Application of Machine Learning On Water Quality

**Ethan Kong**

# ABSTRACT

The Standard Drinking Water Test is considered essential for determining potability(safeness to drink). However, its complexity poses challenges as it requires multiple labs to conduct various tests on the water. Additionally, limited access to resources prevents many countries from implementing this test effectively. This research paper utilizes machine learning models to determine the quality and the potableness of water. pH, hardness, solids, chloramines, sulfates, conductivity, organic carbons, trihalomethanes, and turbidity are used to determine if the sample of water is potable. I trained multiple different machine learning models such as logistic regression, decision tree, random forest, KNN, and XG Boost to analyze the potability of new water sources. The dataset's final, most accurate model is the random forest model, with an accuracy of 71%, precision of 74% and f1-score of 52%. Furthermore, the model has determined that pH is the most important value in determining water quality.

# INTRODUCTION

Water is essential for the survival of every organism on earth, and it is critical that the water we drink is clean and will not cause any adverse effects on our bodies. Ensuring access to clean drinking water for all citizens is a crucial objective, particularly in a developed country like America. However, America still has 7.15 million water borne illnesses per year (CDC 2023). By establishing a more accessible and cost-effective method for determining water potability, the safety of every individual can be safeguarded. In order to solve this problem, we use a numerical dataset from Kaggle that measured concentrations of 9 chemicals in the water and used supervised learning to classify each data point as potable or nonpotable.

**BACKGROUND**

Water quality refers to the chemical, physical, and biological characteristics of water that determine its suitability for various purposes, including human consumption. Water quality is measured through a range of tests and analyses conducted in laboratories. These tests evaluate various factors, such as the levels of bacteria, chemicals, heavy metals, and other substances present in the water. The Standard Drinking Water Test, commonly used as an industry standard, involves a comprehensive analysis that requires multiple tests and specialized equipment. When we refer to potable water, we mean water that is safe for human consumption and poses no health risks when consumed. Multiple other similar research papers have been written about using machine learning models to determine water quality. In particular, a paper published by Menguan Zhu et al evaluated 45 different machine learning models’ abilities to complete water-related tasks, For predicting water quality, the paper used decision tree, random forest, discounted cash flow, and 10 other models to evaluate which model was the most efficient. Their dataset had 33612 values but only 4 features, including pH, DO, CONMn, and NH3–N. Their results found that decision trees, random forests, and discounted cash flow models provided the best predictive performance. (Mengyuan Zhu et al 2022)

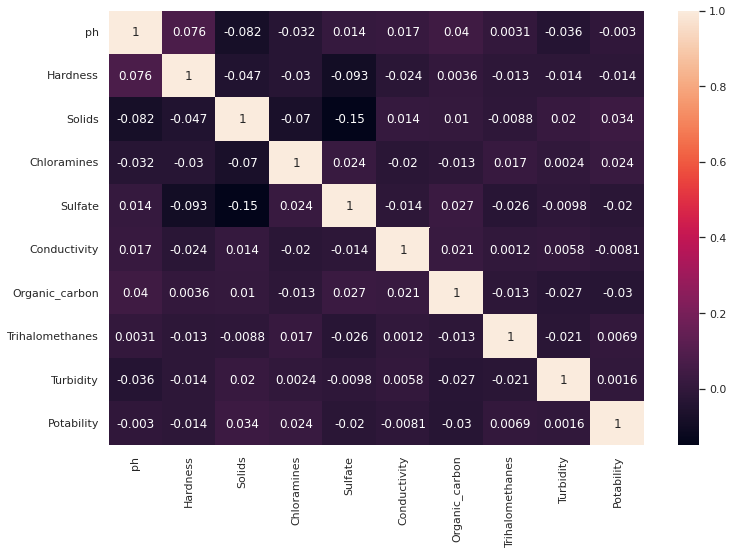
**DATA**

The data used comes from the Water Potability Dataset on Kaggle (Kadiwal). The dataset provides information for 10 features including pH, conductivity, and concentrations of different chemicals. All of the variables in this dataset are numerical, so there is no need to convert any strings into numbers. The 10 features are as included.

