LAPORAN

RENCANA TUGAS MANDIRI (RTM) Ke-5

MATA KULIAH BIG DATA B

"MEMBUAT AUTOMATED SCORING SYSTEM MENGGUNAKAN PySPARK"

DISUSUN OLEH:

Angela Lisanthoni (21083010032)

DOSEN PENGAMPU:

Tresna Maulana Fahrudin S.ST., M.T.

Kartika Maulida Hindrayani, S.Kom, M.Kom

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN" JAWA TIMUR

2023

A. Membuat Session Baru

1. Import Modul yang dibutuhkan

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.ml.recommendation import ALS
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.sql.functions import hash, abs
```

Angela_Big Data B

Berikut adalah modul yang dibutuhkan dalam tugas kali ini diantaranya:

- SparkSession → digunakan untuk menginisialisasi lingkungan spark sehingga dapat melakukan pemrosesan data
- pyspark.sql.types → digunakan untuk mendefinisikan struktur skema data yang digunakan dalam pemrosesan data
- ALS → digunakan untuk membangkitkan algoritma ALS yang digunakan dalam model yang akan dibuat
- RegressionEvaluator → digunakan untuk membangkitkan evaluasi kualitas model regresi yang telah dilatih
- hash → digunakan untuk menghasilkan nilai hash dari suatu kolom dataset. Nilai hash adalah representasi numerik yang unik untuk tiap objek
- abs → digunakan untuk menghitung nilai absolut dari suatu kolom dalam dataset.

2. Membuat Session Baru

```
appName = "Sistem Penskoran Otomatis pada Soal Essay 2"
spark = SparkSession \
    .builder \
    .appName(appName) \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

Angela_Big Data B

```
spark
```

Angela_Big Data B

SparkSession - in-memory
SparkContext

Spark UI
Version
`v3.4.0`
Master
`local[*]`
AppName
`Sistem Penskoran Otomatis pada Soal Essay 2`

Code diatas digunakan untuk membuat session baru, berikut adalah penjelasan tiap barisnya:

- appName → merupakan defnisi variabel yang berisikan nama aplikasi spark yang digunakan dengan tujuan mengidentifikasi aplikasi dalam lingkungan spark.
- spark = SparkSession → membuat objek SparkSession dengan metode builder dari kelas SparkSession
- .builder → memanggil metode builder pada objek
- .appName(appName) → fungsi appName() digunakan untuk mengatur nama aplikasi yang telah didefinisikan sebelumnya
- .config("spark.some.config.option", "some-value") → fungsi config() digunakan untuk mengatur konfigurasi khusus Spark. "spark.some.config.option" adalah parameter yang digunakan

- .getOrCreate() → digunakan untuk membuat SparkSession sesuai konfigurasi yang diatur sebelumnya.

B. Import Dataset

1. Import Dataset

```
df_pyspark=spark.read.csv('training_data_essay.csv', sep=';', inferSchema=True, header=True)
```

```
df_pyspark.show(20)
```

Angela_Big Data B

```
+----------+------------+--------------------+----+-------------+
|       npm|nama_peserta|             jawaban|soal|skor_per_soal|
+----------+------------+--------------------+----+-------------+
|         0|       Admin|Tidak, Hanya memb...|   1|          100|
|         0|       Admin|Biaya dihitung be...|   2|          100|
|         0|       Admin|Hak cipta adalah ...|   3|          100|
|         0|       Admin|Dijelaskan kepada...|   4|          100|
|         0|       Admin|1. Melindungi dan...|   5|          100|
|         0|       Admin|Ruang Komputer, P...|   6|          100|
|         0|       Admin|Aturlah posisi pe...|   7|          100|
|         0|       Admin|Posisi Kepala dan...|   8|          100|
|         0|       Admin|1. Kecocokan soft...|   9|          100|
|         0|       Admin|1. Fokus dan expo...|  10|          100|
|         0|       Admin|1. Peralatan yang...|  11|          100|
|         0|       Admin|1. Dibuat grafik ...|  12|          100|
|1121020033|          AP|tidak, cuma mengi...|   1|         52,7|
|1121020033|          AP|biaya dihitung be...|   2|        42,86|
|1121020033|          AP|hak membuat merup...|   3|        42,16|
|1121020033|          AP|dipaparkan pada k...|   4|        27,19|
|1121020033|          AP|1. mencegah serta...|   5|        44,14|
|1121020033|          AP|ruang komputer, p...|   6|          100|
|1121020033|          AP|aturlah posisi fi...|   7|        57,68|
|1121020033|          AP|posisi kepala ser...|   8|        45,71|
+----------+------------+--------------------+----+-------------+
only showing top 20 rows
```

