LAPORAN UAS PENGOLAHAN DATASHEET KESEHATAN JANTUNG
BIG DATA & DATA MINING

DISUSUN OLEH :

1. **Muhammad Ikram** **5190411092**
2. **Muhammad Abdul Latif** **5190411109**
3. **Vikry Adiyanto** **5190411084**
4. **Arif Ramadhan G** **5190411193**

KELAS F

PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS & TEKNOLOGI
UNIVERSITAS TEKNOLOGI YOGYAKARTA
2021/2022

Soal

1. Lakukan exploratory data analysis terhadap data tersebut.
2. Buatlah model machine learning berdasarkan data tersebut menggunakan Spark. Boleh berupa klasifikasi maupun klastering.
3. Jelaskan proses yang Anda kerjakan pada laporan tertulis. Kumpulkan dalam bentuk file PDF maksimal pada saat ujian di e-learning.
4. Buatlah video presentasi singkat tentang kerjaan Anda. Kumpulkan URL Link dari video Anda di e-learning.

Jawaban

Menampilkan Dataset heart_2020_cleaned.crv dalam bentuk dataframe

tentang dataset

Indikator Kunci Penyakit Jantung Data survei CDC tahunan 2020 dari 400 ribu orang dewasa terkait dengan status kesehatan mereka. Topik apa yang dicakup oleh dataset? Menurut CDC, penyakit jantung adalah salah satu penyebab utama kematian bagi sebagian besar ras di AS (Afrika-Amerika, Indian Amerika dan Penduduk Asli Alaska, dan orang kulit putih). Sekitar setengah dari semua orang Amerika (47%) memiliki setidaknya 1 dari 3 faktor risiko utama penyakit jantung: tekanan darah tinggi, kolesterol tinggi, dan merokok. Indikator kunci lainnya termasuk status diabetes, obesitas (BMI tinggi), tidak cukup aktivitas fisik atau minum terlalu banyak alkohol. Mendeteksi dan mencegah faktor-faktor yang memiliki dampak terbesar pada penyakit jantung sangat penting dalam perawatan kesehatan. Perkembangan komputasi, pada gilirannya, memungkinkan penerapan metode pembelajaran mesin untuk mendeteksi "pola" dari data yang dapat memprediksi kondisi pasien.

1. Explorasi data dan analysis

Import Modules

In [ ]:

```
import pandas as pd
```

load dataset

Import dataset untuk digunakan sebagai bahan visualisasi

In [ ]:

```
dataset_df = pd.read_csv('heart_2020_cleaned.csv')
dataset_df.head()
```

Out[ ]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	Diff
0	No	16.60	Yes	No	No	3.0	30.0	
1	No	20.34	No	No	Yes	0.0	0.0	
2	No	26.58	Yes	No	No	20.0	30.0	
3	No	24.21	No	No	No	0.0	0.0	
4	No	23.71	No	No	No	28.0	0.0	

identity the shape of the dataset

Mencari hasil atau total dari kolom dan record dataset yang digunakan

In [ ]:

```
dataset_df.shape
```

Out[ ]:

```
(306236, 18)
```

Get the list of columns

digunakan untuk mengambil judul dari kolom yang digunakan di dataset tersebut

In [ ]:

```
dataset_df.columns
```

Out[ ]:

```
Index(['HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'Stroke',
       'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategor
y',
       'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime',
       'Asthma', 'KidneyDisease', 'SkinCancer'],
      dtype='object')
```

identyty data types for each column

proses pencarian indentity atau jenis variable yang digunakan pada tabel dataset yang digunakan

