	...	...
127272	ESE	SE	ESE	C38
127273	SSE	SSE	E	C16
127274	NW	N	NW	C17
127275	E	ESE	SE	C11
127276	WNW	NNE	NE	C16

127277 rows × 4 columns

Gambar 5: Sebelum mengubah Categorical menjadi Numerical

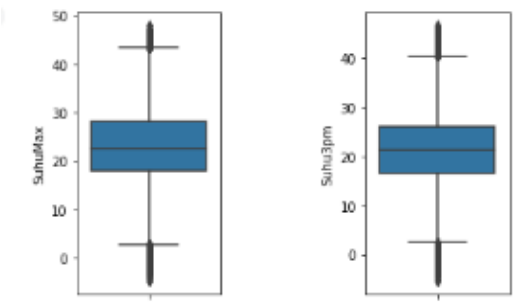
	ArahAnginTerkencang	ArahAngin9am	ArahAngin3pm	KodeLokasi
0	15	13	13	32
1	14	13	7	28
2	11	4	3	9
3	12	0	10	24
4	7	13	14	5
...	...	...	...	...
127272	2	9	2	31
127273	10	10	0	7
127274	7	3	7	8
127275	0	2	9	2
127276	14	5	4	7

127277 rows × 4 columns

Gambar 6: Sesudah mengubah Categorical menjadi Numerical

2.Atribut untuk Modelling

Setelah itu dilanjutkan dengan pemilihan atribut untuk modelling, pemilihan ini dilakukan dengan cara memilih 3 (tiga) kolom yang akan diproses selanjutnya. Kami memilih SuhuMax, Suhu3pm, dan BersaljuBesok. Pada tahap ini kami melakukan mengatasi *outlier*.



Gambar 7: Mengatasi Outlier

3. Normalisasi

Pada program ini kami menggunakan MinMaxScalling yang berguna untuk mengubah data dengan rentang 0 sampai 1. Sebelum itu kami telah memindahkan kolom target ‘BersaljuBesok’ ke variabel lain dimana akan diproses pada saat split data.

	SuhuMax	Suhu3pm
0	27.5	23.6
1	19.9	18.9
2	27.2	26.3
3	27.0	26.4
4	7.9	6.0
...	...	...
127272	23.7	22.1
127273	25.2	24.4
127274	20.4	19.8
127275	29.8	29.2
127276	27.4	23.3

127277 rows × 2 columns

Gambar 8: Sebelum Normalisasi

	SuhuMax	Suhu3pm
0	0.610586	0.556622
1	0.466919	0.466411
2	0.604915	0.608445
3	0.601134	0.610365
4	0.240076	0.218810
...	...	...
127272	0.538752	0.527831
127273	0.567108	0.571977
127274	0.476371	0.483685
127275	0.654064	0.664107
127276	0.608696	0.550864

127277 rows × 2 columns

Gambar 9: Setelah Normalisasi

4. Split Data

Pada Sesi ini dilakukannya Split Data dikarenakan program ini bersifat supervised. Data yang sebelumnya digabungkan dipisah kembali menjadi 2 (dua) bagian, Training set menjadi 2 dan Testing set menjadi 2, membagi data dengan skala 30:70.

x_train dan x_test berisi 2 kolom yang dipilih sebelumnya yakni ‘SuhuMax’ dan ‘Suhu3pm’, disinggung pada Gambar 10. Sedangkan y_train dan y_test berisi 1 kolom saja yakni ‘BersaljuBesok’ dimana sebelumnya telah dipisah pada tahap Normalisasi, disinggung pada nomor 11.

```
Xtrain Value
         SuhuMax   Suhu3pm
89229   0.510397  0.508637
124121  0.474480  0.468330
70125   0.264650  0.259117
38195   0.697543  0.694818
66505   0.689981  0.681382
...          ...       ...
61404   0.455577  0.454894
17730   0.410208  0.416507
28030   0.376181  0.387716
15725   0.534972  0.512476
118270  0.370510  0.341651

[89093 rows x 2 columns]
Xtest Value
         SuhuMax   Suhu3pm
90023   0.434783  0.519910
116093  0.476371  0.477927
43582   0.436673  0.416507
110534  0.587902  0.598848
26645   0.648393  0.652591
...          ...       ...
92966   0.538752  0.529750
650     0.393195  0.389635
91937   0.621928  0.591171
735     0.661626  0.658349
108691  0.366730  0.355086

[38184 rows x 2 columns]
```

Gambar 10: x_train & x_test

```
ytrain Value
89229     1
124121    0
70125     1
38195     0
66505     0
         ..
61404     0
17730     0
28030     0
15725     0
118270    0
Name: BersaljuBesok, Length: 89093, dtype: int64
ytest Value
90023     0
116093    0
43582     0
110534    0
26645     0
         ..
92966     1
650       0
91937     0
735       0
108691    0
Name: BersaljuBesok, Length: 38184, dtype: int64
```

Gambar 11: y_train & y_test

5 .Pemodelan

Pada Program ini kami menggunakan 2 (dua) pemodelan yakni Naive Bayes dan Random Forest yang nantinya akan dibandingkan hasilnya.

Untuk Naive Bayes mendapat hasil akurasi :

0.7750366645715483

Berikut Classification Report:

	precision	recall	f1-score	support
0	0.79	0.98	0.87	29781
1	0.41	0.05	0.09	8403
accuracy			0.78	38184
macro avg	0.60	0.52	0.48	38184
weighted avg	0.70	0.78	0.70	38184

Gambar 12: Classification Report Naive Bayes

Untuk Random Forrest mendapat hasil akurasi :
0.77076786088414

Berikut Classification Report:

	precision	recall	f1-score	support
0	0.80	0.94	0.87	29781
1	0.45	0.17	0.25	8403
accuracy			0.77	38184
macro avg	0.63	0.56	0.56	38184
weighted avg	0.72	0.77	0.73	38184

Gambar 13: Classification Report Random Forest

Evaluasi
Naive Bayes
ACCURACY :
0.7750366645715483 - 78%

PRECISION :
0.412861136999068 - 41%

Random Forest
ACCURACY :
0.7707154829247852 - 77%

PRECISION :
0.4458128078817734 - 45%

Kesimpulan
Dari eksplorasi dan eksperimen yang dilakukan oleh kami, dapat disimpulkan bahwa hasil dari Naive Bayes lebih unggul dari Random Forest.

Program dapat diakses dan dijalankan di :
https://drive.google.com/file/d/1sq_ErIYoe7yA0UOia4tKRcJjCbTMwA6q/view?usp=sharing

Berikut Link Github dari Tugas Kami :
http://github.com/Shibakabul/TubesMalin2/

Referensi

Asisten Dosen Machine Learning. Responsi 3 : Classification. Diakses pada : https://drive.google.com/drive/folders/1-7x1uWzclRq06r0K6tevTrP8Mn2x6ghW

scikit-learn.org. A random forest classifier. Diakses pada : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html