

# **WATER QUALITY ANALYSIS**



# DEFENITION

**Water quality analysis is also called hydrochemical analysis. That is to use chemical and physical methods to determine the content of various chemical components in water. Water quality analysis can be divided into three types: simple analysis, complete analysis and special analysis.**

# PROJECT OBJECTIVES

- **Water quality objectives are designed for the substances or conditions of concern in a watershed so that their attainment will protect the designated uses.**

- **ANALYSIS APPROACH**

- **Water quality parameters are of three types – physical, chemical and biological – and are tested or monitored according to the desired water parameters. Water quality parameters often sampled or monitored include pH, ORP, conductivity, dissolved oxygen, chlorine, salinity, ozone, and corrosion**

# VISUALIZATION TECHNIQUES

- Data visualization is an important tool for communicating science to a broader audience. Whether you are a volunteer community scientist or a professional aquatic ecologist, there are many free tools and low-cost programs that you can use to link the scientific data to actions that can improve water quality. We will walk you through the process with some tips and tricks on how to communicate your results most effectively

# PYTHON AND LIBRARIES

- **Being able to provide enough fresh drinking water is a core requirement. Within the climate change debate, one of the largest challenges is ensuring enough freshwater to survive. Water quality is a big concern that impacts all the specifics. Only about three percent of Earth's water is freshwater. Of that, only 1.2 percent can be used as drinking water, with the remainder locked up in glaciers, ice caps, and permafrost, or buried deep in the ground. Using a data-driven approach to assess the features that impact the water quality could greatly improve our understanding of what makes water drinkable.**

