Project Kaggle

Rain Prediction in Australia

By Pandas Team

DigitalSkola

Table of Contents

Stage 1

Understanding the Data

Understanding the Data

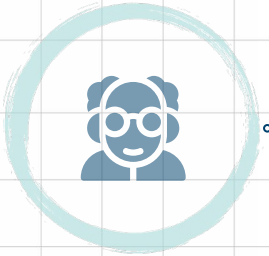
- Dataset pengamatan ini diambil dari sistem biro meteorologi. Sebagian besar data dihasilkan dan ditangani secara otomatis.
- Dataset ini memaparkan data meteorologi (145.460 observasi) dalam beberapa tahun (2008-12-01 s/d 2017-07-01) di sejumlah lokasi (49) di Australia dengan beberapa parameter (variabel) yang berhubungan seperti temperatur, evaporasi, curah hujan, tekanan udara, kelembapan udara, sinar matahari, awan, dan kecepatan angin.
- Dengan beberapa parameter (variabel) ini kita akan memprediksi target variabel yang mana adalah RainTomorrow (Hujan besok).

Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	**Min**	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	**Dirn**	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	**Temp**	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	**Temp**	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

Stage 2

Plan the Activities

Identify Which Activities Should be Done

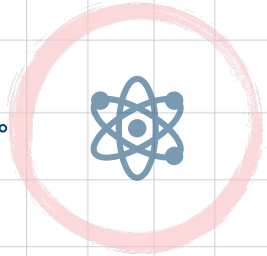


1

Melakukan EDA dan data visualisasi.

2

Melakukan pre proses data



3

Membuat model Machine Learning

Stage 3

Exploratory Data Analysis

Explore the Dataset

Dataset

Fitur
info & duplicated

16 Variabel data numerik

7 Variabel data kategorik

0 Data yang terduplikasi

Check For Missing Data

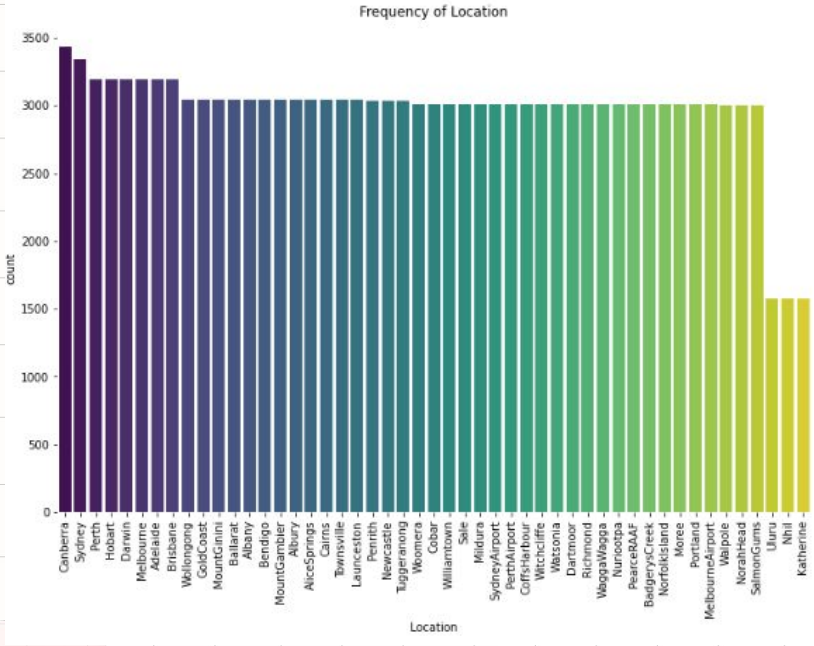
	Total Missing	%
Sunshine	69835	48.0
Evaporation	62790	43.2
Cloud3pm	59358	40.8
Cloud9am	55888	38.4
Pressure9am	15065	10.4
Pressure3pm	15028	10.3
WindDir9am	10566	7.3
WindGustDir	10326	7.1
WindGustSpeed	10263	7.1
Humidity3pm	4507	3.1

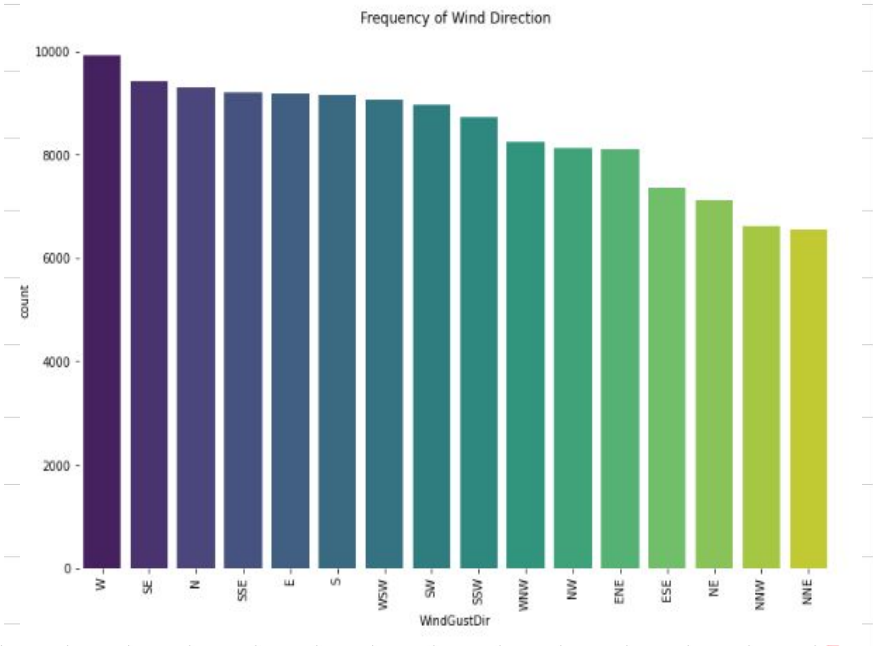
Fitur
isnull & sum

- Variable Date dan Location tidak memiliki missing values.
- 4 besar variabel dengan missing values :
  1. Sunshine
  2. Evaporation
  3. Cloud3pm
  4. Cloud9am