Sintaks diatas digunakan untuk membaca file dataset dan ditampilkan, berikut adalah penjelasannya:

- spark.read.csv() → fungsi yang digunakan untuk membaca file format csv
- sep = ‘;’ → sep adalah separator yang digunakan untuk melakukan pemisahan data berdasarkan tanda ‘;’
- inferSchema=True → parameter opsional untuk menentukan apakah spark harus menginfer skema (struktur kolom) dari data yang dibaca. nilai True berarti spark menentukan tipe data tiap kolom secara otomatis
- header = True → parameter opsional untuk menentukan apakah baris pertama dalam csv berisi header atau bukan. Nilai True berarti baris pertama adalah headernya.
- .show(20) → fungsi yang digunakan untuk menampilkan dataframe sebanyak 20 baris teratas

C. Pre-Processing

1. Mengubah nilai ‘,’ menjadi ‘.’ dalam kolom skor_per_soal

```
# Replace comma (",") with dot (".") in the "score" column
df_pyspark = df_pyspark.withColumn("skor_per_soal", regexp_replace(col("skor_per_soal"), ",", "."))

# Show the DataFrame after replacing the comma with a dot
df_pyspark.show()
```
Angela_Big Data B

```
+----------+------------+--------------------+----+-------------+
|       npm|nama_peserta|             jawaban|soal|skor_per_soal|
+----------+------------+--------------------+----+-------------+
|         0|       Admin|Tidak, Hanya memb...|   1|          100|
|         0|       Admin|Biaya dihitung be...|   2|          100|
|         0|       Admin|Hak cipta adalah ...|   3|          100|
|         0|       Admin|Dijelaskan kepada...|   4|          100|
|         0|       Admin|1. Melindungi dan...|   5|          100|
|         0|       Admin|Ruang Komputer, P...|   6|          100|
|         0|       Admin|Aturlah posisi pe...|   7|          100|
|         0|       Admin|Posisi Kepala dan...|   8|          100|
|         0|       Admin|1. Kecocokan soft...|   9|          100|
|         0|       Admin|1. Fokus dan expo...|  10|          100|
|         0|       Admin|1. Peralatan yang...|  11|          100|
|         0|       Admin|1. Dibuat grafik ...|  12|          100|
|1121020033|          AP|tidak, cuma mengi...|   1|         52.7|
|1121020033|          AP|biaya dihitung be...|   2|        42.86|
|1121020033|          AP|hak membuat merup...|   3|        42.16|
|1121020033|          AP|dipaparkan pada k...|   4|        27.19|
|1121020033|          AP|1. mencegah serta...|   5|        44.14|
|1121020033|          AP|ruang komputer, p...|   6|          100|
|1121020033|          AP|aturlah posisi fi...|   7|        57.68|
|1121020033|          AP|posisi kepala ser...|   8|        45.71|
+----------+------------+--------------------+----+-------------+
only showing top 20 rows
```

Berikut adalah sintaks yang digunakan untuk mengubah tanda koma menjadi tanda titik dalam kolom skor_per_soal dengan tujuan agar bisa merubah jadi float. Berikut adalah penjelasannya per baris:

- withColum() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘skor_per_soal’
- parameter kedua berisi ekspresi yang dilakukan untuk mengubah nilai dalam kolom ‘skor_per_soal’. Dalam studi kasus ini, digunakan fungsi regexp_replace() untuk mengganti koma menjadi titik.
- Kemudian dataframe yang baru akan ditampilkan dengan show()

