WATER QUALITY AND POTABILITY

Final Project

Presented by Mawar Melati

DigitalSkola

DigitalSkola

OUR TEAM

TUBAGUS FATIN K

DIMAS BAGUS C W

ZHARIFAH INDRA P

ISWANTI SIHALOHO

DigitalSkola

Overview

- Data Understanding
- Exploratory Data Analyst (EDA)
- Data Preprocessing
- Data Modeling
- Conclusion

About Dataset

Data ini berharga dalam penilaian kualitas air, perencanaan pengolahan air, dan memastikan keamanan pasokan air minum. Hal ini dapat dimanfaatkan oleh instalasi pengolahan air, lembaga lingkungan hidup, dan peneliti untuk membuat keputusan berdasarkan data mengenai kualitas dan kelayakan air.

Tujuan

Tujuan utama dari dataset ini untuk menilai dan memprediksi kelayakan air berdasarkan parameter kualitas air. Hal ini dapat digunakan dalam mengevaluasi keamanan dan kesesuaian sumber air untuk konsumsi manusia, membuat keputusan mengenai pengolahan air, dan memastikan kepatuhan terhadap standar kualitas air.

Dengan model yang dilatih untuk memprediksi kelayakan air berdasarkan parameter kualitas air yang disediakan. Model ini bertujuan untuk mengklasifikasikan sampel air sebagai layak minum (1) atau tidak layak minum (0).

DigitalSkola

DATA UNDERSTANDING

Data Understanding

Dataset ini didapatkan dari kaggle yang memiliki 3276 baris dan 10 kolom dengan beberapa variabel seperti berikut:

```
print(df.shape)
df.head()
```

```
(3276, 10)
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability

DigitalSkola

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64  
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Dataset Information

Menampilkan informasi detail tentang dataframe, seperti jumlah baris data ,nama-nama kolom beserta jumlah data dan tipe data.

DigitalSkola

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
ph	2785.00	7.08	1.59	0.00	6.09	7.04	8.06	14.00
Hardness	3276.00	196.37	32.88	47.43	176.85	196.97	216.67	323.12
Solids	3276.00	22014.09	8768.57	320.94	15666.69	20927.83	27332.76	61227.20
Chloramines	3276.00	7.12	1.58	0.35	6.13	7.13	8.11	13.13
Sulfate	2495.00	333.78	41.42	129.00	307.70	333.07	359.95	481.03
Conductivity	3276.00	426.21	80.82	181.48	365.73	421.88	481.79	753.34
Organic_carbon	3276.00	14.28	3.31	2.20	12.07	14.22	16.56	28.30
Trihalomethanes	3114.00	66.40	16.18	0.74	55.84	66.62	77.34	124.00
Turbidity	3276.00	3.97	0.78	1.45	3.44	3.96	4.50	6.74
Potability	3276.00	0.39	0.49	0.00	0.00	0.00	1.00	1.00

Describe

Describe berfungsi untuk mengetahui statistika data untuk data numeric seperti **count, mean, standard deviation, maximum, minimum, dan quartile.**

DigitalSkola

