



**LAPORAN ANALISIS DATASET BIG DATA**

“Milk Quality Prediction”

**ANGGOTA KELOMPOK:**

Rayhan Vito (20051397042)  
Gustiansyah  
Muhammad Auliya'ur (20051397066)  
Rahman

PRODI D4 MANAJEMEN INFORMATIKA

FAKULTAS VOKASI

**UNIVERSITAS NEGERI SURABAYA**

## BAB I

### Pendahuluan

#### A. Datasheet

Dataset adalah sekumpulan data yang disusun secara terstruktur. Biasanya, dataset dipresentasikan dalam bentuk tabel, alias baris dan kolom. Tiap baris dan kolom biasanya mewakili variabel tertentu. Dataset dapat digunakan untuk memberikan pemahaman mengenai tema atau konsep tertentu. Mereka menyimpan data yang dibutuhkan oleh aplikasi atau sistem operasi agar dapat berfungsi dengan baik.

#### B. Prediksi Kualitas Susu

Dataset ini dikumpulkan secara manual dari observasi. Ini membantu kami membuat model pembelajaran mesin untuk memprediksi kualitas susu.

Dataset ini terdiri dari 7 variabel bebas yaitu pH, Suhu, Rasa, Bau, Lemak, Kekeruhan, dan Warna.

Umumnya, Grade atau Kualitas susu tergantung pada parameter ini. Parameter ini memainkan peran penting dalam analisis prediksi susu.

Penggunaan Variabel target tidak lain adalah Grade susu. Bisa jadi Target

1. Rendah (Buruk)
2. Sedang (Sedang)
3. Tinggi (Bagus)

Jika Rasa, Bau, Lemak, dan Kekeruhan terpenuhi dengan kondisi optimal maka mereka akan memberikan 1 jika tidak 0.

Suhu dan pH diberikan nilai sebenarnya dalam dataset.

Kita harus melakukan preprocessing data, dan teknik augmentasi data untuk membangun model statistik dan prediktif untuk memprediksi kualitas susu.

## BAB II

### Pembahasan

#### A. Google Collab

Kami menganalisis dataset tersebut menggunakan salah satu platform yang cukup similar yaitu google collab dimana google collab sama seperti jupyter notebook akan tetapi berbasis cloud.

Dataset yang kami analisis ini berisi beberapa aspek yang mempengaruhi untuk Kualitas Susu.

Antara lain :

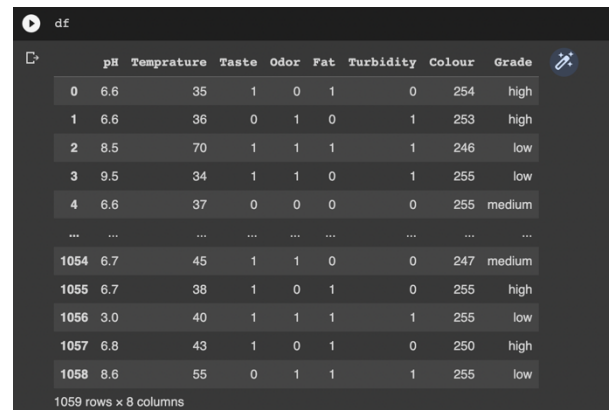
1. Fat Content
2. pH level
3. Temperature
4. Turbidity
5. Odor
6. Colour

#### B. Hasil Analisis

##### a. Input library

```
[1] import pandas as pd
import numpy as np
```

##### b. Melihat data



|      | pH  | Temperature | Taste | Odor | Fat | Turbidity | Colour | Grade  |
|------|-----|-------------|-------|------|-----|-----------|--------|--------|
| 0    | 6.6 | 35          | 1     | 0    | 1   | 0         | 254    | high   |
| 1    | 6.6 | 36          | 0     | 1    | 0   | 1         | 253    | high   |
| 2    | 8.5 | 70          | 1     | 1    | 1   | 1         | 246    | low    |
| 3    | 9.5 | 34          | 1     | 1    | 0   | 1         | 255    | low    |
| 4    | 6.6 | 37          | 0     | 0    | 0   | 0         | 255    | medium |
| ...  | ... | ...         | ...   | ...  | ... | ...       | ...    | ...    |
| 1054 | 6.7 | 45          | 1     | 1    | 0   | 0         | 247    | medium |
| 1055 | 6.7 | 38          | 1     | 0    | 1   | 0         | 255    | high   |
| 1056 | 3.0 | 40          | 1     | 1    | 1   | 1         | 255    | low    |
| 1057 | 6.8 | 43          | 1     | 0    | 1   | 0         | 250    | high   |
| 1058 | 8.6 | 55          | 0     | 1    | 1   | 1         | 255    | low    |

df Terlihat data beserta nama kolom dapat ditampilkan dengan baik Untuk melihat lebih banyak data, perintah head dapat diberiparameter jumlah dataframe yang ingin ditampilkan

