

RATING AND PARTICIPATION PREDICTION

CPD ANALYSIS 2020

JAM CO.



JAFFRINA

197486@student.upm.edu.my

011-20922832



AMMAR

197710@student.upm.edu.my

010-9861415



FAIRUZ

198096@student.upm.edu.my

013-2069295

AGENDA

O1 DATA MINING
APPLICATION

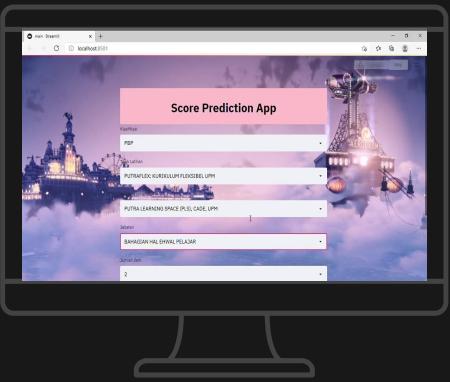
O2 DATA MINING FLOW

O3 COMPARISON

O4 REFLECTION

OI DATA MINING APPLICATION

DEMO



- Build web app by using **Streamlit**
- The function of this web app is to predict the score rating by participant based on the attributes in the dataset.
- **Stakeholders :** Staff & Management

IMPORTANCE TO STAKEHOLDER

MANAGEMENT

- Easy to predict the preferences of participants for the next event
- Collect data and analyze the most preferred activities to improve the rating and gain more participant
- Improvise facilities

PARTICIPANT

- Participants can easily choose what topic of activity that suitable for them to join.

02

DATA
MINING
FLOW

Problem Understanding

We find out the rating score given and number of participants from our CPD activities in 2020 does not achieve our target is High score rate and 10,000 participants. We only achieve Middle score while 5,437 participants. In 2021, we want to make sure the goal will be achieved.

Data Description

Dataset contains analysis of CPD used in computing rating and participation prediction in UPM. Prediction of the rating are based on the activities held. This dataset contains 1318 instances and 7 columns.

DATA PREPROCESSING

DATA CLEANING

**Filter and remove missing
data. Remove unimportant
columns.**

USING VLOOKUP

```
=VLOOKUP(value,table,col_index  
,[range_lookup])
```

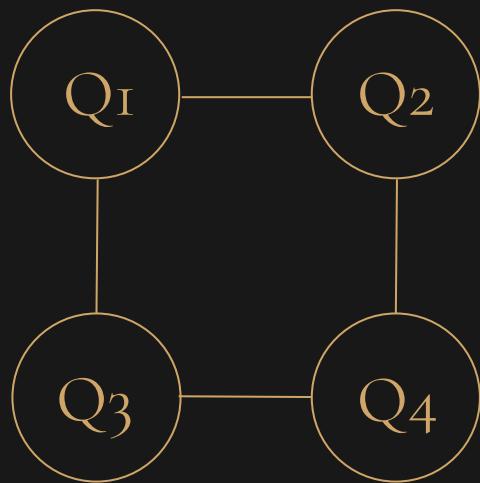
```
=VLOOKUP(value,table,col_index  
,[range_lookup])
```

QUESTIONS

Which classification
has the lowest score
penilaian?

What is the
relationship between
rating and
participants?

Does total hour of
course affecting the
grade score?



What is the total
participants of every
classifications?

Q₁

KAP has the lowest score rating.

Q₂

Yes. Higher total hours,
higher score given.

Q₃

Yes. Higher participants,
higher score given.

