WATER QUALITY ANALYSIS AND PREDICTION

IBM Machine Learning Professional Certificate

Course 03: Supervised Machine Learning: Classification

By Moustafa Abada

Content:

- Dataset Description
- Main objectives of the analysis.
- Applying various classification models.
- Machine learning analysis and findings.
- Models flaws and advanced steps.

DATASET DESCRIPTION

Dataset Description:

- Water quality is a crucial aspect of environmental management, and it is essential to measure various physical, chemical, and biological parameters to monitor it effectively. This dataset of 200 rows contains measurements of six critical water quality parameters widely used in water quality monitoring and analysis. The dataset provides a representative snapshot of water quality and can be used for various research, education, and decision-making purposes. The dataset suits various data science applications such as data visualization, machine learning, and statistical analysis. It can be used to explore and analyze water quality trends, patterns, and relationships and can help researchers and analysts gain insights into the complex dynamics of water quality.

Water Quality Testing Dataset

```
[24] df = pd.read_csv("water_potability.csv")
     df
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

Dataset Description:

The dataset contains the following features:

- pH value: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
- Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
- Solids (Total dissolved solids - TDS): Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

Dataset Description:

- Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
- Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.
- Conductivity: Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

Dataset Description:

- Organic_carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.
- Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
- Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
- Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

DATA ANALYSIS

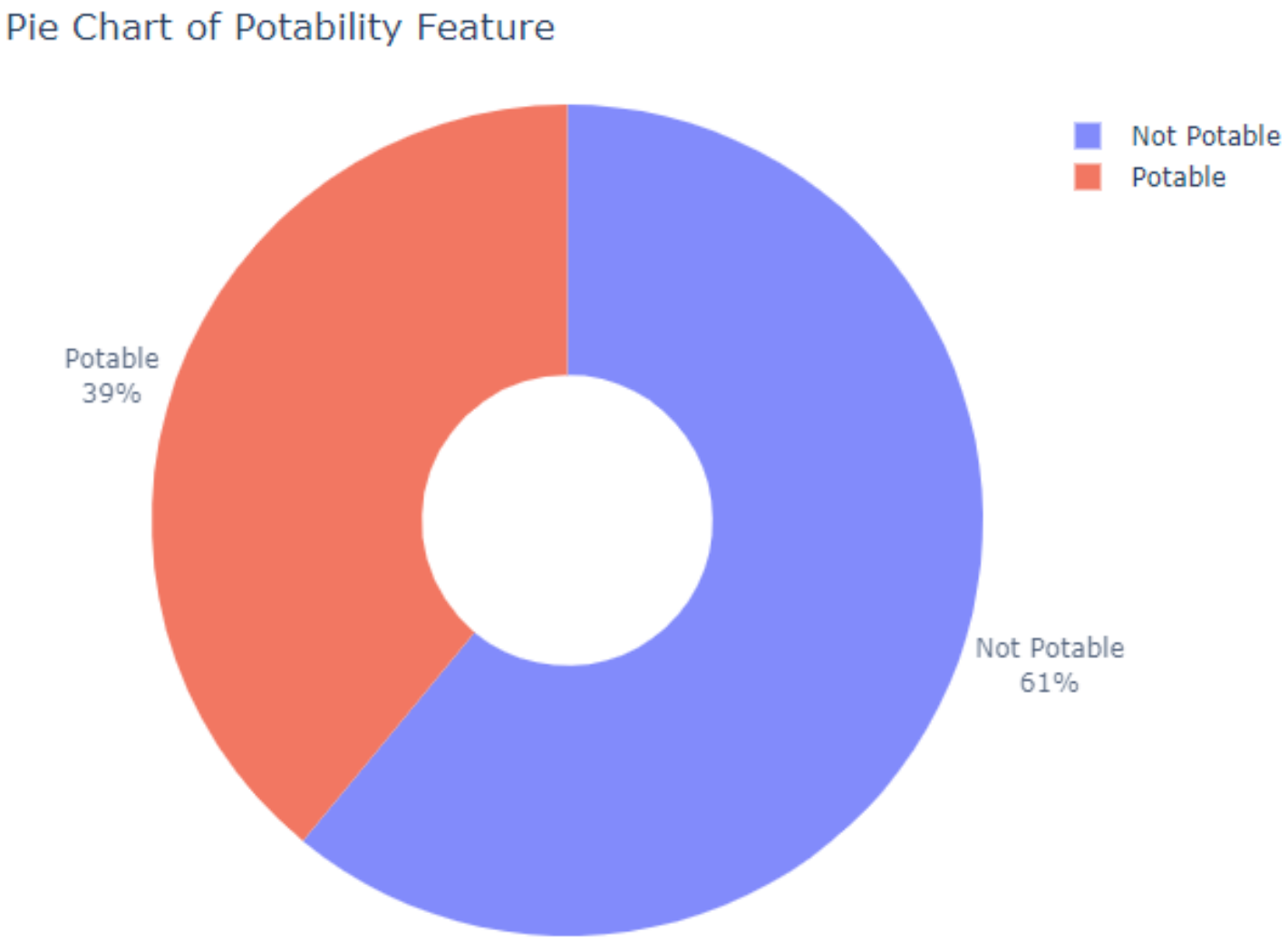
Main objectives of the analysis:

- The main objective of the water quality testing dataset analysis and prediction is to utilize the provided dataset, which includes measurements of various water quality parameters, to gain insights into the quality of water sources. The analysis aims to evaluate and predict the potability of the water based on parameters such as pH value, hardness, total dissolved solids, chloramines, sulfate concentration, conductivity, organic carbon, trihalomethanes, and turbidity. By understanding and predicting water potability, the analysis contributes to environmental management, public health protection, and informed decision-making regarding water usage and treatment.