2. Pengubahan tipe data kolom skor_per_soal

```
df_pyspark.printSchema()
```
Angela_Big Data B

```
root
 |-- npm: integer (nullable = true)
 |-- nama_peserta: string (nullable = true)
 |-- jawaban: string (nullable = true)
 |-- soal: integer (nullable = true)
 |-- skor_per_soal: string (nullable = true)
```

```
# Change the data type of the "id" column to integer
df_pyspark = df_pyspark.withColumn("skor_per_soal", col("skor_per_soal").cast("float"))

# Display the updated DataFrame schema
df_pyspark.printSchema()
df_pyspark.show()
```



```
root
 |-- npm: integer (nullable = true)
 |-- nama_peserta: string (nullable = true)
 |-- jawaban: string (nullable = true)
 |-- soal: integer (nullable = true)
 |-- skor_per_soal: float (nullable = true)

+----------+------------+--------------------+----+-------------+
|       npm|nama_peserta|             jawaban|soal|skor_per_soal|
+----------+------------+--------------------+----+-------------+
|         0|       Admin|Tidak, Hanya memb...|   1|        100.0|
|         0|       Admin|Biaya dihitung be...|   2|        100.0|
|         0|       Admin|Hak cipta adalah ...|   3|        100.0|
|         0|       Admin|Dijelaskan kepada...|   4|        100.0|
|         0|       Admin|1. Melindungi dan...|   5|        100.0|
|         0|       Admin|Ruang Komputer, P...|   6|        100.0|
|         0|       Admin|Aturlah posisi pe...|   7|        100.0|
|         0|       Admin|Posisi Kepala dan...|   8|        100.0|
|         0|       Admin|1. Kecocokan soft...|   9|        100.0|
|         0|       Admin|1. Fokus dan expo...|  10|        100.0|
|         0|       Admin|1. Peralatan yang...|  11|        100.0|
|         0|       Admin|1. Dibuat grafik ...|  12|        100.0|
|1121020033|          AP|tidak, cuma mengi...|   1|         52.7|
|1121020033|          AP|biaya dihitung be...|   2|        42.86|
|1121020033|          AP|hak membuat merup...|   3|        42.16|
|1121020033|          AP|dipaparkan pada k...|   4|        27.19|
|1121020033|          AP|1. mencegah serta...|   5|        44.14|
|1121020033|          AP|ruang komputer, p...|   6|        100.0|
|1121020033|          AP|aturlah posisi fi...|   7|        57.68|
|1121020033|          AP|posisi kepala ser...|   8|        45.71|
+----------+------------+--------------------+----+-------------+
only showing top 20 rows
```

Berikut adalah sintaks yang digunakan untuk mengubah tipe data dalam kolom ‘skor_per_soal’. Berikut adalah penjelasannya per baris:

- printSchema() adalah fungsi yang digunakan untuk menampilkan tipe data dalam dataset. Pada awalnya tipe data kolom skor_per_soal adalah string yang perlu diubah ke float untuk proses regresi
- withColum() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘skor_per_soal’
- parameter kedua berisi ekspresi yang dilakukan untuk mengubah nilai dalam kolom ‘skor_per_soal’. Dalam studi kasus ini, digunakan fungsi cast() yang bertujuan mengubah tope data kolom menjadi tipe data float.
- Kemudian dataframe yang baru akan ditampilkan dengan show() beserta menampilkan tipe data yang terbaru.

D. Menyiapkan Data Untuk Pemodelan

1. Feature Selection

```
data = df_pyspark.select("soal","jawaban",'skor_per_soal')
data.show()
```

Angela_Big Data B

```
+----+--------------------+-------------+
|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
```

Berikut adalah sintaks yang digunakan untuk hanya mengambil kolom yang dibutuhkan. Digunakan fungsi select() yang bertujuan untuk memilih kolom mana saja yang akan dibuat dalam dataframe yang baru. Dalam studi kasus ini ada 3 yaitu ‘soal’, ‘jawaban’, dan ‘skor_per_soal’.

