MODUL

Fundamental Data Analyst

Minggu ke-6

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 6 yang membahas tentang Unsupervised Learning (Clustering), CRISP-DM Model Kmeans 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

COVER i
KATA PENGANTAR ii
DAFTAR ISI iii
PEMBAHASAN 1
1. Unsupervised Learning (Clustering) 1
2. CRISP-DM: Model K-Means Clustering 1
A. Implementasi K-Means dengan Python 2
B. CRISP DM : Business Understanding 2
C. CRISP DM : Data Understanding 2
D. CRISP DM : Data Preparation 4
E. I CRISP DM : Modeling 6
F. CRISP DM : Evaluasi Model 6
G. CRISP DM : Deployment 8
REFERENSI 9

PEMBAHASAN

1. Unsupervised Learning (Clustering)

Clustering merupakan teknik dalam data mining yang digunakan untuk mengelompokkan sekumpulan objek dalam satu kelompok (cluster) yang memiliki kemiripan satu sama lain. Teknik ini sangat berguna dalam mengidentifikasi pola dan struktur dalam data yang tidak diketahui sebelumnya. Adapun tujuan clustering adalah:

a. Mengelompokkan data sehingga data dalam satu kelompok memiliki karakteristik yang mirip.
b. Membantu dalam menemukan struktur yang mendasar dalam data.
c. Digunakan untuk segmentasi data dalam berbagai aplikasi seperti pemasaran, pengelompokan pelanggan, analisis gambar, dan bioinformatika.

Sedangkan peran Clustering dalam Data Mining sebagai berikut:

a. Segmentasi Pasar dapat mengelompokkan pelanggan berdasarkan perilaku pembelian untuk membuat kampanye pemasaran yang ditargetkan.
b. Deteksi Anomali dapat mengidentifikasi pola yang berbeda secara signifikan dari yang lain dalam dataset.
c. Pengelompokan Dokumen dapat mengelompokkan dokumen teks berdasarkan topik yang sama.
d. Analisis Genom dapat mengelompokkan data genetik untuk menemukan keluarga gen yang serupa.

2. CRISP-DM: Model K-Means Clustering

Mempartisi data ke dalam k cluster dengan cara meminimalkan jarak kuadrat total antara titik data dan pusat cluster. Algoritma ini termasuk algoritma iteratif yaitu pilih pusat cluster awal, tetapkan titik data ke cluster terdekat, perbarui pusat cluster berdasarkan anggota cluster baru, dan ulangi sampai konvergen. Kelebihan dari algoritma ini cepat dan sederhana. Sedangkan untuk kekurangannya memerlukan jumlah cluster k yang ditentukan sebelumnya, sensitif terhadap outlier.

A. Implementasi K-Means dengan Python

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

B. CRISP DM : Business Understanding

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

a. Menentukan tujuan bisnis dan kebutuhan yang spesifik.
b. Mengidentifikasi masalah bisnis yang ingin diselesaikan.
c. Menyusun rencana proyek, termasuk sasaran, anggaran, dan waktu.

Tujuan Bisnis (Study Kasus: House Prices):

a. Mengetahui Kelompok Sebaran Fitur terhadap Target (SalePrice)
b. Hasil pengelompokkan tersebut bisa digunakan untuk penentuan penggunaan fitur terbaik dalam pengambilan Keputusan seperti
   Decision Tree.

C. CRISP DM : Data Understanding

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

a. Mengumpulkan data awal yang diperlukan 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

Import Library yang Diperlukan

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

4  from sklearn.cluster import KMeans
5  from sklearn.preprocessing import StandardScaler
6  from sklearn.metrics import silhouette_score,
   davies_bouldin_score, calinski_harabasz_score,
   adjusted_rand_score, normalized_mutual_info_score

7  from sklearn.metrics import silhouette_samples,
   silhouette_score
8  import matplotlib.pyplot as plt
9  import seaborn as sns
```

Keterangan Library :

1. Manipulasi dan analisis data tabular dimjuat dalam bentuk dataframe
2. Manipulasi data dalam bentuk array
3. Bagian dari fungsi library sklearn yaitu untuk Membagi dataset ke dalam data training dan testing
4. Bagian dari fungsi library sklearn yaitu penggunaan model Clustering Kmeans
5. Bagian dari fungsi library sklearn untuk melakukan standardisasi fitur-fitur dalam dataset
6. Bagian dari fungsi library sklearn untuk mengevaluasi kinerja algoritma clustering
7. Bagian dari fungsi library sklearn untuk Menghitung nilai Silhouette pada setiap sampel dalam dataset dan untuk Menghitung nilai rata-rata Silhouette untuk seluruh dataset

8. Library untuk visualisasi data/ grafik plot dalam python
9. Library visualisasi data yang dibangun di atas Matplotlib dan menyediakan antarmuka tingkat tinggi untuk menggambar grafik statistik yang menarik dan informatif.

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
```

D. CRISP DM : Data Preparation

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

a. Memeriksa apakah ada nilai 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())
```

b. Mengisi Nilai yang Hilang 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())
```

c. Memilih Fitur (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’]
```

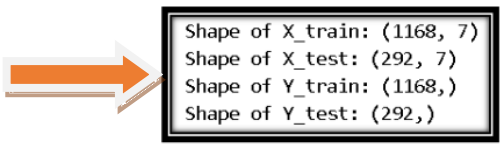
d. Standarisasi fitur

Dalam kasus ini akan menggunakan StandardScaler() dalam konteks clustering untuk menstandarisasi fitur-fitur dalam dataset sebelum melakukan clustering. Standardisasi sangat penting dalam clustering karena banyak algoritma clustering (seperti K-Means) sensitif terhadap skala fitur. Dengan menggunakan StandardScaler, Anda memastikan bahwa semua fitur memiliki skala yang sama sebelum melakukan clustering, yang dapat meningkatkan akurasi dan kinerja dari algoritma clustering.