```
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)
```

DATA ANALYSIS:

Find the data types of all features:



DATA ANALYSIS:

In [11]:
```
df.isnull().sum()
```

Out[11]:

ph	491
Hardness	0
Solids	0
Chloramines	0
Sulfate	781
Conductivity	0
Organic_carbon	0
Trihalomethanes	162
Turbidity	0
Potability	0

dtype: int64

In [12]:
```
# handle missing value with average of features
df["ph"].fillna(value = df["ph"].mean(), inplace = True)
df["Sulfate"].fillna(value = df["Sulfate"].mean(), inplace = True)
df["Trihalomethanes"].fillna(value = df["Trihalomethanes"].mean(), inplace = True)
```

In [13]:
```
df.isnull().sum()
```

Out[13]:

ph	0
Hardness	0
Solids	0
Chloramines	0
Sulfate	0
Conductivity	0
Organic_carbon	0
Trihalomethanes	0
Turbidity	0
Potability	0

dtype: int64

Data Analysis:

- Find the null values and replace them by the average.
- There are Null Values in the following columns:

1. Ph
2. Sulfate
3. trihalomethanes

APPLYING VARIOUS CLASSIFICATION MODELS

Train-Test Split and Normalization

In [14]:
```
X = df.drop("Potability", axis=1).values
y = df["Potability"].values
```

In [15]:
```
# train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3,
random_state = 3)
print("X_train",X_train.shape)
print("X_test",X_test.shape)
print("y_train",y_train.shape)
print("y_test",y_test.shape)
```

```
X_train (2293, 9)
X_test (983, 9)
y_train (2293,)
y_test (983,)
```

In [16]:
```
# min-max normalization
x_train_max = np.max(X_train)
x_train_min = np.min(X_train)
X_train = (X_train - x_train_min)/(x_train_max - x_train_min)
X_test = (X_test - x_train_min)/(x_train_max - x_train_min)
```

Machine Learning Models

- The following Machine Learning algorithms are used:

1. Support Vector Machine (SVM)
2. K-nearest Neighbors Classifier (KNN)
3. Logistic Regression Classifier
4. Naïve Bayes Classifier
5. Decision Tree Classifier
6. Random Forest Classifier
7. AdaBoost Classifier