2. Hash Function

```
# Apply the hash function
hashedData = data.withColumn("hashedValue", hash("jawaban"))

# Show the results
hashedData.select("soal", "hashedValue","skor_per_soal").show(truncate=False)
```

Angela_Big Data B

```
+----+-----------+-------------+
|soal|hashedValue|skor_per_soal|
+----+-----------+-------------+
|1   |-2059296905|100.0        |
|2   |1183180174 |100.0        |
|3   |1232762403 |100.0        |
|4   |-2035408785|100.0        |
|5   |1588395990 |100.0        |
|6   |339970513  |100.0        |
|7   |50850002   |100.0        |
|8   |-945877996 |100.0        |
|9   |1576366224 |100.0        |
|10  |-1905649442|100.0        |
|11  |550139146  |100.0        |
|12  |1727767227 |100.0        |
|1   |1947733435 |52.7         |
|2   |-1139863335|42.86        |
|3   |122676417  |42.16        |
|4   |-1054163002|27.19        |
|5   |1990940339 |44.14        |
|6   |1770907636 |100.0        |
|7   |-463479969 |57.68        |
|8   |-412537011 |45.71        |
+----+-----------+-------------+
only showing top 20 rows
```

Berikut adalah sintaks yang digunakan untuk mengaplikasikan hash function pada kolom ‘jawaban’. Berikut adalah penjelasannya per baris:

- withColum() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘hashedValue’ yaitu kolom yang menyimpan nilai hash
- parameter kedua berisi ekspresi yang dilakukan untuk menambahkan nilai dalam kolom ‘hashedValue’. Dalam studi kasus ini, digunakan fungsi hash() yang bertujuan mendapatkan nilai hash dari kolom ‘jawaban’
- Kemudian dataframe yang baru akan ditampilkan berdasarkan kolom yang dipilih menggunakan fungsi select()

E. Splitting Dataset dan Mendefinisikan Model

1. Splitting Dataset

```
#membagi data, 70% training dan 30% testing
splits = hashedData.randomSplit([0.7, 0.3])
train = splits[0].withColumnRenamed("skor_per_soal", "Label")
test = splits[1].withColumnRenamed("skor_per_soal", "trueLabel")

#menghitung baris data training dan testing
train_rows = train.count()
test_rows = test.count()
print ("Jumlah baris data training:", train_rows,
       ", jumlah baris data testing:", test_rows)
```



```
Jumlah baris data training: 73 , jumlah baris data testing: 47
```

Berikut adalah sintaks yang digunakan untuk membagi data menjadi data training dan data testing. Berikut adalah penjelasannya per baris:

- baris pertama menggunakan randomSplit() pada dataframe hashedData untuk membagi data menjadi dua bagian berdasarkan proporsi yang diberikan. Dalam studi kasus ini, proporsinya adalah 0.7 banding 0.3 artinya 70% menjadi data training dan 30% menjadi data testing. Hal ini dilakukan secara random
- baris kedua digunakan untuk menyimpan bagian pertama dari splits (indeks 0) dalam variabel train yang menjadi data training. Kemudian, digunakan withColumnRenamed() untuk mengubah nama kolom ‘skor_per_soal’ menjadi ‘Label’
- baris ketiga digunakan untuk menyimpan bagian pertama dari splits (indeks 1) dalam variabel test yang menjadi data testing. Kemudian, digunakan withColumnRenamed() untuk mengubah nama kolom ‘skor_per_soal’ menjadi ‘trueLlabel’
- train_rows dan test_rows masing – masing digunakan untuk menghitung jumlah baris dalam dataframe train dan dataframe test. Digunakan count() sebagai fungsinya. kemudian hasilnya ditampilkan

2. Mendefinisikan Model

```
#mendefinisikan algoritma ALS untuk sistem rekomender kita
als = ALS(maxIter=19, regParam=0.01, userCol="soal",
          itemCol="hashedValue", ratingCol="Label")
#mentraining model dengan fungsi ".fit()"
model = als.fit(train)
print("Model telah selesai ditraining!")
```

Angela_Big Data B

```
Model telah selesai ditraining!
```

Berikut adalah sintaks yang digunakan untuk membangun model berdasarkan algoritma ALS. Berikut adalah penjelasannya per baris:

- ALS() mengandung beberapa parameter diantaranya:
  - maxIter adalah parameter untuk menentukan jumlah iterasi maksimum yang dilakukan algoritma ALS. Dalam studi kasus ini dipilih 19
  - regParam adalah parameter untuk mengontrol kekuatan regularisasi algoritma ALS. Nilai yang lebih tinggi menghasilkan regularisasi yang lebih kuat. Dalam studi kasus ini dipilih 0.01
  - userCol adalah parameter untuk menentukan nama kolom yang berisi data pengguna. Dalam studi kasus ini, dipilih kolom 'soal'
  - itemCol adalah parameter untuk menentukan nama kolom yang berisi data item. Dalam studi kasus ini dipilih kolom 'itemCol'
  - ratingCol adalah parameter untuk menentukan nama kolom yang berisi data peringkat. Dalam studi kasus ini, dipilih kolom 'Label'
- kemudian gunakan fit() pada objek ALS untuk melatih model menggunakan dataframe train sehingga menghasilkan model yang telah ditraining
- Kemudian cetak pesana jika proses pelatihan model selesai dan berhasil.

F. Menyiapkan Data Baru

1. Import Dataset

```
data_baru=spark.read.csv('dataset_baru.csv', sep=';', inferSchema=True, header=True)
```

```
data_baru.show()
```

Angela_Big Data B

```
+-----------+------------+--------------------+----+-------------+
|        npm|nama_peserta|            jawaban|soal|skor_per_soal|
+-----------+------------+--------------------+----+-------------+
|21083010032|      Angela|Ya, semakin banya...|   1|         20,5|
|21083010032|      Angela|Jumlah uang yang ...|   2|           45|
|21083010032|      Angela|hak membuat merup...|   3|        43,18|
|21083010032|      Angela|bila graf sangat ...|   4|        24,56|
|21083010032|      Angela|1. mencegah serta...|   5|         46,9|
|21083010032|      Angela|ruang komputer, p...|   6|          100|
|21083010032|      Angela|aturlah posisi  k...|   7|         63,4|
|21083010032|      Angela|posisi kepala ser...|   8|           48|
|21083010032|      Angela|1.kesesuaian apli...|   9|        51,33|
|21083010032|      Angela|fokus serta apa a...|  10|        39,08|
|21083010032|      Angela|1. perlengkapan y...|  11|        39,88|
|21083010032|      Angela|metode artwork 2d...|  12|        25,67|
+-----------+------------+--------------------+----+-------------+
```

Sintaks diatas digunakan untuk membaca file dataset dan ditampilkan, berikut adalah penjelasannya:

- spark.read.csv() → fungsi yang digunakan untuk membaca file format csv
- sep = ‘;’ → sep adalah separator yang digunakan untuk melakukan pemisahan data berdasarkan tanda ‘;’
- inferSchema=True → parameter opsional untuk menentukan apakah spark harus menginfer skema (struktur kolom) dari data yang dibaca. nilai True berarti spark menentukan tipe data tiap kolom secara otomatis
- header = True → parameter opsional untuk menentukan apakah baris pertama dalam csv berisi header atau bukan. Nilai True berarti baris pertama adalah headernya.
- .show() → fungsi yang digunakan untuk menampilkan dataframe

2. Pre-Processing (Pengubahan tipe data)

```
# Replace comma (",") with dot (".") in the "score" column
data_baru = data_baru.withColumn("skor_per_soal", regexp_replace(col("skor_per_soal"), ",", "."))

# Show the DataFrame after replacing the comma with a dot
data_baru.show()
```

Angela_Big Data B

```
+-----------+------------+--------------------+----+-------------+
|        npm|nama_peserta|             jawaban|soal|skor_per_soal|
+-----------+------------+--------------------+----+-------------+
|21083010032|      Angela|Ya, semakin banya...|   1|         20.5|
|21083010032|      Angela|Jumlah uang yang ...|   2|           45|
|21083010032|      Angela|hak membuat merup...|   3|        43.18|
|21083010032|      Angela|bila graf sangat ...|   4|        24.56|
|21083010032|      Angela|1. mencegah serta...|   5|         46.9|
|21083010032|      Angela|ruang komputer, p...|   6|          100|
|21083010032|      Angela|aturlah posisi  k...|   7|         63.4|
|21083010032|      Angela|posisi kepala ser...|   8|           48|
|21083010032|      Angela|1.kesesuaian apli...|   9|        51.33|
|21083010032|      Angela|fokus serta apa a...|  10|        39.08|
|21083010032|      Angela|1. perlengkapan y...|  11|        39.88|
|21083010032|      Angela|metode artwork 2d...|  12|        25.67|
+-----------+------------+--------------------+----+-------------+
```