```
df.isnull().sum()
```

```
ph                 491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

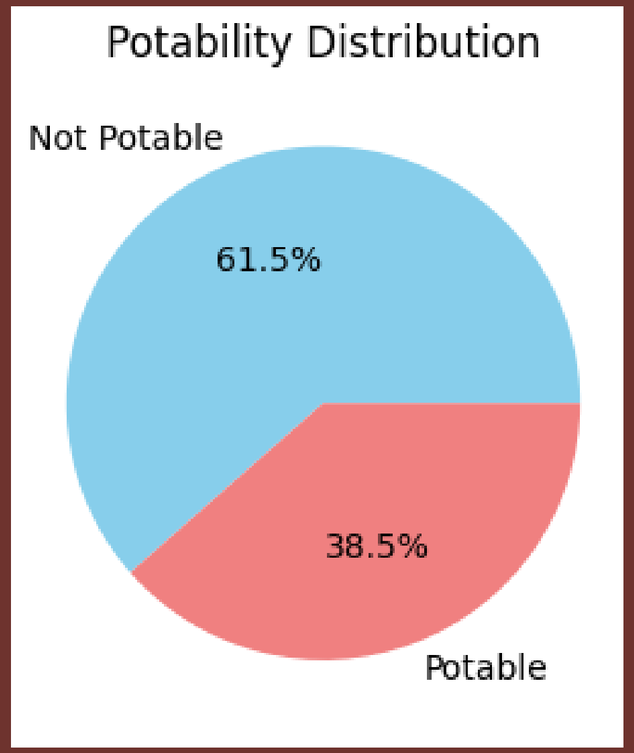
Missing Value

Untuk menangani missing value pada kolom pH, Sulfate, dan Trihalomethanes kita dapat melakukan pengecekan apakah kolom tersebut memiliki distribusi normal untuk datanya. Jika Iya maka baris data yang kosong dapat diisi dengan mean dari seluruh data kolom masing-masing.

DigitalSkola

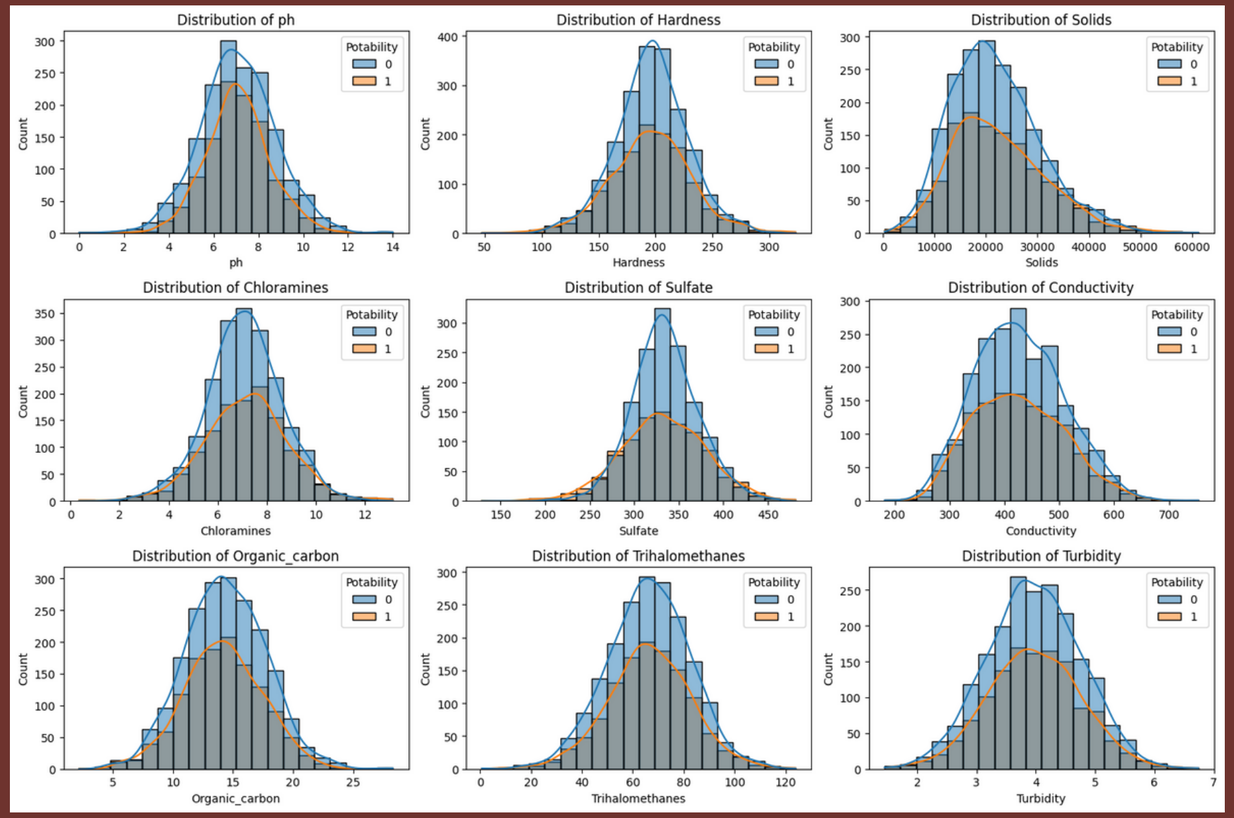
EXPLORATORY DATA ANALYST (EDA)

Satisfaction



Distribusi dari target label atas kolom satisfaction. Terlihat untuk **kelayakan air dapat diminum** masih dibawah **50%**.