In [ ]:

```
dataset_df.dtypes
```

Out[ ]:

```
HeartDisease         object
BMI                 float64
Smoking              object
AlcoholDrinking      object
Stroke               object
PhysicalHealth      float64
MentalHealth        float64
DiffWalking          object
Sex                  object
AgeCategory          object
Race                 object
Diabetic             object
PhysicalActivity     object
GenHealth            object
SleepTime           float64
Asthma               object
KidneyDisease        object
SkinCancer           object
dtype: object
```

getbasic dataset information

proses untuk melihat sebgaian informasi data yang ada di dalam record dataset

In [ ]:

```
dataset_df.info
```

Out[ ]:

```
<bound method DataFrame.info of        HeartDisease    BMI Smoking Alcohol
Drinking Stroke  PhysicalHealth  \
0                No  16.60     Yes             No     No             3.0
1                No  20.34      No             No    Yes             0.0
2                No  26.58     Yes             No     No            20.0
3                No  24.21      No             No     No             0.0
4                No  23.71      No             No     No            28.0
...             ...    ...     ...            ...    ...             ...
306231           No  22.67     Yes             No    Yes             0.0
306232           No  21.97      No             No     No             0.0
306233           No  22.40      No             No     No             0.0
306234          Yes  28.82      No             No     No             0.0
306235           No  27.60      No              N    NaN             NaN

        MentalHealth DiffWalking     Sex  AgeCategory   Race Diabetic  \
0               30.0          No  Female        55-59  White      Yes
1                0.0          No  Female  80 or older  White       No
2               30.0          No    Male        65-69  White      Yes
3                0.0          No  Female        75-79  White       No
4                0.0         Yes  Female        40-44  White       No
...              ...         ...     ...          ...    ...      ...
306231           0.0          No  Female        70-74  White       No
306232           2.0          No  Female        75-79  White       No
306233           0.0          No  Female        65-69  White       No
306234           7.0         Yes  Female        70-74  White       No
306235           NaN         NaN     NaN          NaN    NaN      NaN

       PhysicalActivity  GenHealth  SleepTime Asthma KidneyDisease SkinCan
cer
0                   Yes  Very good        5.0    Yes            No
Yes
1                   Yes  Very good        7.0     No            No
No
2                   Yes       Fair        8.0    Yes            No
No
3                    No       Good        6.0     No            No
Yes
4                   Yes  Very good        8.0     No            No
No
...                 ...        ...        ...    ...           ...
...
306231              Yes  Excellent        7.0     No            No
No
306232              Yes  Very good        6.0     No            No
No
306233              Yes  Very good        7.0     No            No
No
306234               No       Good        6.0     No            No
No
306235              NaN        NaN        NaN    NaN           NaN
NaN

[306236 rows x 18 columns]>
```

identiti missing value

In [ ]:

```
dataset_df.isna().values.any()
```

Out[ ]:

```
True
```

identify duplicate entries/rows

proses pencarian data atau record yang terdapat duplikasi di dalam tabel atau record dataset

In [ ]:

```
dataset_df[dataset_df.duplicated()]
```

Out[ ]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth
2182	No	19.85	No	No	No	0.0	0.0
3182	No	28.19	No	No	No	0.0	0.0
3397	No	26.54	No	No	No	0.0	0.0
3650	No	32.89	Yes	No	No	2.0	1.0
4061	No	25.84	No	No	No	0.0	0.0
...	...	...	...	...	...	...	...
306192	No	29.84	Yes	No	No	0.0	0.0
306198	No	23.63	No	No	No	0.0	0.0
306204	No	29.26	No	No	No	0.0	0.0
306221	No	22.60	No	No	No	0.0	0.0
306228	No	22.24	Yes	No	No	0.0	0.0

17064 rows × 18 columns

di bawah ini hasil dari pencarian dataset atau record yang memiliki duplikasi yaitu dengan jumlah seperti di bawah

In [ ]:

```
dataset_df.duplicated().value_counts()
```

Out[ ]:

```
False    289172
True      17064
dtype: int64
```

Drop duplicated enntries/rows

setelah didapat hasil reecord yang terdapat duplikasi maka untuk mempersingkat data atau mengefisienkan data maka akan dilakukan penghapusan data yang terdapat duplikasi tersebut

In [ ]:

```
dataset_df.drop_duplicates(inplace=True)
dataset_df.shape
```

Out[ ]:

```
(289172, 18)
```

Describe the dataset

proses untuk memeriksa tipedata yang digunakan pada beberapa kolom yang terdapat pada dataset, di bawah ini tertara ada meas, std, min, dll.

pengecekan tipe data ini dilakukan supaya nantinya dimudahkan ketika akan memvisualisasikan data sesuai dengan tipe datanya.