```
[23]
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Define the models
models = [("SVM", SVC(probability=True)),
          ("KNN", KNeighborsClassifier()),
          ("LR", LogisticRegression(solver='liblinear', random_state=0)),
          ("NB", GaussianNB()),
          ("DTC", DecisionTreeClassifier(max_depth = 3)),
          ("RF", RandomForestClassifier()),
          ("AdaBoost", AdaBoostClassifier())]

finalResults = []
cmList = []

# Train and evaluate each model
for name, model in models:
    model.fit(X_train, y_train)
    model_result = model.predict(X_test)
    score = accuracy_score(y_test, model_result)
    cm = confusion_matrix(y_test, model_result)
    finalResults.append((name, score))
    cmList.append((name, cm))

finalResults
```

```
[('SVM', 0.6052899287894201),
 ('KNN', 0.5279755849440488),
 ('LR', 0.6052899287894201),
 ('NB', 0.6256358087487284),
 ('DTC', 0.62970498474059),
 ('RF', 0.6581892166836215),
 ('AdaBoost', 0.6073245167853509)]
```

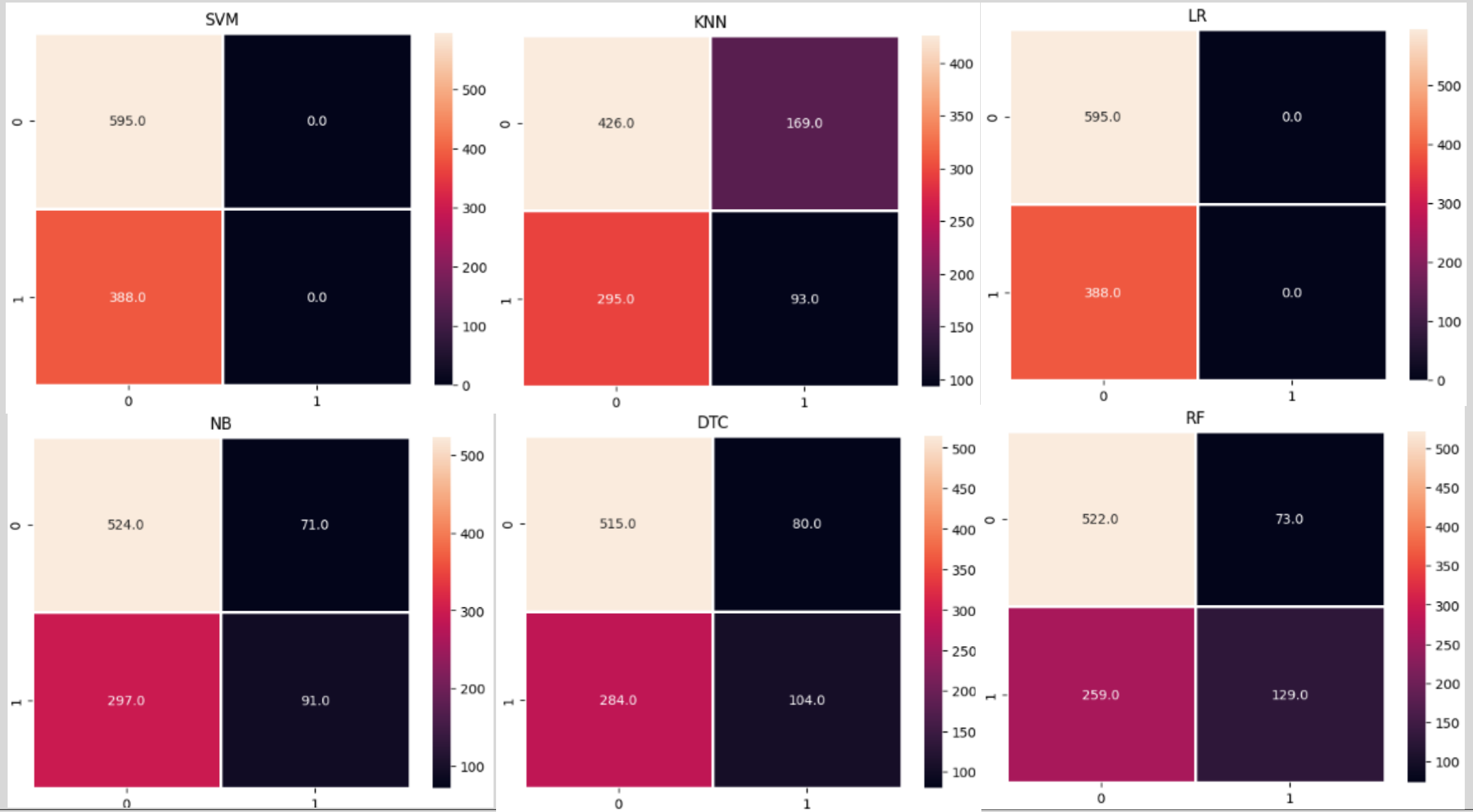
MACHINE LEARNING MODELS

MACHINE LEARNING ANALYSIS AND FINDINGS

Machine Learning Analysis and Findings:

- The machine learning analysis was performed on the water quality testing dataset, and several algorithms were evaluated for their performance in predicting water potability. The algorithms assessed included Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes (NB), Decision Tree Classifier (DTC), Random Forest (RF), and AdaBoost. The findings revealed varying levels of accuracy among the algorithms. RF achieved the highest accuracy with 65.82%, followed by DTC with 62.97%, NB with 62.56%, and AdaBoost with 60.73%. SVM and LR both achieved an accuracy of 60.53%, while KNN lagged behind with 52.80%. These findings suggest that RF, DTC, NB, and AdaBoost exhibit promising performance in predicting water potability based on the provided dataset. Further exploration and optimization of these algorithms can potentially contribute to more accurate predictions and informed decision-making regarding water quality management and safety.

CONFUSION MATRIX OF ALL USED CLASSIFICATIONS:


SVM
595.0
0.0
388.0
0.0
KNN
426.0
169.0
295.0
93.0
LR
595.0
0.0
388.0
0.0
NB
524.0
71.0
297.0
91.0