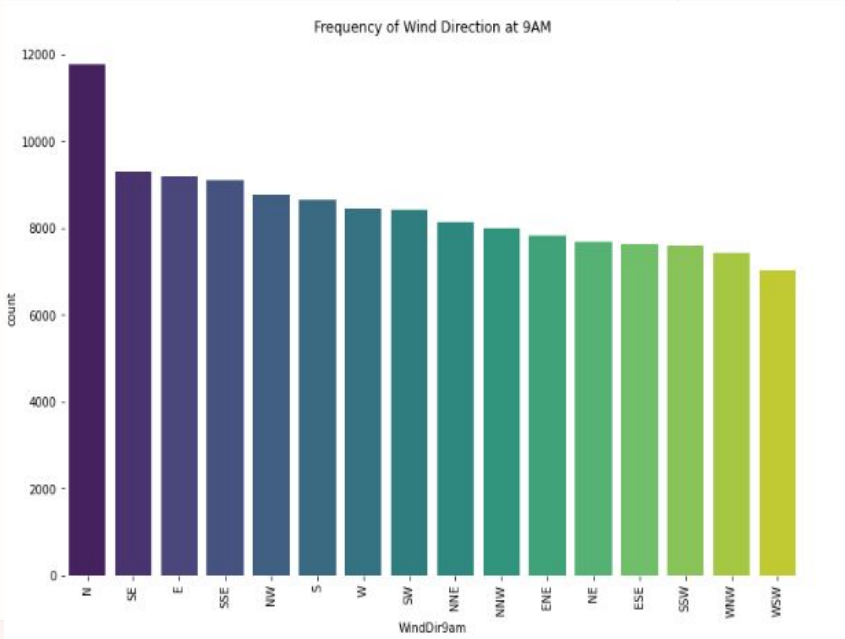
Analyze the Dataset with Visualization

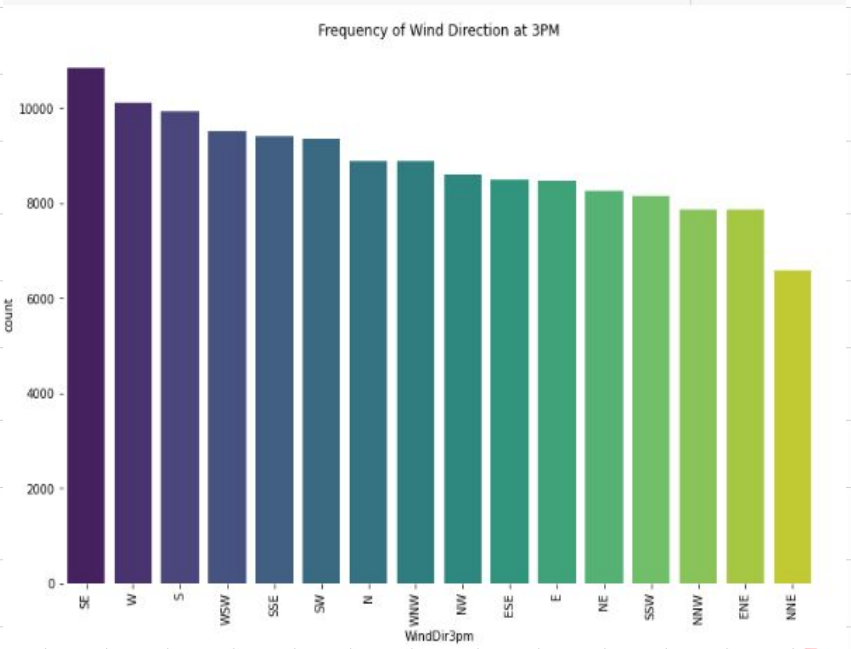
Find the Frequency of Data



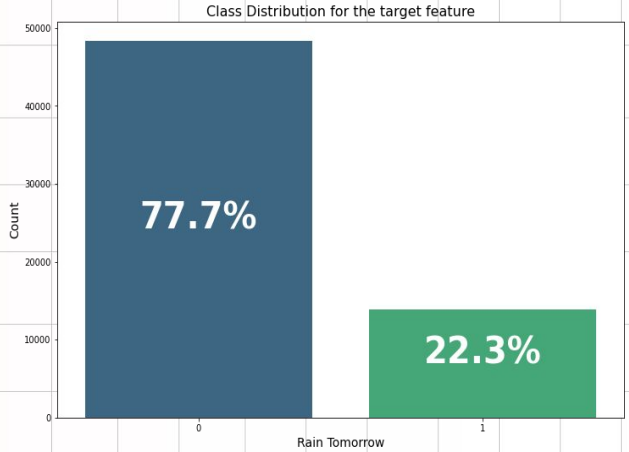


Find the Frequency of Data





Find the Frequency of Data

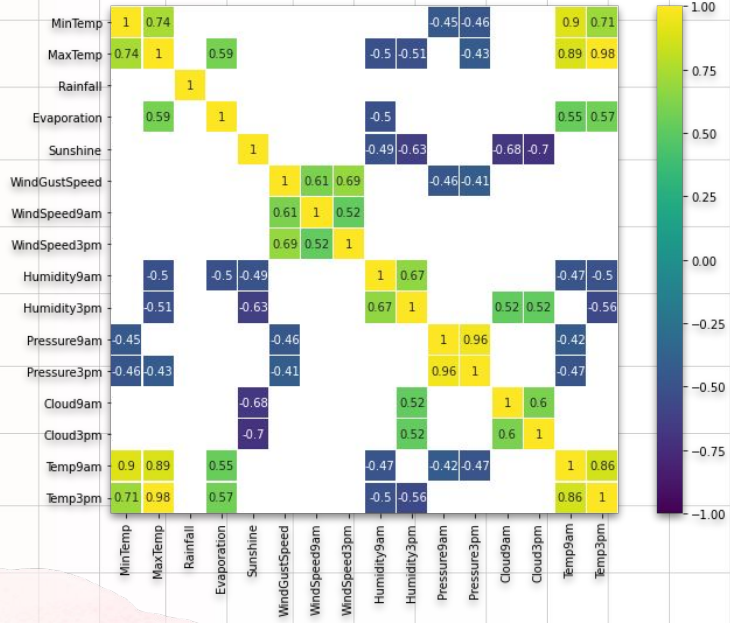


Countplot

Identifikasi freukensi variabel kategorik dengan countplot:

- Top 5 lokasi dengan frekuensi terbanyak : Canberra, Sydney, Darwin, Melbourne, Perth.
- Top 5 wind direction dengan frekuensi terbanyak : W, SE, N, SSE, E.
- Top 5 windgustdir 9am dengan frekuensi terbanyak : N, SE, E, SSE, NW.
- Top 5 windgustdir 3pm dengan frekuensi terbanyak : SE, W, S, WSW, SSE.
- Persentase tidak hujan (77.7%) lebih besar daripada hujan (22.3%) untuk variabel RainTomorrow.

Find the Correlation of Data



Fitur Corr Pearsonr

Nilai	Korelasi	Variabel
> 0.9	Positif	Pressure9am - Pressure3pm
> 0.8	Positif	MinTemp - MaxTemp - Temp9am - Temp3pm
< -0.7	Negatif	Cloud9am, Cloud3pm - Humidity3pm
> 0.6	Positif	WindGustSpeed - WindSpeed9am - WindSpeed3pm

Find the Distribution of Data

Histplot

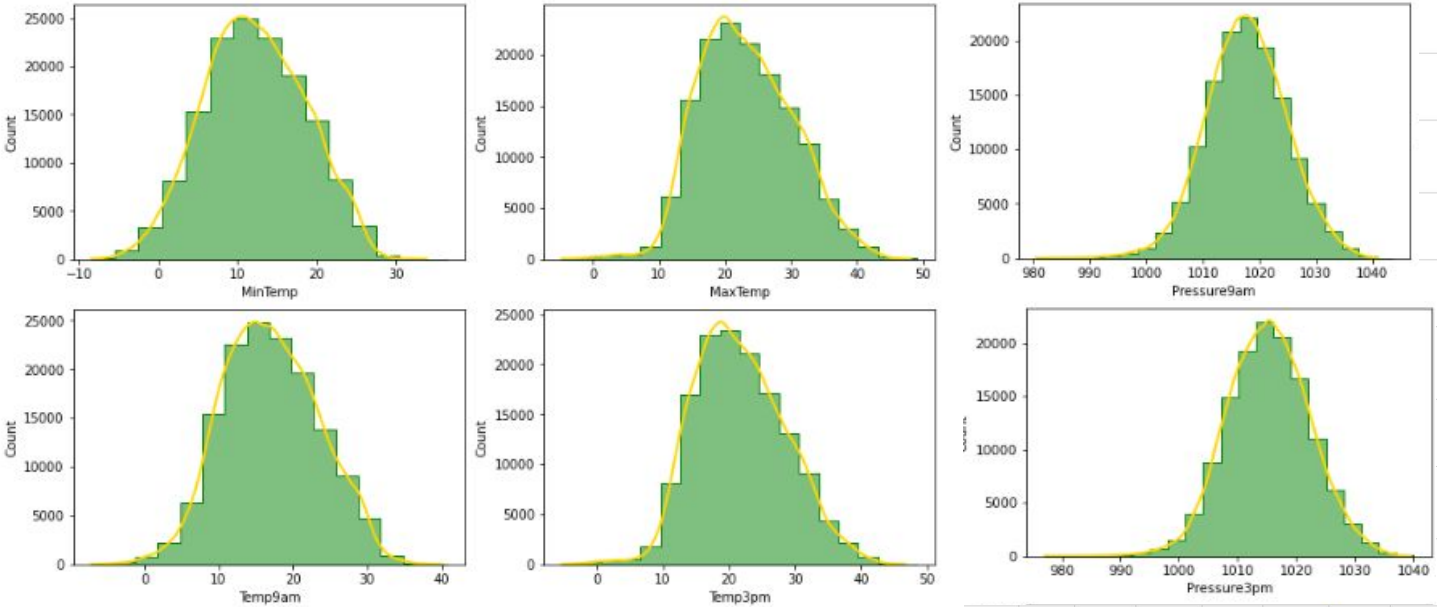
6 Variabel data numerik dengan distribusi normal

5 Variabel data numerik dengan distribusi skew positif

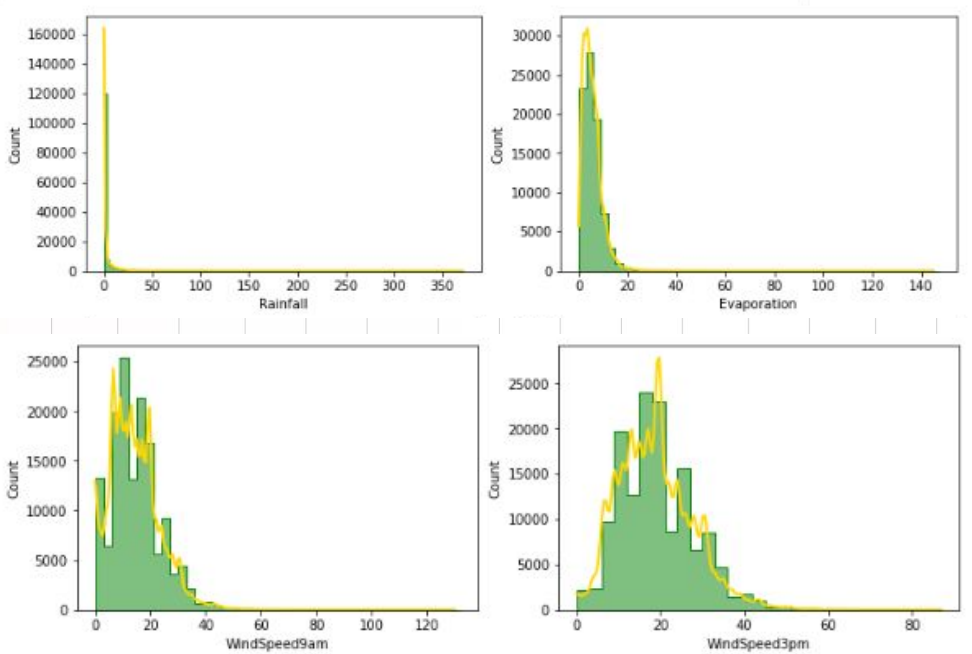
3 Variabel data numerik dengan skew negatif

