| **Project Title** | **Water Safety Classification** |
| --- | --- |
| **Skills take away From This Project** | **Data Preprocessing and Cleaning, Exploratory Data Analysis (EDA), Feature Engineering, Machine Learning Model Development, Model Evaluation, Streamlit Application Development** |
| **Domain** | **Water Quality and Environmental Health** |

**Problem Statement:**

Ensuring access to safe drinking water is a critical public health concern worldwide. Water contamination from natural or anthropogenic sources poses severe risks to health, agriculture, and industry. Traditional methods of water quality testing are often time-consuming, expensive, and limited by geographic accessibility.

This project aims to develop a machine learning-based classification model to predict whether a water sample is safe for consumption based on its chemical properties. Using a dataset that includes various water quality indicators such as pH, dissolved oxygen, electrical conductivity, and concentrations of harmful substances (e.g., arsenic, lead, and nitrates), learners will create a robust model to classify water samples as **safe** or **unsafe**.

The project will also include building an interactive **Streamlit application** to allow stakeholders (e.g., municipal authorities, NGOs, or environmental agencies) to upload new data for real-time predictions and visualize water quality trends across different regions.

**Business Use Cases:**

* **Government Monitoring:**Governments can deploy this model to monitor and classify water sources, focusing on unsafe samples for immediate remediation.
* **Public Health Advisory:**Public health departments can use the tool to issue timely warnings in areas where water safety is compromised, minimizing waterborne diseases.
* **Industry Applications:**
  + Water purification companies can use insights from the model to identify the pollutants requiring targeted treatment.
  + Industries reliant on water (e.g., agriculture, food production) can assess water suitability before usage.
* **Disaster Management:**During floods or industrial spills, this tool can identify contaminated water sources for focused recovery efforts.
* **Environmental Research:**Supports researchers in understanding the patterns and trends of water pollution in various geographic regions.

**Approach:**

**Data Cleaning:**

* Inspect and clean the ammonia column for inconsistencies and convert it to numeric.
* Handle missing or invalid data if detected in other columns.

**Exploratory Data Analysis (EDA):**

* Visualize distributions of chemical concentrations and their relationship with the target variable is\_safe.
* Examine correlations between features and identify any outliers or anomalies.

**Address Data Imbalance:**

* Analyze the distribution of the target variable.
* Apply techniques like oversampling (SMOTE) or class weighting in models to handle any imbalance.

**Feature Engineering:**

* Scale numerical features to standardize their ranges for ML algorithms.
* Analyze feature importance and eliminate redundant or low-impact features.

**Modeling:**

* Train and compare multiple classification models, including Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting.
* Use hyperparameter tuning methods (e.g., GridSearchCV) to optimize models.

**Evaluation:**

* Measure model performance using accuracy, precision, recall, F1-score, and ROC-AUC.
* Compare class-specific metrics to ensure balanced performance.

**Deployment:**

* Develop a **Streamlit application** for real-time water safety prediction

**Results:**

By the end of the project, learners should aim to achieve the following:

1. **Clean and Preprocessed Dataset:**
   * A cleaned dataset with all missing values handled, data types corrected, and features normalized or scaled.
2. **Exploratory Data Analysis (EDA):**
   * Insightful visualizations and statistical summaries highlighting the relationships between chemical concentrations and water safety.
3. **Trained Classification Model:**
   * A machine learning model capable of accurately predicting water safety (is\_safe) based on the chemical composition of the water.
4. **Model Performance Metrics:**
   * Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC demonstrating the model's effectiveness, particularly in handling any class imbalance.
5. **Interactive Streamlit Application:**
   * A user-friendly web app where users can upload datasets, preprocess data, and receive predictions on water safety in real time.
6. **Documentation:**
   * A well-documented report detailing the project approach, key findings, model evaluation, and deployment process.

**Project Evaluation metrics:**

* Model Performance Metrics: Accuracy, Precision , Recall , F1-Score, ROC-AUC
* Data Imbalance Handling: Performance on both majority and minority classes will be reviewed using class-specific precision, recall, and F1-scores.
* Deployment Evaluation: Functionality and usability of the Streamlit app
* Documentation and Reporting: Completeness and clarity of the final report

**Technical Tags:**

* **Machine Learning**
* **Classification**
* **Water Quality Prediction**
* **Exploratory Data Analysis (EDA)**
* **Imbalanced Data Handling**
* **Model Evaluation**
* **Streamlit Application**
* **Python**
* **Scikit-learn**

**Data Set:**

Data Set: [water-quality-data](https://www.kaggle.com/datasets/mssmartypants/water-quality/data)

**Project Deliverables:**

* **Source Code**: Python scripts for data processing, model training, and the Streamlit app.
* **Model Files**: Trained models and tuning configurations.
* **Data**: Cleaned dataset.
* **Documentation**: A brief report detailing the approach, model evaluation, and deployment instructions.
* **Performance Metrics**: Key model evaluation metrics (accuracy, precision, recall, etc.).

**Project Guidelines:**

**Coding Standards**:

* Write clean, readable code with descriptive names.
* Follow PEP 8 for formatting and style.
* Use functions and modules to organize code.

**Version Control**:

* Use Git for version control and clear commit messages.

**Testing**:

* Test data preprocessing, model evaluation, and the Streamlit app thoroughly.

**Documentation**:

* Include a README.md for setup and usage instructions.
* Provide a project report detailing approach and results.

**Deployment**:

* Ensure the Streamlit app is deployable and provide deployment instructions.

**Timeline:**

The project must be completed and submitted within **7 days from the assigned date.**

**PROJECT DOUBT CLARIFICATION SESSION ( PROJECT AND CLASS DOUBTS)**

**About Session:** The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

**Note: Book the slot at least before 12:00 Pm on the same day**

**Timing: Monday-Saturday (4:00PM to 5:00PM)**

**Booking link :**[**https://forms.gle/XC553oSbMJ2Gcfug9**](https://forms.gle/XC553oSbMJ2Gcfug9)

**LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)**

**About Session:** The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

**Note: This form will Open only on Saturday (after 2 PM ) and Sunday on Every Week**

**Timing: Monday-Saturday (05:30PM to 07:00PM)**

**Booking link :** [**https://forms.gle/1m2Gsro41fLtZurRA**](https://forms.gle/1m2Gsro41fLtZurRA)