Berikut adalah sintaks yang digunakan untuk mengubah tanda koma menjadi tanda titik dalam kolom skor_per_soal dengan tujuan agar bisa merubah jadi float. Berikut adalah penjelasannya per baris:

- withColum() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘skor_per_soal’
- parameter kedua berisi ekspresi yang dilakukan untuk mengubah nilai dalam kolom ‘skor_per_soal’. Dalam studi kasus ini, digunakan fungsi regexp_replace() untuk mengganti koma menjadi titik.
- Kemudian dataframe yang baru akan ditampilkan dengan show()

```
# Change the data type of the "id" column to integer
data_baru_2 = data_baru.withColumn("skor_per_soal", col("skor_per_soal").cast("float"))

# Display the updated DataFrame schema
data_baru_2.printSchema()
data_baru_2.show()
```

Angela_Big Data B

```
root
 |-- npm: long (nullable = true)
 |-- nama_peserta: string (nullable = true)
 |-- jawaban: string (nullable = true)
 |-- soal: integer (nullable = true)
 |-- skor_per_soal: float (nullable = true)

+-----------+------------+--------------------+----+-------------+
|        npm|nama_peserta|             jawaban|soal|skor_per_soal|
+-----------+------------+--------------------+----+-------------+
|21083010032|      Angela|Ya, semakin banya...|   1|         20.5|
|21083010032|      Angela|Jumlah uang yang ...|   2|         45.0|
|21083010032|      Angela|hak membuat merup...|   3|        43.18|
|21083010032|      Angela|bila graf sangat ...|   4|        24.56|
|21083010032|      Angela|1. mencegah serta...|   5|         46.9|
|21083010032|      Angela|ruang komputer, p...|   6|        100.0|
|21083010032|      Angela|aturlah posisi  k...|   7|         63.4|
|21083010032|      Angela|posisi kepala ser...|   8|         48.0|
|21083010032|      Angela|1.kesesuaian apli...|   9|        51.33|
|21083010032|      Angela|fokus serta apa a...|  10|        39.08|
|21083010032|      Angela|1. perlengkapan y...|  11|        39.88|
|21083010032|      Angela|metode artwork 2d...|  12|        25.67|
+-----------+------------+--------------------+----+-------------+
```

Berikut adalah sintaks yang digunakan untuk mengubah tipe data dalam kolom ‘skor_per_soal’. Berikut adalah penjelasannya per baris:

- printSchema() adalah fungsi yang digunakan untuk menampilkan tipe data dalam dataset. Pada awalnya tipe data kolom skor_per_soal adalah string yang perlu diubah ke float untuk proses regresi
- withColum() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘skor_per_soal’
- parameter kedua berisi ekspresi yang dilakukan untuk mengubah nilai dalam kolom ‘skor_per_soal’. Dalam studi kasus ini, digunakan fungsi cast() yang bertujuan mengubah tope data kolom menjadi tipe data float.
- Kemudian dataframe yang baru akan ditampilkan dengan show() beserta menampilkan tipe data yang terbaru.

3. Feature Selection dan Hash Function

```
data2 = data_baru_2.select("soal","jawaban",'skor_per_soal')
data2.show()
```

Angela_Big Data B

```
+----+--------------------+-------------+
|soal|             jawaban|skor_per_soal|
+----+--------------------+-------------+
|   1|Ya, semakin banya...|         20.5|
|   2|Jumlah uang yang ...|         45.0|
|   3|hak membuat merup...|        43.18|
|   4|bila graf sangat ...|        24.56|
|   5|1. mencegah serta...|         46.9|
|   6|ruang komputer, p...|        100.0|
|   7|aturlah posisi  k...|         63.4|
|   8|posisi kepala ser...|         48.0|
|   9|1.kesesuaian apli...|        51.33|
|  10|fokus serta apa a...|        39.08|
|  11|1. perlengkapan y...|        39.88|
|  12|metode artwork 2d...|        25.67|
+----+--------------------+-------------+
```

