

LAPORAN
RENCANA TUGAS MAHASISWA (RTM) Ke-5
MATA KULIAH BIG DATA
“AUTOMATED SCORING SYSTEM MENGGUNAKAN PYSPARK”



DISUSUN OLEH:

Rahmah Lidya Nastiti (22083010102)

DOSEN PENGAMPU:

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

PROGRAM STUDI SAINS DATA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR

2024

- Menginstal PySpark

```
!pip install pyspark
!pip install install-jdk

Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
    317.0/317.0 MB 3.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=0887fb97dab832000e702c39e
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38ddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
Collecting install-jdk
  Downloading install_jdk-1.1.0-py3-none-any.whl (15 kB)
Installing collected packages: install-jdk
Successfully installed install-jdk-1.1.0
```

- Mengimport Spark Sql, library machine learning, dan membuat session

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.ml.feature import HashingTF, Tokenizer, StopWordsRemover, IDF
from pyspark.ml import Pipeline
from pyspark.ml.recommendation import ALS
from pyspark.ml.evaluation import RegressionEvaluator

#membuat session
appName = 'Automated Scoring Using pySpark'
spark = SparkSession \
    .builder \
    .appName(appName) \
    .config('spark.some.config.option','some-value') \
    .getOrCreate()
```

- Membuat data dari file

```
import pandas as pd
auto_score = pd.read_excel("/content/drive/MyDrive/training_data_essay.xlsx")
auto_score
```

	npm	nama_peserta		jawaban	soal	skor_per_soal
0	0	Admin	Tidak, Hanya membutuhkan satu karena satu soft...		1	100.00
1	0	Admin	Biaya dihitung berdasarkan waktu pengerjaan da...		2	100.00
2	0	Admin	Hak cipta adalah hak eksklusif bagi pencipta a...		3	100.00
3	0	Admin	Dijelaskan kepada klien jika huruf terlalu bes...		4	100.00
4	0	Admin	1. Melindungi dan menjamin pekerja atas hak ke...		5	100.00
...
115	1121020035	BS	posisi kepala dan leher harus tegak lurus deng...		8	48.41
116	1121020035	BS	kecocokan software camera dengan system hardwa...		9	65.89
117	1121020035	BS	okus dan exposure dipertimbangkan dalam operas...		10	48.51
118	1121020035	BS	peralatan yang dibutuhkan untuk transfer image...		11	78.41
119	1121020035	BS	dibuat grafik yang menggunakan prinsip desain ...		12	64.29

120 rows × 7 columns

```
df = spark.createDataFrame(auto_score)
df = df.select('soal', 'jawaban', 'skor_per_soal')
df.show()
```

```
+-----+-----+-----+
|soal|          jawaban|skor_per_soal|
+-----+-----+-----+
| 1|Tidak, Hanya memb...|      100.0|
| 2|Biaya dihitung be...|      100.0|
| 3|Hak cipta adalah ...|      100.0|
| 4|Dijelaskan kepada...|      100.0|
| 5|1. Melindungi dan...|      100.0|
| 6|Ruang Komputer, P...|      100.0|
| 7|Aturlah posisi pe...|      100.0|
| 8|Posisi Kepala dan...|      100.0|
| 9|1. Kecocokan soft...|      100.0|
|10|1. Fokus dan expo...|      100.0|
|11|1. Peralatan yang...|      100.0|
|12|1. Dibuat grafik ...|      100.0|
| 1|tidak, cuma mengi...|       52.7|
| 2|biaya dihitung be...|      42.86|
| 3|hak membuat merup...|      42.16|
| 4|dipaparkan pada k...|      27.19|
| 5|1. mencegah serta...|      44.14|
| 6|ruang komputer, p...|      100.0|
| 7|aturlah posisi fi...|      57.68|
| 8|posisi kepala ser...|      45.71|
+-----+-----+-----+
```

only showing top 20 rows

```
regex_pattern = '^[a-zA-Z\\s]+|\\d+'
cleanningTrain = df.withColumn('jawaban_v2', regexp_replace('jawaban', regex_pattern, ''))
cleanningTrain.show()
```

