Miftah Shidqi Rabbani – 1301184371

Muhammad Tegar Zharfan Humam Setiabakti - 1301184354

LAPORAN TUGAS BESAR ML ADVANCE

1. Pendahuluan

Machine Learning (ML) adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah. *Machine Learning* pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar *Machine Learning* dan konsepnya. Sejak saat itu banyak yang mengembangkan *Machine Learning*.

Masalah yang akan kita selesai kan menggunakan *Automated Machine Learning*, *Automated Machine Learning* (AutoML) adalah proses mengotomatisasi tugas penerapan pembelajaran mesin ke masalah dunia nyata. AutoML mencakup alur lengkap dari set data mentah hingga model pembelajaran mesin yang dapat diterapkan. AutoML diusulkan sebagai solusi berbasis kecerdasan buatan untuk tantangan penerapan pembelajaran mesin yang terus berkembang. Otomatisasi tingkat tinggi di AutoML memungkinkan non-ahli untuk menggunakan model dan teknik pembelajaran mesin tanpa mengharuskan mereka menjadi ahli dalam pembelajaran mesin.

Kita ditugas ini menggunakan dataset weatherAUS.csv yaitu untuk memprediksi apakah hari ini dan atau besok akan turun hujan atau tidak. Hal ini didasari oleh variable-variabel yang terdapat dalam Dataset tersebut, seperti kecepatan angin dan kelembapan pada hari itu.

Tujuan dari tugas ini adalah menyelesaikan sebuah task terhadap dataset weatherAUS.csv dengan menerapkan Automated Machine Learning. Machine Learning Automation Tools yang kita gunakan adalah TPOT

2. Data Eksploration dan Data Cleansing

Pada tahap ini dataset akan dieksplor, mulai dari menghitung jumlah record (data), identifikasi atribut-atribut yang tersedia, analisis missing value dan outlier pada masing-masing atribut, dan analisis hubungan antar atribut.

A. Data Eksporation.

Gambaran dari dataset yang akan digunakan pada bagian eksplorasi data.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0	24.0	71.0
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0	22.0	44.0
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0	26.0	38.0
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0	9.0	45.0
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0	20.0	82.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
145455	2017-06-21	Uluru	2.8	23.4	0.0	NaN	NaN	E	31.0	SE	ENE	13.0	11.0	51.0
145456	2017-06-22	Uluru	3.6	25.3	0.0	NaN	NaN	NNW	22.0	SE	N	13.0	9.0	56.0
145457	2017-06-23	Uluru	5.4	26.9	0.0	NaN	NaN	N	37.0	SE	WNW	9.0	9.0	53.0
145458	2017-06-24	Uluru	7.8	27.0	0.0	NaN	NaN	SE	28.0	SSE	N	13.0	7.0	51.0
145459	2017-06-25	Uluru	14.9	NaN	0.0	NaN	NaN	NaN	NaN	ESE	ESE	17.0	17.0	62.0

145460 rows × 23 columns

Terdapat 145.460 data dengan 23 kolom pada dataset weatherAUS.csv

Berikut adalah dataset weatherAUS.csv beserta keterangannya

- Date : Objek tanggal
- Location : Lokasi
- MinTemp : Suhu minumun
- MaxTemp : Suhu maksimal
- Rainfal : Banyaknya curah hujan
- Evaporation : Penguapan
- Sunshine : Sinar Matahari
- WindGustDir : Arah angin
- WindGustSpeed : Kecepatan angin
- WindDir9am : Arah angin jam 9 pagi
- WindDir3pm : Arah angin jam 3 sore
- WindSpeed9am : Kecepatan angin jam 9 pagi
- WindSpeed3pm : Kecepatan angin jam 3 sore
- Humidity9am : Kelembaban jam 9 pagi
- Humidity3pm : Kelembaban jam 3 sore
- Pressure9am : Tekanan atmosfer jam 9 pagi
- Pressure3pm : Tekanan atmosfer jam 3 sore
- Cloud9am : Langit tertutup awan jam 9 pagi

