أكاديمية سدايا
SDAIA Academy

Applying Different Classification Algorithms to Predict the Water Potability

Raghad Aloraini
Raghad Alawad
2 Oct 2021

Introduction

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

We plan to make a prediction to diagnose whether the water is safe to drink or not, based on certain diagnostic metrics included in the data set.

Dataset

The datasets we will be using are taken from the kaggle website (https://www.kaggle.com/adityakadiwal/water-potability). By using this dataset, we are trying to make predictions of the water potability, using features such as;

- ph: pH of 1. water (0 to 14).
- Hardness: Capacity of water to precipitate soap in mg/L.
- Solids: Total dissolved solids in ppm.
- Chloramines: Amount of Chloramines in ppm.
- Sulfate: Amount of Sulfates dissolved in mg/L.
- Conductivity: Electrical conductivity of water in μS/cm.
- Organic_carbon: Amount of organic carbon in ppm.
- Trihalomethanes: Amount of Trihalomethanes in μg/L.
- Turbidity: Measure of light emiting property of water in NTU.
- Potability: Indicates if water is safe for human consumption. Potable: 1, not potable: 0

Algorithms

For data cleaning and pre-processing, we will start by deleting the duplicate records and replace NaN values with medians of those columns.Also, and then we will do Visualization for the data to understand it more. Also, Splitting data into train and test sets, and standardizing the data.

Also, we will apply different classification models such as: (Logistic Regression Model, K-nearest neighbors, Decision Trees, Random Forest Classifier) we will compare the results(accuracy) of each model.

Tools

To predict the water potability, we will be using different tools such as Jupyter notebook, Excel. Also, we will use different libraries with python such as pandas, sklearn, numpy.

Conclusion

To conclude, we expect the classification model will predict the property (outcome), this research helps people with investments in water supply and sanitation.