```
+-----+-----+-----+-----+
|soal|          jawaban|skor_per_soal|          jawaban_v2|
+-----+-----+-----+-----+
| 1|Tidak, Hanya memb...|      100.0|Tidak Hanya membu...|
| 2|Biaya dihitung be...|      100.0|Biaya dihitung be...|
| 3|Hak cipta adalah ...|      100.0|Hak cipta adalah ...|
| 4|Dijelaskan kepada...|      100.0|Dijelaskan kepada...|
| 5|1. Melindungi dan...|      100.0|Melindungi dan m...|
| 6|Ruang Komputer, P...|      100.0|Ruang Komputer Po...|
| 7|Aturlah posisi pe...|      100.0|Aturlah posisi pe...|
| 8|Posisi Kepala dan...|      100.0|Posisi Kepala dan...|
| 9|1. Kecocokan soft...|      100.0|Kecocokan softwa...|
|10|1. Fokus dan expo...|      100.0|Fokus dan exposu...|
|11|1. Peralatan yang...|      100.0|Peralatan yang d...|
|12|1. Dibuat grafik ...|      100.0|Dibuat grafik ya...|
| 1|tidak, cuma mengi...|       52.7|tidak cuma mengin...|
| 2|biaya dihitung be...|      42.86|biaya dihitung be...|
| 3|hak membuat merup...|      42.16|hak membuat merup...|
| 4|dipaparkan pada k...|      27.19|dipaparkan pada k...|
| 5|1. mencegah serta...|      44.14|mencegah serta m...|
| 6|ruang komputer, p...|      100.0|ruang komputer po...|
| 7|aturlah posisi fi...|      57.68|aturlah posisi fi...|
| 8|posisi kepala ser...|      45.71|posisi kepala ser...|
+-----+-----+-----+-----+
```

only showing top 20 rows

kode diatas untuk membersihkan teks dalam kolom 'jawaban' dengan menghapus semua karakter yang tidak diinginkan (non-huruf dan digit) dan menyimpan hasil yang telah dibersihkan dalam kolom baru 'jawaban_v2'.

- Menyiapkan data training

```

) tokenizer = Tokenizer(inputCol = 'jawaban_v2', outputCol = 'Tokenize_jawaban')
tokenizedTrain = tokenizer.transform(cleanningTrain)
tokenizedTrain.show()

```

```

+-----+-----+-----+-----+-----+
|soal|      jawaban|skor_per_soal|      jawaban_v2|      Tokenize_jawaban|
+-----+-----+-----+-----+-----+
| 1|Tidak, Hanya memb...|      100.0|Tidak Hanya membu...|[tidak, hanya, me...|
| 2|Biaya dihitung be...|      100.0|Biaya dihitung be...|[biaya, dihitung,...|
| 3|Hak cipta adalah ...|      100.0|Hak cipta adalah ...|[hak, cipta, adal...|
| 4|Dijelaskan kepada...|      100.0|Dijelaskan kepada...|[dijelaskan, kepa...|
| 5|1. Melindungi dan...|      100.0|Melindungi dan m...|[, melindungi, da...|
| 6|Ruang Komputer, P...|      100.0|Ruang Komputer Po...|[ruang, komputer,...|
| 7|Aturlah posisi pe...|      100.0|Aturlah posisi pe...|[aturlah, posisi,...|
| 8|Posisi Kepala dan...|      100.0|Posisi Kepala dan...|[posisi, kepala, ...|
| 9|1. Kecocokan soft...|      100.0|Kecocokan softwa...|[, kecocokan, sof...|
|10|1. Fokus dan expo...|      100.0|Fokus dan exposu...|[, fokus, dan, ex...|
|11|1. Peralatan yang...|      100.0|Peralatan yang d...|[, peralatan, yan...|
|12|1. Dibuat grafik ...|      100.0|Dibuat grafik ya...|[, dibuat, grafik...|
| 1|tidak, cuma mengi...|      52.7|tidak cuma mengin...|[tidak, cuma, men...|
| 2|biaya dihitung be...|      42.86|biaya dihitung be...|[biaya, dihitung,...|
| 3|hak membuat merup...|      42.16|hak membuat merup...|[hak, membuat, me...|
| 4|dipaparkan pada k...|      27.19|dipaparkan pada k...|[dipaparkan, pada...|
| 5|1. mencegah serta...|      44.14|mencegah serta m...|[, mencegah, sert...|
| 6|ruang komputer, p...|      100.0|ruang komputer po...|[ruang, komputer,...|
| 7|aturlah posisi fi...|      57.68|aturlah posisi fi...|[aturlah, posisi,...|
| 8|posisi kepala ser...|      45.71|posisi kepala ser...|[posisi, kepala, ...|
+-----+-----+-----+-----+-----+

```

only showing top 20 rows

kode diatas untuk memecah teks yang ada di kolom 'jawaban_v2' menjadi kata-kata individu dan menyimpannya dalam kolom baru 'Tokenize_jawaban'.

