Praktikum 2
Pemrosesan Awal Data

Tujuan

Setelah mengikuti praktikum ini, mahasiswa diharapkan mampu:

1. Mengetahui permasalahan-permasalahan yang mungkin terdapat pada data mentah.
2. Mengimplementasikan teknik imputasi data.
3. Mengimplementasikan teknik normalisasi data.

Dasar Teori

Data yang tersedia mungkin memiliki beberapa permasalahan sehingga tidak dapat langsung digunakan oleh metode pembelajaran mesin (machine learning). Permasalahan-permasalahan yang sering dijumpai pada data, diantaranya:

1. Missing value

Missing value merupakan kondisi dimana sebuah data memiliki nilai yang hilang pada satu atau beberapa fiturnya. Hilangnya nilai pada suatu fitur tertentu pada sebuah record (baris) data menyebabkan record tersebut tidak dapat digunakan untuk proses pembelajaran mesin.

2. Rentang data tidak sama

Sebuah data pada umumnya akan memiliki rentang data yang berbeda-beda pada setiap fiturnya. Hal ini dapat menimbulkan efek negatif pada metode pembelajaran mesin yang menggunakannya. Permasalahan timbul ketika suatu metode pembelajaran mesin menggunakan perhitungan jarak pada algoritmanya, karena rentang data yang lebih lebar akan mendominasi perhitungan jarak. Dengan kata lain, fitur yang memiliki rentang yang lebih sempit menjadi tidak terlalu berguna

Permasalahan pada data dapat diatasi pada tahapan preprocessing atau pemrosesan awal. Pada praktikum ini Anda akan mempelajari dua metode preprocessing, yaitu imputasi dan normalisasi. Imputasi digunakan untuk mengatasi permasalahan missing value. Terdapat beberapa teknik untuk *mengganti* nilai yang hilang. Salah satunya adalah dengan mengisi nilai kosong dengan rata-rata data pada fitur dan kelas yang sama.

Normalisasi digunakan untuk mengatasi permasalahan rentang data yang berbeda-beda. Agar rentang data pada masing-masing fitur menjadi sama atau hampir sama, dilakukanlah normalisasi. Terdapat dua metode normalisasi yang Anda pelajari

pada praktikum ini, yaitu normalisasi MinMax dan Normalisasi Zscore. Persamaan normalisasi MinMax dan Zscore adalah sebagai berikut:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \qquad z = \frac{x - \mu}{\sigma}$$

Keterangan

x'	: Nilai ternormalisasi menggunakan minmax
x	: Nilai asal
$min(x)$	: Nilai minimum pada suatu fitur
$max(x)$	: Nilai maksimum pada suatu fitur
z	: Nilai ternormalisasi menggunakan z score
μ	: Rata-rata pada suatu fitur
σ	: Standar deviasi pada suatu fitur

Perlu Anda perhatikan bahwa perhitungan normalisasi dilakukan per fitur. Dengan demikian, nilai min, max, rata-rata, dan standar deviasi dihitung per fitur, bukan keseluruhan data.

Praktikum

1. **Import Data**

Unduh dataset yang akan digunakan pada praktikum kali ini. Anda dapat menggunakan aplikasi **wget** untuk mendowload dataset dan menyimpannya dalam Google Colab. Jalankan cell di bawah ini untuk mengunduh dataset

```
! wget https://dataset-ppm.s3.amazonaws.com/iris_missing.csv
```

Setelah dataset berhasil diunduh, langkah berikutnya adalah membaca dataset dengan memanfaatkan fungsi readcsv dari library pandas. Lakukan pembacaan berkas csv menggunakan fungsi **readcsv**. Jangan lupa untuk melakukan import library pandas terlebih dahulu

```
import pandas as pd
import numpy as np
data = pd.read_csv('iris_missing.csv')
```

Tampilkan beberapa baris dari dataset untuk mendapatkan informasi singkat mengenai isi data. Gunakan fungsi **head()** untuk menampilkan 5 data pertama.

```
data.head()
```

Berdasarkan informasi dari fungsi head(), data iris yang digunakan mempunyai 4 fitur sebagai berikut :

1. sepal length
2. sepal width
3. petal length
4. petal width

2. MIssing Value dan Imputasi Data

Jika Anda perhatikan dengan seksama, data pada baris ke-3 (index 2) pada fitur sepal length memiliki nilai 0.0. Hal ini menandakan adanya missing value pada data. Jalankan cell di bawah ini untuk mendapatkan semua data yang mengandung missing value. Pencarian data yang mengandung missing value dilakukan dengan tahapan sebagai berikut:

1. Membuat filter untuk mencari data dengan sepal_length = 0, sepal_width = 0, petal_length = 0, petal_width =0
2. Mencari data yang memenuhi kondisi1 **atau** kondisi2 **atau** kondisi3 **atau** kondisi4 menggunakan property loc pada dataframe

```
kondisi1 = data['sepal_length']==0.0
kondisi2 = data['sepal_width']==0.0
kondisi3 = data['petal_length']==0.0
kondisi4 = data['petal_width']==0.0
data.loc[kondisi1 | kondisi2 | kondisi3 | kondisi4]
```

Penanganan missing value pada Pandas akan lebih mudah apabila data yang hilang (bernilai 0.0) diganti dengan NaN (Not A Number). Gunakan properti **replace** pada dataframe untuk menggganti 0.0 menjadi NaN

```
data = data.replace(0.0,np.NaN)
```

Terdapat beberapa cara untuk mengatasi permasalahan missing value pada data. Salah satu cara yang sederhana adalah dengan mengganti nilai NaN pada suatu fitur dengan rata-rata nilai fitur tersebut pada data lain yang bernilai bukan Nan. Perhatikan bahwa Anda harus mengganti nilai NaN dengan rata-rata data lain yang memiliki kategori (species) yang sama.