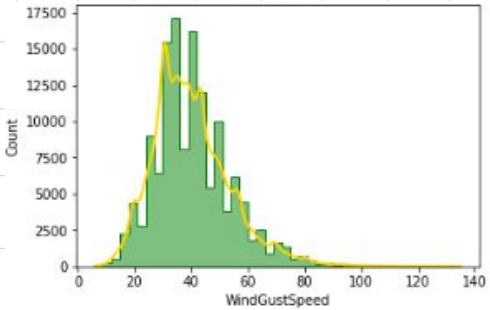
2 Variabel data numerik dengan distribusi uniform

Find the Distribution of Data

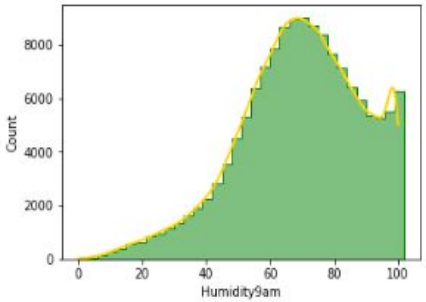


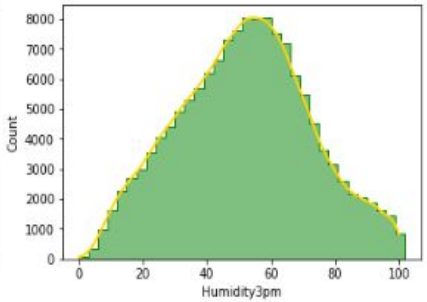
Find the Distribution of Data

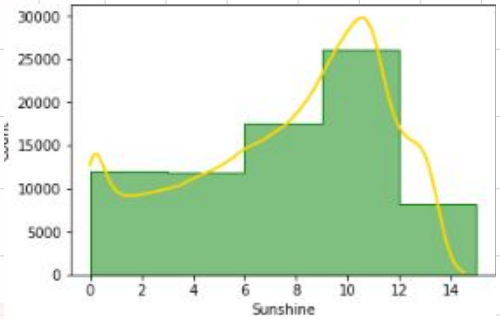


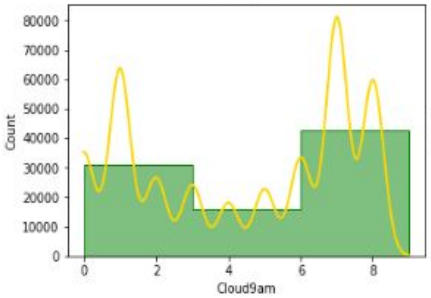


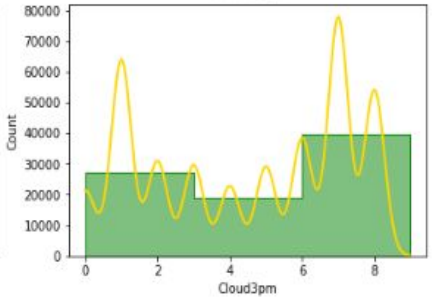
Find the Distribution of Data



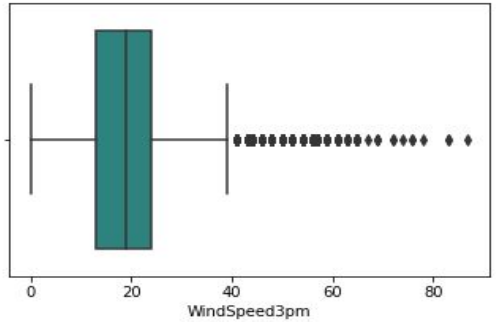








Find the Outliers of Data



BoxPlot

16 Variabel data numerik

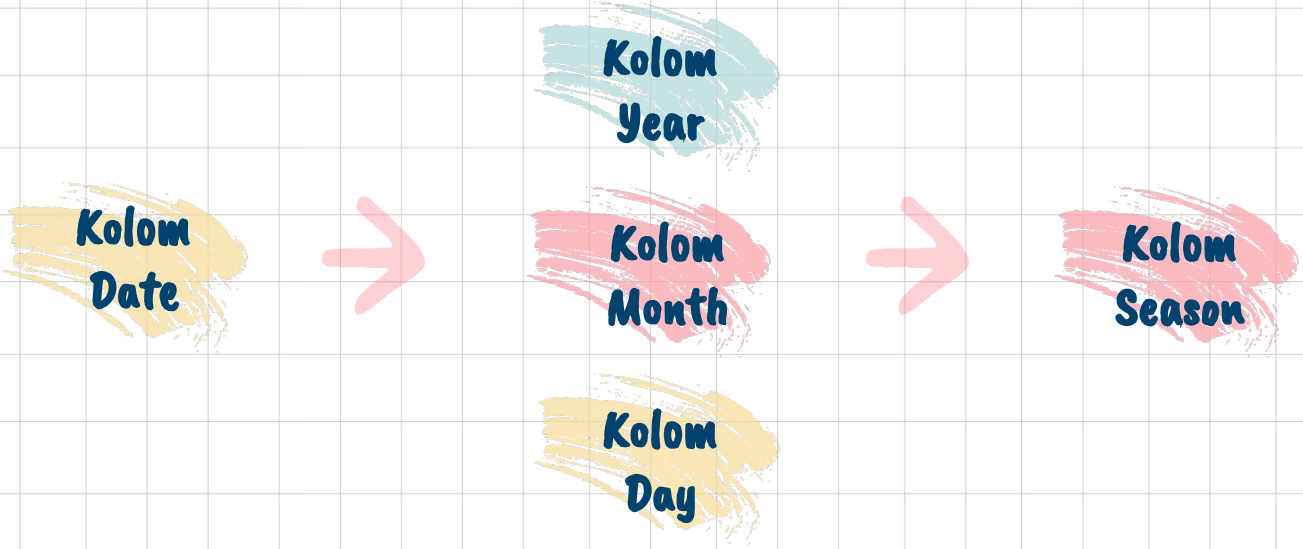
12 Variabel data numerik memiliki outlier.