```

) from pyspark.ml.feature import StopWordsRemover

# Read the stopwords from the file
with open('/content/drive/MyDrive/id.stopwords.02.01.2016.txt') as f:
    stopwords = f.read().splitlines()

# Initialize the StopWordsRemover without the stopwords
stopword_remover = StopWordsRemover(inputCol='Tokenize_jawaban', outputCol='kata_bermakna')

# Set the custom stopwords
stopword_remover = stopword_remover.setStopWords(stopwords)

# Apply the stopword remover to the DataFrame
SwRemovedTrain = stopword_remover.transform(tokenizedTrain)
SwRemovedTrain.show()

```

```

+-----+-----+-----+-----+-----+-----+
|soal|      jawaban|skor_per_soal|      jawaban_v2|      Tokenize_jawaban|      kata_bermakna|
+-----+-----+-----+-----+-----+-----+
| 1|Tidak, Hanya memb...|      100.0|Tidak Hanya membu...|[tidak, hanya, me...|[membutuhkan, sof...|
| 2|Biaya dihitung be...|      100.0|Biaya dihitung be...|[biaya, dihitung,...|[biaya, dihitung,...|
| 3|Hak cipta adalah ...|      100.0|Hak cipta adalah ...|[hak, cipta, adal...|[hak, cipta, hak,...|
| 4|Dijelaskan kepada...|      100.0|Dijelaskan kepada...|[dijelaskan, kepa...|[klien, huruf, me...|
| 5|1. Melindungi dan...|      100.0|Melindungi dan m...|[, melindungi, da...|[, melindungi, me...|
| 6|Ruang Komputer, P...|      100.0|Ruang Komputer Po...|[ruang, komputer,...|[ruang, komputer,...|
| 7|Aturlah posisi pe...|      100.0|Aturlah posisi pe...|[aturlah, posisi,...|[aturlah, posisi,...|
| 8|Posisi Kepala dan...|      100.0|Posisi Kepala dan...|[posisi, kepala, ...|[posisi, kepala, ...|
| 9|1. Kecocokan soft...|      100.0|Kecocokan softwa...|[, kecocokan, sof...|[, kecocokan, sof...|
|10|1. Fokus dan expo...|      100.0|Fokus dan exposu...|[, fokus, dan, ex...|[, fokus, exposur...|
|11|1. Peralatan yang...|      100.0|Peralatan yang d...|[, peralatan, yan...|[, peralatan, dib...|
|12|1. Dibuat grafik ...|      100.0|Dibuat grafik ya...|[, dibuat, grafik...|[, grafik, prinsi...|
| 1|tidak, cuma mengi...|      52.7|tidak cuma mengin...|[tidak, cuma, men...|[analikasi, cocok...|
+-----+-----+-----+-----+-----+-----+

```

kode diatas untuk membersihkan teks yang sudah ditokenisasi dengan menghapus kata-kata yang tidak memiliki makna (stopwords) dan menyimpan hasil yang telah dibersihkan dalam kolom baru 'kata_bermakna'