- Cloud3pm : Langit tertutup awan jam 3 sore
- Temp9am : Suhu jam 9 pagi
- Temp3pm : Suhu jam 3 sore
- RainToday : Hujan hari ini
- RainTomorrow : Hujan besok

Mencari dan menghapus Missing Value :

```
data_cuaca.isnull().any()
```

```
Date             False
Location         False
MinTemp           True
MaxTemp           True
Rainfall          True
Evaporation       True
Sunshine          True
WindGustDir       True
WindGustSpeed     True
WindDir9am        True
WindDir3pm        True
WindSpeed9am      True
WindSpeed3pm      True
Humidity9am       True
Humidity3pm       True
Pressure9am       True
Pressure3pm       True
Cloud9am          True
Cloud3pm          True
Temp9am           True
Temp3pm           True
RainToday         True
RainTomorrow      True
dtype: bool
```

```
Date             0
Location         0
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
RainTomorrow     0
dtype: int64
```

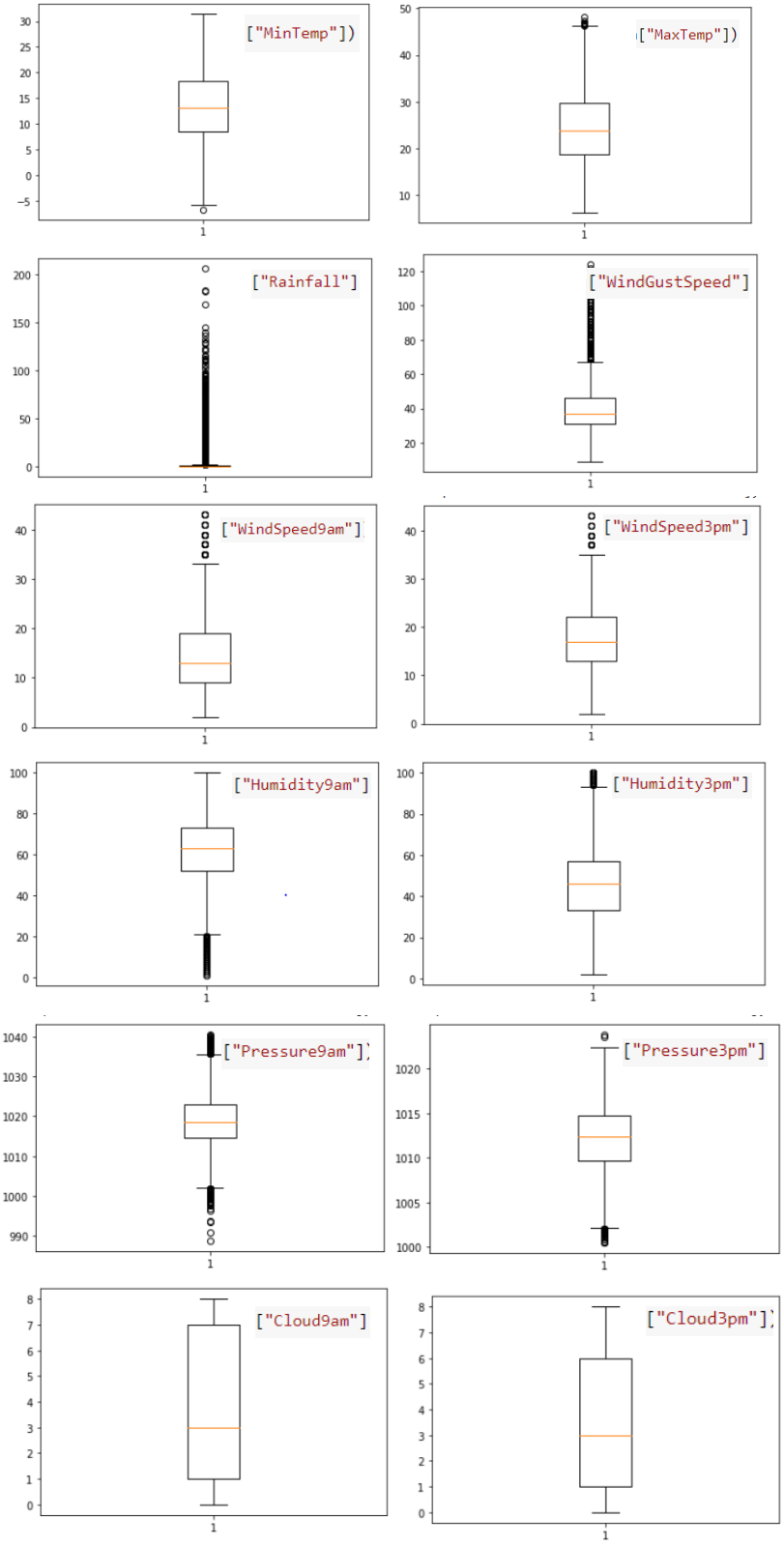
Cek data apakah terdapat missing value >>> setelah di hapus missing value

