**Kanyiswa Ngalwa**

**Capstone Project- Water Analysis**

**1.Importing Libraries:**

* + numpy as np: Used for numerical operations, especially with arrays.
  + pandas as pd: Essential for data manipulation and analysis, particularly with dataframes.
  + csv: Provides functionality to read and write CSV files.
  + seaborn as sns: A statistical data visualization library based on Matplotlib.
  + os: Allows interaction with the operating system, such as file and directory operations.
  + matplotlib.pyplot as plt: A plotting library for creating static, animated, and interactive visualizations.
  + %matplotlib inline: Ensures that Matplotlib plots are displayed inline in Jupyter Notebooks.

1. **Plotly for Interactive Plots:**
   * plotly.graph\_objs as go: Used for creating interactive plots.
   * plotly.offline as py: Enables Plotly to work in offline mode.
   * init\_notebook\_mode(connected=True): Initializes Plotly to work in offline mode within Jupyter Notebooks.
   * from plotly import tools: Provides various utilities for Plotly.
2. **Time Series Analysis:**
   * from statsmodels.tsa.arima.model import ARIMA: Used for time series analysis and forecasting with ARIMA models.
3. **Handling Warnings:**
   * import warnings: Allows control over warning messages.
   * warnings.filterwarnings('ignore'): Ignores all warnings.
   * warnings.filterwarnings("ignore", category=DeprecationWarning): Specifically ignores deprecation warnings.

**Purpose:**

This setup is commonly used for data analysis and visualization. It includes libraries for handling data, creating both static and interactive plots, and performing time series analysis. The warnings are suppressed to keep the output clean.

**DATA CLEANING**

* Column names: Whitespace removed
* Date format: Converted to proper datetime
* Missing values:
  + TSS (mL sed/L) and Turbidity (NTU) filled with their mean values
* Date filter: Only entries from June 1 to June 30, 2023 were retained

**EDA**

**Histogram**

<blob:https://m365.cloud.microsoft/871dfe06-ea01-482a-ad8f-e0896c2b8c31>

This histogram with a KDE (Kernel Density Estimate) curve helps you understand how pH values are spread across the samples. Most values appear to cluster around neutral to slightly alkaline levels.

**Boxplot**

This visualization helps identify:

* **Median EC levels** at each location
* **Variability** and **outliers** in measurements
* Differences in water conductivity between sites

**Scatter Plot of TDS vs. TSS**

* This plot helps identify how TDS and TSS vary together.
* Clusters or trends by sampling point may indicate differences in water quality across locations**.**

**Correlation Heat Map**

* Correlation coefficients between columns A, B, and C.
* Values close to 1 or -1 indicate strong positive or negative relationships.
* Values near 0 suggest weak or no correlation.

**Line Plot of DO Over Time**

* This visualization helps track how DO levels fluctuate over time.
* Comparing locations reveals differences in oxygenation, which may indicate varying water quality or environmental conditions.

**MODELLING**

* The model perfectly classified both classes of water hardness (BLANDA and SEMIDURA) in the test set.
* Cross-validation confirms the model's robustness across different data splits.

**Modelling Evaluation**

* The model's perfect performance might indicate overfitting, especially if the dataset is small or imbalanced.
* The warnings about in place=True suggest that future versions of pandas may not support this usage

**Conclusion and Future Work**

**This project aimed to analyze river water quality data and develop a predictive model for classifying water hardness.** Through exploratory data analysis (EDA) and machine learning, several key insights were uncovered:

* Water Quality Trends: Parameters such as pH, EC, TDS, and DO varied across sampling points and dates, reflecting environmental and seasonal influences.
* Strong Correlations: EC and TDS showed a strong positive correlation, as expected in mineral-rich water bodies.
* Hardness Classification: Using a Random Forest Classifier, the model achieved:
  + 100% accuracy on the test set
  + 99.1% average cross-validation score
  + Perfect precision, recall, and F1-scores for both BLANDA and SEMIDURA classes

These results demonstrate the model’s robustness and the predictive power of the selected features

**Insight Gained**

* Feature Importance: Parameters like EC, TDS, and hardness levels are highly indicative of water classification.
* Data Quality: The dataset was generally clean, with minimal missing values that were easily imputed.
* Model Reliability: The Random Forest model performed exceptionally well, suggesting that the classification task is well-suited to this algorithm.

**Limitations**

* Overfitting Risk: The perfect accuracy may indicate overfitting, especially given the relatively small dataset.
* Class Imbalance: The dataset had more SEMIDURA samples than BLANDA, which could bias the model.
* Temporal Scope: The data spans only a few months, limiting long-term trend analysis.
* Geographic Scope: Only a few sampling points were included, which may not generalize to other regions or water bodies.