```
hash = SwRemovedTrain.withColumn('hash_value', hash('kata_bermakna'))
hash.show()
```

soal	jawaban	skor_per_soal	jawaban_v2	Tokenize_jawaban	kata_bermakna	hash_value
1 Tidak, Hanya memb...	100.0	Tidak Hanya membu...	[tidak, hanya, me...	[membutuhkan, sof...		
2 Biaya dihitung be...	100.0	Biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
3 Hak cipta adalah ...	100.0	Hak cipta adalah ...	[hak, cipta, adal...	[hak, cipta, hak,...		
4 Dijelaskan kepada ...	100.0	Dijelaskan kepada...	[dijelaskan, kepa...	[klien, huruf, me...		
5 1. Melindungi dan...	100.0	Melindungi dan m...	[, melindungi, da...	[, melindungi, me...		
6 Ruang komputer, P...	100.0	Ruang Komputer Po...	[ruang, komputer,...	[ruang, komputer,...		
7 Aturlah posisi pe...	100.0	Aturlah posisi pe...	[aturlah, posisi,...	[aturlah, posisi,...		
8 Posisi Kepala dan...	100.0	Posisi Kepala dan...	[posisi, kepala, ...	[posisi, kepala, ...		
9 1. Kecocokan soft...	100.0	Kecocokan softwa...	[, kecocokan, sof...	[, kecocokan, sof...		
10 1. Fokus dan expo...	100.0	Fokus dan exposu...	[, fokus, dan, ex...	[, fokus, exposur...		
11 1. Peralatan yang...	100.0	Peralatan yang d...	[, peralatan, yan...	[, peralatan, dib...		
12 1. Dibuat grafik ...	100.0	Dibuat grafik ya...	[, dibuat, grafik...	[, grafik, prinsi...		
1 tidak, cuma mengi...	52.7	tidak cuma mengin...	[tidak, cuma, men...	[aplikasi, cocok,...		
2 biaya dihitung be...	42.86	biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
3 hak membuat merup...	42.16	hak membuat merup...	[hak, membuat, me...	[hak, hak, khusus...		
4 dipaparkan pada k...	27.19	dipaparkan pada k...	[dipaparkan, pada...	[dipaparkan, kons...		
5 1. mencegah serta...	44.14	mencegah serta m...	[, mencegah, sert...	[, mencegah, menj...		
6 ruang komputer, p...	100.0	ruang komputer po...	[ruang, komputer,...	[ruang, komputer,...		
7 aturlah posisi fi...	57.68	aturlah posisi fi...	[aturlah, posisi,...	[aturlah, posisi,...		
8 posisi kepala ser...	45.71	posisi kepala ser...	[posisi, kepala, ...	[posisi, kepala, ...		

only showing top 20 rows

Kode di atas digunakan untuk menambahkan kolom baru bernama 'hash_value'. kolom 'hash_value' berisi nilai hash yang dihitung dari kolom 'kata_bermakna'.

- Memisahkan data training dan data testing

```
✓ [9] split = hash.randomSplit([0.7, 0.3])
os train = split[0].withColumnRenamed('skor_per_soal', 'skor')
test = split[1].withColumnRenamed('skor_per_soal', 'true_skor')
```

```
✓ [10] row_train = train.count()
ss row_test = test.count()
print('data training', row_train)
print('data testing', row_test)
```

```
data training 82
data testing 38
```

Kode di atas untuk membagi DataFrame yang telah diproses (hash) menjadi dua bagian yaitu data training dan data testing. Selain itu, kolom 'skor_per_soal' dalam masing-masing DataFrame diubah namanya untuk membedakan antara skor sebenarnya dan skor yang diprediksi nanti.