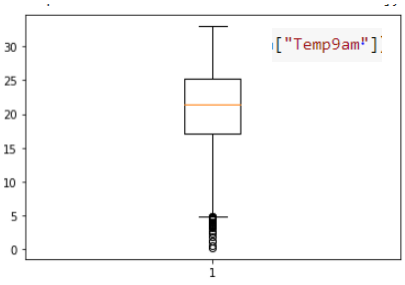
Melihat Statistic Summary dari setiap atribut

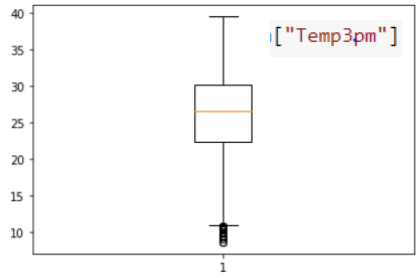
	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm
count	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000
mean	13.464770	24.219206	2.130397	5.503135	7.735626	40.877366	15.667228	19.786778
std	6.416689	6.970676	7.014822	3.696282	3.758153	13.335232	8.317005	8.510180
min	-6.700000	4.100000	0.000000	0.000000	0.000000	9.000000	2.000000	2.000000
25%	8.600000	18.700000	0.000000	2.800000	5.000000	31.000000	9.000000	13.000000
50%	13.200000	23.900000	0.000000	5.000000	8.600000	39.000000	15.000000	19.000000
75%	18.400000	29.700000	0.600000	7.400000	10.700000	48.000000	20.000000	26.000000
max	31.400000	48.100000	206.200000	81.200000	14.500000	124.000000	67.000000	76.000000

B. Data Cleansing

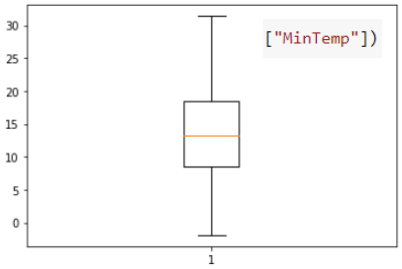
Mencari dan menghapus data outlier :

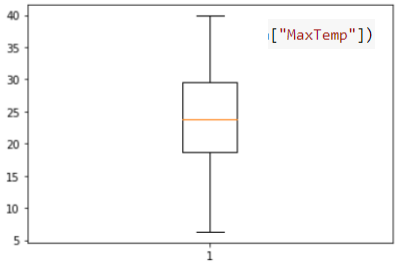


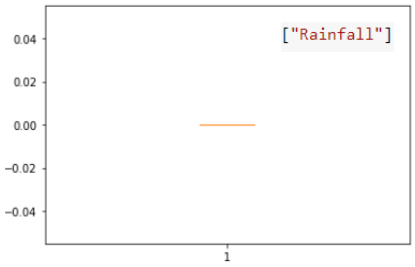


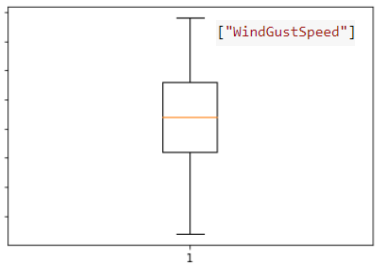


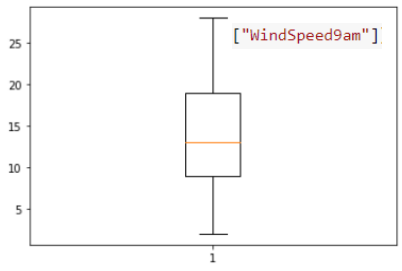
Menghapus outlier :