Berikut adalah sintaks yang digunakan untuk hanya mengambil kolom yang dibutuhkan. Digunakan fungsi select() yang bertujuan untuk memilih kolom mana saja yang akan dibuat dalam dataframe yang baru. Dalam studi kasus ini ada 3 yaitu ‘soal’, ‘jawaban’, dan ‘skor_per_soal’.

```
# Apply the hash function
hashedData2 = data2.withColumn("hashedValue", hash("jawaban"))

# Show the results
hashedData2.select("soal", "hashedValue","skor_per_soal").show(truncate=False)
```

Angela_Big Data B

```
+----+-----------+-------------+
|soal|hashedValue|skor_per_soal|
+----+-----------+-------------+
|1   |1019100933 |20.5         |
|2   |1481524314 |45.0         |
|3   |122676417  |43.18        |
|4   |76487259   |24.56        |
|5   |1990940339 |46.9         |
|6   |1770907636 |100.0        |
|7   |400780623  |63.4         |
|8   |-412537011 |48.0         |
|9   |-55989520  |51.33        |
|10  |670920752  |39.08        |
|11  |723150141  |39.88        |
|12  |343114756  |25.67        |
+----+-----------+-------------+
```

Berikut adalah sintaks yang digunakan untuk mengaplikasikan hash function pada kolom ‘jawaban’. Berikut adalah penjelasannya per baris:

- withColumn() berfungsi untuk memilih sebuah kolom yang ingin ditambakan atau mengganti nilainya. Dalam studi kasus ini adalah kolom ‘hashedValue2’ yaitu kolom yang menyimpan nilai hash

- parameter kedua berisi ekspresi yang dilakukan untuk menambahkan nilai dalam kolom ‘hashedValue2’. Dalam studi kasus ini, digunakan fungsi hash() yang bertujuan mendapatkan nilai hash dari kolom ‘jawaban’
- Kemudian dataframe yang baru akan ditampilkan berdasarkan kolom yang dipilih menggunakan fungsi select()

G. Melakukan Prediksi dengan Dataset Baru

1. Hasil Prediksi Dataset Baru

```
predictions2 = model.transform(hashedData2)
predictions2.show()
```

Angela_Big Data B

```
+----+--------------------+-------------+-----------+----------+
|soal|             jawaban|skor_per_soal|hashedValue|prediction|
+----+--------------------+-------------+-----------+----------+
|   1|Ya, semakin banya...|         20.5| 1019100933|       NaN|
|   5|1. mencegah serta...|         46.9| 1990940339| 44.139896|
|   2|Jumlah uang yang ...|         45.0| 1481524314|       NaN|
|   9|1.kesesuaian apli...|        51.33|  -55989520|       NaN|
|   7|aturlah posisi  k...|         63.4|  400780623|       NaN|
|   3|hak membuat merup...|        43.18|  122676417| 42.159916|
|  10|fokus serta apa a...|        39.08|  670920752|       NaN|
|  12|metode artwork 2d...|        25.67|  343114756|       NaN|
|  11|1. perlengkapan y...|        39.88|  723150141|       NaN|
|   6|ruang komputer, p...|        100.0| 1770907636|     100.0|
|   8|posisi kepala ser...|         48.0| -412537011|       NaN|
|   4|bila graf sangat ...|        24.56|   76487259|       NaN|
+----+--------------------+-------------+-----------+----------+
```

Berikut adalah sintaks yang digunakan untuk mengaplikasikan model terhadap dataset yang baru untuk melakukan prediksi. Berikut adalah penjelasannya per baris:

- transform() digunakan untuk menghasilkan dataframe yang baru berisi prediksi yang dibuat oleh model. Dataframe yang akan diprediksi adalah dataframe yang baru
- kemudian tampilkan hasil prediksi dengan show()