### c. Describe data

```
df.describe()
```

|       | pH          | Temperature | Taste       | Odor        | Fat         | Turbidity   | Colour      |
|-------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| count | 1059.000000 | 1059.000000 | 1059.000000 | 1059.000000 | 1059.000000 | 1059.000000 | 1059.000000 |
| mean  | 6.630123    | 44.226629   | 0.546742    | 0.432483    | 0.671388    | 0.491029    | 251.840415  |
| std   | 1.399679    | 10.098364   | 0.498046    | 0.495655    | 0.469930    | 0.500156    | 4.307424    |
| min   | 3.000000    | 34.000000   | 0.000000    | 0.000000    | 0.000000    | 0.000000    | 240.000000  |
| 25%   | 6.500000    | 38.000000   | 0.000000    | 0.000000    | 0.000000    | 0.000000    | 250.000000  |
| 50%   | 6.700000    | 41.000000   | 1.000000    | 0.000000    | 1.000000    | 0.000000    | 255.000000  |
| 75%   | 6.800000    | 45.000000   | 1.000000    | 1.000000    | 1.000000    | 1.000000    | 255.000000  |
| max   | 9.500000    | 90.000000   | 1.000000    | 1.000000    | 1.000000    | 1.000000    | 255.000000  |

describe () akan menampilkan jumlah data di setiap kolom (count), rata-rata nilai tiap kolom (mean), standar deviasi (std), nilai minimum (min), nilai maksimum (max), serta batas nilai dari masing-masing kuartil (25%, 50%, 75%).

Hasil analisis menampilkan sejumlah data dengan persentase dan ada nilai max dan min

### d. Membuat perubahan kerangka data

```
df[df['Grade'] == 'high'].describe()
```

|       | pH         | Temperature | Taste      | Odor       | Fat        | Turbidity  | Colour     |
|-------|------------|-------------|------------|------------|------------|------------|------------|
| count | 256.000000 | 256.000000  | 256.000000 | 256.000000 | 256.000000 | 256.000000 | 256.000000 |
| mean  | 6.692578   | 40.648438   | 0.664062   | 0.750000   | 0.996094   | 0.632812   | 252.536    |
| std   | 0.108752   | 3.739749    | 0.473242   | 0.433861   | 0.062500   | 0.482982   | 3.711      |
| min   | 6.500000   | 35.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 245.000    |
| 25%   | 6.600000   | 37.000000   | 0.000000   | 0.750000   | 1.000000   | 0.000000   | 250.000    |
| 50%   | 6.700000   | 40.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |
| 75%   | 6.800000   | 45.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |
| max   | 6.800000   | 45.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |

```
df[df['Grade'] == 'medium'].describe()
```

|       | pH         | Temperature | Taste      | Odor       | Fat        | Turbidity  | Colour     |
|-------|------------|-------------|------------|------------|------------|------------|------------|
| count | 374.000000 | 374.000000  | 374.000000 | 374.000000 | 374.000000 | 374.000000 | 374.000000 |
| mean  | 6.635027   | 39.721925   | 0.414439   | 0.163102   | 0.385027   | 0.125668   | 250.336    |
| std   | 0.130899   | 3.613751    | 0.493285   | 0.369953   | 0.487253   | 0.331919   | 5.262      |
| min   | 6.400000   | 34.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 240.000    |
| 25%   | 6.500000   | 37.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 245.000    |
| 50%   | 6.600000   | 38.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 255.000    |
| 75%   | 6.800000   | 45.000000   | 1.000000   | 0.000000   | 1.000000   | 0.000000   | 255.000    |
| max   | 6.800000   | 45.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |

```
df[df['Grade'] == 'low'].describe()
```

|       | pH         | Temperature | Taste      | Odor       | Fat        | Turbidity  | Colour     |
|-------|------------|-------------|------------|------------|------------|------------|------------|
| count | 429.000000 | 429.000000  | 429.000000 | 429.000000 | 429.000000 | 429.000000 | 429.000000 |
| mean  | 6.588578   | 50.289044   | 0.592075   | 0.477855   | 0.727273   | 0.724942   | 252.734    |
| std   | 2.194712   | 13.043794   | 0.492023   | 0.500093   | 0.445882   | 0.447065   | 3.233      |
| min   | 3.000000   | 34.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 245.000    |
| 25%   | 4.700000   | 40.000000   | 0.000000   | 0.000000   | 0.000000   | 0.000000   | 250.000    |
| 50%   | 6.800000   | 45.000000   | 1.000000   | 0.000000   | 1.000000   | 1.000000   | 255.000    |
| 75%   | 8.000000   | 55.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |
| max   | 9.500000   | 90.000000   | 1.000000   | 1.000000   | 1.000000   | 1.000000   | 255.000    |