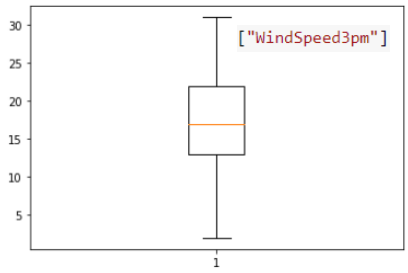


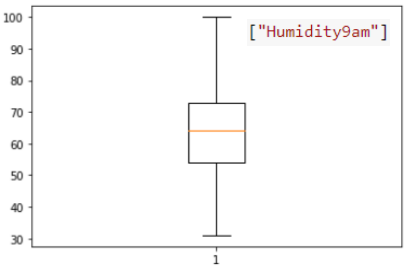


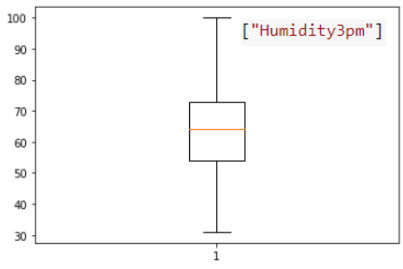


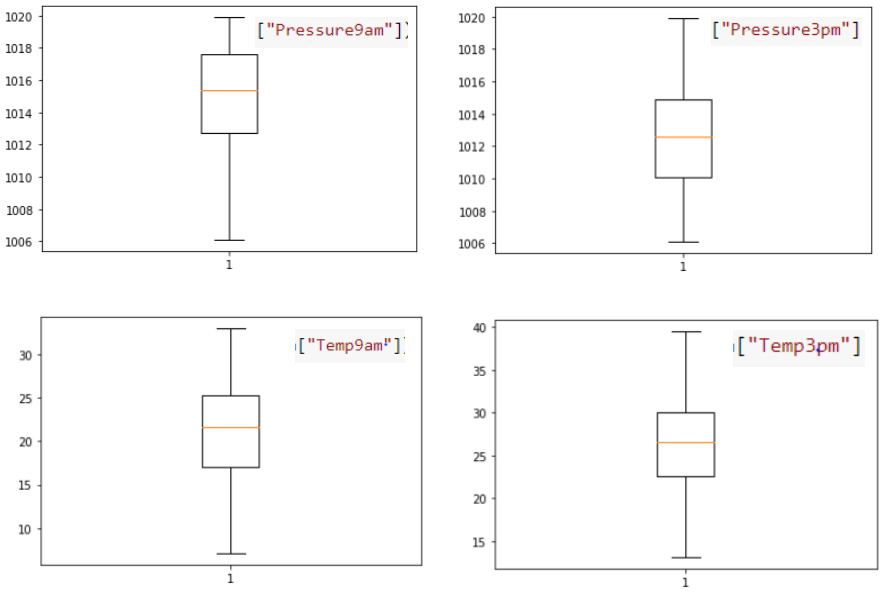












	Date	Location	MinTemp	MaxTemp	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
6053	2009-01-05	Cobar	21.9	38.4	11.4	12.2	WNW	31.0	WNW	WSW	6.0	6.0	37.0	22.0
6056	2009-01-08	Cobar	23.3	34.0	9.8	12.6	SSW	41.0	S	SSE	17.0	19.0	33.0	15.0
6061	2009-01-13	Cobar	23.9	39.1	13.8	12.1	ENE	39.0	NE	N	24.0	9.0	40.0	15.0
6067	2009-01-19	Cobar	21.4	37.5	14.8	6.9	NNE	43.0	ENE	NNE	26.0	9.0	34.0	29.0
6069	2009-01-21	Cobar	25.4	33.5	13.6	3.7	N	46.0	NW	N	9.0	28.0	46.0	52.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
142298	2017-06-20	Darwin	19.3	33.4	6.0	11.0	ENE	35.0	SE	NE	9.0	20.0	63.0	32.0
142299	2017-06-21	Darwin	21.2	32.6	7.6	8.6	E	37.0	SE	SE	13.0	11.0	56.0	28.0
142300	2017-06-22	Darwin	20.7	32.8	5.6	11.0	E	33.0	E	W	17.0	11.0	46.0	23.0
142301	2017-06-23	Darwin	19.5	31.8	6.2	10.6	ESE	26.0	SE	NNW	9.0	17.0	62.0	58.0
142302	2017-06-24	Darwin	20.2	31.7	5.6	10.7	ENE	30.0	ENE	NNW	15.0	7.0	73.0	32.0

15196 rows × 22 columns

Dataset weatherAUS.csv setelah di menghapus missing value dan oulier

C. Feature Engineering

Pada Feature Engineering ini melakukan Categorial Enconding yaitu merubah data categorial menjadi numerik.

```
# cek variable yang akan diubah data kategorial menjadi data numerik

print(data_cuaca.WindGustDir.unique())