Q₄

CPD- 287
KAP- 94
OLCPD- 3666
PBP- 937
PicTL- 453

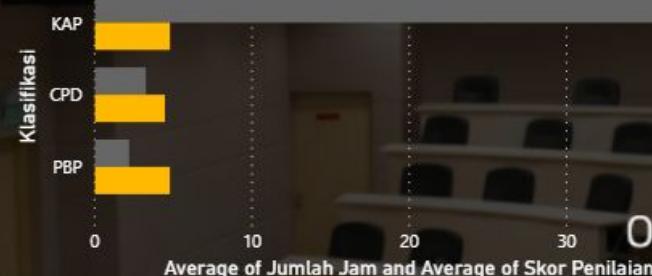
DASHBOARD OF CPD ANALYSIS

Skor Penilaian by Jabatan

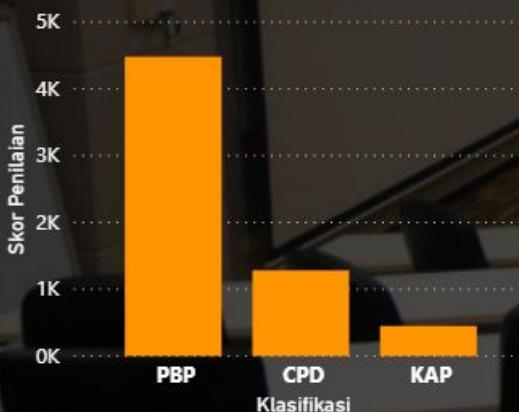


Average of Jumlah Jam and Average of Skor Penilaian by Klasifikasi

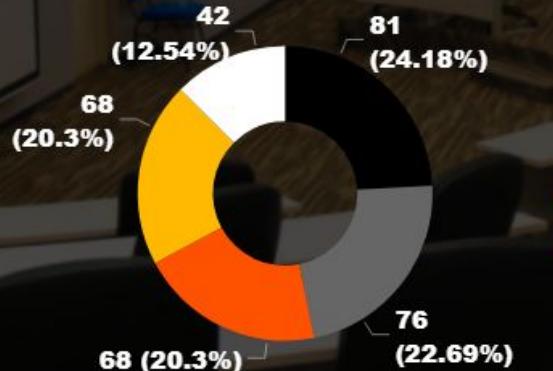
● Average of Jumlah Jam ● Average of Skor Penilaian



Skor Penilaian by Klasifikasi



Jumlah Jam by Tajuk Latihan and Tempat Latihan



Skor Penilaian by Jabatan and Klasifikasi



Klasifikasi	Average of Jumlah Jam	Average of Skor Penilaian
CPD	3.26	4.47
PBP	2.19	4.78
KAP	37.12	4.79
Total	4.91	4.71

Classify the score of the events into three classes
which are Low, Middle and High.

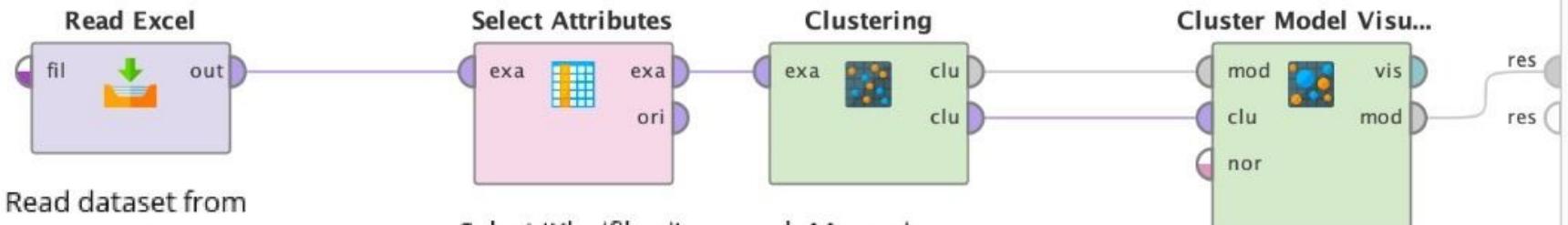
Skor Klasifikasi	Range
Low	Less than 4.377
Middle	$4.377 < \text{Score} < 4.772$
High	More than 4.772

	A	B	C	D	E	F	G
1	Klasifikasi	Tajuk Latihan	Tempat Latihan	Jumlah Jam	Jabatan	Skor Penilaian	Skor Klasifikasi
1						FAKULTI SAINS KOMPUTER DAN TEKNOLOGI	
2	PBP	WACANA - INSTRUC PLATFORM DALAM T.		3	MAKLUMAT	4.776	High
3	PBP	WACANA - INSTRUC PLATFORM DALAM T.		3	SEKOLAH PERNIAGAAN DAN EKONOMI	4.776	High
4	PBP	WACANA - INSTRUC PLATFORM DALAM T.		3	FAKULTI EKOLOGI MANUSA	4.776	High

DESCRIPTIVE DATA MINING

Objective : Cluster Klasifikasi and Tajuk Latihan according to the Skor Penilaian which are
High , Medium, Low