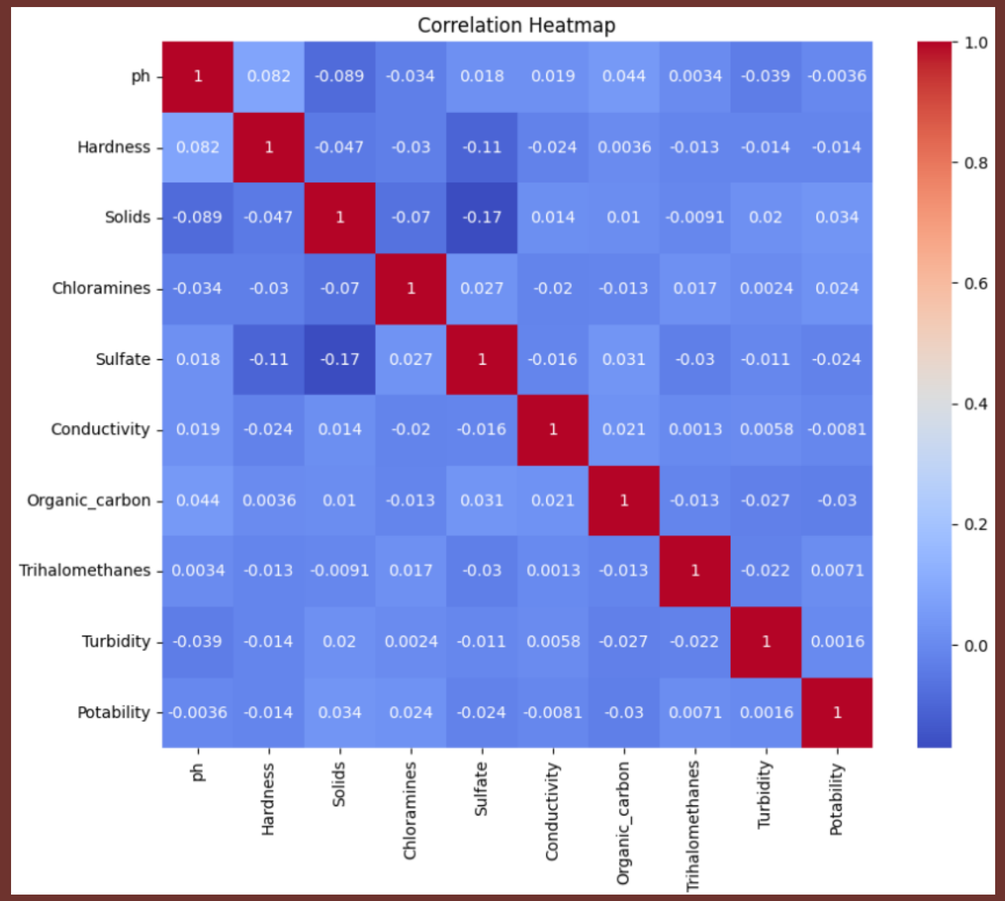
DigitalSkola



Visualization of Numerical Columns

Dari hasil histogram, setiap kolom memiliki data yang mendekati distribusi normal, sehingga nantinya penanganan missing value pada pH, Sulfate, dan Trihalomethanes menggunakan nilai mean dari setiap kolom. Penggunaan nilai mean untuk mengatasi missing value agar statistik deskriptif data yang ada, seperti rata-rata, median, dan standar deviasi. Hal ini dapat membantu mencegah distorsi terhadap data.