print(data_cuaca.WindDir9am.unique())
print(data_cuaca.WindDir3pm.unique())
print(data_cuaca.RainToday.unique())
print(data_cuaca.RainTomorrow.unique())
```

```
['WNW' 'SSW' 'ENE' 'NNE' 'N' 'S' 'E' 'ESE' 'SW' 'SSE' 'NE' 'WSW' 'SE'
 'NNW' 'NW' 'W']
['WNW' 'S' 'NE' 'ENE' 'NW' 'NNE' 'SSW' 'E' 'WSW' 'N' 'SW' 'ESE' 'SSE' 'SE'
 'W' 'NNW']
['WSW' 'SSE' 'N' 'NNE' 'E' 'ESE' 'ENE' 'NE' 'S' 'SE' 'SSW' 'NNW' 'WNW'
 'NW' 'SW' 'W']
['No']
['No' 'Yes']
```

```
data_num = {"WindGustDir" : {"W": 0,"WNW": 1,"NW": 2,"NNW": 3,"N": 4,"NNE": 5,"NE": 6,"ENE": 7,
                             "E": 8,"ESE": 9,"SE": 10,"SSE": 11,"S": 12,"SSW": 13,"SW": 14,"WSW": 15},
            "WindDir9am"  : {"W": 0,"WNW": 1,"NW": 2,"NNW": 3,"N": 4,"NNE": 5,"NE": 6,"ENE": 7,
                             "E": 8,"ESE": 9,"SE": 10,"SSE": 11,"S": 12,"SSW": 13,"SW": 14,"WSW": 15},
            "WindDir3pm"  : {"W": 0,"WNW": 1,"NW": 2,"NNW": 3,"N": 4,"NNE": 5,"NE": 6,"ENE": 7,
                             "E": 8,"ESE": 9,"SE": 10,"SSE": 11,"S": 12,"SSW": 13,"SW": 14,"WSW": 15},
            "RainToday"   : {"Yes": 1, "No": 0},
            "RainTomorrow": {"Yes": 1, "No": 0}}
data_cuaca = data_cuaca.replace(data_num)
```

Pada tahap ini atribut yang bertipe kategorial akan diubah ke dalam tipe numerik, sehingga atribut – atribut kategorial tersebut dapat dianalisa. Data yang dirumah yaitu pada artibut WindGustDir, WindDir9am, WindDir3pm, RainToday, RaindTomorrow.