# DATA SOURCE

- <https://www.kaggle.com/datasets/adi>




# SOURCE CODE

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as ex
import plotly.graph_objs as go
import plotly.offline as pyo
import scipy.stats as stats
import pymc3 as pm
import theano.tensor as tt
```

Numpy- NumPy is the fundamental package needed for scientific computing with Python.



- 
- **Numpy** - NumPy is the fundamental package needed for scientific computing with Python. Pandas - Python library used to analyze data.
  - **Matplotlib** - Most of the Matplotlib utilities lies under the pyplot submodule.
  - **Seaborn** - An open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis.
  - **Plotly** - provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.
  - **Scikit-learn** - tool for predictive data analysis built on numpy, scipy and matplotlib.

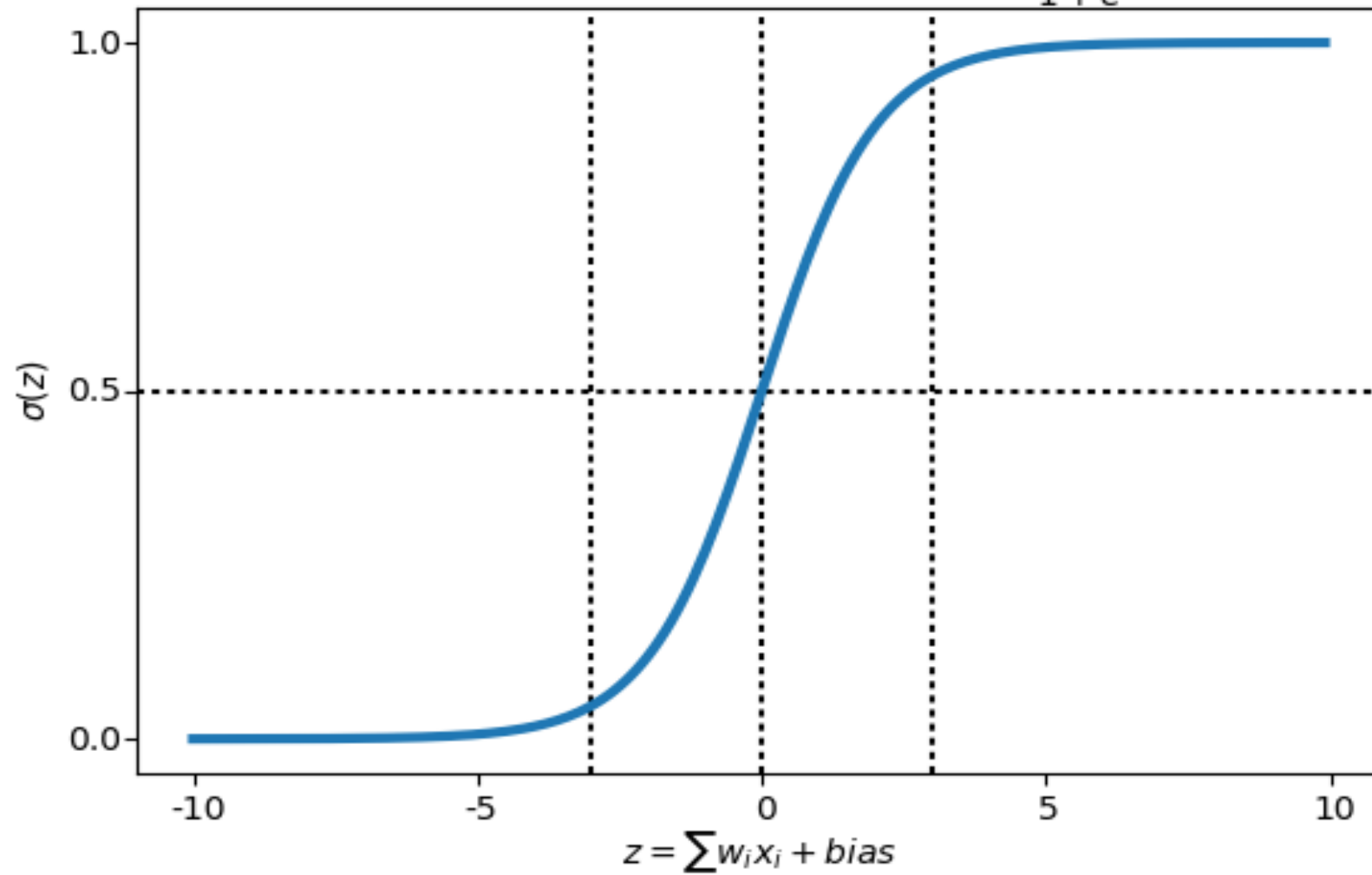