```
# Standarisasi fitur [lihat materi minggu ke-1]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

e. 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

E. CRISP DM : Modeling

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

a. Memilih teknik pemodelan yang sesuai dengan masalah bisnis dan jenis data.

b. Melatih model menggunakan data pelatihan.

Dalam tahap pembelajaran ini algoritma yang akan digunakan adalah Clustering (Kmeans).

```
# Inisialisasi model K-Means dengan jumlah cluster
kmeans = KMeans(n_clusters=4, random_state=42)
# Latih model menggunakan data
kmeans.fit(X_scaled)
# Dapatkan label cluster
labels = kmeans.labels_
```

Keterangan:

y_pred : Variabel untuk menampung hasil melatih model

F. CRISP DM : Evaluasi Model

Langkah kelima melibatkan evaluasi model untuk memastikan model memenuhi tujuan bisnis dan persyaratan proyek, Adapun Evaluasi model Clustering mencakup :

a. Silhouette Score

b. Davies-Bouldin Index

c. Calinski-Harabasz Index

d. Adjusted Rand Index

e. Normalized Mutual Information

f. Scatter Plot (Visualisasi)

Visualisasi hasil clustering


```
# Tambahkan label cluster ke data asli
data['Cluster'] = labels
# Visualisasi hasil clustering
plt.figure(figsize=(10, 6))
sns.scatterplot(data=data, x='GrLivArea', y='SalePrice',
hue='Cluster', palette='viridis')
plt.title('K-Means Clustering of House Prices')
plt.show()
```

Evaluasi Metric (Silhouette, Davies-Bouldin, Calinski, ARI & NMI)

```
# Calculate evaluation metrics
sil_score = silhouette_score(X_scaled, labels)
db_index = davies_bouldin_score(X_scaled, labels)
ch_index = calinski_harabasz_score(X_scaled, labels)
ari = adjusted_rand_score(y, labels)
nmi = normalized_mutual_info_score(y, labels)

# Print evaluation metrics
print("Silhouette Score: ", sil_score)
print("Davies-Bouldin Index:",db_index )
print("Calinski-Harabasz Index:",ch_index )
print("Adjusted Rand Index: ",ari )
print("Normalized Mutual Information: ",nmi )
```

```
Silhouette Score:  0.26525371019478705
Davies-Bouldin Index: 1.4472384245093164
Calinski-Harabasz Index: 674.6522184471318
Adjusted Rand Index:  0.0027789357904749843
Normalized Mutual Information:  0.2013225052945592
```

Catatan:

Interpretasi Hasil Evaluasi Model dijelaskan pada Minggu Ke 9

Visualisasi Sebaran kluster Feature terhadap SalePrice

```
# Tambahkan hasil klaster dan Silhouette Scores ke dalam DataFrame
data['Cluster'] = labels
# features = ['OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea',
'TotalBsmtSF', 'FullBath', 'YearBuilt']
fig, axes = plt.subplots(2, 4, figsize=(30, 10))
fig.suptitle('Visualisasi Sebaran kluster Feature terhadap SalePrice')

sns.scatterplot(ax=axes[0, 0], x='OverallQual', y='SalePrice', hue='Cluster', data=data,palette='Set2')
sns.scatterplot(ax=axes[0, 1], x='GrLivArea', y='SalePrice', hue='Cluster', data=data, palette='Set2')
sns.scatterplot(ax=axes[0, 2], x='GarageCars', y='SalePrice', hue='Cluster', data=data, palette='Set2')
sns.scatterplot(ax=axes[1, 3], x='GarageArea', y='SalePrice', hue='Cluster', data=data, palette='Set2')
sns.scatterplot(ax=axes[1, 0], x='TotalBsmtSF', y='SalePrice', hue='Cluster', data=data,palette='Set2')
sns.scatterplot(ax=axes[1, 1], x='FullBath', y='SalePrice', hue='Cluster', data=data, palette='Set2')
sns.scatterplot(ax=axes[1, 2], x='YearBuilt', y='SalePrice', hue='Cluster', data=data, palette='Set2')
```

Detail Hasil Visualisasi Langsung di Text Editor

Catatan:

Interpretasi Hasil Evaluasi Model dijelaskan pada Minggu Ke 9

Link full script:

https://drive.google.com/file/d/1QyDH2thyjtecRh3sdAj6QYwmJ_vPetKt/view?usp=sharing

G. CRISP DM : Deployment

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

a. Merencanakan dan menjalankan implementasi model dalam sistem produksi.

b. Memantau dan memelihara model untuk memastikan kinerjanya tetap optimal.

c. Mengkomunikasikan hasil dan manfaat model kepada pemangku kepentingan.

d. Dalam Perkuliahan ini tidak membahas tahap Deployment

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