### e. Data eksploratif

Fungsi apakah ada data yang null atau

```
[9] df.nunique()
```

```
pH          16
Temperature 17
Taste        2
Odor          2
Fat           2
Turbidity     2
Colour        9
Grade         3
dtype: int64
```

kosong , hasilnya tidak ada data yang kosong

### f. Encoding data

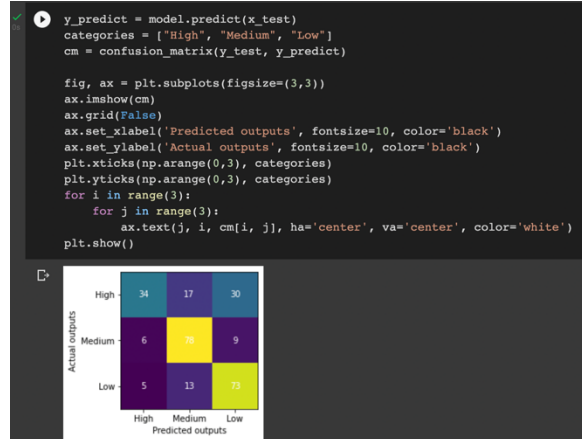
```
from tensorflow.keras.utils import plot_model, to_categorical
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
```

### g. Analisi fitur

```
{13} grades = {'high': 0, 'medium': 1, 'low': 2}
a = []
for row in df.index:
    a.append(grades[df['Grade']][row])
y = a
x = df.drop(['Grade'], axis=1)
x = x.drop(['Taste'], axis=1)
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=0, train_size= 0.75)
```

Terdapat beberapa fungsi untuk menampilkan sebuah prediksi dimana yang ditampilkan yaitu “high, 'medium', 'low’”

### h. Menampilkan gambar



### i. Menampilkan Report

```
[19] print(classification_report(y_test, y_predict, target_names = categories))
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| High         | 0.76      | 0.42   | 0.54     | 81      |
| Medium       | 0.72      | 0.84   | 0.78     | 93      |
| Low          | 0.65      | 0.80   | 0.72     | 91      |
| accuracy     |           |        | 0.70     | 265     |
| macro avg    | 0.71      | 0.69   | 0.68     | 265     |
| weighted avg | 0.71      | 0.70   | 0.68     | 265     |

### BAB III

#### Kesimpulan

Berdasarkan hasil mean kita dapat mengambil kesimpulan sebagai berikut:

- Susu dengan kadar *tinggi* hampir pasti mengandung lemak
- Susu dengan kadar *sedang* mungkin mengandung lemak atau tidak
- Susu dengan kadar *rendah* mungkin mengandung lemak tetapi cenderung tidak mengandung lemak dibandingkan dengan kadar *tinggi*
- Tampaknya rasa tidak masalah di salah satu nilai karena tampaknya hampir sempurna acak dengan standar deviasi tinggi pada masing-masing.
- Dengan pH, tampaknya masalah dengan kadar *rendah* memiliki tingkat pH sekitar 6,59, kadar *sedang* sekitar 6,64, dan kadar *tinggi* sekitar 6,69 dengan sedikit penyimpangan sekitar 0,1 kecuali untuk kadar *rendah* dengan standar 2,19 deviasi.
- Suhu juga tampaknya menjadi masalah antara susu kelas *tinggi* dan *sedang* dengan standar deviasi rendah dibandingkan dengan susu kelas *rendah* dengan standar deviasi sangat tinggi, tampaknya juga susu kelas *tinggi* dan *sedang* hampir memiliki suhu yang sama dibandingkan dengan susu kelas *rendah*.
- Untuk bau, susu kelas *tinggi* kemungkinan akan memiliki bau, susu kelas *sedang* tidak akan berbau, dan untuk susu kelas *rendah*, bau tidak menjadi masalah.
- Untuk kekeruhan, susu tingkat *sedang* kemungkinan tidak akan memilikinya, sedangkan susu tingkat *rendah* dan *tinggi* akan memiliki satu atau tidak, di mana susu tingkat *rendah* lebih cenderung memiliki kekeruhan.
- Untuk warna, baik susu *rendah* dan *tinggi* memiliki warna yang hampir sama, sedangkan kadar *sedang* berbeda dari keduanya.
- Dengan wawasan berikut, kita dapat dengan aman berasumsi bahwa rasa tidak diperlukan untuk prediksi kadar susu dan dapat menghilangkannya. Dengan demikian indikator berikut yang akan digunakan adalah:
  1. Fat Content
  2. pH level
  3. Temperature
  4. Turbidity
  5. Odor
  6. Colour