# IMPORTING DATASET

- **`df=pd.read_csv('water_potability.csv')`**
- If your notebook and csv files are in different places you can write the whole path to import the file.
- **`df=pd.read_csv('../input/water-potability/water_potability.csv')`**

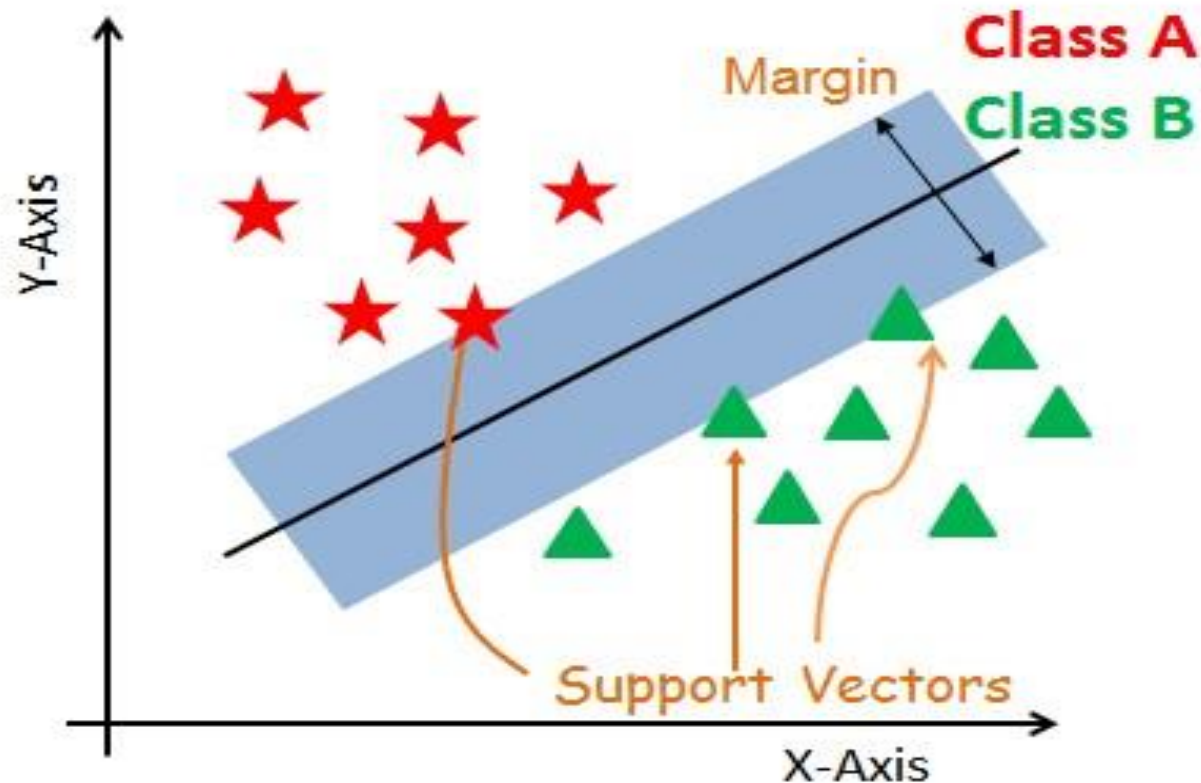
# MODEL USED FOR TRAINING


- Logistic Regression - Logistic Regression is named for the function used at the core of the method, the logistic function.
- The [logistic function](#), also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Sigmoid Function  $\sigma(z) = \frac{1}{1 + e^{-z}}$



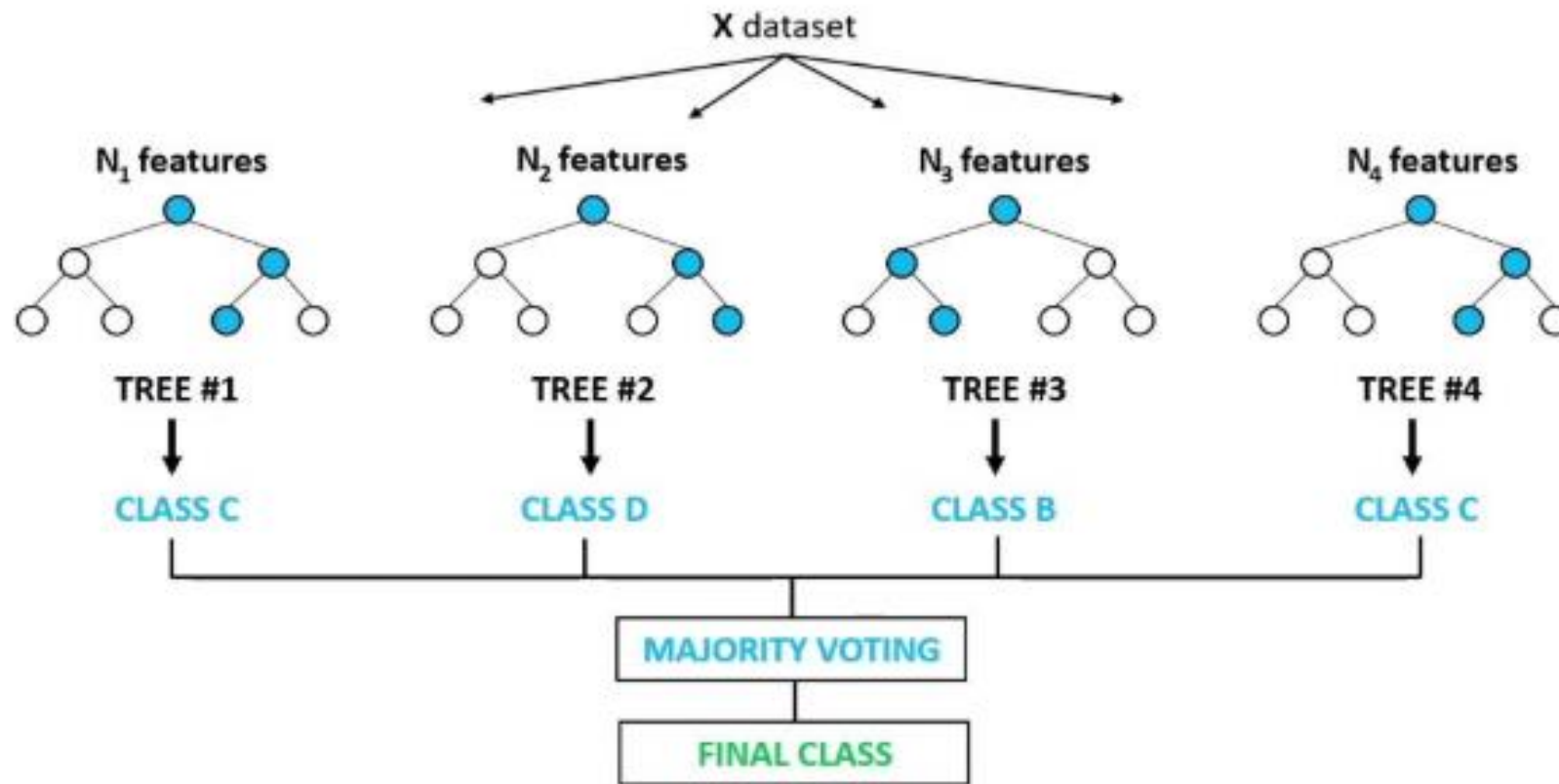
Support Vector Classifier - The objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides, or categorizes your data.






**Random Forest Classifier - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.**

# Random Forest Classifier





- 
- **XGBoost** - XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM).