Laporan Tugas Big Data Analysis

Oleh :

Imaddudin Muhammad Fadhil (2301212043)

Program Studi S2 Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2022

DAFTAR ISI

PENJELASAN ISI LAPORAN

A. FORMULASI MASALAH

Formulasi masalah yang kami selesaikan dalam tugas ini adalah untuk mencari insight dari data South German Credit dengan cara melakukan Customer Segmentation untuk mengetahui karakteristik dari setiap segmentasi user yang ada dan mengetahui cara pakai dari library PySpark

Sistem yang dibuat menggunakan metode clustering dengan model K-Means untuk mencari ada berapa segmentasi pada user yang mengambil pinjaman dan menggali insight yang didapat dari data dengan menggunakan library PrSpark

B. EKSPLORASI DAN PERSIAPAN DATA

Data yang digunakan pada tugas ini adalah data South German Credit yang bersifat deskriptif dengan format .csv dan tahap pertama dari pembuatan sistem ini adalah pengenalan dan eksplorasi data.

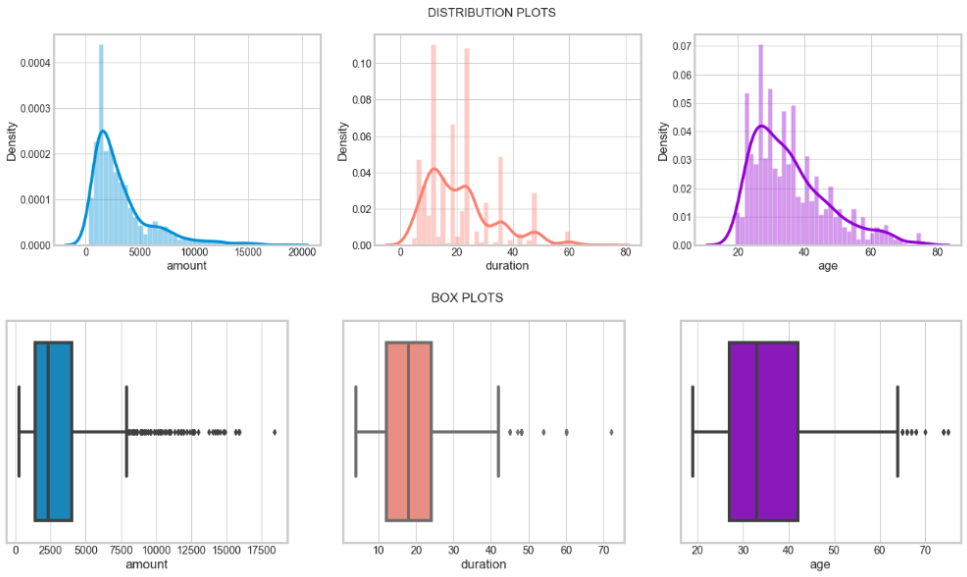
1. Missing Value

   Untuk semua kolom tidak ada satupun yang mempunyai missing value.

#	Column	Non-Null Count	Dtype
0	_c0	1000 non-null	int32
1	status	1000 non-null	object
2	duration	1000 non-null	int32
3	credit_history	1000 non-null	object
4	purpose	1000 non-null	object
5	amount	1000 non-null	int32
6	savings	1000 non-null	object
7	employment_duration	1000 non-null	object
8	installment_rate	1000 non-null	object
9	personal_status_sex	1000 non-null	object
10	other_debtors	1000 non-null	object
11	present_residence	1000 non-null	object
12	property	1000 non-null	object
13	age	1000 non-null	int32
14	other_installment_plans	1000 non-null	object
15	housing	1000 non-null	object
16	number_credits	1000 non-null	object
17	job	1000 non-null	object
18	people_liable	1000 non-null	object
19	telephone	1000 non-null	object
20	foreign_worker	1000 non-null	object
21	credit_risk	1000 non-null	object

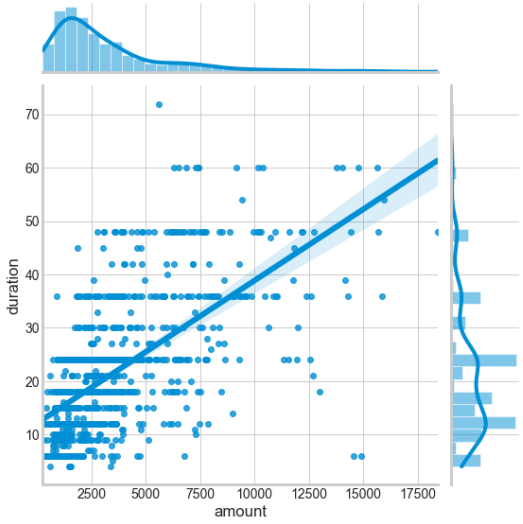
2. Data Distribution

   Kebanyakan Credit Card yang diajukan memiliki Amount sebesar 1500 – 4000 dengan distribusi data dari Credit Amount adalah positif.



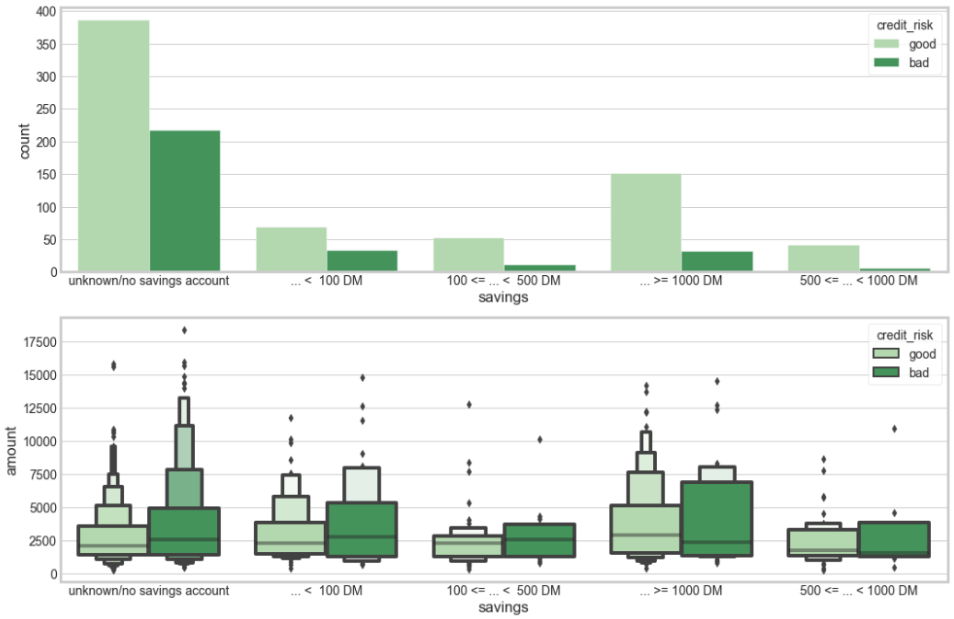
3. Credit Amount by Duration Analysis

   Semakin besar credit yang diambil maka, semakin lama pula periode pembayarannya.



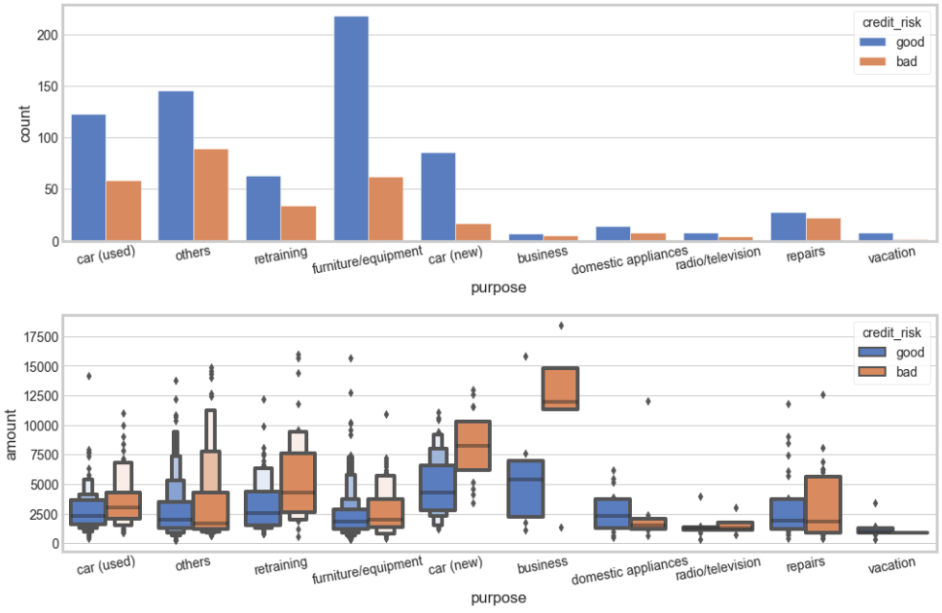
4. Saving Analysis

   Kebanyakan orang yang mengambil credit tidak mempunyai akun tabungan dan untuk setiap kategori tabungan, orang yang mempunyai credit risk buruk cenderung mengambil credit amount yang lebih besar.



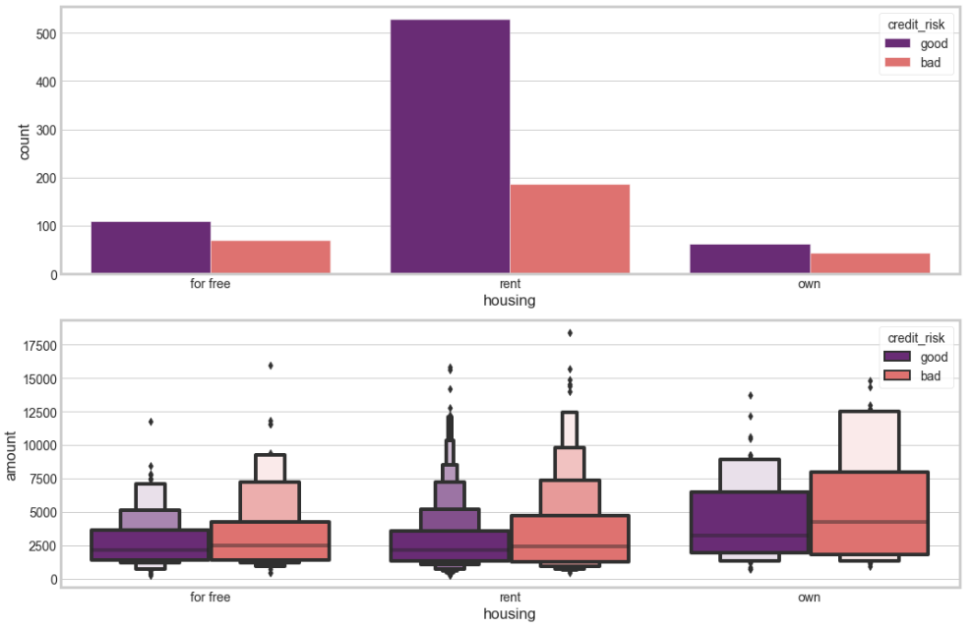
5. Credit Card Purpose Analysis

Orang yang mengambil credit dengan tujuan untuk membeli mobil, retrainging, dan furniture cenderung memiliki credit risk yang baik sedangkan orang yang mengambil pinjaman besar untuk membeli mobil baru dan membuka bisnis cenderung memiliki credit risk yang buruk



6. Credit Risk by Housing Analysis

Kebanyakan orang yang mengambil pinjaman adalah orang yang tinggal di tempat sewaan dan jika urutan membesar dari jumlah pinjaman yang diambil adalah free housing, rent housing dan own housing.

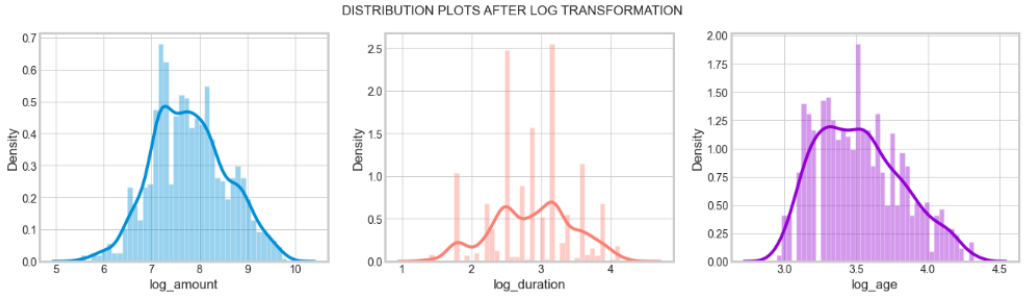


C. PRAPROSES DATA

Data yang sudah dilakukan eksplorasi sekarang dilakukan praproses data sebelum data tersebut digunakan untuk tahap selanjutnya, pada tahap ini ada beberapa proses yang harus dilakukan, proses yang dilakukan pada sistem ini adalah sebagai berikut:

1. Log Transformation untuk Numerikal Feature

   Transformasi log perlu dilakukan untuk fitur dengan tipe data numerikan untuk memperbaiki distribusi data yang sebelumnya condong positif ke normal.



2. Standart Scalling untuk Numerikal Feature

   Setelah distribusi data sudah benar, dilakukan scalling pada setiap fitur yang akan digunakan untuk membantu model belajar lebih cepat dan meningkatkan nilai evaluasi model.

Standart Scalling

```
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.feature import StandardScaler
```

```
#menyatukan kolom input (log_duration, log_age, log_amount) menjadi satu kolom vektor yang dinamakan feature
features = [ele for ele in num_sdf.columns if ele not in numerical_cols]

assemble=VectorAssembler(inputCols=features, outputCol='features')
assembled_num_sdf=assemble.transform(num_sdf)
assembled_num_sdf.show(2)
```

```
+--------+---+------+------------------+-----------------+-----------------+--------------------+
|duration|age|amount|      log_duration|          log_age|       log_amount|            features|
+--------+---+------+------------------+-----------------+-----------------+--------------------+
|      18| 21|  1049|2.8903717578961645|3.044522437723423|6.955592608396297|[2.89037175789616...|
|       9| 36|  2799|2.1972245773362196| 3.58351893845611|7.937017489515453|[2.19722457733621...|
+--------+---+------+------------------+-----------------+-----------------+--------------------+
only showing top 2 rows
```

```
#normalisasi skala dengan Standart Scalling
scale=StandardScaler(inputCol='features',outputCol='standardized')
data_scale=scale.fit(assembled_num_sdf)
data_scale_output=data_scale.transform(assembled_num_sdf)
data_scale_output.show(2)
```

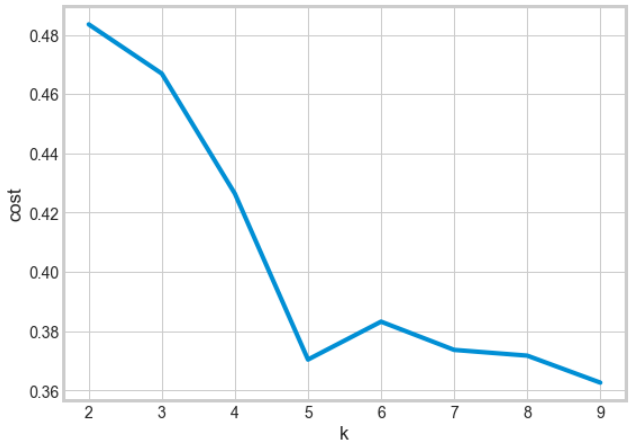
```
+--------+---+------+------------------+-----------------+-----------------+--------------------+--------------------+
|duration|age|amount|      log_duration|          log_age|       log_amount|            features|        standardized|
+--------+---+------+------------------+-----------------+-----------------+--------------------+--------------------+
|      18| 21|  1049|2.8903717578961645|3.044522437723423|6.955592608396297|[2.89037175789616...|[4.96422717493093...|
|       9| 36|  2799|2.1972245773362196| 3.58351893845611|7.937017489515453|[2.19722457733621...|[3.77374361150620...|
+--------+---+------+------------------+-----------------+-----------------+--------------------+--------------------+
only showing top 2 rows
```

