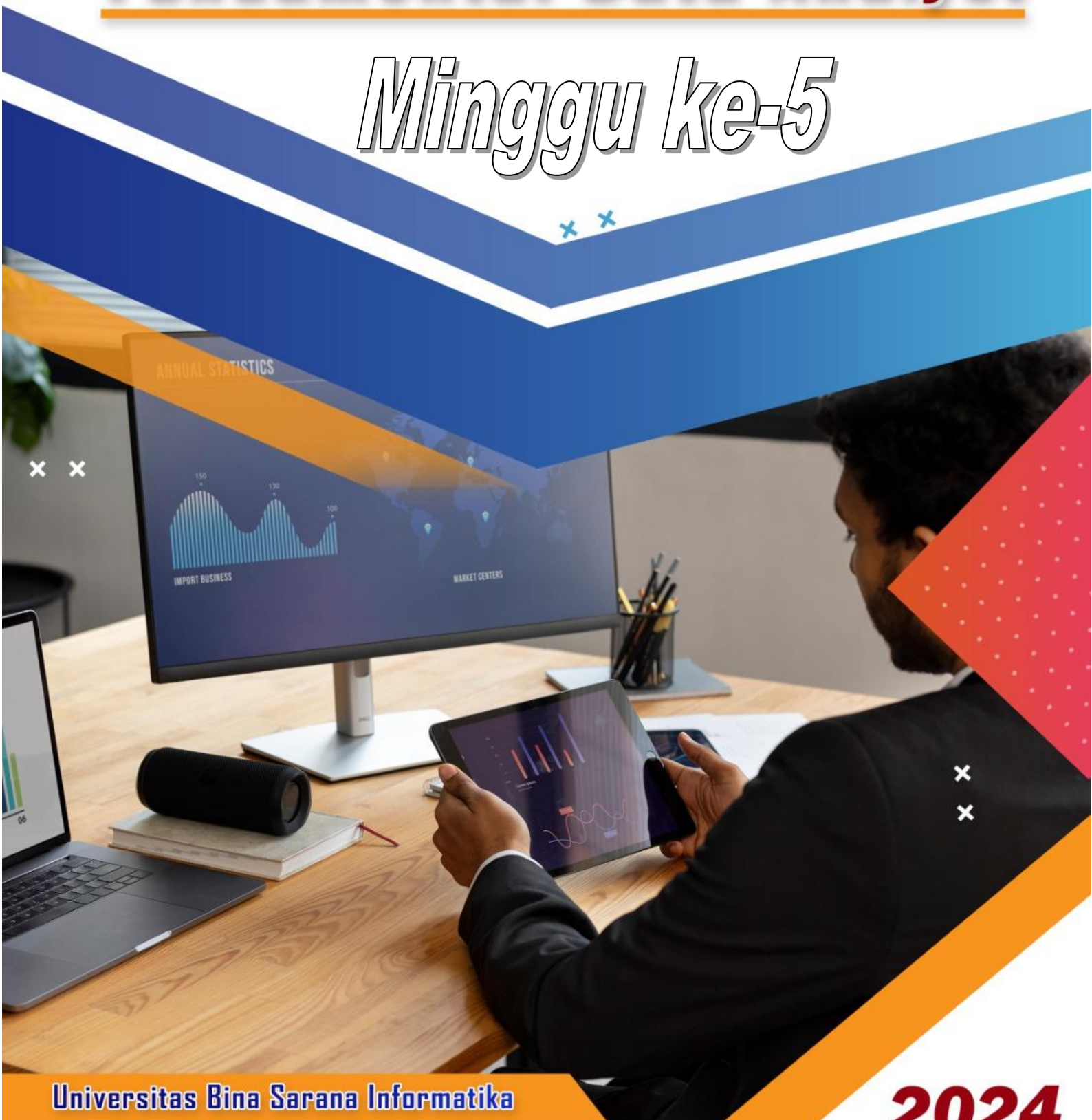
MODUL

Fundamental Data Analyst

Minggu ke-5



Universitas Bina Sarana Informatika

Sekolah Tinggi Informatika & Komputer Indonesia

2024

KATA PENGANTAR

Puji syukur Alhamdulillah, penulis panjatkan kehadirat Allah ta'ala, yang telah melimpahkan Rahmat dan Karunia-Nya sehingga pada akhirnya penulis dapat menyelesaikan modul ini dengan baik. Dimana modul ini penulis sajikan dalam bentuk modul yang sederhana. Adapun modul ini penulis buat untuk menambah wawasan para pembaca pada umumnya dan untuk menambah bahan materi untuk mata kuliah Fundamental Data Analyst bagi mahasiswa prodi Sistem Informasi Universitas Bina Sarana Informatika.