SET UP FROM RAPIDMINER- K-Means



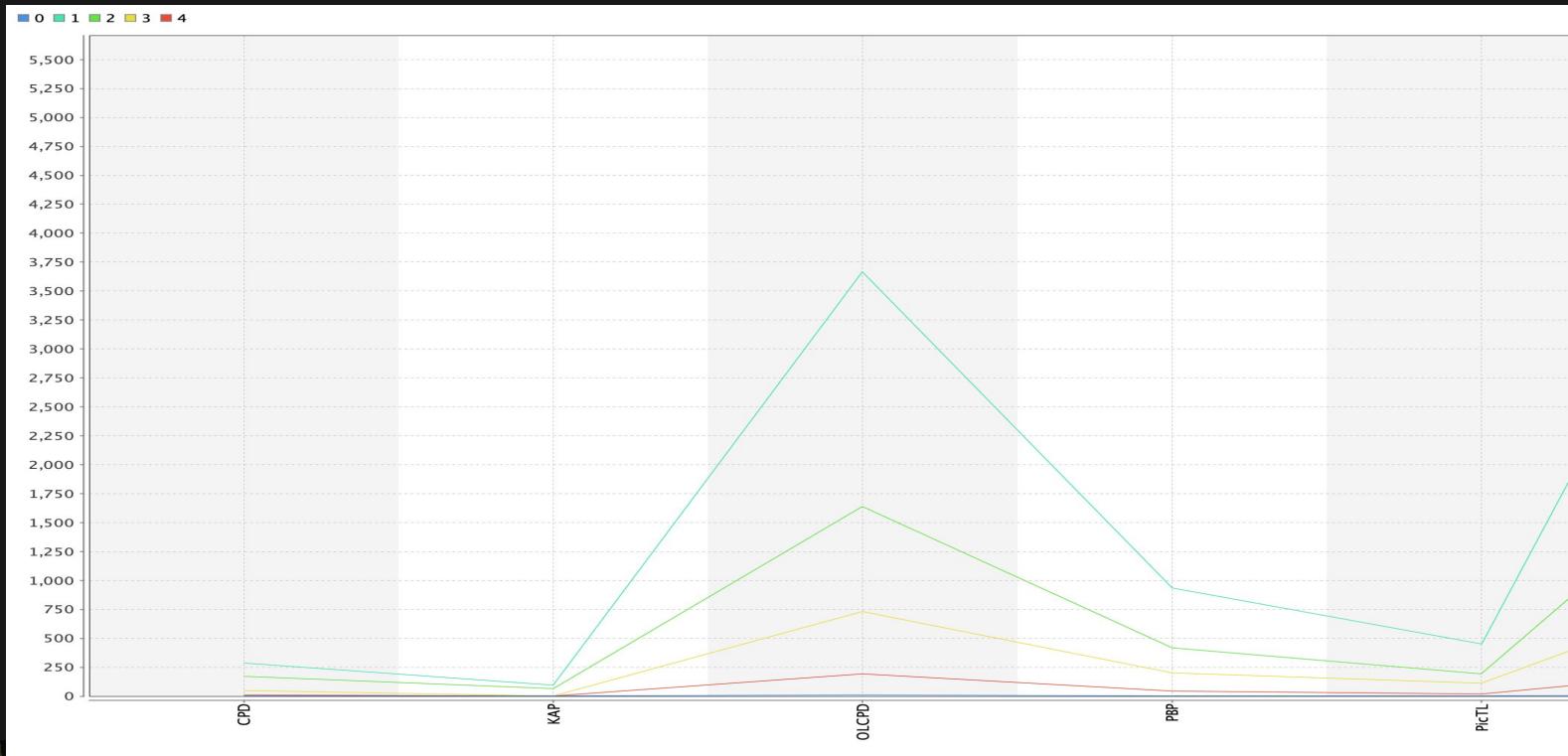
Read dataset from
Excel

Select 'Klasifikasi'
attributes to
count each
participants

k-Means is a
Clustering Models

To show the
visualization

K-means RapidMiner



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
CPD	0.620	287	171	48	9.250
KAP	0.300	94	65	1	3.250
OLCPD	10.520	3666	1641	730	192.250
PBP	2.840	937	417	202	44
PicTL	1.280	453	193	113	20.750
Grand Total	15.560	5437	2487	1094	269.500