4 Variabel data numerik tidak memiliki outlier.

Stage 4

Data Pre-processing

Feature Engineering



Missing Values Handling

Column	Total	%	Column	Total	%
Sunshine	69835	48.0	RainToday	3261	2.2
Evaporation	62790	43.2	Rainfall	3261	2.2
Cloud3pm	59358	40.8	WindSpeed3pm	3062	2.1
Cloud9am	55888	38.4	Humidity9am	2654	1.8
Pressure9am	15065	10.4	Temp9am	1767	1.2
Pressure3pm	15028	10.3	WindSpeed9am	1767	1.2
WindDir9am	10566	7.3	MinTemp	1485	1.0
WindGustDir	10326	7.1	MaxTemp	1261	0.9
WindGustSpeed	10263	7.1	Season	0	0.0
Humidity3pm	4507	3.1	Day	0	0.0
WindDir3pm	4228	2.9	Year	0	0.0
Temp3pm	3609	2.5	Month	0	0.0
RainTomorrow	3267	2.2	Location	0	0.0

1. Kolom >30% -> Drop.
2. Data kategorik -> Imputasi dengan Metode Mode.
3. Data numerik -> Imputasi Metode Median.

Outliers Values Handling

Column	Total	%	Column	Total	%
RainToday	13940	22.4	MinTemp	3	0.0
RainTomorrow	13909	22.3	Day	0	0.0
Rainfall	11148	17.9	Cloud9am	0	0.0
WindGustSpeed	1392	2.2	Location	0	0.0
Evaporation	1254	2.0	Humidity3pm	0	0.0
WindSpeed9am	1180	1.9	Year	0	0.0
WindSpeed3pm	1132	1.8	Month	0	0.0
Humidity9am	994	1.6	Season	0	0.0
Pressure9am	646	1.0	Sunshine	0	0.0
Pressure3pm	485	0.8	WindDir3pm	0	0.0
Temp3pm	32	0.1	WindDir9am	0	0.0
MaxTemp	13	0.0	WindGustDir	0	0.0
Temp9am	6	0.0	Cloud3pm	0	0.0

1. Kolom >15% -> Tetap.
2. Kolom <15% -> Imputasi dengan Metode Median.

Label Encoding

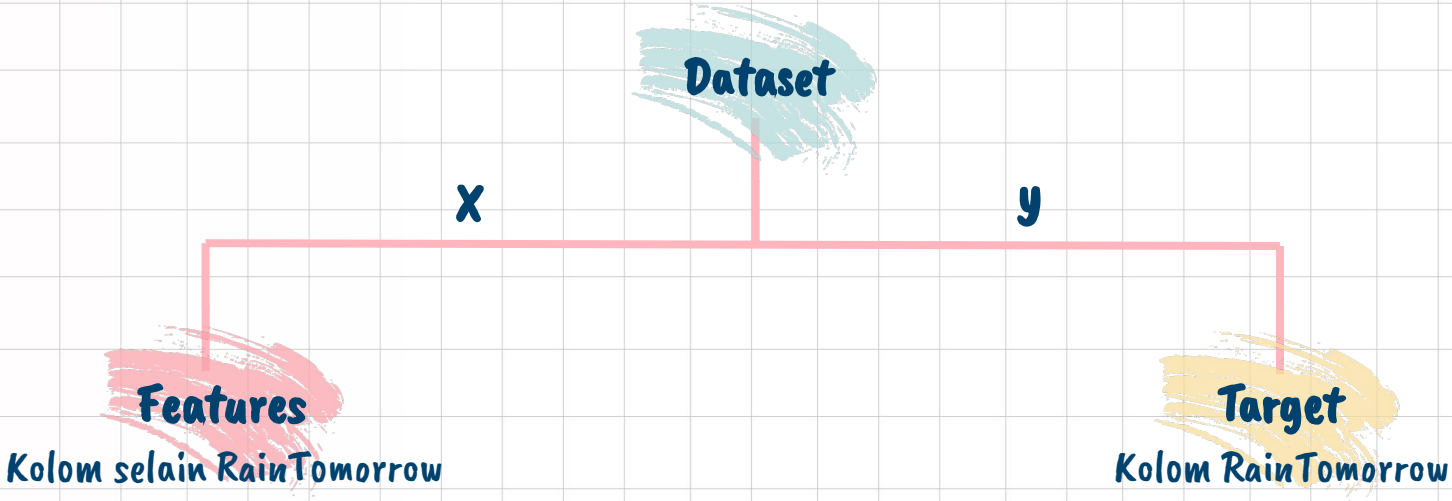
Nilai Data Kategorik

Fitur
Label Encoder

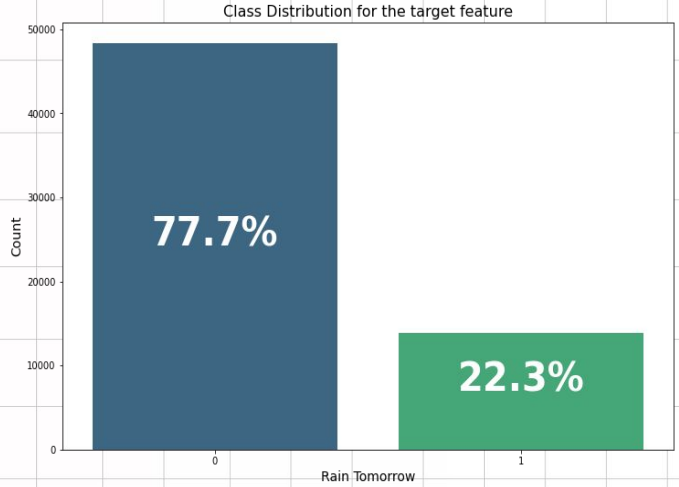
Nilai Data Numerik

Variabel:
WindGustDir, WindDir9am, WindDir3pm,
Location, RainToday, RainTomorrow.