DTC
515.0
80.0
284.0
104.0
RF
522.0
73.0
259.0
129.0

MODELS FLAWS AND ADVANCED STEPS

Model Flaws and Advanced Steps:

- Every model used in the analysis of water quality testing datasets may have certain flaws that should be considered. Firstly, it's important to recognize that the accuracy achieved by the machine learning algorithms may not be sufficient for all practical applications. The reported accuracies, while informative, do not necessarily capture other important metrics such as precision, recall, or F1 score, which provide a more comprehensive evaluation of model performance. Additionally, the dataset itself may have limitations, such as insufficient data points, imbalanced classes, or missing values, which can impact the reliability and generalizability of the models. Furthermore, the chosen set of features may not encompass all relevant variables, potentially leading to incomplete representations of water quality.

Model Flaws and Advanced Steps:

- Advanced steps can be taken:

1. Employ cross-validation techniques to assess model performance across multiple folds of the dataset. This can help detect any overfitting or generalization issues.
2. Conduct hyperparameter tuning to optimize the algorithms' settings and improve their predictive capabilities.
3. Apply feature engineering approaches, such as creating new derived variables or incorporating domain knowledge, to enhance the representation of the data and potentially improve model accuracy.
4. Explore ensemble methods, such as stacking or boosting, to combine the strengths of multiple models and mitigate individual model weaknesses.

Model Flaws and Advanced Steps:

- Advanced steps can be taken:

1. Collecting more data points, especially from diverse sources and locations, to provide a broader and more representative sample.
2. Deal with imbalanced classes through techniques such as oversampling, undersampling, or the use of appropriate evaluation metrics for imbalanced datasets.
3. Apply imputation methods to handle missing values, ensuring that the dataset is as complete as possible.

THANK YOU!

IBM Machine Learning Professional Certificate

Course 03: Supervised Machine Learning: Classification

By Moustafa Abada