```
) train.show()
```

soal	jawaban	skor	jawaban_v2	Tokenize_jawaban	kata_bermakna	hash_value
1 tidak, cuma mengi...	52.7	tidak cuma mengin...	[tidak, cuma, men...	[aplikasi, cocok,...		
1 tidak, hanya memb...	100.0	tidak hanya membu...	[tidak, hanya, me...	[membutuhkan, sof...		
1 tidak, hanya memb...	100.0	tidak hanya membu...	[tidak, hanya, me...	[membutuhkan, sof...		
1 tidak, hanya memb...	100.0	tidak hanya membu...	[tidak, hanya, me...	[membutuhkan, sof...		
2 Biaya dihitung be...	100.0	Biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
2 biaya dihitung be...	100.0	biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
2 biaya dihitung be...	100.0	biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
2 biaya dihitung be...	42.86	biaya dihitung be...	[biaya, dihitung,...	[biaya, dihitung,...		
3 Hak cipta adalah ...	100.0	Hak cipta adalah ...	[hak, cipta, adal...	[hak, cipta, hak,...		
3 hak cipta adalah ...	91.71	hak cipta adalah ...	[hak, cipta, adal...	[hak, cipta, hak,...		
3 hak cipta adalah ...	91.71	hak cipta adalah ...	[hak, cipta, adal...	[hak, cipta, hak,...		
3 hak membuat merup...	42.16	hak membuat merup...	[hak, membuat, me...	[hak, hak, khusus...		
4 Dijelaskan kepada...	100.0	Dijelaskan kepada...	[dijelaskan, kepa...	[klien, huruf, me...		
4 dijelaskan kepada...	72.06	dijelaskan kepada...	[dijelaskan, kepa...	[klien, huruf, me...		
4 dipaparkan pada k...	27.19	dipaparkan pada k...	[dipaparkan, pada...	[dipaparkan, kons...		
5 1. Melindungi dan...	100.0	Melindungi dan m...	[, melindungi, da...	[, melindungi, me...		
5 melindungi dan me...	74.73	melindungi dan me...	[melindungi, dan,...	[melindungi, menj...		
5 melindungi dan me...	74.73	melindungi dan me...	[melindungi, dan,...	[melindungi, menj...		
6 ruang komputer, p...	100.0	ruang komputer po...	[ruang, komputer,...	[ruang, komputer,...		
6 ruang komputer, p...	100.0	ruang komputer po...	[ruang, komputer,...	[ruang, komputer,...		

only showing top 20 rows

```
test.show()
```

soal	jawaban	true_skor	jawaban_v2	Tokenize_jawaban	kata_bermakna	hash_value
1	Tidak, Hanya memb...	100.0	Tidak Hanya membu...	[[tidak, hanya, me...	[membutuhkan, sof...	-1678823266]
2	biaya dihitung be...	100.0	biaya dihitung be...	[[biaya, dihitung,...	[biaya, dihitung,...	868079823]
3	hak cipta adalah ...	91.71	hak cipta adalah ...	[[hak, cipta, adal...	[hak, cipta, hak,...	-348474399]
4	dijelaskan kepada...	72.06	dijelaskan kepada...	[[dijelaskan, kepa...	[klien, huruf, me...	636521394]
4	dijelaskan kepada...	72.06	dijelaskan kepada...	[[dijelaskan, kepa...	[klien, huruf, me...	636521394]
5	1. mencegah serta...	44.14	mencegah serta m...	[[, mencegah, sert...	[, mencegah, menj...	1658505624]
5	melindungi dan me...	90.27	melindungi dan me...	[[melindungi, dan,...	[melindungi, menj...	-1211017882]
6	Ruang Komputer, P...	100.0	Ruang Komputer Po...	[[ruang, komputer,...	[ruang, komputer,...	829849610]
6	ruang komputer, p...	100.0	ruang komputer po...	[[ruang, komputer,...	[ruang, komputer,...	829849610]
6	ruang komputer, p...	100.0	ruang komputer po...	[[ruang, komputer,...	[ruang, komputer,...	829849610]
7	aturlah posisi pe...	86.22	aturlah posisi pe...	[[aturlah, posisi,...	[aturlah, posisi,...	-1894676254]
7	aturlah posisi pe...	86.22	aturlah posisi pe...	[[aturlah, posisi,...	[aturlah, posisi,...	-1894676254]
9	1. Kecocokan soft...	100.0	Kecocokan softwa...	[[, kecocokan, sof...	[, kecocokan, sof...	-2081253499]
9	kecocokan softwar...	65.89	kecocokan softwa...	[[kecocokan, softw...	[kecocokan, softw...	1900032256]
10	fokus dan exposur...	100.0	fokus dan exposur...	[[fokus, dan, expo...	[fokus, exposure,...	-2119091795]
11	1. Peralatan yang...	100.0	Peralatan yang d...	[[, peralatan, yan...	[, peralatan, dib...	-521292198]
11	peralatan yang di...	77.63	peralatan yang di...	[[peralatan, yang,...	[peralatan, dibut...	-1374699892]
11	peralatan yang di...	100.0	peralatan yang di...	[[peralatan, yang,...	[peralatan, dibut...	-383074426]
12	1. Dibuat grafik ...	100.0	Dibuat grafik ya...	[[, dibuat, grafik...	[, grafik, prinsi...	481493723]
1	tidak, hanya memb...	100.0	tidak hanya membu...	[[tidak, hanya, me...	[membutuhkan, sof...	-1678823266]

only showing top 20 rows

- Mendefinisikan model dan mentrainingnya