H. Evaluasi Model

1. Evaluasi Model menggunakan RMSE

```
evaluator = RegressionEvaluator(
    labelCol="skor_per_soal", predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions2)
print ("Root Mean Square Error (RMSE):", rmse)
```

Angela_Big Data B

```
Root Mean Square Error (RMSE): nan
```

Berikut adalah sintaks yang digunakan untuk mengevaluasi kualitas prediksi berdasarkan model yang didefinisikan sebelumnya. Berikut adalah penjelasannya per baris:

- Digunakan objek ‘RegressionEvaluator’ yang terdiri dari berbagai parameter, diantaranya:
  - labelCol digunakan untuk menentukan nama kolom yang berisi nilai target yaitu kolom ‘skor_per_soal’
  - predictionCol digunakan untuk menentukan nama kolom yang berisi nilai prediksi yaitu kolom ‘prediction’
  - metricName digunakan untuk menentukan metrik evaluasi yang digunakan yaitu RMSE.

- kemudian digunakan evaluate() pada objek evaluator untuk mengevaluasi kualitas prediksi yang dihasilkan oleh model pada dataframe ‘predictions2’. Metode ini menghitung metrik evaluasi yaitu RMSE. kemudian tampilkan hasilnya RMSEnya
- Ternyata hasilnya adalah nan, hal ini menunjukkan bahwa masih ada hasil prediksi yang bernilai nan

```
a = predictions2.count()
print("jumlah baris sebelum di hapus data kosong: ", a)
cleanPred = predictions2.dropna(how="any", subset=["prediction"])
b = cleanPred.count()
print("jumlah baris setelah di hapus data kosong: ", b)
print("jumlah baris data kosong: ", a-b)
```

Angela_Big Data B

```
jumlah baris sebelum di hapus data kosong:  12
jumlah baris setelah di hapus data kosong:  3
jumlah baris data kosong:  9
```

Berikut adalah sintaks yang digunakan untuk menghitung berapa banyak baris yang berisi data kosong sebelum dan sesudah dihapus. Berikut adalah penjelasannya per baris:

- baris pertama digunakan untuk menghitung jumlah baris sebelum dilakukan penghapusan data kosong dengan count() dan baris kedua digunakan untuk menampilkan hasilnya
- baris ketiga digunakan untuk menghapus baris yang memiliki nilai kosong dalam kolom ‘prediction’ menggunakan fungsi dropna(). Parameter how=’any’ berarti baris akan dihapus jika setidaknya terdapat satu nilai kosong dalam subset kolom ang digunakan. kemudian disimpan dalam dataframe Bernama CleanPred
- baris keempat digunakan untuk menghitung jumlah baris sesudah dilakukan penghapusan data kosong dengan count(). Kemudian baris kelima dan keenam untuk menampilkan hasilnya beserta selisih sebelum dan sesudah penghapusan.

```
cleanPred.show()
```

Angela_Big Data B

```
+----+--------------------+------------+-----------+----------+
|soal|             jawaban|skor_per_soal|hashedValue|prediction|
+----+--------------------+------------+-----------+----------+
|   5|1. mencegah serta...|        46.9| 1990940339| 44.139896|
|   3|hak membuat merup...|       43.18|  122676417| 42.159916|
|   6|ruang komputer, p...|       100.0| 1770907636|     100.0|
+----+--------------------+------------+-----------+----------+
```

Dataframe yang sudah dilakukan penghapus data null ditampilkan dengan fungsi show()

```
rmse = evaluator.evaluate(cleanPred)
print ("Root Mean Square Error (RMSE):", rmse)
```

Angela_Big Data B

```
Root Mean Square Error (RMSE): 1.6988969450888771
```

Berikut adalah sintaks yang digunakan untuk mengulangi evaluasi model. Berikut adalah penjelasannya per baris:

- Digunakan evaluate() pada objek evaluator untuk mengevaluasi kualitas prediksi yang dihasilkan oleh model pada dataframe ‘predictions2’. Metode ini menghitung metrik evaluasi yaitu RMSE. kemudian tampilkan hasilnya RMSEnya. Hasilnya adalah 1.6988 yang cukup kecil namun, ternyata prediksi awal banyak nilai nan.