Cluster Model

Cluster 0: 50 items

Cluster 1: 1 items

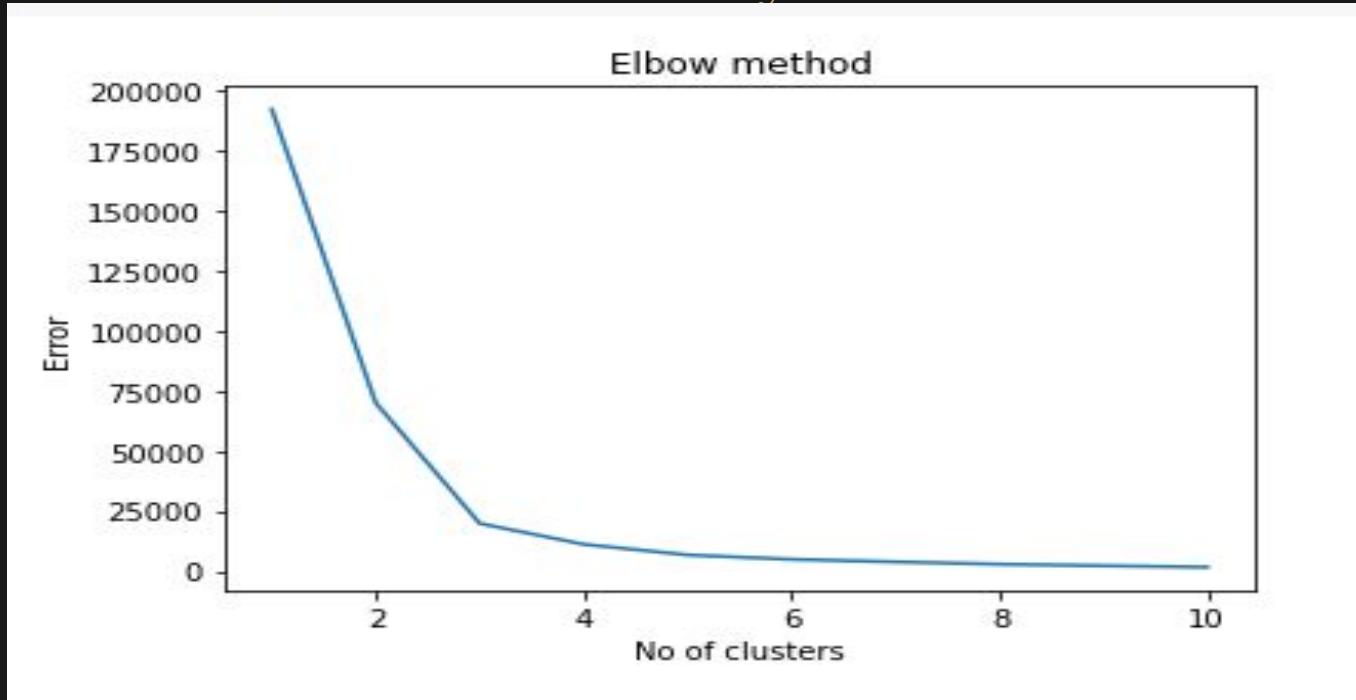
Cluster 2: 1 items

Cluster 3: 1 items

Cluster 4: 4 items

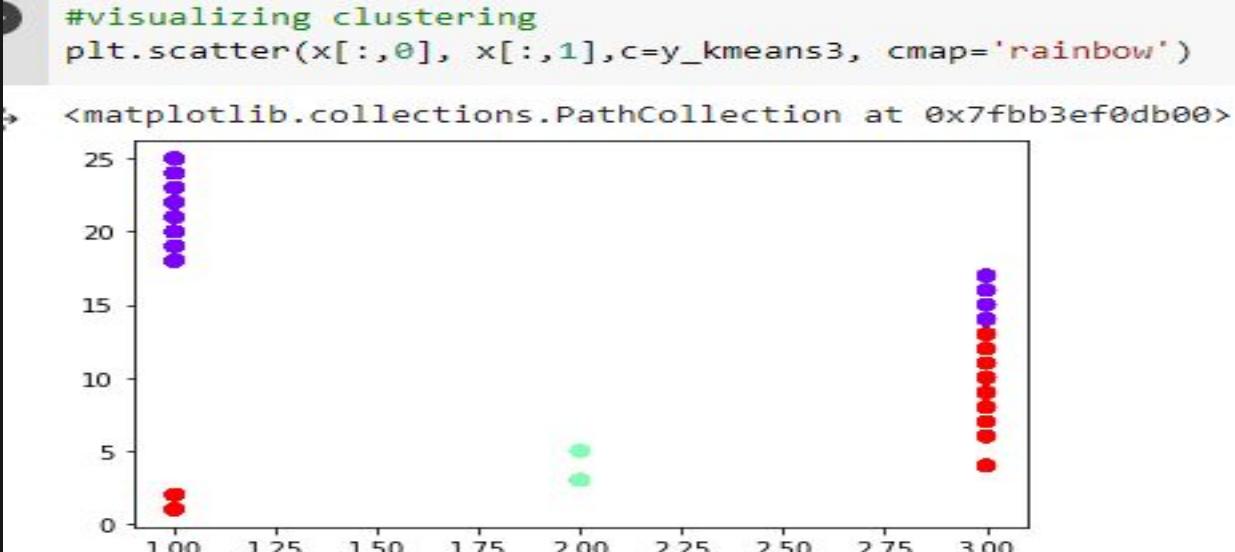
Total number of items: 57

K-means Python