Fungsi-fungsi pada Pandas yang dapat Anda manfaatkan:

1. **transform** untuk mengaplikasikan fungsi tertentu pada dataframe, pada permasalahan ini fungsi yang digunakan adalah fungsi **mean**
2. **groupby** untuk mengelompokkan dataframe berdasarkan nilai kolom tertentu, pada permasalahan ini kolom yang digunakan adalah **species**

3. **fillna** untuk mengganti nilai NaN dengan nilai yang telah ditentukan

Buatlah fungsi bernama **imputasi** yang melakukan tahapan berikut:

1. Menghitung rata-rata masing-masing kolom berdasarkan kelasnya
2. Mengisi kolom yang berisi NaN dengan rata-rata kelas yang sesuai

```
def imputasi(df_input):
  list_columns = df_input.columns #mendapatkan daftar kolom pada
dataframe
  class_column = list_columns[-1] #kolom terakhir merupakan kolom
kelas
  for column in list_columns[:-1]:
    df_input[column] =
df_input[column].fillna(df_input.groupby(class_column)[column].tra
nsform('mean')) # Penggantian nilai NaN dilakukan per kolom
  return df_input
```

Buatlah sebuah dataframe baru bernama **data_imputasi** yang berisi dataset dengan nilai NaN yang sudah diganti dengan cara memanggil fungsi **imputasi**.

```
data_imputasi = imputasi(data)
```

Cek apakah masih terdapat nilai NaN pada dataframe.

```
data_imputasi.isnull().values.any()
```

3. Normalisasi MinMax

Normalisasi bertujuan menyamakan rentang nilai pada setiap fitur. Beberapa metode pembelajaran mesin memiliki kinerja yang buruk apabila rentang nilai tiap variabel berbeda jauh.

Sebelum melakukan normalisasi, cek terlebih dahulu rentang (nilai max - nilai min) pada masing masing fitur

```
def cetak_rentang(df_input):
  list_fitur = df_input.columns[:-1]#mengambil nama kolom, kecuali
yang terakhir (kelas)
  for fitur in list_fitur:
    max = df_input[fitur].max()
    min = df_input[fitur].min()
    print("Rentang fitur ",fitur," adalah ",max-min)


cetak_rentang(data_imputasi)
```

Terlihat bahwa masing-masing fitur memiliki rentang yang berbeda, meskipun tidak terlalu signifikan. Pada praktikum ini Anda akan mengimplementasikan normalisasi MinMax untuk menyamakan rentang setiap fitur menjadi satu (1). Persamaan dari fungsi normalisasi MinMax adalah sebagai berikut:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

Implementasikan metode normalisasi MinMax pada fungsi bernama **minmax**.

```
def minmax(df_input):
  list_fitur = df_input.columns[:-1]
  for fitur in list_fitur:
    max = df_input[fitur].max()
    min = df_input[fitur].min()
    df_input[fitur] = (df_input[fitur]-min)/(max-min)
  return df_input
```

Buatlah sebuah dataframe baru bernama **data_normal** yang berisi hasil dari metode minmax dengan input **data_imputasi**

```
data_normal = minmax(data_imputasi)
```

Cek 5 baris pertama **data_normal**

```
data_normal.head()
```

Tampilkan rentang masing-masing fitur menggunakan fungsi **cetak_rentang** yang telah dibuat

```
cetak_rentang(data_normal)
```

Tugas

1. Implementasikan metode normalisasi Z-score dengan cara membuat fungsi bernama **zscore**.
2. Normalisasikan dataframe **data_imputasi** menggunakan fungsi **zscore**. Simpan hasilnya pada dataframe bernama **data_zscore**.
3. Jelaskan perbedaan hasil normalisasi **MinMax** dan **ZScore**. Petunjuk : cek nilai rentang, rata-rata dan standar deviasi dari **data_zscore**.

Petunjuk pengerjaan soal:

1. Klik link berikut, pastikan Anda login menggunakan **akun student UB**. https://colab.research.google.com/drive/1hflrZnvQCXMBkXqZkRd0OEIxVZvtMUAu?usp=sharing
2. Klik tombol Copy to Drive

3. Beri nama file Praktikum 1 - Nama - NIM
4. Isilah cell yang kosong
5. Download file *.ipynb dengan cara klik **File -> Download .ipynb**
6. Kumpulkan file *.ipynb ke asisten