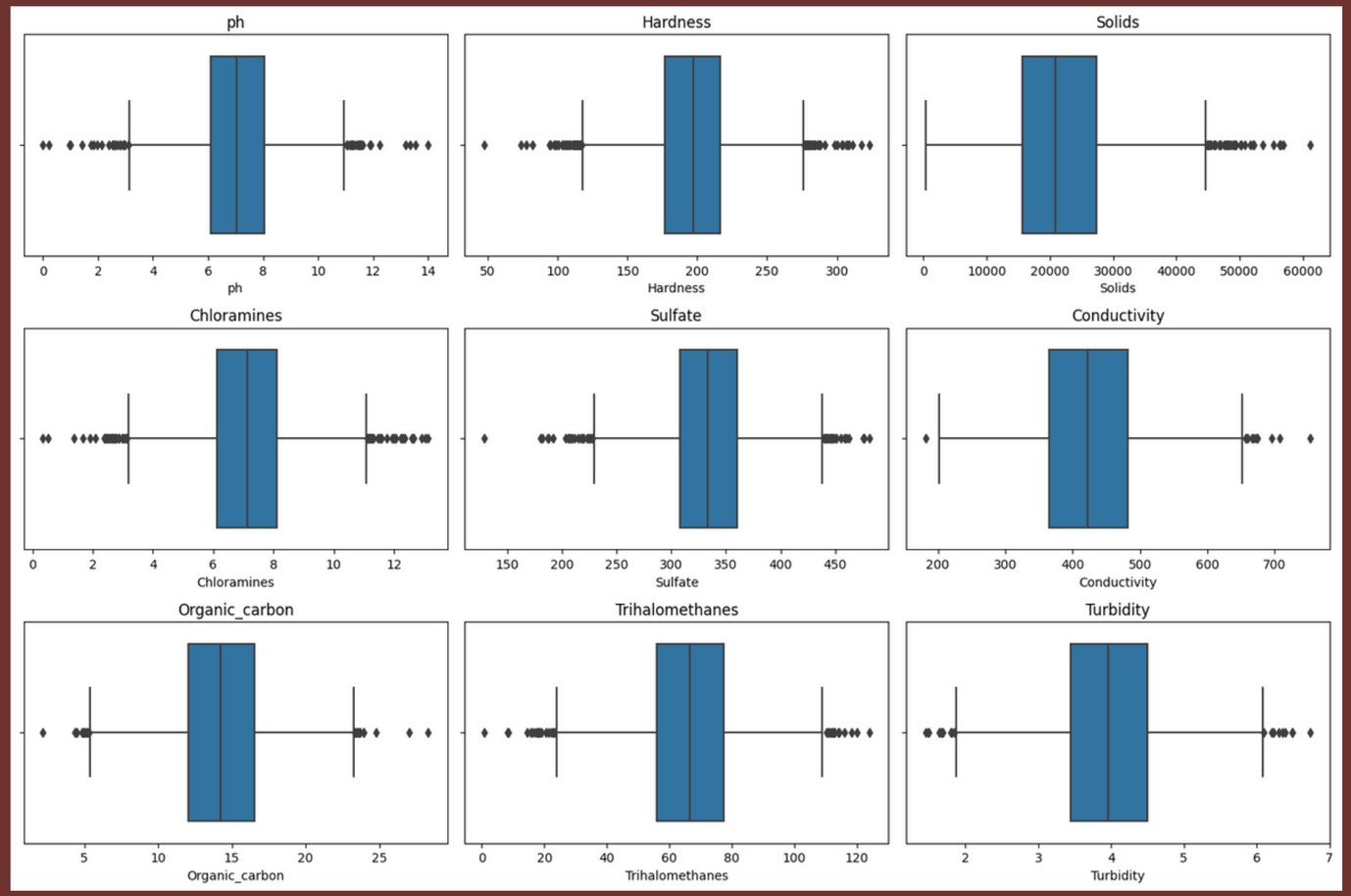
DigitalSkola



Correlation Heatmap

Correlation heatmap menunjukkan kepada kita seberapa berkorelasi variabel yang ada. Korelasi yang rendah antara variabel dependen (Potability) dan independen dapat mengindikasikan bahwa variabel independen mungkin tidak memiliki hubungan linear yang kuat dengan variabel dependen,artinya kita dapat mengggunakan semua variabel independen sebagai sumber informasi untuk proses klasifikasi. Akan tetapi, model klasifikasi mungkin memiliki kesulitan dalam memahami pola yang ada dalam data. Berikut adalah 5 data dengan korelasi tertinggi dengan kolom potability

DigitalSkola



Boxplot

Dapat dilihat pada hasil boxplot distribusi setiap kolom terdapat data outlier. Outlier merupakan data penting dikarenakan model klasifikasi yang sensitif terhadap outlier untuk menentukan hasil klasifikasi.

DigitalSkola

Missing Value Handling

```
df['ph'].fillna(df['ph'].mean(),axis=0, inplace=True)
df['Sulfate'].fillna(df['Sulfate'].mean(),axis=0, inplace=True)
df['Trihalomethanes'].fillna(df['Trihalomethanes'].mean(),axis=0, inplace=True)
```

```
df.isnull().sum()
```

```
ph                0
Hardness          0
Solids            0
Chloramines       0
Sulfate           0
Conductivity      0
Organic_carbon    0
Trihalomethanes   0
Turbidity         0
Potability        0
dtype: int64
```

Berdasarkan hasil Data Understanding sebelumnya terdapat tiga kolom yang memiliki Missing Value, Yaitu pH, Sulfate, dan Trihalomethanes.

Dikarenakan distribusi setiap kolom mendekati distribusi normal, maka data akan diisi dengan nilai rata-rata atau mean.

DATA PREPROCESSING

DigitalSkola

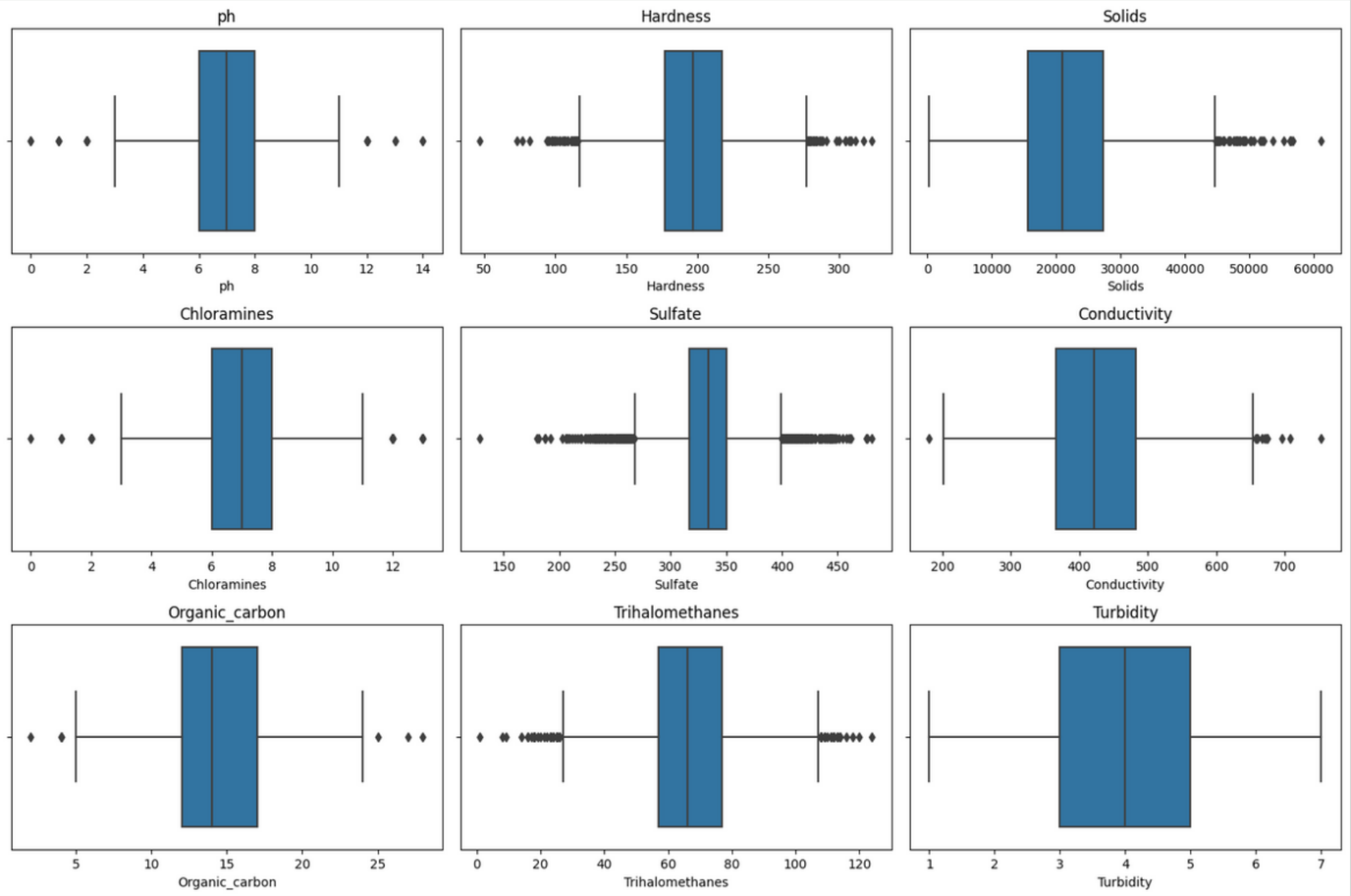
Feature Engineering

Rounding the Data

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	3276.00	3276.00	3276.00	3276.00	3276.00	3276.00	3276.00	3276.00	3276.00	3276.00
mean	7.07	196.37	22014.09	7.12	333.82	426.20	14.29	66.38	3.97	0.39
std	1.49	32.88	8768.57	1.61	36.15	80.82	3.31	15.77	0.83	0.49
min	0.00	47.00	321.00	0.00	129.00	181.00	2.00	1.00	1.00	0.00
25%	6.00	177.00	15666.50	6.00	317.00	366.00	12.00	57.00	3.00	0.00
50%	7.00	197.00	20928.00	7.00	334.00	422.00	14.00	66.00	4.00	0.00
75%	8.00	217.00	27332.50	8.00	350.00	482.00	17.00	77.00	5.00	1.00
max	14.00	323.00	61227.00	13.00	481.00	753.00	28.00	124.00	7.00	1.00

Pembulatan data dilakukan untuk membulatkan nilai numerik menjadi angka yang lebih mudah dibaca atau dipahami, sekaligus mengubah DataType yang sebelumnya Float, menjadi Integer.

Feature Engineering



Outliers

Dapat dilihat dari boxplot berikut, banyak fitur-fitur yang memiliki outliers.

DigitalSkola

Feature Engineering

Removing Outliers

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	3200.00	3200.00	3200.00	3200.00	3200.00	3200.00	3200.00	3200.00	3200.00	3200.00
mean	7.06	196.53	21949.21	7.12	333.66	426.09	14.31	66.33	3.97	0.39
std	1.47	29.99	8714.65	1.61	35.77	80.85	3.31	15.79	0.83	0.49
min	0.00	117.00	321.00	0.00	129.00	181.00	2.00	1.00	1.00	0.00
25%	6.00	177.75	15651.00	6.00	317.00	365.75	12.00	57.00	3.00	0.00
50%	7.00	197.00	20882.50	7.00	334.00	422.00	14.00	66.00	4.00	0.00
75%	8.00	216.00	27250.75	8.00	350.00	481.00	17.00	77.00	5.00	1.00
max	14.00	277.00	61227.00	13.00	481.00	753.00	28.00	124.00	7.00	1.00

Dari sekian banyak fitur yang memiliki outliers, kita hanya menghapus outliers di fitur Hardness saja. Karena seperti yang dikatakan di awal, bahwa Outliers merupakan data penting karena dapat mewakili kejadian nyata dalam data.

Outliers pada Hardness di hapus untuk upaya meningkatkan akurasi dari model.

DigitalSkola

Feature Engineering

Scaling Data

```
scaler = StandardScaler()

X = df.drop(columns="Potability")
y = df["Potability"]

scaled_data = scaler.fit_transform(X)
```

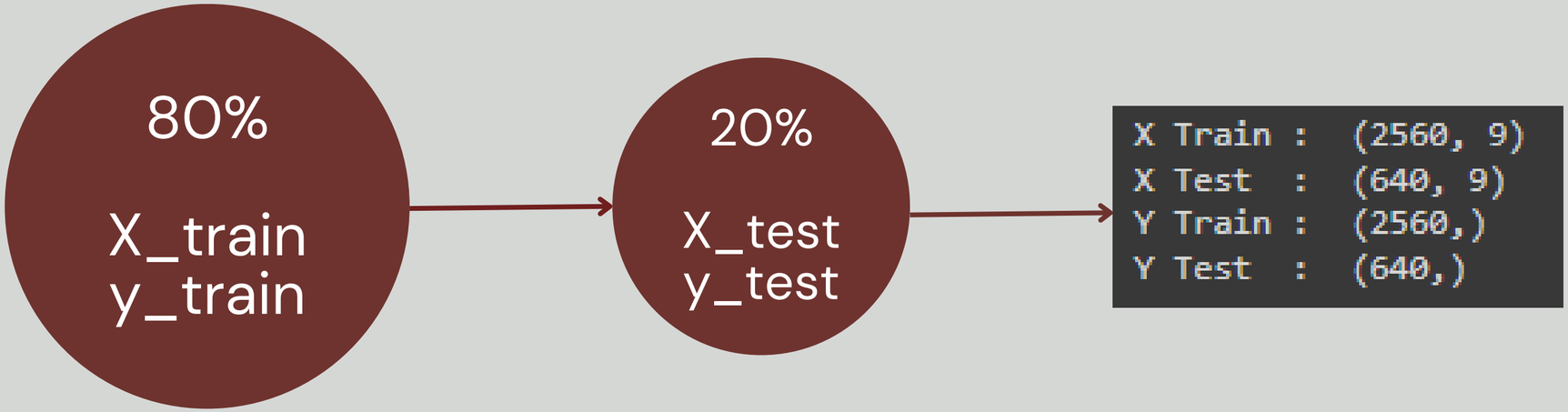
Scaling data dilakukan menggunakan:
StandardScaler()

DigitalSkola

Feature Engineering

Data Split

```
X_train, X_test, y_train, y_test = train_test_split(
    scaled_data,
    y,
    test_size=0.20,
    random_state=42
)
```




```
X Train :  (2560, 9)
X Test  :  (640, 9)
Y Train :  (2560,)
Y Test  :  (640,)
```

DigitalSkola

DATA MODELING

DigitalSkola

Lazy Classifier

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
QuadraticDiscriminantAnalysis	0.72	0.64	0.64	0.69	0.05
LGBMClassifier	0.70	0.64	0.64	0.68	0.42
XGBClassifier	0.68	0.63	0.63	0.67	0.53
SVC	0.73	0.63	0.63	0.69	0.69
ExtraTreesClassifier	0.70	0.62	0.62	0.68	2.93
RandomForestClassifier	0.70	0.62	0.62	0.68	1.58

Dengan tingkat akurasi tertinggi dan waktu pengambilan yang sangat rendah, maka kami memutuskan untuk melakukan modelling menggunakan Quadratic Discriminant Analysis (QDA)

DigitalSkola

```
Training Accuracy: 0.67
Testing Accuracy: 0.72
              precision    recall  f1-score   support

           0       0.73      0.91      0.81       420
           1       0.68      0.36      0.47       220

    accuracy                           0.72       640
   macro avg       0.70      0.64      0.64       640
weighted avg       0.71      0.72      0.69       640
```