BY USING ELBOW METHOD, WE GET THE $K = 3$ BECAUSE WE CAN SEE IN THE GRAPH AT 3 IT ALMOST LINEAR

K-means Python

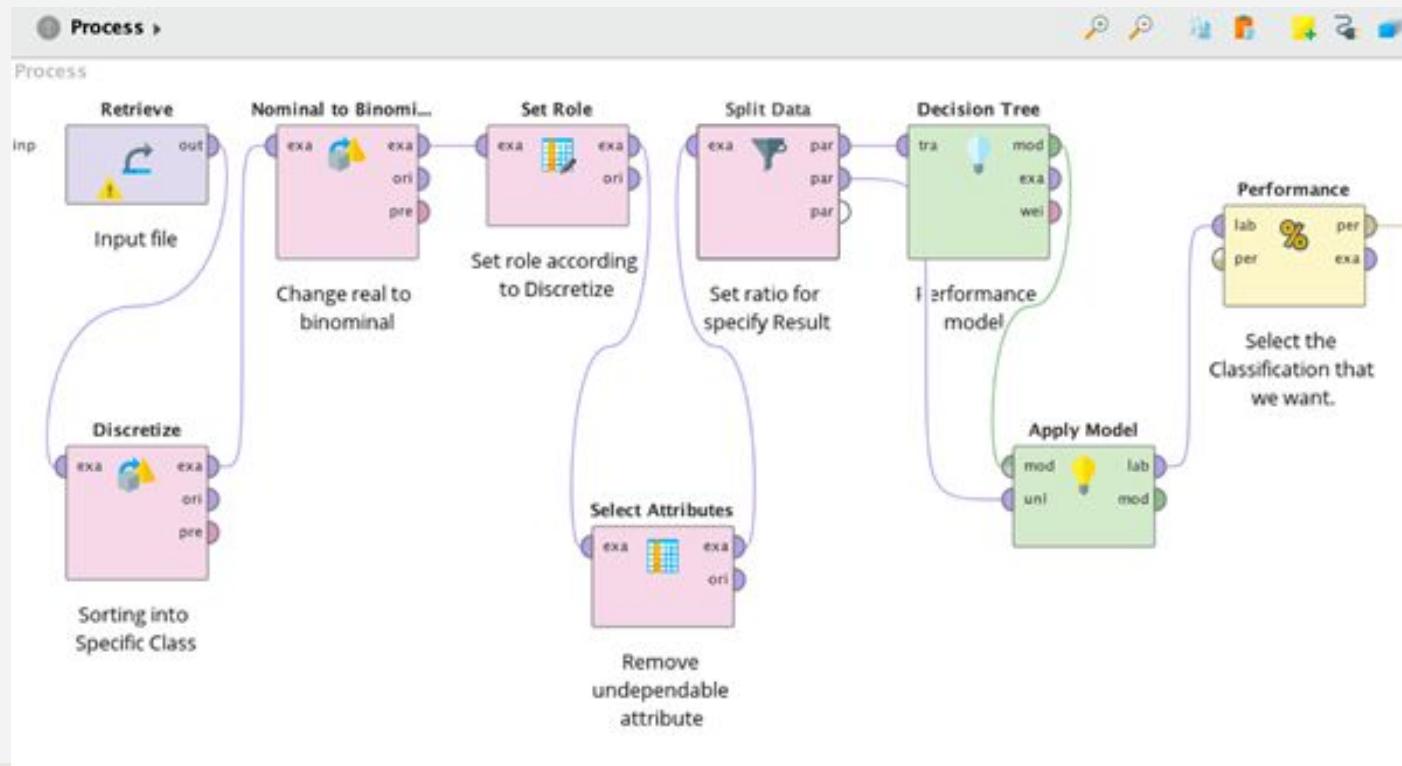


The three colours in the visualization show the Skor Klasifikasi which is High, Middle and Low

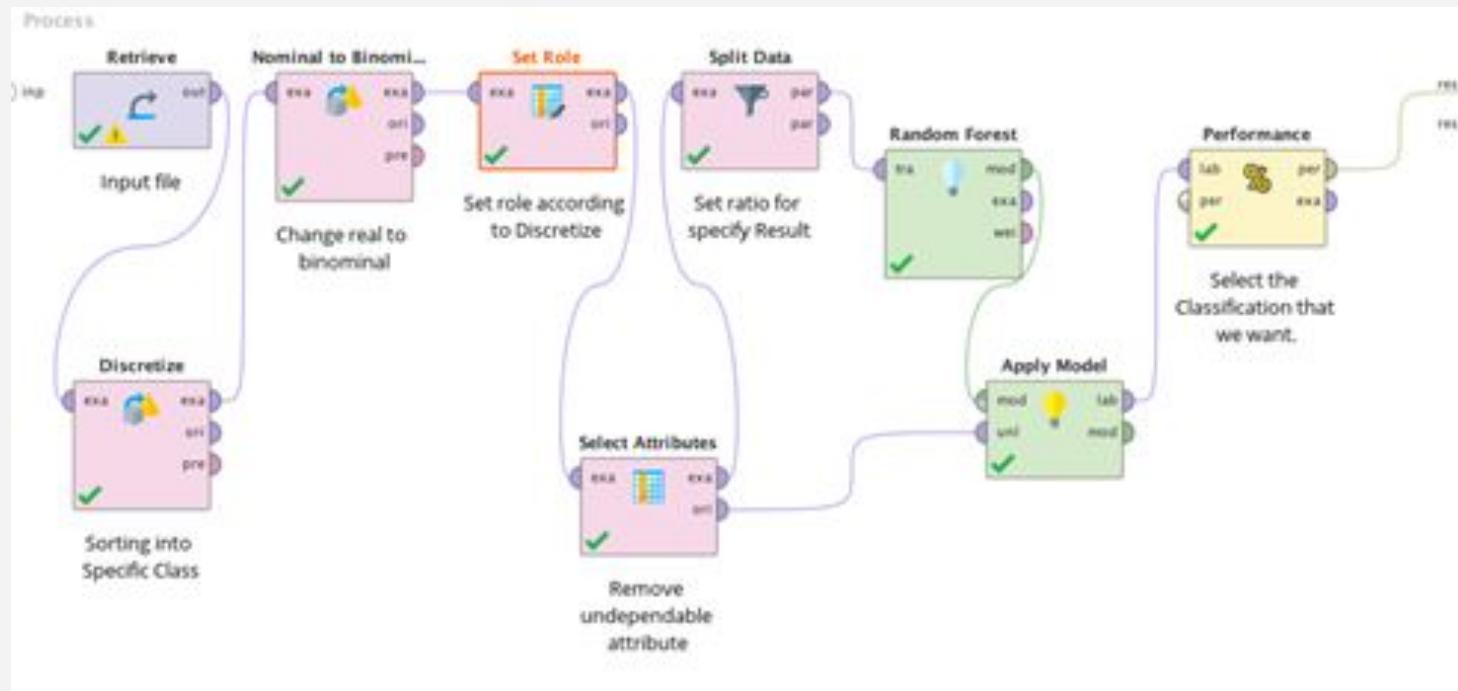
PREDICTIVE DATA MINING

Objective : Able to predict Skor Klasifikasi based on Tajuk Latihan and Klasifikasi

SET UP FROM RAPIDMINER- Decision Tree



INTERFACE FROM RAPIDMINER- Random Forest



Decision Tree - Python

```
[25] mod_dt = DecisionTreeClassifier(max_depth = 2, random_state = 1)
    mod_dt.fit(X_train,y_train)
    y_pred=mod_dt.predict(X_test)
    print('Decision Tree:')
    print("Accuracy = {:.2f}%".format(metrics.accuracy_score(y_pred,y_test)*100))
```