	WindGustDir	WindDir9am	WindDir3pm	RainToday	RainTomorrow
6053	1	1	15	0	0
6056	13	12	11	0	0
6061	7	6	4	0	0
6067	5	7	5	0	1
6069	4	2	4	0	0
...	...	...	...	...	...
142298	7	10	6	0	0
142299	8	10	10	0	0
142300	8	8	0	0	0
142301	9	10	3	0	0
142302	7	7	3	0	0

15196 rows × 5 columns

Berikut adalah hasil setelah melakukan encoding

D. Penentuan Fitur dan Target

Setelah dilakukan data cleansing selanjutnya adalah menentukan fitur beserta targetnya. Kami menentukan yang menjadi fitur adalah mulai dari kolom ke 3, yaitu MinTemp sampai kolom ke 20, yaitu Temp3pm. Lalu untuk targetnya kami bagi menjadi 2, yaitu Target_Today untuk label RainToday dan Target_Tomorrow untuk label RainTomorrow.

```
#Feature and Target
Feature = data_cuaca.iloc[:,[2,3,4,5,6,7,8,9,19,11,12,13,14,15,16,17,18,19]].values
Target_Today = data_cuaca.iloc[:,20].values
Target_Tomorrow = data_cuaca.iloc[:,21].values
```

E. Data Scaling

Kami melakukan scaling untuk mengubah semua value yang terdapat didalam fitur menjadi berada di range 0 sampai 1. Oleh karena itu kami menggunakan function min_max_scaler untuk dilakukannya scaling.

```
#Data Scaling
min_max_scaler = preprocessing.MinMaxScaler()
Feature = min_max_scaler.fit_transform(Feature)
```

3. Pemodelan

Untuk pemodelan kali ini kita menggunakan salah satu tools AutoML, yaitu TPOT. TPOT dapat melakukan optimasi machine learning pipelines menggunakan Genetic Algorithm (GA). Untuk dapat menggunakan TPOT, Pertama-tama kita import TPOT dengan ketik "from tpot import TPOTClassifier".

```
from tpot import TPOTClassifier
```

Setelah itu kita definisikan tpot, untuk hal tersebut saya menggunakan 2 tpot dengan cross validation yang berbeda.Untuk yang pertama kami menggunakan 5 fold cross validation.

```
#tpot1
tpot1 = TPOTClassifier(generations=5, population_size=20, verbosity=2, cv=5, n_jobs=-2)
```

Dan untuk yang kedua menggunakan repeated stratisfied k fold dengan dengan 10 fold.

```
# define model evaluation using stratisfied kfold
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1)

#tpot2
tpot2 = TPOTClassifier(generations=5, population_size=20, verbosity=2, cv=cv, n_jobs=-2)
```

4. Eksperimen

Untuk eksperimen yang kami lakukan adalah melakukan proses AutoML TPOT sebanyak 4 kali. Untuk 2 yang pertama, kami lakukan tpot dengan 5 generasi, 20 populasi, dan untuk model validationnya kami menggunakan 5-fold cross validation.

Lalu untuk 2 percobaan terakhir kami lakukan tpot dengan 5 generasi, 20 populasi, dan untuk model validationnya kami menggunakan Repeated Stratified KFold dengan sebanyak 10 fold.

5. Evaluasi

Untuk hasil yang didapat dari percobaan diatas, antara menggunakan 5-fold cross validation dan Repeated Stratified KFold, mendapatkan akurasi yang tidak jauh berbeda. Hanya saja terdapat perbedaan waktu prosesnya, Untuk tpot yang menggunakan Repeated Stratified KFold membutuhkan waktu yang lebih lama dari 5-fold cross validation. Hal tersebut karena melakukan repeat cross validation procedure berulang kali (kami set n-repeats nya sebanyak 3 kali) lalu mengeluarkan rata-rata hasil di tiap fold dari tiap run.

6. Kesimpulan

Setelah melalui proses-proses diatas, dapat kami simpulkan bahwa TPOT merupakan tools AutoML yang cukup baik dengan hanya membutuhkan kerja keras manusia dalam hal data cleansing, feature engineering, data scaling, dan model validation. Untuk sisanya seperti pemilihan model yang terbaik, proses genetic algorithm, dll dilakukan oleh TPOT secara otomatis. Dari hasil yang didapat TPOT berhasil melakukan pemodelan Machine Learning dengan akurasi 100% untuk label RainToday dan 89% untuk label RainTomorrow.