

BAB IV HASIL DAN ANALISIS PENELITIAN

4.1 Analisa dan Perancangan Sistem

Pada bagian ini akan menjelaskan perancangan sistem yang akan dikembangkan dengan menerapkan salah satu algoritma pada *data mining* yaitu *K-Nearest Neighbor* (KNN). Pengolahan *data mining* dalam penelitian ini, mengikuti tahapan dalam *Knowledge Discovery in Database* (KDD), yang bertujuan untuk menghasilkan informasi sesuai dengan urutan yang sudah ditentukan, berikut tahapan-tahapannya:

4.1.1. Data Selection

Data yang digunakan dalam penelitian ini adalah Riwayat data transaksi penjualan pada PT. Joanne Studio, atribut-atribut yang terdapat dalam data transaksi penjualan, diantaranya adalah field *Transaction Number*, *Service Name*, *Quantity*, *Price*, *Total Transaction*, *Transaction Date Time*. Selanjutnya akan dilakukan penyeleksian data dari semua data yang ada. Adapun jumlah datanya adalah 8.936 data. Pada penelitian ini yang dibutuhkan hanya 3 (tiga) field diantaranya adalah:

1. *Field Service Name*

Field ini merupakan sebuah atribut yang terdapat di dalam tabel data transaksi penjualan, dan berisi tentang informasi nama-nama produk.

2. *Field Quantity*

Field ini merupakan sebuah atribut yang terdapat di dalam tabel data transaksi penjualan dan berisi tentang informasi jumlah produk terjual setiap ada transaksi penjualan.

3. *Field Transaction Date Time*

Field ini merupakan sebuah atribut yang terdapat di dalam tabel data transaksi penjualan dan berisi tentang informasi hari, bulan, tahun, dan jam setiap transaksi penjualan, pada atribut ini yang akan diambil hanya bulan dan tahun.

Adapun *field-field* tersebut dapat dilihat pada tabel 4.1

Tabel 4. 1 Tabel Transaksi

No	Service Name	Qty	Transaction Date
1	Clean Lash	1	2022-01-08 13:00:00
2	Nails Extensions	1	2022-01-13 16:00:00
3	Double Lash	1	2022-05-31 21:03:59
4	Clean Lash	1	2022-01-13 16:00:00
5	Single Lash	1	2022-01-13 16:00:00
6	Single Lash	1	2022-01-14 10:00:00
7	Double Lash	1	2022-01-13 14:30:35
8	Bottom Single	1	2022-01-13 14:30:35
9	Manicure Only	1	2022-01-13 14:30:35
10	Pedicure Only	1	2022-01-13 14:30:35
11	Remove Gel Polish	8	2022-01-13 14:30:35
12	Single Gel Polish	20	2022-01-13 14:30:35
13	Double Lash	1	2022-01-13 13:00:00
14	Manicure Spa	1	2022-01-13 13:00:00
15	Single Gel Polish	10	2022-01-13 14:38:50
...
8936	Clean Lash	1	2022-05-31 21:03:59

4.1.2. *Preprocessing*

Tahap ini adalah untuk melakukan pengelompokan data berdasarkan jenis produk, jumlah transaksi penjualan setiap bulan. Pengelompokan tersebut berguna untuk mempermudah saat melakukan proses perhitungan akurasi prediksi, kemudian selanjutnya adalah melakukan *cleaning* data yang bertujuan untuk menghasilkan dataset yang bersih, contohnya tidak ada data yang bernilai null, tujuannya agar lebih ketika masuk ke dalam proses perhitungan prediksi. Adapun setelah dilakukan proses *cleaning* data, terdapat perubahan pada jumlah

data yaitu menjadi 4.365 data dari 20 produk yang nantinya dijadikan sebagai data input dan label sebagai acuan untuk mengetahui kategori produk mana yang paling diminati. Label terdiri dari angka 1 dan angka 2 yang dimana angka 1 berarti produk dengan kategori 1 yang paling banyak diminati, begitupun sebaliknya jika angka 2 berarti produk dengan kategori 2 yang paling banyak diminati oleh konsumen. Adapun produk kategori 1 berkaitan dengan produk single lash, clean lash, double lash, mega volume, bottom single. Lalu pada kategori 2 berkaitan dengan produk manicure only, manicure spa, pedicure only, pedicure spa, single gel polish. Keseluruhan kategori tersebut didapat berdasarkan data dari PT. Joanne Studio dan sudah sesuai ketentuan perusahaan.

1. Pengelompokan Data

Pengelompokan data dapat dilihat pada tabel 4.2

Tabel 4.2 Pengelompokan Data Berdasarkan Jenis Service Produk

No	Single Lash	Clean Lash	Label
1	24	5	2
2	16	8	2
3	8	8	2
4	18	3	1
5	7	4	2
6	3	14	2
7	15	1	1
8	24	4	1
9	21	7	1
10	12	12	2
....
4365	20	5	1

2. Data Cleaning

Tahap ini adalah proses membersihkan data yang bersifat tidak relevan, *missing value*, ataupun *redundant*. Kemudian jika terdapat atribut-atribut tidak berisi nilai atau *null* maka dapat

dikategorikan data tersebut termasuk data bersifat *missing value*, selanjutnya apabila di dalam 1 dataset ada lebih dari 1 record berisi nilai yang sama, maka hal tersebut dapat dikategorikan sebagai data yang bersifat *redundant*. Sebagai contoh dapat dilihat pada gambar 4.1 di bawah ini:

ServiceName	Qty	TransactionDatetime
NULL	NULL	2022-01-07 18:18:01
NULL	NULL	2022-01-14 16:00:00
NULL	NULL	2022-01-29 13:23:43
NULL	NULL	2022-02-14 15:25:19
NULL	NULL	2022-03-10 14:27:45
NULL	NULL	2022-04-05 18:44:04
NULL	NULL	2022-04-24 11:03:22
NULL	NULL	2022-04-30 18:00:00
NULL	1	2022-05-04 11:00:00
NULL	1	2022-05-07 14:00:00

Gambar 4.1 Cleaning Data missing value, redundant

4.1.3. Transformation

Hasil dari pengelompokan data pada tahap *preprocessing* akan digunakan sebagai data *training*. Proses pembentukan data training berdasarkan data yang sudah ada, kemudian dalam menentukan data target, maka data harus diseleksi terlebih dahulu. Data target merupakan atribut yang dapat mempengaruhi penjualan produk.

Pada penelitian ini, data terbagi menjadi 2 bagian yang digunakan sebagai data training dan data test. Adapun dalam penelitian ini, data training sebesar 75% dari keseluruhan data, selanjutnya data test sebesar 25% dari keseluruhan data. Selain itu data training juga dikelompokkan menjadi 2 antara lain data yang digunakan sebagai input dan data yang digunakan sebagai target, yang dimana data input adalah data jenis produk berjumlah 20 jenis, sedangkan data target adalah label daripada dataset yang digunakan.