Decision Tree:
Accuracy = 89.27%

AFTER TUNING

BEFORE TUNING

```
[28] mod_dt = DecisionTreeClassifier(max_depth = 6, random_state = 1, criterion="gini")
    mod_dt.fit(X_train,y_train)
    y_pred=mod_dt.predict(X_test)
    print('Decision Tree:')
    print("Accuracy = {:.2f}%".format(metrics.accuracy_score(y_pred,y_test)*100))
```

Decision Tree:
Accuracy = 97.18%

Random Forest - Python

```
[56] from sklearn.ensemble import RandomForestClassifier  
RandomF = RandomForestClassifier(max_depth = 2, random_state = 1)  
RandomF.fit(X_train, y_train)  
y_pred = RandomF.predict(X_test)  
print('Random Forest Model:')  
print("Accuracy = {:.2f}%".format(metrics.accuracy_score(y_pred,y_test)*100))
```

Random Forest Model:
Accuracy = 95.34%

AFTER TUNING

BEFORE TUNING

```
mod_dt = RandomForestClassifier(max_depth = 4, random_state = 1, criterion="gini")  
mod_dt.fit(X_train,y_train)  
y_pred=mod_dt.predict(X_test)  
print('Random Forest Model:')  
print("Accuracy = {:.2f}%".format(metrics.accuracy_score(y_pred,y_test)*100))
```

#lets check the accuracy of the model

Random Forest Model:
Accuracy = 96.32%

03

COMPARISON

COMPARISON IN RAPIDMINER

70 : 30	Decision Tree		Random Forest	
	Accuracy	Classification Error	Accuracy	Classification Error
LOW	91.14%	8.86%	91.05%	8.95%
MIDDLE	83.04%	16.96%	82.85%	17.15%
HIGH	81.27%	18.73%	83.31%	16.69%
AVERAGE	85.15%	14.85%	85.74%	14.26%

COMPARISON IN RAPIDMINER

50 : 50	Decision Tree		Random Forest	
	Accuracy	Classification Error	Accuracy	Classification Error
LOW	91.05%	8.95%	91.05%	8.95%
MIDDLE	82.09%	17.91%	82.09%	17.91%
HIGH	83.31%	16.69%	83.31%	16.69%
AVERAGE	85.43%	14.51%	85.43%	14.51%

COMPARISON IN PYTHON

	Model	Decision Tree	Random Forest
70 : 30	Before Tuning	89.27%	95.34%
	After Tuning	97.18%	96.42%
50 : 50	Before Tuning	89.38%	93.93%
	After Tuning	96.66%	96.51%

INSIGHT

Cluster	Metric	Quantity
High	Klasifikasi	940
	Tajuk Latihan	940
	Tempat Latihan	940
	Jumlah Jam	5428
	Jabatan	940

INSIGHT

Cluster	Metric	Quantity
Middle	Klasifikasi	260
	Tajuk Latihan	280
	Tempat Latihan	280
	Jumlah Jam	691
	Jabatan	260

INSIGHT

Cluster	Metric	Quantity
Low	Klasifikasi	118
	Tajuk Latihan	118
	Tempat Latihan	118
	Jumlah Jam	354
	Jabatan	118

CONCLUSION

- ❑ All the model show increase in accuracy after hyperparameter tuning
- ❑ The best model goes to Decision Tree in predictive data mining with the accuracy of 97.18%
- ❑ We get accuracy near 100% because the dataset is overfit and easy to predict

04

REFLECTION

REFLECTIONS

- Learn to visualize our hypothesis into interactive visualization
- Experienced in using software tool : PowerBI, Tableau, RapidMiner. Google Colaboratory-Python
- Exploratory Data Analysis is a vital step in a data science project. The main pillars of EDA are data cleaning, data preparation, data exploration, and data visualization.
- Learn the scope of work of data scientist
- Such an incredible journey throughout the one semester learning Data Mining with Dr Fadh and do all the tasks and project
- Thank you Dr Fadh for all the knowledge and effort in teaching us



Group Members

OI Ammar Rozman bin
Mohd Azman
197710

O2 Jaffrina Umaira binti Jafni
197486

O3 Fairuz Syamimi binti
Ehsanuddin
198096



A woman with long dark hair is holding four large, shiny gold-colored foil balloons. The balloons are shaped to spell out the words "THANK YOU". The letters are partially inflated, creating a 3D effect. The background is dark, making the gold balloons stand out. The woman's hands are visible at the bottom, holding the balloons. The overall composition is a close-up shot focusing on the balloons and the text they form.

THANK YOU