Splitting Dataset into Features and Target



Imbalanced Class Handling



- Kelas kami katakan tidak seimbang ketika kelas minoritas hanya 5-10% dari keseluruhan.
- Tetapi untuk pelatihan kali ini kami mencoba metode random resampling.

Splitting Dataset Into Training and Test Set

Feature Scaling

Data Features

Metode Standardisation

1. Normalisasi data X_Train
2. Normalisasi data X_Test

Stage 5

Data Modelling

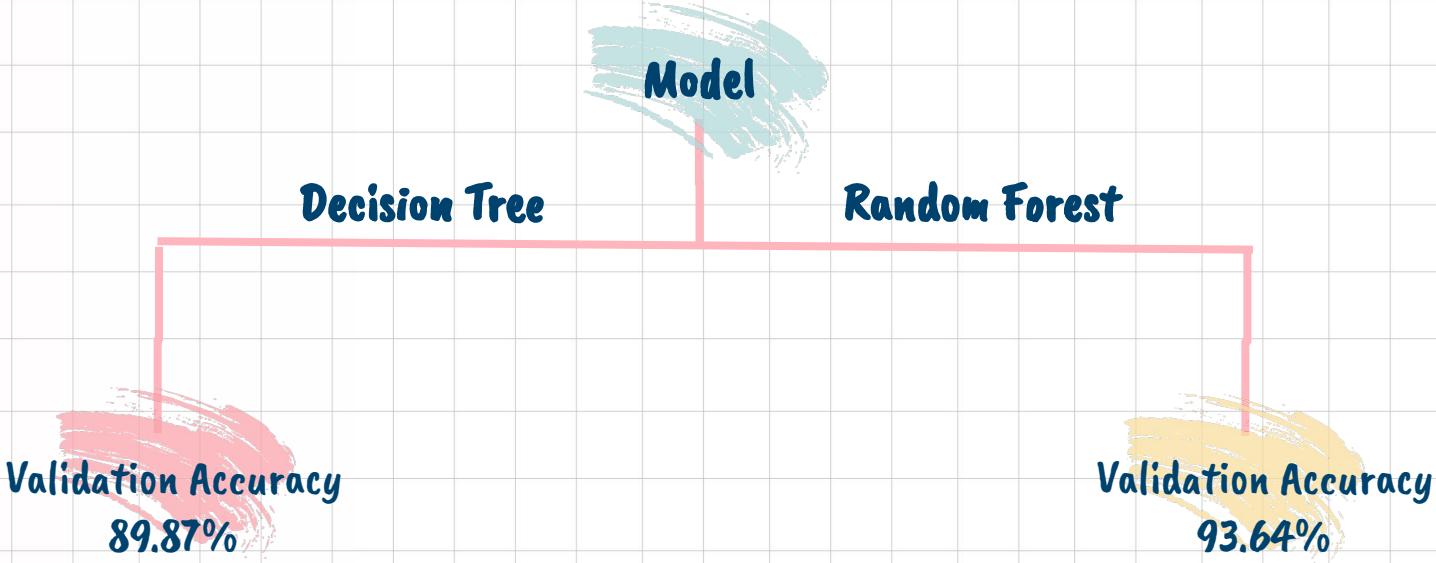
Model Selection

Score	Model
100.00	Random Forest
100.00	Decision Tree
89.14	KNN
79.53	Support Vector Machines
79.52	Logistic Regression

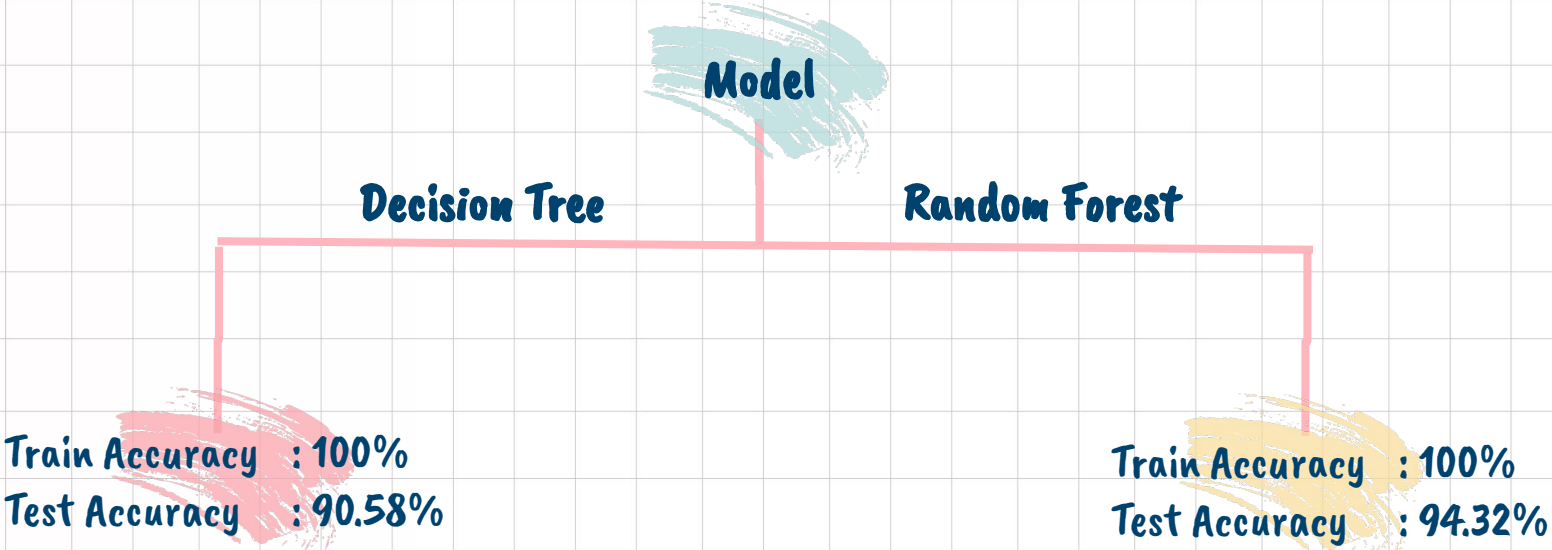
Berdasarkan hasil Score dari masing-masing Model Selection kami melakukan percobaan dengan:

1. Decision Tree
2. Random Forest

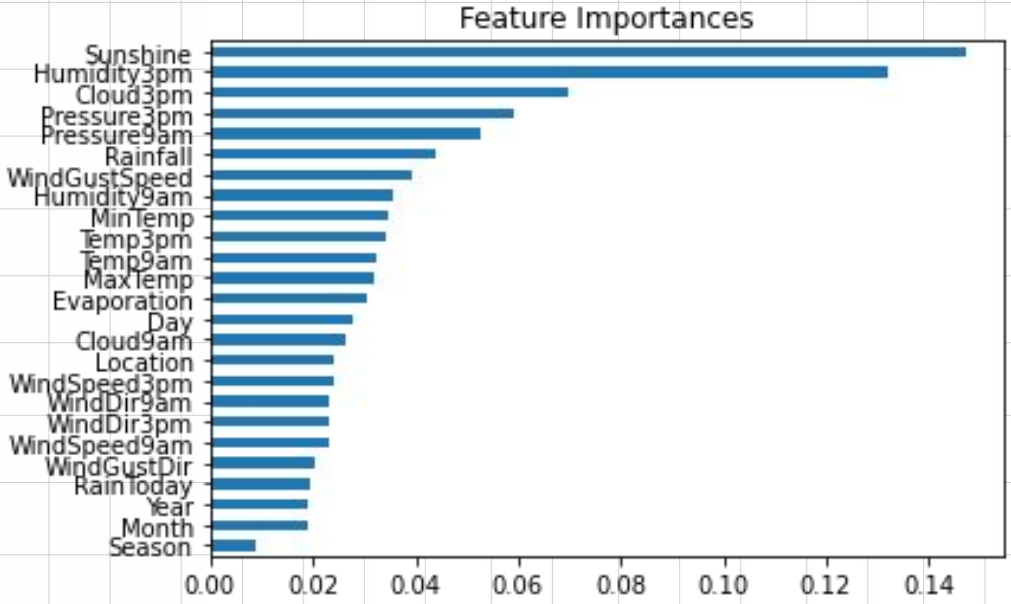
Cross Validation using K-Fold



Original Features



Feature Importance - Random Forest



Hyper Parameter

Model
Random Forest

Original Features

Train Accuracy : 100%
Test Accuracy : 94.32%

- Auto feature selection.
- Akurasi sedikit turun.
- Dapat mengurangi dimensional feature.