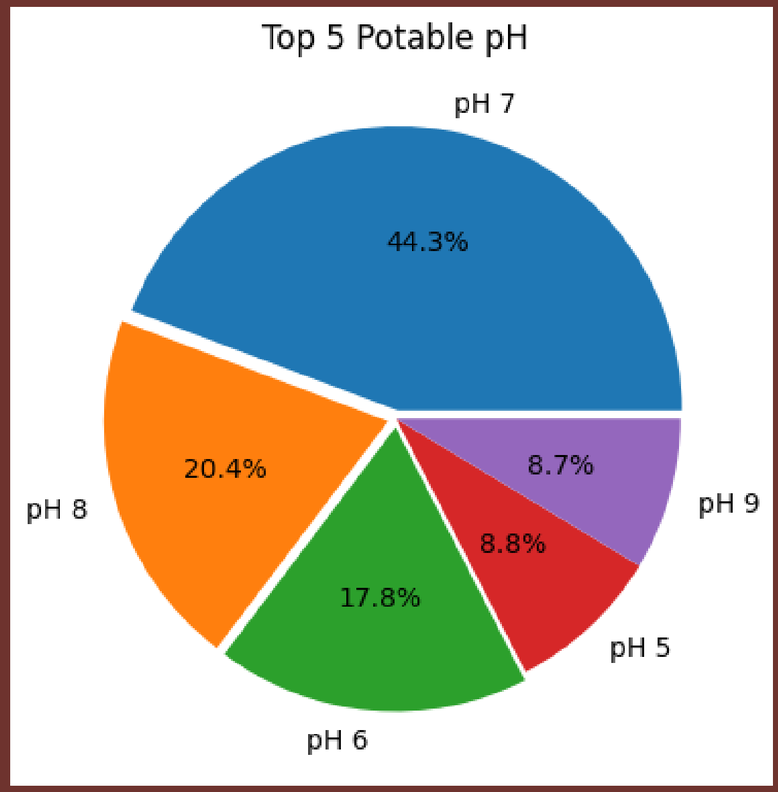
Quadratic Discriminant Analysis

Conclusion

- Dari Dataset, kita mengetahui bahwa pH terbaik untuk Air Minum adalah sekitar pH 6 - pH 8.

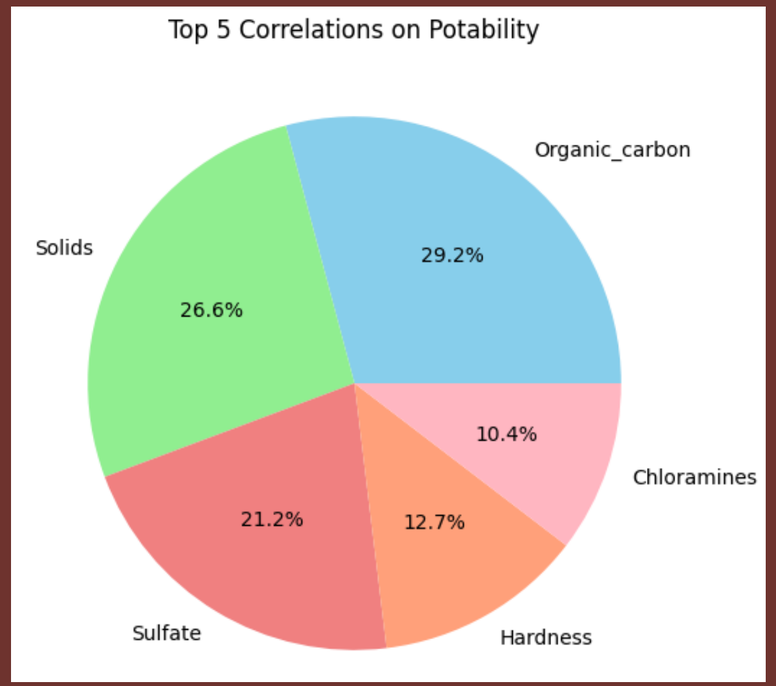
- The Environmental Protection Agency (EPA) di Amerika Serikat merekomendasikan bahwa pH air minum harus berada antara 6,5 dan 8,5 untuk konsumsi yang aman. Air dalam rentang ini tidak akan berbahaya dan dapat digunakan untuk memastikan hidrasi yang sehat.

- Peraturan Kementerian Kesehatan Indonesia Nomor 32 tahun 2017 menetapkan standar ideal untuk hasil uji kualitas air (pH) antara 6,5 dan 8,5.



Conclusion

Dataset ini menunjukkan bahwa Karbon Organik, Padatan, Sulfat, Kekerasan, dan Kloramina adalah fitur-fitur yang memiliki korelasi tertinggi dengan potabilitas.



DigitalSkola

THANK YOU

MAWAR MELATI