**Project Report: Water Quality Prediction & Analysis**

**Objective:**  
To build a machine learning model that can predict whether water is safe to drink (potable) based on various chemical attributes.

**Dataset Overview:**

* Source: data.csv file
* Features: ph, Temperature, Turbidity
* Target: Potability (0 = Not drinkable, 1 = Drinkable)

**Tools & Libraries Used:**

* Python
* Pandas, NumPy (data manipulation)
* Seaborn, Matplotlib (visualization)
* Scikit-learn (modeling and evaluation)

**Key Steps:**

1. **Data Loading & Inspection**
   * Loaded using pandas.read\_csv()
   * Checked data types and missing values
2. **Data Cleaning**
   * Removed rows with missing values using .dropna()
   * Renamed columns for consistency and readability
3. **Exploratory Data Analysis (EDA)**
   * Used histograms and count plots to explore distributions
   * Correlation matrix plotted to identify feature relationships
4. **Feature Selection & Splitting**
   * Selected all features except Potability as input variables
   * Split data into training and testing sets (80-20 ratio)
5. **Model Building**
   * Used Random Forest Classifier
   * Trained model using fit() method
6. **Model Evaluation**
   * Evaluated using accuracy, confusion matrix, and classification report
7. **Export for Visualization**
   * Cleaned dataset exported to cleaned\_water\_quality.csv for Power BI dashboard

**Results:**

* Achieved reasonable accuracy in classification
* Model performed better after dropping missing values

**Next Steps:**

* Improve model with hyperparameter tuning
* Apply feature scaling or engineering
* Deploy model as an API for real-time predictions
* Integrate with Power BI for interactive dashboards

**Conclusion:**  
The project successfully demonstrated the use of Python for predictive analytics in water quality. It included a complete pipeline from raw data to a working model, with clean exports ready for visualization.

**Prepared by:** Jay Kamavisdar