D. CLUSTERING

Setelah dilakukan preprocessing maka tahap selanjutnya adalah melakukan clustering menggunakan model K-Means untuk customer segmentation. Tahapan yang dilakukan untuk mengimplementasi metode tersebut pada tugas ini adalah sebagai berikut.

1. Silhouette Method

Elbow method digunakan untuk mencari nilai K atau jumlah segment yang paling optimal, dengan melakukan iterasi evaluasi training menggunakan silhouette score untuk setiap K = {2,..,10} dan mencari local maxima dari hasil yang didapatkan.



Setelah iterasi selesai dapat kita lihat bahwa K = 3 adalah local maxima.

2. Train Model

Setelah mengetahui nilai K paling optimum untuk data ini, maka diputuskan untuk menggunakan nilai K = 3 untuk traning dan evaluasi insight yang didapan dari hasil clustering.

3. Test Model Rekomendasi

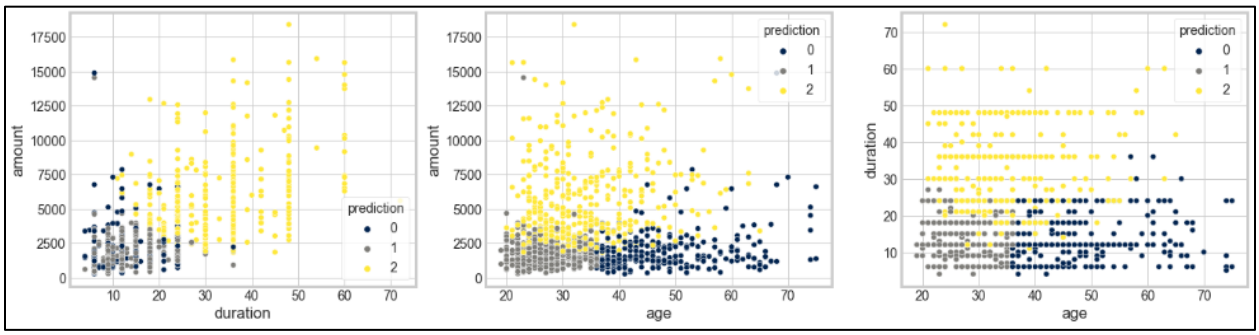
Setelah data selesai di latih, dilakukan pengecekan apakah sistem rekomendasi berhasil memberikan 5 rekomendasi dengan implementasi sebagai berikut.

```
#Train dengan K=3
KMeans_algo=KMeans(featuresCol='standardized', k=3)
KMeans_fit=KMeans_algo.fit(data_scale_output)
KMeans_result=KMeans_fit.transform(data_scale_output)
```

E. EVALUASI

Jika dilihat dari grafik yang disediakan, juga rata2 setiap fitur untuk seitap cluster maka dapat disimpulkan bahwa:

- Cluster 0: orang yang mempunyai usia menengah, cenderung mengambil pinjaman besar dengan durasi yang lama
- Cluster 1: orang yang mempunyai usia muda, cenderung mengambil pinjaman kecil-menengah dengan durasi yang sebentar
- Cluster 2: orang yang mempunyai usia tua, cenderung mengambil pinjaman menengah dengan durasi yang sebentar



prediction	age mean	duration mean	amount mean
0	48.511450	13.843511	1967.038168
1	27.710383	14.390710	1747.090164
2	34.112903	32.282258	5689.379032