Sebagai bahan penulisan diambil berdasarkan pencarian di beberapa sumber, seperti buku, internet dan masih banyak lagi yang lainnya. Dalam modul ini menjelaskan materi Fundamental Data Analyst pertemuan 5 yang membahas tentang Supervised Learning (Linear Regression), CRISP-DM : Model Linear Regression dan Evaluasi. Penulis menyadari bahwa tanpa bimbingan dan dorongan dari semua pihak, maka penulisan dan pembuatan modul ini tidak akan berjalan dengan lancar.

Penulis mengucapkan terima kasih kepada tim sehingga bisa menyelesaikan penyusunan modul ini. Semoga modul ini berguna bagi para pembaca baik mahasiswa ataupun siapapun yang bisa dijadikan bahan referensi untuk pembelajaran.

Agustus 2024

Tim Penyusun

Unit Pengembangan Akademik
Program Studi Sistem Informasi

DAFTAR ISI

PEMBAHASAN

1. Regresi (Linear Regression)

Dalam konteks data mining, regresi digunakan untuk memodelkan dan menganalisis hubungan antara variabel-variabel dalam data, serta untuk membuat prediksi tentang nilai-nilai variabel tertentu berdasarkan variabel lainnya. Regresi dalam data mining memiliki beberapa aplikasi penting dan sering digunakan dalam berbagai domain bisnis, ilmiah, dan teknologi

Dalam regresi, output atau label adalah nilai kontinu. Misalnya, memprediksi harga rumah berdasarkan fitur-fitur seperti ukuran, lokasi, dan jumlah kamar, atau memprediksi suhu berdasarkan data historis dan lainnya

2. Peran Regresi dalam Data Mining

a. Prediksi

Regresi digunakan untuk memprediksi nilai variabel dependen berdasarkan satu atau lebih variabel independen. Misalnya, memprediksi harga rumah berdasarkan ukuran, lokasi, dan jumlah kamar.

b. Pemodelan Hubungan

Membantu memahami dan memodelkan hubungan antara variabel dalam dataset. Hal ini bisa digunakan untuk mengidentifikasi variabel mana yang memiliki pengaruh signifikan terhadap variabel target

c. Deteksi Anomali

Regresi dapat digunakan untuk mendeteksi outlier atau anomali dalam data dengan memeriksa nilai residual yang besar, yaitu perbedaan antara nilai yang diamati dan nilai yang diprediksi oleh model regresi.

d. Pengoptimalan

Dalam konteks bisnis, regresi bisa digunakan untuk mengoptimalkan proses dan pengambilan keputusan, seperti menentukan faktor-faktor yang paling berpengaruh terhadap penjualan atau profit.

3. Langkah-langkah dalam Melakukan Regresi pada Data Mining

a. Pengumpulan dan Pra-pemrosesan Data

Mengumpulkan data yang relevan dan membersihkannya dari noise atau missing values**.**

b. Eksplorasi Data

Menganalisis data untuk memahami distribusi, hubungan antar variabel, dan karakteristik data lainnya

c. Pemilihan Model

Memilih jenis model regresi yang sesuai dengan sifat data dan tujuan analisis

d. Pelatihan Model

Melatih model regresi pada data pelatihan untuk menentukan koefisien regresi.

e. Evaluasi Model

Mengevaluasi kinerja model menggunakan metrik seperti Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2) dan Mean Absolute Percentage Error (MAPE)

Langkah-Langkah diatas merupakan tahapan pada proses model CRISP DM yang telah dibahas sebelumnya.