In [ ]:

```
dataset_df.describe()
```

Out[ ]:

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	289172.000000	289171.000000	289171.000000	289171.000000
mean	28.439439	3.579325	4.145509	7.088398
std	6.479322	8.143421	8.138156	1.466211
min	12.020000	0.000000	0.000000	1.000000
25%	24.020000	0.000000	0.000000	6.000000
50%	27.400000	0.000000	0.000000	7.000000
75%	31.650000	2.000000	4.000000	8.000000
max	94.850000	30.000000	30.000000	24.000000

Correlation Matrix

correlatioon matrix yang menunjukan korelasi sederhana antara kemungkinan variabel yang nantinya akan dilibatkan dalam visualisai data

In [ ]:

```
dataset_df.corr()
```

Out[ ]:

	BMI	PhysicalHealth	MentalHealth	SleepTime
BMI	1.000000	0.104680	0.056757	-0.048225
PhysicalHealth	0.104680	1.000000	0.279828	-0.056900
MentalHealth	0.056757	0.279828	1.000000	-0.116706
SleepTime	-0.048225	-0.056900	-0.116706	1.000000

Iris Dataset:Data Visualisation

Import Module

In [ ]:

```
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

Heatmap

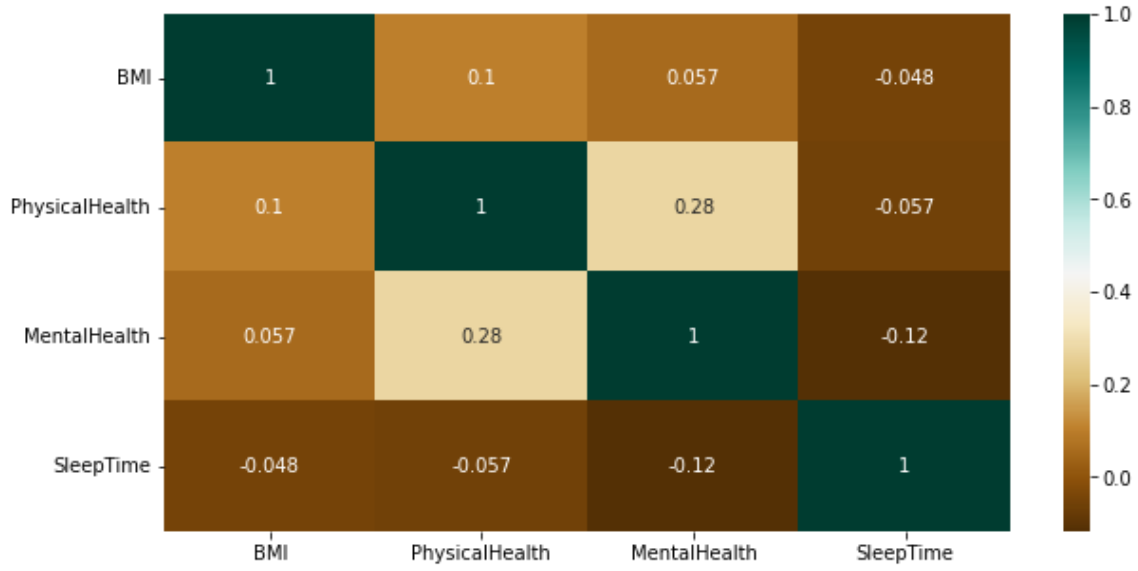
visualisasi heatmap biasanya pada heatmap semakin tinggi angka suatu kelompok data maka warnanya akan semakin gelap di sini disimbolkan dengan warna berwarna hijau tua.

In [ ]:

```
plt.figure(figsize=(10,5))
c= dataset_df.corr()
sns.heatmap(c,cmap="BrBG",annot=True)
c
```

Out[ ]:

	BMI	PhysicalHealth	MentalHealth	SleepTime
BMI	1.000000	0.104680	0.056757	-0.048225
PhysicalHealth	0.104680	1.000000	0.279828	-0.056900
MentalHealth	0.056757	0.279828	1.000000	-0.116706
SleepTime	-0.048225	-0.056900	-0.116706	1.000000



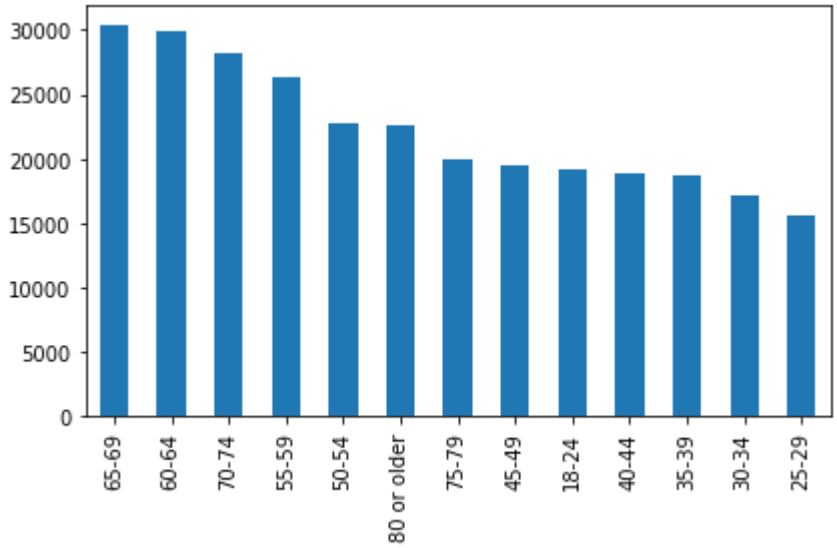
Bar Plot

Pada Bar plot kami membuat visualisasi range umur yang melakukan pencucian hati/jantung, agar dapat melihat rata-rata pasien yang melakukan tersebut paling banyak di range umur berapa.

Dilihat dari visualisasi di bawah terdapat kurang lebih 30.000 pasien yang melakukan pencucian hati/jantung berada di range umur 60-69 pada dua chart teratas terdapat perbedaan yang sedikit yang dapat disimpulkan bahwa pada umur 60-69 tahun memiliki jumlah pasien terbanyak yang melakukan pencucian hati/jantung.

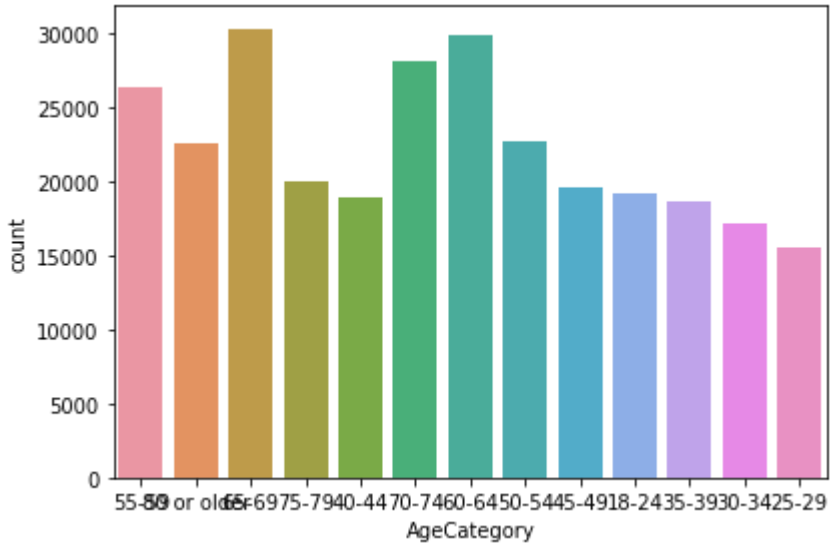
In [ ]:

```
dataset_df['AgeCategory'].value_counts().plot.bar()
plt.tight_layout()
plt.show()
```



In [ ]:

```
sns.countplot(data=dataset_df, x='AgeCategory')
plt.tight_layout()
```



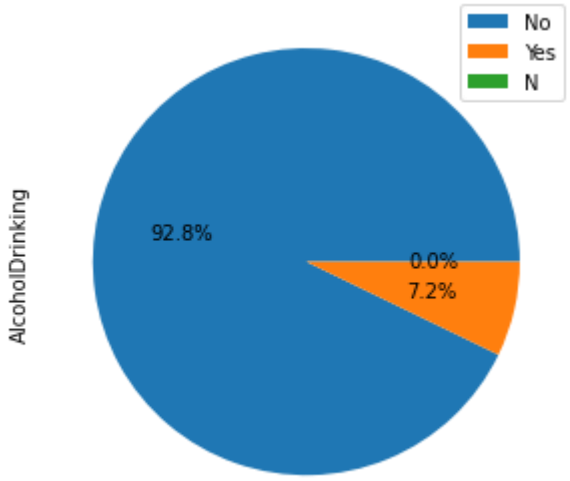
Pie Chart

Pada bagian Pie char kami mengambil data atau kolom 'AlcholDringkin' untuk mengukur apakah orang-orang yang melakukan pencucian hati/jantung itu berasal dari pola hidup yang tidak sehat dengan sering mengkonsumsi alkohol.

Namun dapat dilihat pada pie chart di bawah menunjukan bahwa sebagian besar yang melakukan pencucian hati/jantung tidak berasal dari peminum alkohol. jadi dapat disimpulkan kemungkinan orang-orang yang melakukan proses pencucian hati/jantung tidak meminum alkohol tetapi bisa juga terdapat dari penyakit bawaan lainnya.

In [ ]:

```
dataset_df['AlcoholDrinking'].value_counts().plot.pie(autopct='%1.1f%%', labels=None, l
egend=True)
plt.tight_layout()
```



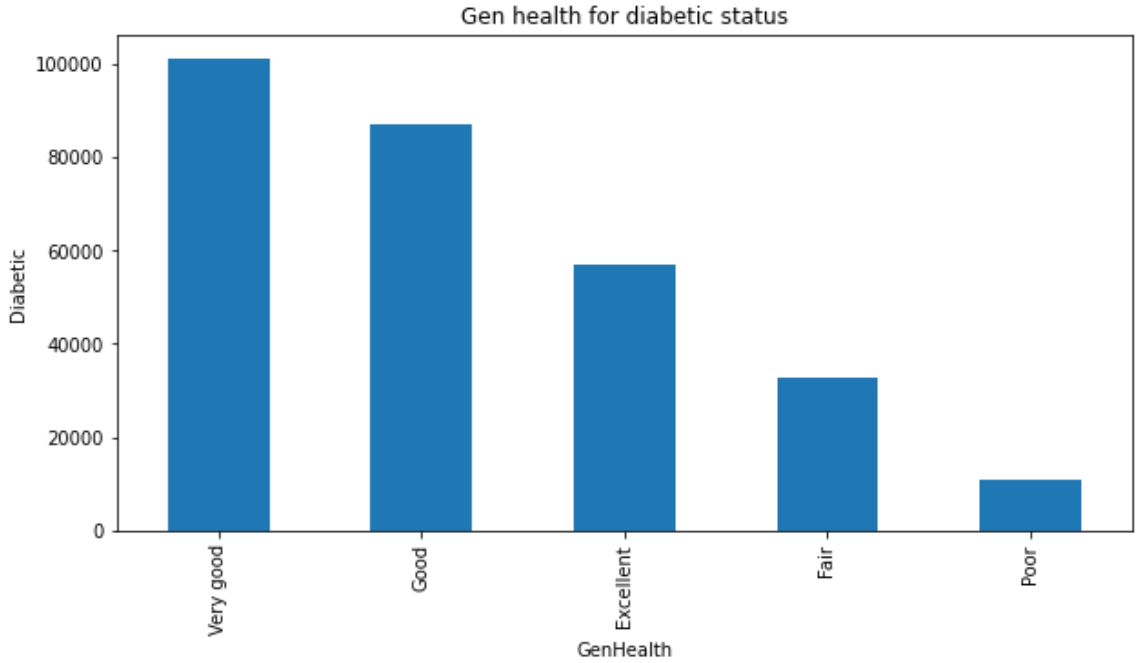
Histogram

Pada Histogram kami membuat visualisasi tentang orang-orang yang memiliki Gen health yang bagus apakah masih dapat terkena diabetes atau tidak.

Hasil di bawah ini menunjukan bahwa mayoritas yang memiliki gen health 'Very good' masih dapat terkena diabetes. Sedangkan di sini yang memiliki Gen health Poor malah menghasilkan visualisasi yang paling sedikit terkena diabetic.

In [ ]:

```
dataset_df.GenHealth.value_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title("Gen health for diabetic status")
plt.ylabel('Diabetic')
plt.xlabel('GenHealth');
```



2. Mengelola Data Klastering Menggunakan Spark

Konfigurasi Spark

In [2]:

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null #install java development ki
t
!wget -q https://dlcdn.apache.org/spark/spark-3.3.0/spark-3.3.0-bin-hadoop3.tgz #ins
tall spark
!tar xf spark-3.3.0-bin-hadoop3.tgz #unzip spark
!pip install -q findspark #install findspark
```

In [10]:

```
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.3.0-bin-hadoop3"
```

In [11]:

```
import findspark
findspark.init()
import pyspark
```

In [24]:

```
##Import Module yang dibutuhkan
from pyspark.ml.clustering import KMeans
from pyspark.ml.feature import VectorAssembler
from pyspark.sql import SparkSession
from pyspark.sql import SQLContext
```

In [27]:

```
#Membuat Session
appName = "Klastering Di Spark"
spark = SparkSession.builder.appName(appName).config("spark.some.config.option", "some-value").getOrCreate()
```

Memuat data kesehatan jantung dari dataset yang ada

Data kesehatan jantung diberikan file " heart_2020_cleaned.csv " di folder "dataset". Adapun data di file tersebut memiliki kolom sebagai berikut: HeartDisease, BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory Race, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer.