Hyper Parameter

Train Accuracy : 99.26%
Test Accuracy : 93.19%

Evaluation Matrix

01

Training & Test Accuracy

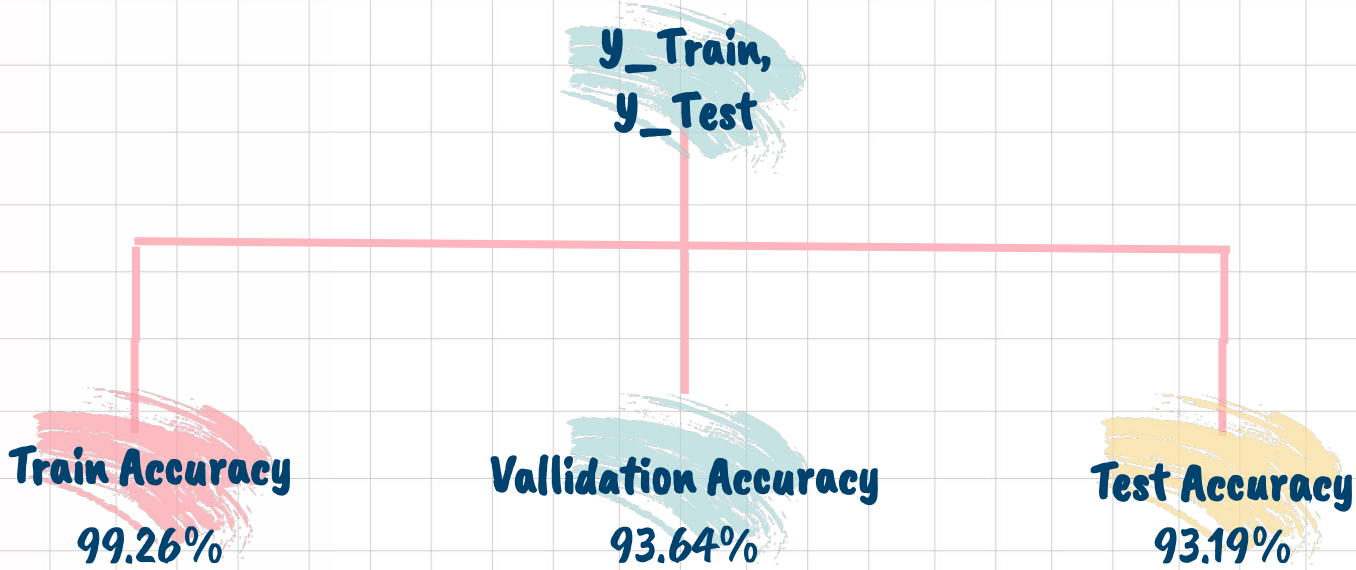
02

Precission, Recall & F-1 Score

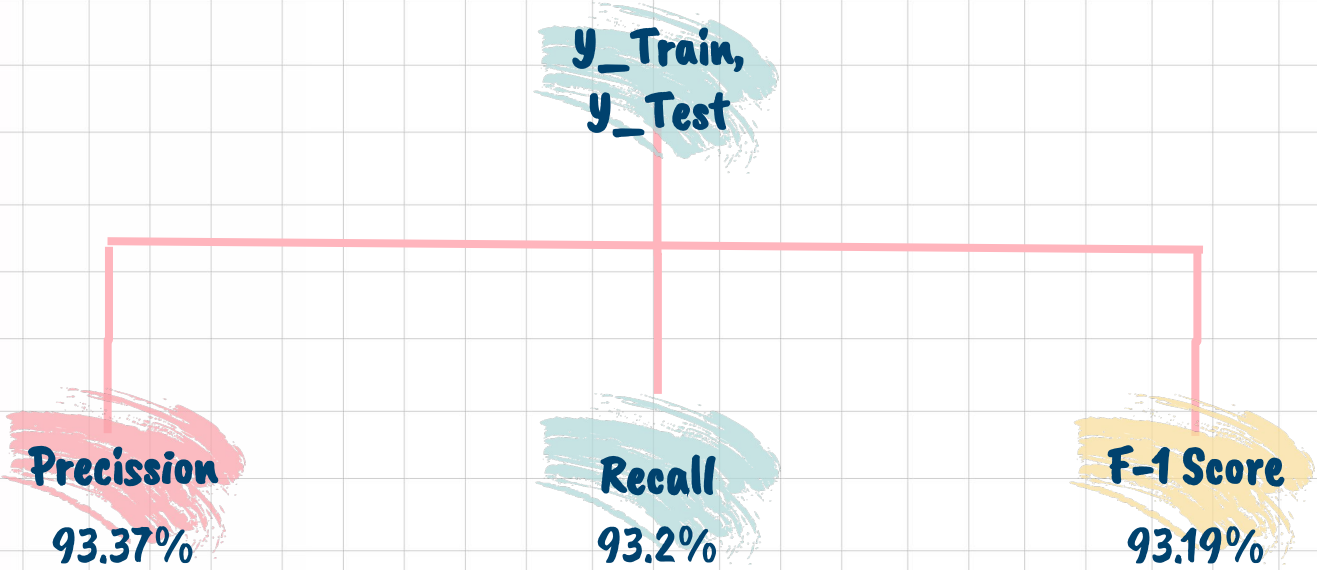
03

Confusion Marix

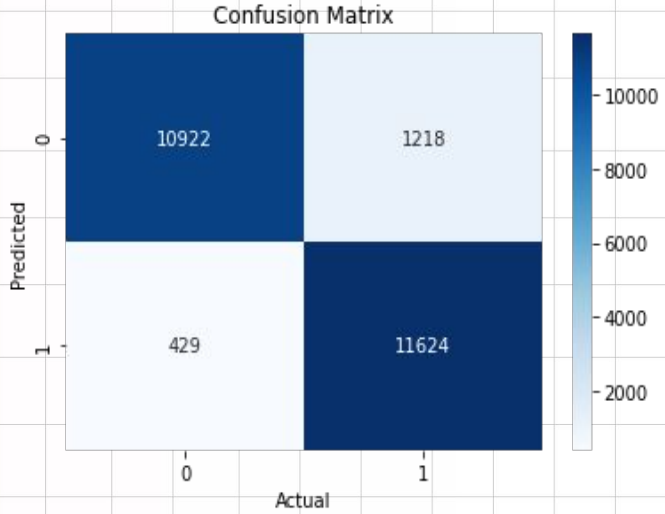
Train & Test Accuracy



Precission, Recall & F-1 Score



Confussion Matrix



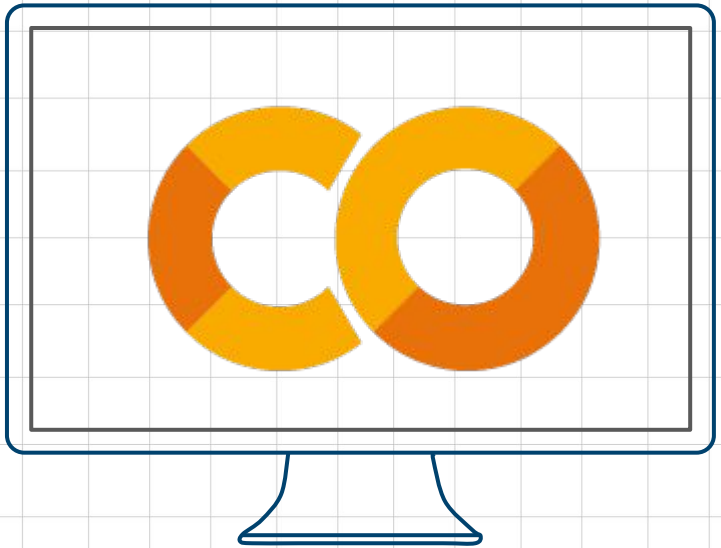
Fitur
Confusion Matrix

- Prediksi hujan yang sebenernya benar hujan adalah 10922
- Prediksi tidak hujan yang sebenernya tidak hujan adalah 11624
- Prediksi tidak hujan yang sebenernya benar hujan adalah 1218
- Prediksi hujan yang sebenernya tidak hujan adalah 429

Appendix

Link Notebook

Google Colab



Thanks!

Do you have any questions?
addyouremail@freepik.com
+91 620 421 838 yourcompany.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik.**

Please keep this slide for attribution.