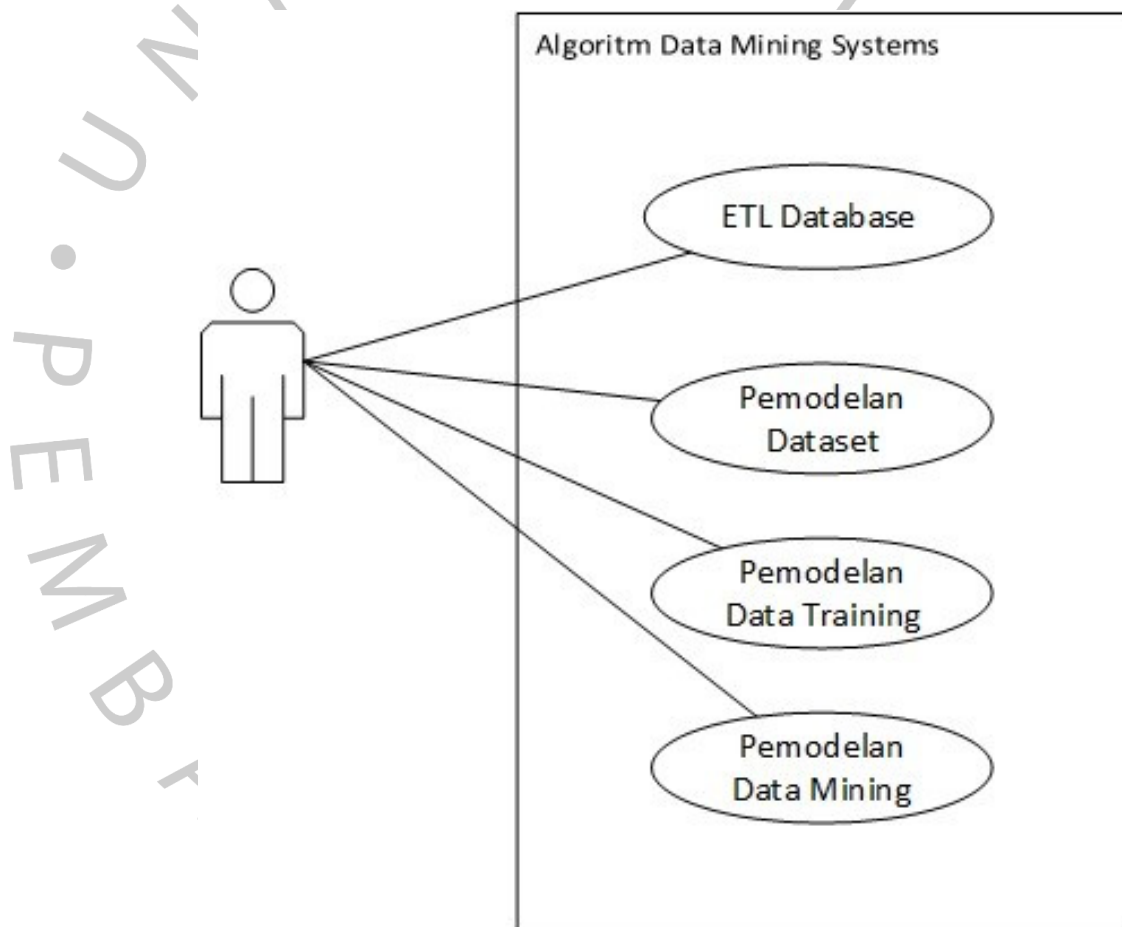
4.2 Perancangan Diagram Sistem Usulan

Dalam menyusun penelitian ini, akan dijelaskan bentuk diagram-diagram yang mendukung dalam pengolahan data transaksi penjualan pada PT.

Joanne Studio. Diagram-diagram yang digunakan dalam penelitian ini yaitu *use case* diagram, *class* diagram, *activity* diagram, dan *sequence* diagram.

4.2.1. Use Case Diagram

Pada penelitian ini *use case diagram* digunakan untuk menjelaskan apa yang akan dilakukan sistem dan aktor-aktor pada PT. Joanne Studio, yang berhubungan dengan dengan proses-proses pada sistem penjualan. Pada prosesnya pertama-tama actor melakukan proses ETL database, lalu Setelah aktor-aktor teridentifikasi maka dapat dilakukan permodelan *Use Case Diagram*, seperti gambar di bawah 4.2 yaitu sebagai berikut:

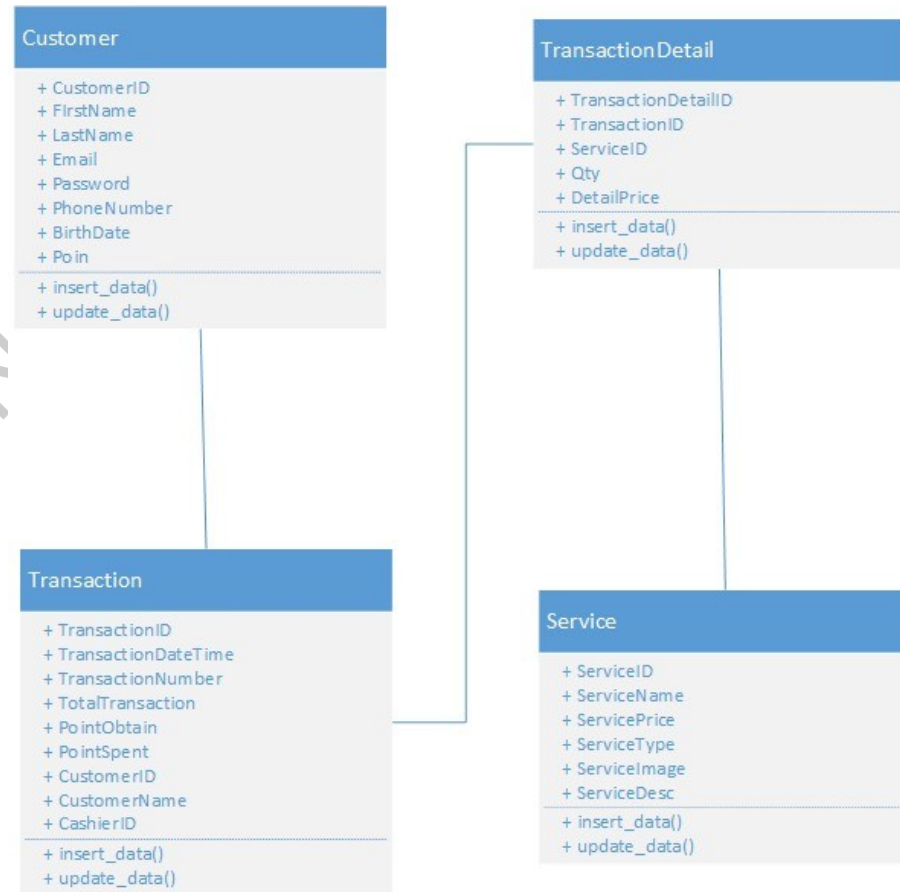


Gambar 4.2 Use Case diagram

4.2.2. Class Diagram

Dalam penelitian ini ada 4 *class* utama yang saling berhubungan, diantaranya adalah class *Customer*, *Transaction*, *Transaction Detail*, *Cashier*, dan *Service*. Pada implementasinya *class-class* tersebut bisa dikatakan sebagai *entity* yang tersimpan pada *database*. Salah satu fungsi *class diagram* disini adalah dapat memberikan gambaran

mengenai sistem atau perangkat lunak serta relasi-relasi yang terkandung di dalamnya. Adapun permodelan class diagram pada penelitian ini dapat dilihat pada gambar 4.3 yaitu sebagai berikut:

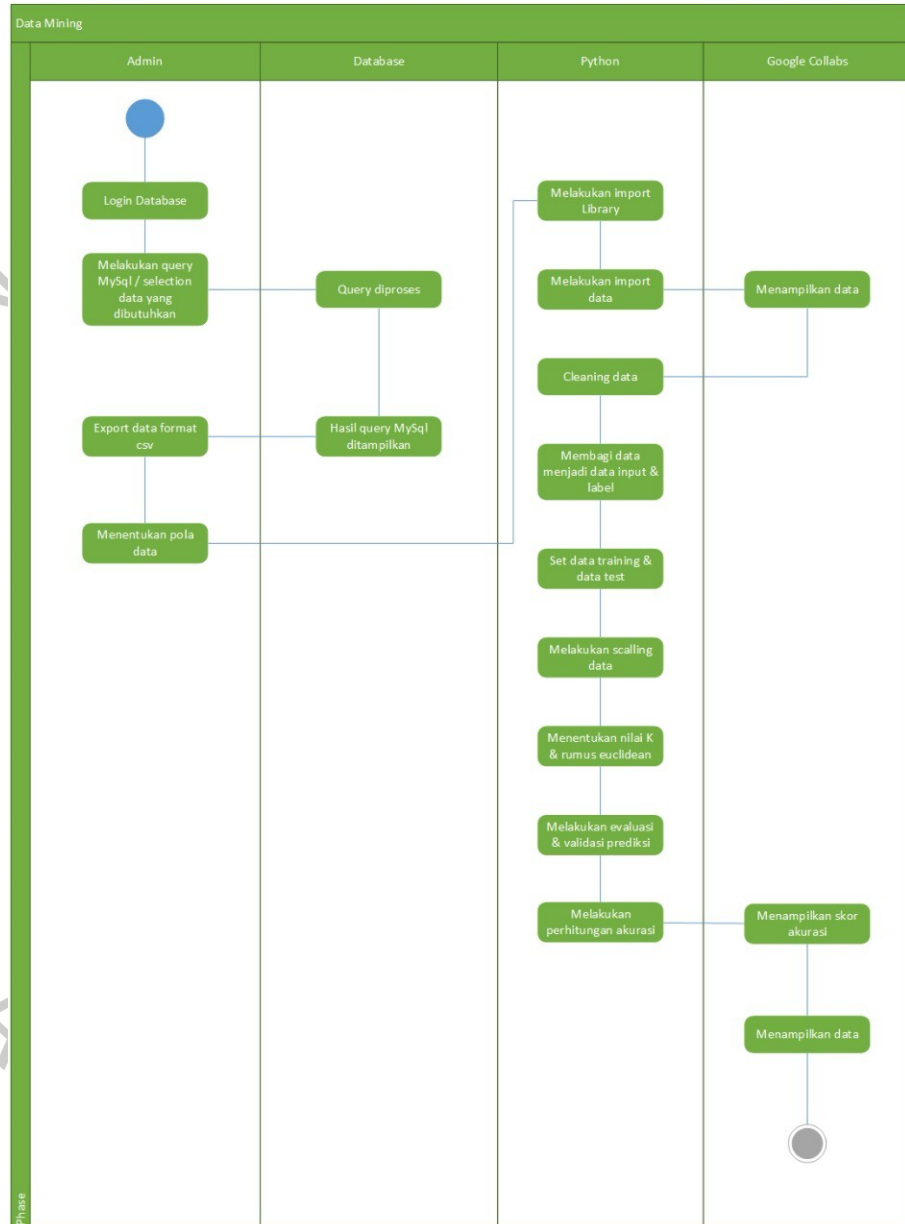


Gambar 4.3 Class Diagram

4.2.3. Activity Diagram

Pada aktivitasnya *Activity* diagram menggambarkan tentang alur sistem dalam penerapan algoritma *K-Nearest Neighbor* di penelitian ini. Adapun awal prosesnya adalah seorang administrator melakukan seleksi data transaksi yang dibutuhkan untuk penelitian dengan menerapkan query MySQL, lalu mengekspor data transaksi tersebut menjadi sebuah file dengan format csv. Proses selanjutnya file yang berisi data transaksi tersebut diimport ke dalam sebuah tools untuk melakukan mining data, yaitu *Google Collabs*. Setelah data berhasil diimport maka dilanjutkan dengan proses mining data seperti melakukan *cleansing* data, membagi data menjadi data input dan label, membagi data menjadi data training dan data test, melakukan *scaling* data, menentukan nilai K yang terbaik,