In [37]:

```
#menampilkan data
dff = spark.read.csv(
    'heart_2020_cleaned.csv', inferSchema=True, header=True)
dff.show()
```

```
+------------+-----+-------+--------------+------+-------------+-------
----+-----------+------+-----------+-----+--------------------+----------
-----+---------+---------+------+------------+----------+
|HeartDisease|  BMI|Smoking|AlcoholDrinking|Stroke|PhysicalHealth|MentalHe
alth|DiffWalking|   Sex|AgeCategory| Race|            Diabetic|PhysicalAct
ivity|GenHealth|SleepTime|Asthma|KidneyDisease|SkinCancer|
+------------+-----+-------+--------------+------+-------------+-------
----+-----------+------+-----------+-----+--------------------+----------
-----+---------+---------+------+------------+----------+
|          No| 16.6|    Yes|            No|    No|          3.0|
30.0|         No|Female|55-59|White|           Yes|
Yes|Very good|      5.0|   Yes|           No|       Yes|
|          No|20.34|     No|            No|   Yes|          0.0|
0.0|         No|Female|80 or older|White|      No|
Yes|Very good|      7.0|    No|           No|        No|
|          No|26.58|    Yes|            No|    No|         20.0|
30.0|         No|  Male|65-69|White|           Yes|
Yes|     Fair|      8.0|   Yes|           No|        No|
|          No|24.21|     No|            No|    No|          0.0|
0.0|         No|Female|      75-79|White|       No|
No|     Good|      6.0|    No|           No|       Yes|
|          No|23.71|     No|            No|    No|         28.0|
0.0|        Yes|Female|      40-44|White|       No|
Yes|Very good|      8.0|    No|           No|        No|
|         Yes|28.87|    Yes|            No|    No|          6.0|
0.0|        Yes|Female|      75-79|Black|       No|
No|     Fair|     12.0|    No|           No|        No|
|          No|21.63|     No|            No|    No|         15.0|
0.0|         No|Female|      70-74|White|       No|
Yes|     Fair|      4.0|   Yes|           No|       Yes|
|          No|31.64|    Yes|            No|    No|          5.0|
0.0|        Yes|Female|80 or older|White|       Yes|
No|     Good|      9.0|   Yes|           No|        No|
|          No|26.45|     No|            No|    No|          0.0|
0.0|         No|Female|80 or older|White|No, borderline di...|
No|     Fair|      5.0|    No|          Yes|        No|
|          No|40.69|     No|            No|    No|          0.0|
0.0|        Yes|  Male|      65-69|White|       No|
Yes|     Good|     10.0|    No|           No|        No|
|         Yes| 34.3|    Yes|            No|    No|         30.0|
0.0|        Yes|  Male|      60-64|White|       Yes|
No|     Poor|     15.0|   Yes|           No|        No|
|          No|28.71|    Yes|            No|    No|          0.0|
0.0|         No|Female|      55-59|White|       No|
Yes|Very good|      5.0|    No|           No|        No|
|          No|28.37|    Yes|            No|    No|          0.0|
0.0|        Yes|  Male|      75-79|White|       Yes|
Yes|Very good|      8.0|    No|           No|        No|
|          No|28.15|     No|            No|    No|          7.0|
0.0|        Yes|Female|80 or older|White|       No|
No|     Good|      7.0|    No|           No|        No|
|          No|29.29|    Yes|            No|    No|          0.0|
30.0|        Yes|Female|60-64|White|           No|
No|     Good|      5.0|    No|           No|        No|
|          No|29.18|     No|            No|    No|          1.0|
0.0|         No|Female|      50-54|White|       No|
Yes|Very good|      6.0|    No|           No|        No|
|          No|26.26|     No|            No|    No|          5.0|
2.0|         No|Female|      70-74|White|       No|
No|Very good|     10.0|    No|           No|        No|
|          No|22.59|    Yes|            No|    No|          0.0|
```

```
30.0|        Yes|  Male|      70-74|White|No, borderline di...|
Yes|     Good|      8.0|    No|           No|        No|
|          No|29.86|    Yes|             No|    No|           0.0|
0.0|        Yes|Female|      75-79|Black|                 Yes|
No|     Fair|      5.0|    No|          Yes|        No|
|          No|18.13|     No|             No|    No|           0.0|
0.0|         No|  Male|80 or older|White|                  No|
Yes|Excellent|      8.0|    No|           No|       Yes|
+------------+-----+------+---------------+-----+-------------+--------
----+-----------+------+-----------+-----+--------------------+-----------
-----+---------+---------+------+------------+----------+
only showing top 20 rows
```

Menyiapkan Data Training

Saat menyiapkan data training kami ambil beberapa atribut seperti BMI, PhysicalHealth, MentalHealth, SleepTime. Data tersebut kami akan tampilkan bersebelahan dengan atribut HeartDiesease dan data training diberi nama train.