```
13] als = ALS(maxIter = 5, regParam = 0.01, userCol = 'hash_value', itemCol = 'soal', ratingCol = 'skor')
model = als.fit(train)
```

Kode di atas untuk membuat dan melatih model rekomendasi menggunakan algoritma Alternating Least Squares (ALS).

- Melakukan prediksi dengan model rekomenr yang telah di training

```
14] prediction = model.transform(test)
```

```
prediction.select('soal', 'hash_value', 'true_skor', 'prediction').show()
```

```
15] +-----+
|soal| hash_value|true_skor|prediction|
+-----+
| 12| -1557371155| 64.29| 64.28996|
| 12| 481493723| 100.0| NaN|
| 1| -1678823266| 100.0| 99.99991|
| 1| -1678823266| 100.0| 99.99991|
| 6| 829849610| 100.0| 99.99992|
| 3| -348474399| 91.71| 91.70997|
| 5| -1211017882| 90.27| NaN|
| 9| -2081253499| 100.0| NaN|
| 9| -720994262| 84.88| 84.87993|
| 4| 636521394| 72.06| 72.63663|
| 4| 636521394| 72.06| 72.63663|
| 4| 636521394| 72.06| 72.63663|
| 8| -1861440661| 100.0| 99.99996|
| 7| -1894676254| 86.22| 86.21996|
| 7| -1894676254| 86.22| 86.21996|
| 7| -1894676254| 86.22| 86.21996|
| 7| 624824635| 59.87| NaN|
| 10| -2119091795| 100.0| NaN|
| 11| -1374699892| 77.63| NaN|
| 11| -521292198| 100.0| NaN|
+-----+
only showing top 20 rows
```

kode diatas untuk melakukan prediksi pada data uji, dan menampilkan beberapa kolom penting dari hasil prediksi tersebut, yaitu 'soal', 'hash_value', 'true_skor', dan 'prediction'.

- Mengevaluasi seberapa akurat system rekomenr

```

16] evaluator = RegressionEvaluator(metricName='rmse', labelCol='true_skor', predictionCol='prediction')
rmse = evaluator.evaluate(prediction)
print('Root Mean Square Error (RMSE):', rmse)

```

```

Root Mean Square Error (RMSE): nan

```

```

evaluator = RegressionEvaluator(metricName='rmse', labelCol='true_skor', predictionCol='prediction')

# Evaluate the predictions
mse = evaluator.evaluate(prediction)

# Print the RMSE
print(' Mean Square Error (RMSE):', mse)

```

```

Mean Square Error (RMSE): nan

```

kode diatas untuk menghitung dan mencetak Root Mean Squared Error (RMSE) dan Mean Squared Error (MSE) dari model regresi berdasarkan prediksinya terhadap data.

```

18] prediction.count()
a = prediction.count()
print('sebelum missing value dihapus:', a, 'baris')
cleanPred = prediction.dropna(how = 'any', subset = ['prediction'])
b = cleanPred.count()
print('setelah missing value dihapus:', b, 'baris')
print('missing value:', a-b, 'baris')

```

```

sebelum missing value dihapus: 38 baris
setelah missing value dihapus: 26 baris
missing value: 12 baris

```

kode diatas untuk menghitung jumlah baris dalam dataframe, menghapus baris yang memiliki nilai yang hilang dalam kolom 'prediction'.

```

9] mse = evaluator.evaluate(cleanPred)

# Print the MSE
print('Mean Squared Error (MSE):', mse)

```

```

Mean Squared Error (MSE): 0.19587052775507344

```

```

0] rmse = evaluator.evaluate(cleanPred)
print('root mean squard error:', rmse)

```

```

root mean squard error: 0.19587052775507344

```

Kode di atas menghitung Mean Squared Error (MSE) dan Root ean quard rror (RMSE) dari model evaluasi data yang telah dibersihkan dari nilai yang hilang (cleanPred).