melakukan evaluasi dan validasi terhadap perhitungan yang telah dilakukan, melakukan perhitungan akurasi dan yang terakhir menampilkan skor akurasi serta visualisasi berupa grafik dari hasil prediksi. Gambaran proses *activity*. diagram pada penelitian ini dapat dilihat pada gambar 4.4 yaitu sebagai berikut:

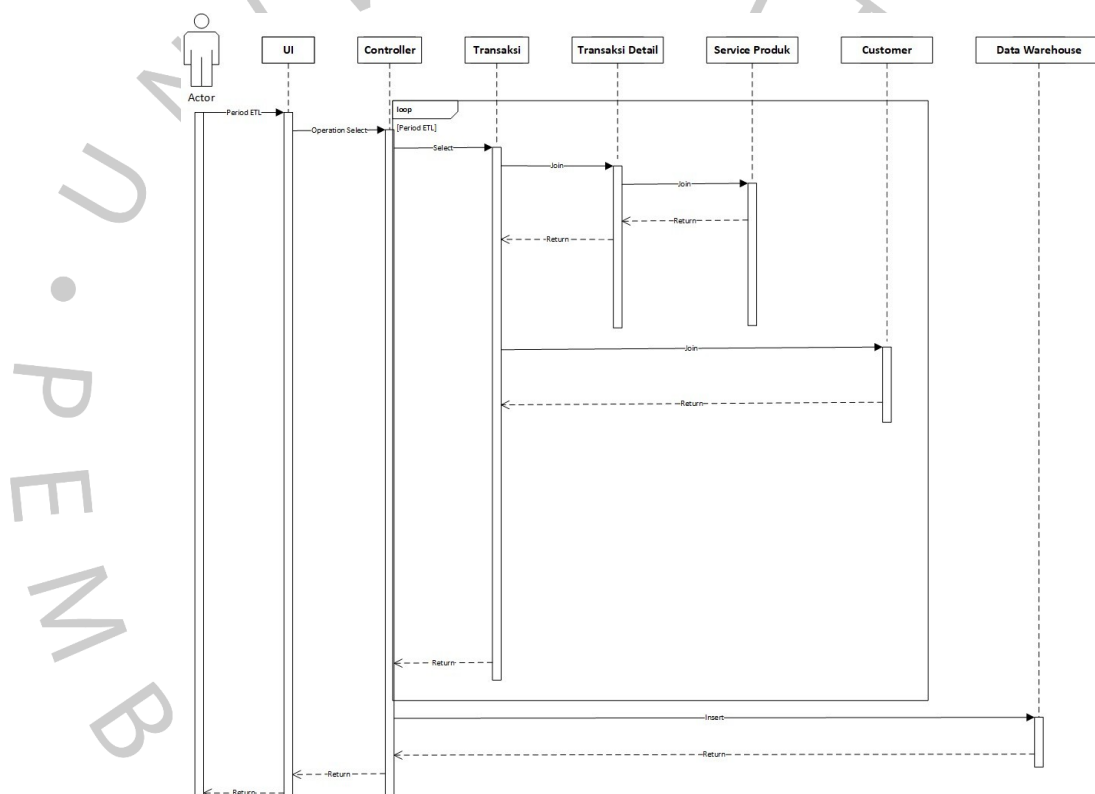


Gambar 4.4 Activity Diagram

4.2.4. Sequence Diagram

Pada penelitian ini *sequence* diagram digunakan untuk menjelaskan dan menampilkan interaksi antar objek-objek pada sistem secara terperinci. Pada prosesnya terdapat 1 (satu) aktor dan 7 (tujuh) objek yaitu UI, Controller, Transaksi, Transaksi Detail, *Service* Produk,

Customer, dan Data Warehouse. Pertama-tama aktor melakukan proses period ETL ke dalam tampilan UI, lalu aktor akan melakukan operation select di dalam sebuah controller, selanjutnya melakukan perulangan beserta melakukan join pada 4 (empat) tabel pada database yaitu, tabel transaksi, tabel transaksi detail, tabel *service* produk, dan tabel customer. Kemudian sistem akan menerima input tersebut dan menampilkan pesan jumlah data serta detail dari data-data yang dibutuhkan. Proses *sequence* diagram pada penelitian ini dapat dilihat pada gambar 4.5 yaitu sebagai berikut:



Gambar 4.5 Sequence Diagram

4.3 Perancangan Antar Muka Pengguna

Perancangan antar muka dalam penelitian ini berupa grafik visualisasi, yang ditampilkan berdasarkan hasil dari perhitungan yang sudah dilakukan. Dalam penerapannya grafik visualisasi tersebut dapat ditampilkan menggunakan sebuah *tools* yaitu *google collabs*. *Google collabs* adalah sebuah code editor yang berfungsi sebagai wadah atau tempat untuk mengimplementasikan syntax-syntax yang ingin ditampilkan, dalam hal ini syntax-syntax tersebut berfungsi untuk melakukan perhitungan terhadap data transaksi penjualan pada PT. Joanne Studio dan dapat memvisualisasikan hasil dari perhitungan tersebut berupa

grafik, yang tujuannya untuk memprediksi penjualan produk dimasa yang akan datang berdasarkan *history* data penjualan yang ada saat ini. Adapun bahasa yang digunakan pada penelitian ini adalah bahasa pemrogram python.

4.4 Perancangan Implementasi

Perancangan implementasi dalam penelitian ini merupakan tahapan dalam menerapkan algoritma KNN untuk melakukan perhitungan yang bertujuan untuk memprediksi penjualan dimasa yang akan datang. Adapun dalam penerapan-nya, dalam melakukan perhitungan menggunakan bahasa pemrograman python dan sebuah tools yaitu google collabs. Proses perancangan implementasi pada penelitian ini adalah sebagai berikut:

4.4.1. Import Library

Pada penelitian ini, langkah pertama dalam data mining adalah melakukan import library. Beberapa library yang digunakan meliputi:

1. Numpy

Numpy atau *numerical python* digunakan untuk komputasi matriks. *Numpy* memiliki kemampuan untuk membentuk objek N-dimensional *array*, yang mirip dengan *list* pada Python. Keunggulan *Numpy array* dibandingkan dengan *list* pada *python* adalah konsumsi *memory* yang lebih kecil serta *runtime* yang lebih cepat. *Numpy* juga memudahkan penelitian ini pada aljabar linear, terutama operasi pada Vector (1-d *array*) dan Matrix (2-d *array*).

2. Matplotlib

Matplotlib digunakan untuk memvisualisasikan data. *Matplotlib* dapat digunakan untuk memvisualisasikan data secara 2D maupun 3D dan menghasilkan gambar berkualitas, bahkan dapat disimpan dalam berbagai format gambar, seperti JPEG dan PNG.

3. Pandas

Pandas adalah sebuah *library* di *python* yang berlisensi BSD dan *open source* yang menyediakan struktur data dan analisis data yang mudah digunakan. *Pandas* biasa digunakan untuk membuat tabel, mengubah dimensi data, mengecek data, dan lain sebagainya. Struktur data dasar pada *pandas* dinamakan *Data Frame*, yang

memudahkan dalam membaca sebuah file dengan banyak jenis format seperti file .txt, .csv, dan .tsv. Fitur ini akan menjadikannya tabel dan juga dapat mengolah suatu data dengan menggunakan operasi seperti *join*, *distinct*, *group by*, agregasi, dan teknik lainnya yang terdapat pada SQL.

4. Sklearn

Merupakan sebuah module dari bahasa pemrograman *python* yang dibangun berdasarkan *numpy*, *scipy*, dan *matplotlib*.

```
[236] import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sb
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from matplotlib.colors import ListedColormap
from sklearn.metrics import f1_score
```

Gambar 4. 6 Import Library Python

4.4.2. Import Dataset

Tahap ini adalah melakukan import dan input dataset yang telah diseleksi ke dalam google collabs, dataset yang digunakan pada penelitian ini adalah data transaksi penjualan 6 bulan terakhir pada PT. Joanne Studio. Dataset tersebut berupa dokumen excel, adapun syntax dalam melakukan input dan menampilkan dataset, dapat dilihat pada gambar:

```

Input Dataset

[ ] dataset = pd.read_excel('Data_Transaksi.xlsx')

Menampilkan Dataset 5 teratas

[ ] dataset.head()

```

	SINGLE_LASH	CLEAN_LASH	DOUBLE_LASH	MEGA_VOLUME	BOTTOM_SINGLE	MANICURE_ONLY	MANICURE_SPA	PEDICURE_ONLY	PEDICURE_S
0	24	14	16	14	7	24	16	19	
1	14	11	15	8	4	29	3	7	
2	6	2	12	4	6	9	8	3	
3	17	8	14	18	7	25	17	2	
4	13	11	2	7	9	6	30	10	

```

[ ] dataset.duplicated().sum()
0

[ ] dataset.isnull().sum()