4. Langkah-langkah dalam Melakukan Regresi pada Data Mining

Studi Kasus : House Prices
Dataset : https://s.id/DatasetFDA
Total Variable : 81 Variable
Sample Varible yang akan digunakan:

7 Variabel Prediktor : OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, FullBath, YearBuilt.

1 Variabel Target : SalePrice

5. CRISP-DM

A. Business Understanding

Langkah pertama ini berfokus pada pemahaman tujuan bisnis dan persyaratan proyek. Ini melibatkan:

a. Menentukantujuanbisnis dan kebutuhan yang spesifik.
b. Mengidentifikasi masalah bisnis yang ingin diselesaikan.
c. Menyusunrencanaproyek, termasuk sasaran, anggaran, dan waktu.

Tujuan Bisnis (Study Kasus: House Prices): Memprediksi harga rumah berdasarkan fitur yang dipilih.

B. Data Understanding

Langkah kedua melibatkan pengumpulan data dan familiarisasi dengan data yang tersedia. Hal Ini mencakup:

a. Mengumpulkandata awalyangdiperlukan untuk analisis.
b. Menjelajahi data untuk memahami strukturnya, kualitas, dan pola yang ada.
c. Mengidentifikasi masalah kualitas data seperti nilai yang hilang atau data yang tidak konsisten.

Studi Kasus	: House Prices
Dataset	: https://s.id/DatasetFDA
Deskripsi	: dataset house prices memiliki 81 fitur, diantaranya seperti OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, FullBath, YearBuilt dan lainnya, serta Sale Prices yang akan membantu peneliti misalkan membutuhkan sebagai target dalam penelitiannya
Prediktor	: OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, FullBath, YearBuilt
Target	: SalePrice

*Sumber Dataset: https://www.kaggle.com

Import library yang dibutuhkan

```
1   import pandas as pd
2   import numpy as np
3   from sklearn.model_selection import train_test_split

4   from sklearn.linear_model import LinearRegression
5   from sklearn.metrics import mean_squared_error,
    r2_score, mean_absolute_percentage_error

6   import matplotlib.pyplot as plt
```

Keterangan Library:

a. Manipulasi dan analisis data tabular dimjuat dalam bentuk dataframe
b. Manipulasi data dalam bentuk array
c. Bagian dari fungsi library sklearn yaitu untuk Membagi dataset ke dalam data training dan testing
d. Bagian dari fungsi library sklearn yaitu penggunaan model Linear Regression
e. Bagiandari fungsi library sklearn yaitu Evaluasi model
f. Library untuk visualisasi data/ grafik plot dalam python

Memuat Dataset

```
data = pd.read_csv('HousePriceTrain.csv')
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1460 non-null   int64
 1   MSSubClass     1460 non-null   int64
 2   MSZoning       1460 non-null   object
 3   LotFrontage    1201 non-null   float64
 4   LotArea        1460 non-null   int64
 5   Street         1460 non-null   object
 6   Alley          91 non-null     object
 7   LotShape       1460 non-null   object
 8   LandContour    1460 non-null   object
 9   Utilities      1460 non-null   object
 10  LotConfig      1460 non-null   object
 11  LandSlope      1460 non-null   object
 12  Neighborhood   1460 non-null   object
 13  Condition1     1460 non-null   object
 14  Condition2     1460 non-null   object
 15  BldgType       1460 non-null   object
 16  HouseStyle     1460 non-null   object
 17  OverallQual    1460 non-null   int64
 18  OverallCond    1460 non-null   int64
 19  YearBuilt      1460 non-null   int64
...
 79  SaleCondition  1460 non-null   object
 80  SalePrice      1460 non-null   int64
dtypes: float64(3), int64(35), object(43)
memory usage: 924.0+ KB
```

C. Data Preparation

Langkah ketiga ini melibatkan pembersihan dan transformasi data agar siap untuk pemodelan. Hal Ini termasuk :

1. Memeriksa apakah adanilai dalam dataset yang “Kosong”atang “NaN”

```
# Memeriksa apakah ada nilai dalam dataset yang "kosong" atau "NaN"
print(data.isnull().values.any())
```

```
Apabila ingin melakukan pemeriksaan data isnull perkolom gunakan
script berikut:
```

```
print(data.isnull().sum())
```

2. MengisiNilai yang Hlang Missing Values

Dalam kasus ini akan dilakukan pengisian data yang kosong untuk kolom dengan type numeric saja dan akan diisi dengan nilai rata-rata yang ada pada variable tersebut menggunakan script berikut :

```
data = data.fillna(data.mean(numeric_only=True))
print(data.isnull().sum())
```

3. MemilihFitur (Feature) dan Target

Seperti pada pertemuan sebelumnya Dari total 81 variable yang ada, dalam kasus ini akan menggunakan beberapa variable saja yang ditentukan untuk fitur dan target:

Fitur : ‘OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea', 'TotalBsmtSF', 'FullBath', 'YearBuilt'

Target : SalePrice

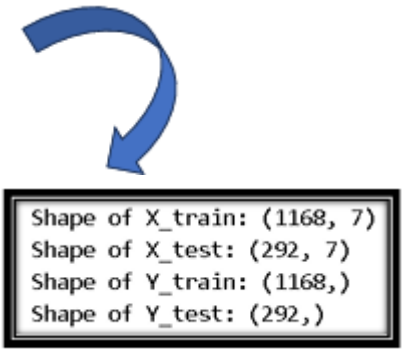
Script :

```
# Memilih fitur (features) dan target
features = ['OverallQual', 'GrLivArea', 'GarageCars','GarageArea',
'TotalBsmtSF', 'FullBath', 'YearBuilt’]
X = data[features]
y = data['SalePrice’]
```

4. Membagi dataset menjadi data pelatihan dan pengujian

```
# Bagi dataset menjadi data latih dan data uji
X_train, X_test, y_train, y_test = train_test_split(X,
```

```
y_binned, test_size=0.2, random_state=42)
print("Shape of X_train:", X_train.shape)
print("Shape of X_test:", X_test.shape)
print("Shape of Y_train:", y_train.shape)
print("Shape of Y_test:", y_test.shape)
```



Keterangan:

1. X_train → random state dibuat dalam 1 baris script
2. Dataset dibagi menjadi 20% Test dan 70%Train
3. Nilai random state memungkinkan untuk dirubah tergantu metode acak yang akan digunakan, contoh RS=0

D. Modelling

Langkah keempat adalah membangun model menggunakan teknik data mining. Hal Ini mencakup:

1. Memilih teknik pemodelan yang sesuai dengan masalah bisnis dan jenis data.
2. Melatih model menggunakandata pelatihan.

Dalam tahap pembelajaran ini algoritma yang akan digunakan adalah Regresi Linier

```
# Inisialisasi model
model = LinearRegression()
# Latih model menggunakan data latih
model.fit(X_train, y_train)
# Buat prediksi menggunakan data uji
y_pred = model.predict(X_test)
```

Keterangan:

y_pred : Variabel untuk menampung hasil melatih model

E. Evaluasi Model

Langkah kelima melibatkan evaluasi model untuk memastikan model memenuhi tujuan bisnis dan persyaratan proyek. Evaluasi yang digunakan pada algoritma Regresi berbeda dengan algoritma Decission Tree, Adapun Evaluasi model Regresi mencakup :

1. **Mean Squared Error (MSE) :** Rata-rata dari kuadrat selisih antara nilai aktual dan nilai prediksi. Fungsi MSE memperbesar kesalahan yang lebih besar karena adanya pemangkatan kuadrat, sehingga lebih sensitif terhadap outlier. Nilai Mean Squared Error (MSE) yang baik adalah yang mendekati 0 (nol).
2. **Root Mean Squared Error (RMSE) :** Akar kuadrat dari MSE, memberikan interpretasi dalam satuan yang sama dengan target. **Kegunaan**: Mengembalikan metrik kesalahan ke unit yang sama dengan target variabel, sehingga lebih mudah untuk diinterpretasikan. Semakin kecil nilai RMSE (mendekati 0), semakin akurat model dalam memprediksi nilai-nilai sebenarnya
3. **R-squared (R^2) :** Mengukur proporsi variansi dalam target yang dapat dijelaskan oleh fitur. Fungsi R^2 adalah untuk Mengukur seberapa baik model regresi linier cocok dengan data aktual. Nilai R^2 berkisar antara 0 dan 1, dengan nilai yang lebih tinggi menunjukkan model yang lebih baik.
4. **Mean Absolute Percentage Error (MAPE) :** mengukur kesalahan rata rata dalam bentuk persentase, memberikan gambaran tentang seberapa besar kesalahan prediksi dalam proporsi terhadap nilai sebenarnya. Semakin kecil nilai MAPE (mendekati 0), semakin kecil kesalahan prediksi-nya.
    a. Evaluasi Model Regresi: MAE, RMSE, R^2, MAPE

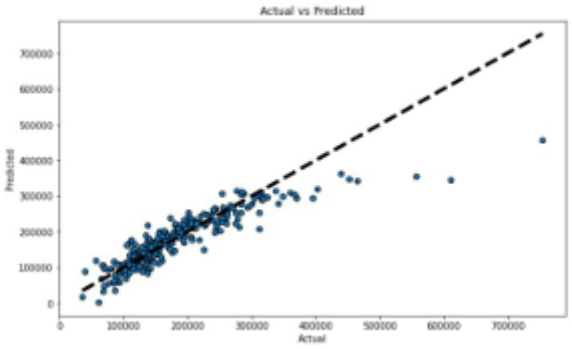
```
# Hitung metrik evaluasi
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)
mape = mean_absolute_percentage_error(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)
print("Root Mean Squared Error (RMSE):", rmse)
print("R-squared (R2):", r2)
print("Mean Absolute Percentage Error :", mape)
```

```
Mean Squared Error (MSE): 1572343803.577848
Root Mean Squared Error (RMSE): 39652.79061526248
R-squared (R2): 0.7950095261783579
Mean Absolute Percentage Error : 0.15160491195301382
```

    b. Visualisasi hasil prediksi vs. nilai actual

```
# Visualisasi hasil prediksi vs. nilai aktual
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, edgecolors=(0, 0, 0))
plt.plot([y_test.min(), y_test.max()], [y_test.min(),
```

```
y_test.max()], 'k--', lw=4)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted')
plt.show()
```



Catatan: Interpretasi Hasil Evaluasi Model dijelaskan pada Minggu Ke 9 Link Script: https://drive.google.com/file/d/1MJ33G2rscV2BB5ePFaqG6aiidmOG2OnM/view?usp=sharing

F. Deployment

Langkah terakhir adalah mengimplementasikan model ke dalam lingkungan operasional. Hal Ini mencakup:

1. Merencanakan dan menjalankan implementasi model dalam sistem produksi.
2. Memantau dan memelihara model untuk memastikan kinerjanya tetap optimal.
3. Mengkomunikasikan hasil dan manfaat model kepada pemangku kepentingan.

Dalam Perkuliahan ini tidak membahas tahap Deployment.

Penjelasan Tugas 2 Untuk penilaian CPMK-2

Langkah selanjutnya dalam menerapkan model pembelajaran pada dataset, mahasiswa diwajibkan membuat laporan terkait dengan dataset yang telah digunakan pada Tugas 1 dengan object laporan Tugas 2:

1. DataPreparation
2. Modelling

Adapun detail format penulisan pedoman Laporan dapat dilihat [disini] Waktu Pengumpulan Tugas 2: Minggu Ke- 8

REFERENSI

Steele, B., Chandler, J., Reddy, S. (2016). Algorithms for Data Science. Jerman: Springer International Publishing.

Abdussomad, dkk. (2021). Dasar Pemrograman Python. Yogyakarta: Teknosain.

Saeful Bahri, dkk (2019), Data mining : algoritma klasifikasi & penerapannya dalam aplikasi, Grha Ilmu

Segmentasi Pelanggan Menggunakan Python. (n.d.). (n.p.): Kreatif

Data Mining Menggunakan Android, Weka, dan SPSS. (2020). (n.p.): Airlangga University Press.

Abdussomad, A., Kurniawan, I., & Wibowo, A. (2023). Implementation of the Decission Tree Algorithm to Determine Credit Worthiness. *Compiler, 12*(2), 103-108