

Final Project Progress

3D - DIGITAL SKOLA

PROJECT GOAL

To predict tomorrow will be rain or not

- •49 different locations in Australia from 1 November 2007 until 25 June 2017
- •23 columns with approximately 145K rows
- •The number of data collected is **not equal for** all locations
- •1 dependent variable ('RainTomorrow'), which called target, and 22 independent variables, which called features.

DATA UNDERSTANDING

DATA INFO

Data shape: 145460 rows and 23 columns

Top columns contain null:

- 1. Sunshine (69835)
- 2. Evaporation (62790)
- 3. Cloud3pm (59358)
- 4. Cloud9am (55888)

Duplicated data: not found

Total Location = 49 locations

Date = 3436 days

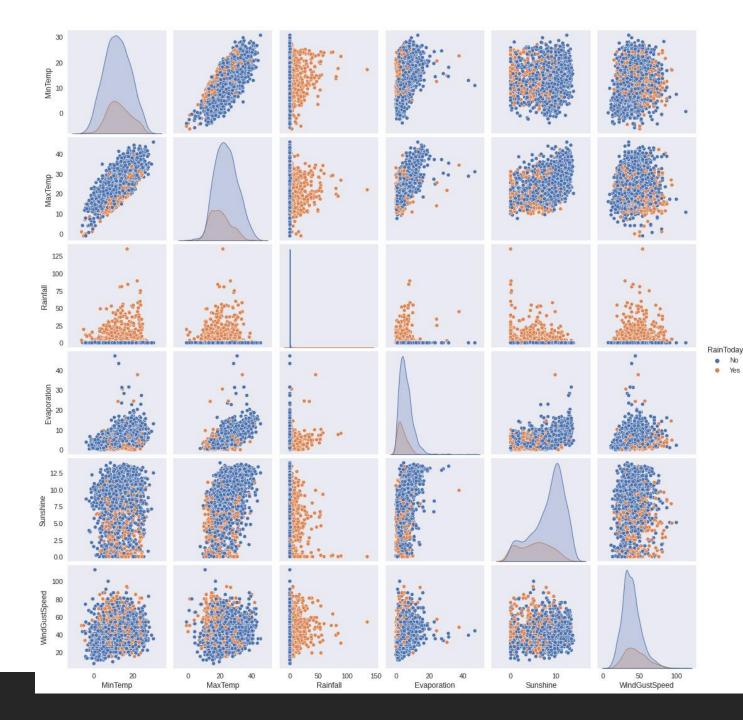
- 1. Berapa besar risiko turun hujan pada setiap tanggal dan setiap bulan sepanjang tahun?
- 2. Berapa besar risiko turun hujan pada setiap kota dalam satu tahun?
- 3. Bagaimana perubahan intensitas hujan dari tahun ke tahun
- 4. Bagaimana distribusi data dari masing-masing kolumn terhadap hujan/tidak hujan?
- 5. Bagaimana korelasi antara fitur numerik?
- 6. Berapa persen jumlah mising value dari masing-masing feature?
- 7. Bagaimana perbandingan antara RainTomorrow = Yes dan RainTomorrow = No?.

EXPLORATORY DATA ANALYSIS

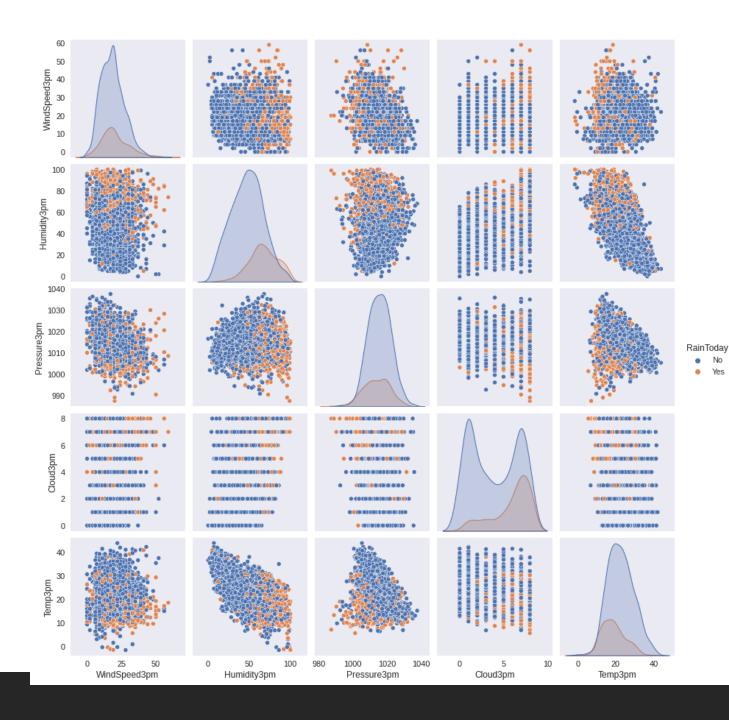
Insight:

- Months with high rain probability are June, July, and August
- Locations with high rain probability are Portland (36.4%),
 Walpole (33.5%), and Cairns (31,7%).
- 3. Yearly rain probability is around 20% and there is no significant change from 2009 to 2016
- 4. Target class distribution is 77% is NO RAIN, 23% RAIN.
- 5. Strong correlation between Temp3pm and MaxTemp,
 Pressure3pm and Pressure9am, Temp9am and MinTemp,
 Temp9am and MaxTemp, Temp3pm and Temp9am.

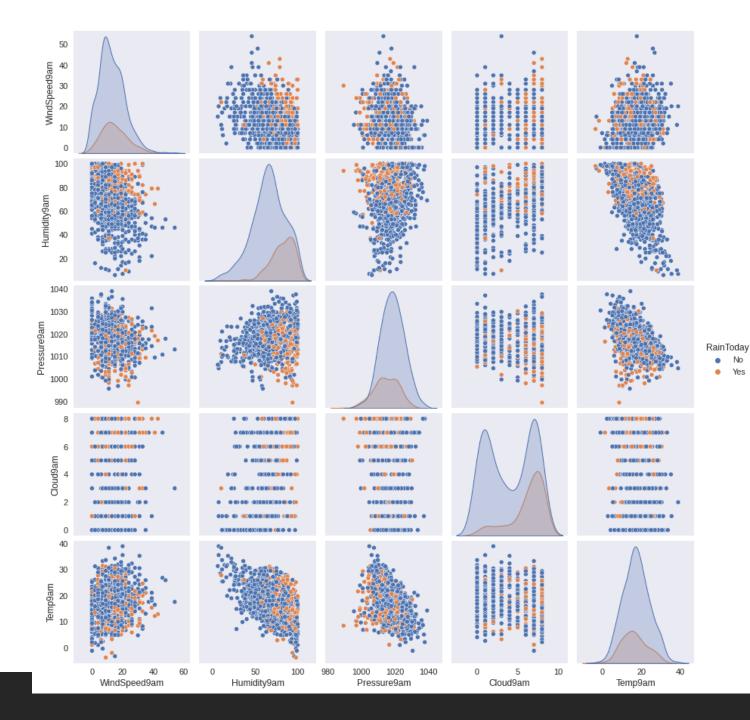
INSIGHT FROM EXPLORATORY DATA ANALYSIS



DATA DISTRIBUTION FROM EACH FEATURE



DATA DISTRIBUTION FROM EACH FEATURE



DATA DISTRIBUTION FROM EACH FEATURE

ı	Carra	lation	hetween	th.o.	facture	
1	erro. J	ianon	nerween	Ine	reamire	16

	Correlation between the reatures															
MinTemp	100%	74%	10%	18%	18%	18%	-23%	1%	-45%	-46%	90%	71%	7%	8%	2%	47%
MaxTemp	74%	100%	-7%	7%	1%	5%	-51%	-51%	-33%	-43%	89%	98%	47%	-29%	-28%	59%
Rainfall	10%	-7%	100%	13%	9%	6%	22%	26%	-17%	-13%	1%	-8%	-23%	20%	17%	-6%
WindGustSpeed	18%	7%	13%	100%	60%	69%	-22%	-3%	-46%	-41%	15%	3%	-3%	7%	11%	20%
WindSpeed9am	18%	1%	99%	60%	100%	52%	-27%	-3%	-23%	-17%	13%	1%	1%	2%	5%	19%
WindSpeed3pm	18%	5%	6%	69%	52%	100%	-15%	2%	-30%	-25%	16%	3%	6%	5%	3%	13%
Humidity9am	-23%	-51%	22%	-22%	-27%	-15%	100%	67%	14%	19%	-47%	-50%	-49%	45%	36%	-51%
Humidity3pm	1%	-51%	26%	-3%	-3%	2%	67%	100%	-3%	5%	-22%	-56%	-63%	52%	52%	-39%
Pressure9am	-45%	-33%	-17%	-46%	-23%	-30%	14%	-3%	100%	96%	-42%	-29%	4%	-13%	-15%	-27%
Pressure3pm	-46%	-43%	-13%	-41%	-17%	-25%	19%	5%	96%	100%	-47%	-39%	-2%	-6%	-8%	-29%
Temp9am	90%	89%	1%	15%	13%	16%	-47%	-22%	-42%	-47%	100%	86%	29%	-14%	-13%	55%
Temp3pm	71%	98%	-8%	3%	1%	3%	-50%	-56%	-29%	-39%	86%	100%	49%	-30%	-32%	57%
Sunshine	7%	47%	-23%	-3%	1%	6%	-49%	-63%	4%	-2%	29%	49%	100%	-68%	-70%	37%
Cloud9am	8%	-29%	20%	7%	2%	5%	45%	52%	-13%	-6%	-14%	-30%	-68%	100%	60%	-19%
Cloud3pm	2%	-28%	17%	11%	5%	3%	36%	52%	-15%	-8%	-13%	-32%	-70%	60%	100%	-18%
Evaporation	47%	59%	-6%	20%	19%	13%	-51%	-39%	-27%	-29%	55%	57%	37%	-19%	-18%	100%
	MinTemp	МахТетр	Rainfall	MndGustSpeed	MndSpeed9am	MindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Тетр9ат	ТетрЗрт	Sunshine	Cloud9am	Cloud3pm	Evaporation

CORRELATION
BETWEEN FEATURES

-0.2

-0.4

Removing Multicollinearity

- Temp9am and Temp3pm will be dropped
- MaxTemp and MinTemp will be combined into new variable named AvgTemp which equals to (MaxTemp+MinTemp)/2
- Pressure9am and Pressure3pm will be combined into new variable named AvgPressure which equals to (Pressure9am+Pressure3pm)/2

Handling Missing Values:

- Missing values in Rainfall, WindGustSpeed,
 WindSpeed9am, WindSpeed3pm, AvgPressure, AvgTemp,
 Evaporation, Humidity9am columns will be imputed with
 median value
- Missing values in Sunshine, Cloud9am, Cloud3pm, Humidity3pm columns will be imputed with mean value
- Missing values in WindDir9am, WindDir3pm, WindGustDir will be imputed with mode value

Encoding Categorical Data:

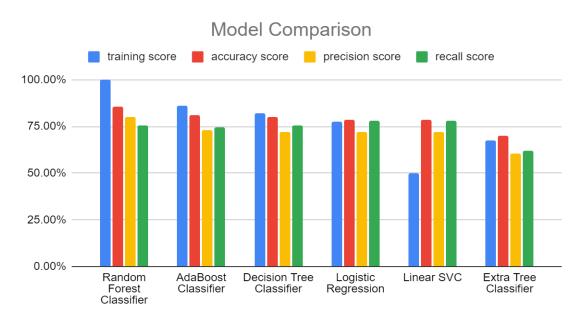
- For location and date will be encoded using pivot table method.
- For wind direction, encoded using one-hot encoding method

DATA PREPROCESSING

Training_Score Accuracy_Score Precision_Score Recall_Score

Model

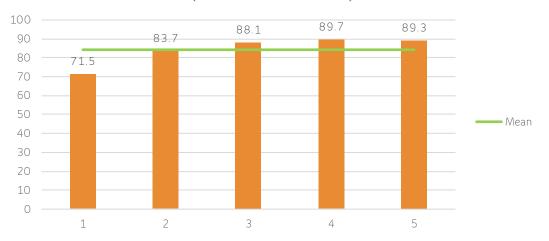
Random Forest Classifier	100.00	0.856111	0.803630	0.758062
AdaBoostClassifier	85.76	0.811993	0.730967	0.744607
Decision Tree Classifier	82.55	0.795959	0.717664	0.756292
Logistic Regression	77.62	0.784191	0.715988	0.776283
Linear SVC	50.00	0.783933	0.718054	0.782123



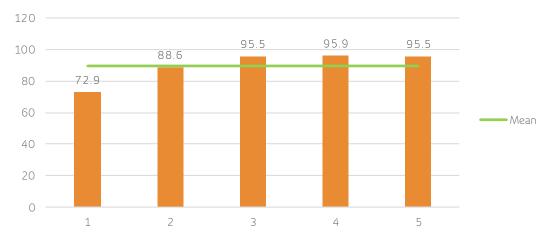
model

MODELS DEVELOPMENT

Cross Validation Score (Ada Boost Classifier)

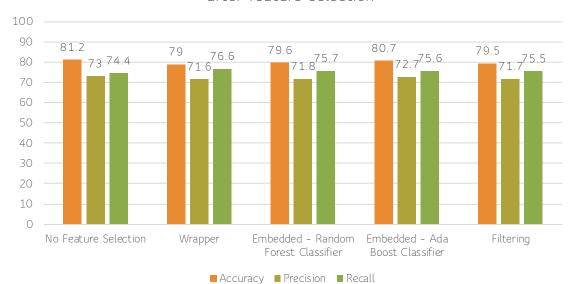


Cross Validation Score (Random Forest Classifier)

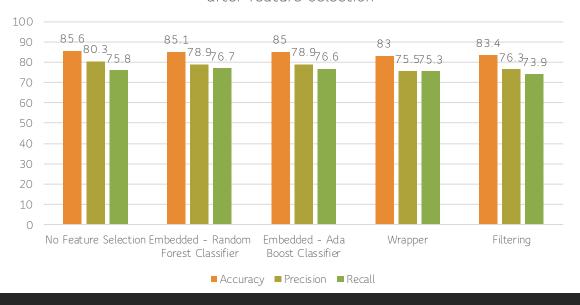


MODELS CROSS VALIDATION

Ada Boost Classifier after feature selection



Random Forest Classifier after feature selection



FEATURE SELECTION

Grid Search CV

The best parameters are:

- bootstrap : True
- max_depth : 4
- max_features : auto
- min_samples_leaf : 1
- min_samples_split : 2
- n_estimators: 72

Random Search CV

The best parameters are:

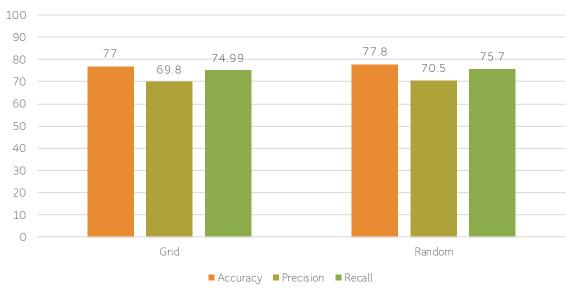
- bootstrap : False
- max_depth : 4
- max_features : auto
- min_samples_leaf : 1
- min_samples_split : 5
- n_estimators: 33

HYPER PARAMETER TUNING Random Forest Classifier

Feature selection:

Embedded Method - Random Forest Classifier

Random Forest Classifier after feature selection and hyper parameter tuning



RANDOM FOREST CLASSIFIER after feature selection and hyper parameter tuning

The best model to predict tomorrow will be rain or not is Random Forest Classifier after feature selecting using embedded method with accuracy, precision, and recall is 85.1%, 78.9%, 76.7% respectively.

Conclusion