```

Gambar 4. 7 Import Dataset Transaksi Penjualan

4.4.3. Menentukan atau Membagi Data

Pada tahap ini adalah tahapan dimana dataset yang sudah terinput dibagi dan ditentukan ke dalam masing-masing variabel yaitu variabel x dan y. Dimana variabel x disini adalah variabel yang memuat atribut, sedangkan variabel y merupakan variabel yang memuat label. Pada penelitian ini data yang menjadi bagian dari variabel x adalah data yang berada pada index ke-1 sampai dengan index ke-20, dan data yang menjadi bagian dari variabel y adalah data yang berada pada index -1 atau data yang terakhir pada dataset. Perhitungan index pada dataset sama seperti array yang dimulai dengan index ke-0, sehingga kolom service name adalah index ke-0 yang tidak termasuk ke dalam bagian variabel x dan y, dikarenakan value dari kolom tersebut tidak berupa numeric. Seperti yang sudah dijelaskan sebelumnya dalam melakukan perhitungan menggunakan algoritma KNN, data hanya bisa diproses apabila berupa numeric atau angka. Penerapan dalam tahap ini dapat dilihat pada gambar:

```

[ ] x = dataset.drop(["Kategori 1", "Kategori 2", "Label"], axis=1)
    y = dataset["Label"]

```

Gambar 4. 8 Pembagian data input dan label

4.4.4. Menentukan Data Training dan Data Testing

Tahap ini adalah tahap membagi dataset kedalam bagian dari data training atau data latih dan data testing atau data uji. Langkah awal dalam tahap ini adalah membuat variabel baru yaitu `x_train`, `x_test`, `y_train`, dan `y_test` yang bertujuan untuk menampung *value* dari variabel `x` dan `y` dimana nantinya data tersebut menjadi data training dan data testing. Pada penelitian ini data testing yang digunakan sebesar 25% dari keseluruhan dataset yang ada. Proses yang dilakukan pada tahap ini dapat dilihat pada gambar:

```
[ ] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state = 0)
[ ] len(x_train)
3273
[ ] len(x)
4364
[ ] len(x_test)
1091
```

Gambar 4.9 Pembagian Data Training dan Data Testing

4.4.5. Scalling Dataset

Scaling data digunakan bertujuan agar jarak antar value data tidak terlalu jauh. Adapun pada proses scaling data dapat dilihat pada gambar 4.10 yaitu sebagai berikut:

```
[ ] myScaling = StandardScaler()
   x_train = myScaling.fit_transform(x_train)
   x_test = myScaling.transform(x_test)
```

```
▶ print(x_train)
```

```
[ ] print(x_test)
```

Gambar 4.10 Scaling Data

4.4.6. Penerapan Fungsi Algoritma KNN

Langkah awal pada tahap ini adalah melakukan import library `sklearn neighbors`. Kemudian menentukan nilai `k`, nilai `k` yang digunakan dalam penelitian ini adalah 42, selanjutnya pada bagian `metric` menggunakan `euclidean` yang digunakan untuk menghitung jarak

nilai yang terdekat. Proses tahap ini dapat dilihat pada gambar:

```
K = 42
classifier = KNeighborsClassifier(n_neighbors = K, metric = 'euclidean', p = 2)

[ ] classifier.fit(x_train, y_train)

KNeighborsClassifier(metric='euclidean', n_neighbors=42)

[ ] y_pred = classifier.predict(x_test)
```

Gambar 4.11 Menentukan nilai K dan rumus Euclidean

4.4.7. Menentukan Prediksi

Tahap ini merupakan tahap untuk melakukan tes pada prediksi yang telah dibuat, dengan melakukan input value data produk.

4.4.8. Evaluasi dan Validasi Prediksi

Setelah melakukan prediksi pada tahap sebelumnya, maka pada tahap ini digunakan untuk memeriksa seberapa akurat hasil dari data mining yang sudah dilakukan dengan cara mengevaluasi dan validasi. Adapun proses pada tahap ini dapat dilihat pada gambar 4.12. Dapat dijelaskan terdapat data yang bernilai benar 387 dan 641, sedangkan data yang salah sejumlah 16 data dan yang terakhir data yang kurang tepat sejumlah 47 data.

```
conf_matrix = confusion_matrix(y_test,y_pred)
print(conf_matrix)
print(f1_score(y_test,y_pred) * 100, '%')

[[387  47]
 [ 16 641]]
92.47311827956989 %
```

Gambar 4.12 Evaluasi dan Validasi Data

4.4.9. Klasifikasi Report

Tahap ini untuk melihat akurasi dari perhitungan secara detail yaitu mendapat skor akurasi sebesar 94%, dan nilai support sebesar 1091.

```
[ ] akurasi = classification_report(y_test, y_pred)
    print(akurasi)
```

	precision	recall	f1-score	support
1	0.96	0.89	0.92	434
2	0.93	0.98	0.95	657
accuracy			0.94	1091
macro avg	0.95	0.93	0.94	1091
weighted avg	0.94	0.94	0.94	1091

Gambar 4. 13 Classification Report

4.4.10. Akurasi Score

Agar lebih mudah dipahami mengenai tingkat keakuratan pada data mining yang dilakukan, maka pada tahap ini menggunakan akurasi *score*, tujuannya untuk melihat seberapa akurat hasil prediksi yang telah dilakukan. Adapun hasil dari prediksi yang telah dilakukan dapat dilihat pada gambar 4.14. Artinya prediksi terhadap tingkat penjualan produk dimasa yang akan datang mempunyai nilai keakuratan sebesar 94%, karena dari hasil evaluasi dan validasi terdapat 63 data yang kurang tepat.

```
akurasi = accuracy_score(y_test, y_pred)
print("Tingkat akurasi : %d persen"%(akurasi * 100))
```

Tingkat akurasi : 94 persen

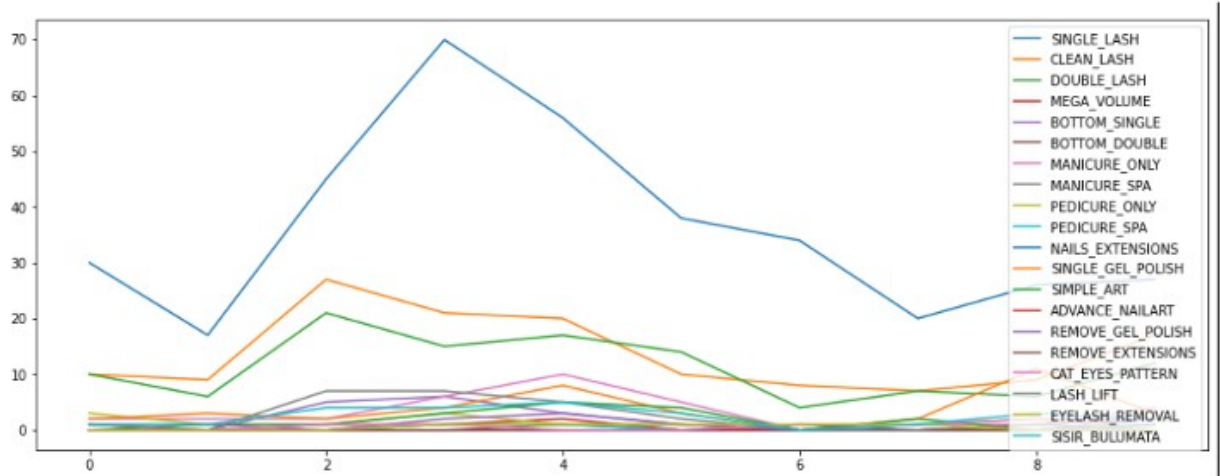
Gambar 4. 14 Accuracy score

4.4.11. Visualisasi Data (Grafik)

Pada tahap ini merupakan tahap menampilkan visualisasi data dari data mining yang telah dilakukan, visualisasi tersebut berupa grafik. Dimana sumbu X memuat data peringkat transaksi penjualan, dan dimana sumbu Y memuat data jumlah transaksi penjualan dari 20 produk. Adapun warna pada visualisasi adalah warna dari jenis produk yaitu sebagai berikut:

1. Biru: Single Lash
2. Orange: Clean Lash
3. Hijau: Double Lash
4. Merah: Mega Volume
5. Ungu: Bottom Single
6. Cokelat: Bottom Double

7. Merah Muda: Manicure Only
8. Abu-abu: Manicure Spa
9. Hijau Muda: Pedicure Only
10. Biru Muda: Pedicure Spa
11. Biru Tua: Nails Extensions
12. Kuning Tua: Single Gel Polish
13. Hijau Tua: Simple Art
14. Merah Tua: Advance Nailart
15. Ungu Muda: Remove Gel Polish
16. Cokelat Tua: Remove Extensions
17. Cokelat Muda: Cat Eyes Pattern
18. Hitam: Lash Lift
19. Hijau Muda: Eyelash Removal
20. Hijau Tosca: Sisir Bulu Mata



Gambar 4.15 Visualisasi Data

Pada grafik tersebut menunjukkan terdapat 3 dari 20 produk yang berpotensi banyak diminati oleh konsumen dimasa yang akan datang, produk-produk tersebut antara lain: single lash, clean lash, dan double lash.