In [38]:

```
#membuat assembler untuk mengubah fitur menjadi satu kolom
assembler = VectorAssembler(inputCols=[
"BMI","PhysicalHealth","MentalHealth", "SleepTime"
], outputCol="Specification")
train = assembler.transform(dff).select('HeartDisease', 'Specification')
train.show(truncate = False)
```

```
+------------+---------------------+
|HeartDisease|Specification        |
+------------+---------------------+
|No          |[16.6,3.0,30.0,5.0]  |
|No          |[20.34,0.0,0.0,7.0]  |
|No          |[26.58,20.0,30.0,8.0]|
|No          |[24.21,0.0,0.0,6.0]  |
|No          |[23.71,28.0,0.0,8.0] |
|Yes         |[28.87,6.0,0.0,12.0] |
|No          |[21.63,15.0,0.0,4.0] |
|No          |[31.64,5.0,0.0,9.0]  |
|No          |[26.45,0.0,0.0,5.0]  |
|No          |[40.69,0.0,0.0,10.0] |
|Yes         |[34.3,30.0,0.0,15.0] |
|No          |[28.71,0.0,0.0,5.0]  |
|No          |[28.37,0.0,0.0,8.0]  |
|No          |[28.15,7.0,0.0,7.0]  |
|No          |[29.29,0.0,30.0,5.0] |
|No          |[29.18,1.0,0.0,6.0]  |
|No          |[26.26,5.0,2.0,10.0] |
|No          |[22.59,0.0,30.0,8.0] |
|No          |[29.86,0.0,0.0,5.0]  |
|No          |[18.13,0.0,0.0,8.0]  |
+------------+---------------------+
only showing top 20 rows
```

Membuat Model K-Means Klastering

Dalam membuat model k-Means ini kami hanya memanggil data training yang bernama train.

In [40]:

```
##mendifinisikan algoritma klastering
kmeans = KMeans(
    featuresCol=assembler.getOutputCol(), predictionCol="cluster",k=5, seed=0
)
#menghitung model dengan perintah fit()
model = kmeans.fit(train)
print("berhasil dibuat")
```

```
berhasil dibuat
```

Mencari Nilai Titik Tengah dari Setiap Klaster

Saat mencari nilai titik tengah dari setiap klister kami hanya perlu print("Cluster Centers: ").

In [42]:

```
centers = model.clusterCenters()
print("Cluster Centers: ")
for center in centers:
  print(center)
```

```
Cluster Centers:
[35.70160675  1.07175077  1.39030189  7.06854433]
[25.13033116  0.66482898  1.10776435  7.19987356]
[28.26753418  2.18627842 22.49882148  6.74794213]
[31.01185544 26.42989015 25.70082156  6.55183236]
[29.55286917 25.01433641  1.83203217  7.02361013]
```

Memprediksi Klaster

Saat memprediksi klaster kami menggunakan orderBy dan groupBy. OrderBy yaitu pengurutan data seperti pada field cluster. Sedangkan, groupBy yaitu untuk mengelompokkan baris- baris data pada field Model dan cluster.

In [43]:

```
prediction = model.transform(train)#melakukan prediksi klaster
prediction.groupBy("cluster").count().orderBy("cluster").show()
prediction.select('HeartDisease','cluster').show() #menampilkan hasil prediksi
```

```
+-------+------+
|cluster| count|
+-------+------+
|      0| 73602|
|      1|186648|
|      2| 27577|
|      3| 10833|
|      4| 21135|
+-------+------+

+------------+-------+
|HeartDisease|cluster|
+------------+-------+
|          No|      2|
|          No|      1|
|          No|      3|
|          No|      1|
|          No|      4|
|         Yes|      1|
|          No|      4|
|          No|      0|
|          No|      1|
|          No|      0|
|         Yes|      4|
|          No|      1|
|          No|      1|
|          No|      1|
|          No|      2|
|          No|      1|
|          No|      1|
|          No|      2|
|          No|      1|
|          No|      1|
+------------+-------+
only showing top 20 rows
```