* pH describes how acidic or basic the water is. It isrecommended not to drink water outside the pH range of 6.5-8.5. (WHO)
* Hardness measures the concentration of magnesium and calcium in the water. 0 to 60 mg/L (milligrams per liter) is classified as soft; 61 to 120 mg/L as moderately hard; 121 to 180 mg/L as hard; and more than 180 mg/L as very hard. In general, hardness from 80 to 100 mg/L are ideal (USGS).
* Solids describe how many minerals or salts such as sodium, potassium, etc are dissolved in the water. The typical lower limit for solids is 500 mg/L and the upper limit is 1000 mg/L (WHO).
* Chloramines measure the concentration of chlorine in water and should not exceed 4 mg/L (CDC).
* Sulfate looks at the concentration of the ion SO4-2 which are naturally occurring substances formed when sulfuric acid dissolves in water. In freshwater, the concentrations of sulfates range from 3 to 30 mg/L (EPA).
* Conductivity tells us how many charged ions positive or negative are dissolved in the water which affects how well the water can conduct electricity. In general, the conductivity value should not exceed 400 micro siemens per centimeter (Jones).
* Organic Carbon measures the amount of carbon from organic compounds like plants. According to the United States Environmental Protection Agency drinking water should have less than 2mg/L of organic carbon (Moore).
* Trihalomethanes are chemicals found in water treated with chlorine and should not exceed 80 mg/L (Florida Department of Health).
* Turbidity is a measure of how much light is reflected off the water. The WHO recommends a turbidity value of 5.00 NTU (EDA).
* Potability indicates whether the water is safe to drink, and instead of using strings like yes and no the dataset uses 1 to represent the water being potable and 0 as not potable.

## FEATURE CORRELATION

Below is a correlation heatmap that is used to quickly gauge what features are expected to have the largest effect on the potability. The values represent how much correlation is between each feature so for example, as ph increases hardness increases as well, but solids decrease.

****

**Figure 1:** Correlation heatmap between features

Looking at this correlation heatmap, there does not seem to be any one feature that directly contributes to the potability score, but the largest correlator seems to be solids and we can see how this compares to the final feature importance graph.

**METHODOLOGY**

## DATA PRE-PROCESSING

Before running any models, the dataset must be pre-processed. First, there were quite a few values that were blank, specifically for the Sulfates, Trihalomethanes, and pH. To fix this problem, I deleted entire data points if they were missing 2 or more values but for those only missing 1 value, the median values for the feature were used to replace the null. Here are the median values for each feature:

* pH: 7.0368
* Sulfates: 333.0735
* Trihalomethanes: 66.6225

## SPLITTING THE DATASET

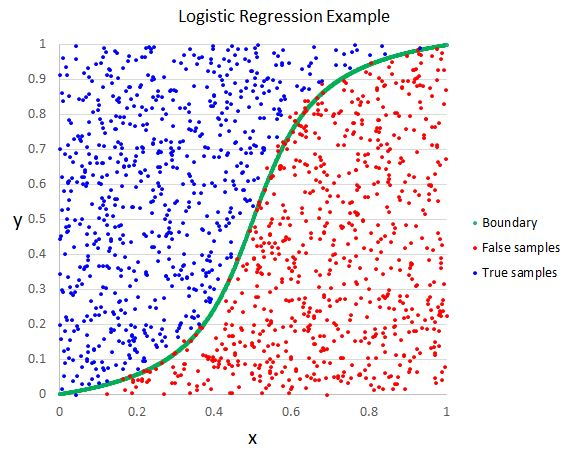
The dataset must also be split into a training set, a validation set, and a testing set. The reason for this is that if the same set of data is used for both training and testing the model, the model would just memorize all the datasets and the answers instead of genuinely predicting the potability based on the data. Using the train\_test\_split function, the dataset was split in a 0.70 to 0.15 to 0.15 ratio to allow for a maximum amount of training data while allowing for a decently high amount of data to calculate the accuracy. The size of each split are 70% (2293 samples) train , 15% (492 samples) validation and 15% (491 samples) in test.

## HYPERPARAMETER TUNING

In order to maximize the accuracy of each model, hyperparameters of each model must be tuned. When training any machine learning models, there are always a plethora of parameters which are settings that control how the model looks at and learns from the data it is given, such as how much weight to put on a feature. Using the results from the research papers I read, I put more time and training into developing the decision tree and random forest models, as according to the research papers they were more accurate. In order to make sure the optimal hyperparameters were used for each model, I first researched what the most important hyperparameters for each model were because I wanted to make sure I didn’t tune a parameter that wouldn’t even affect the model’s performance. After deciding on the hyperparameters to tune, we used grid search to optimize the values of each parameter. Grid search takes in values for each parameter and cross-searches all possible combinations of parameter values to find the value combinations with the highest accuracy and outputs the results.

## LOGISTIC REGRESSION

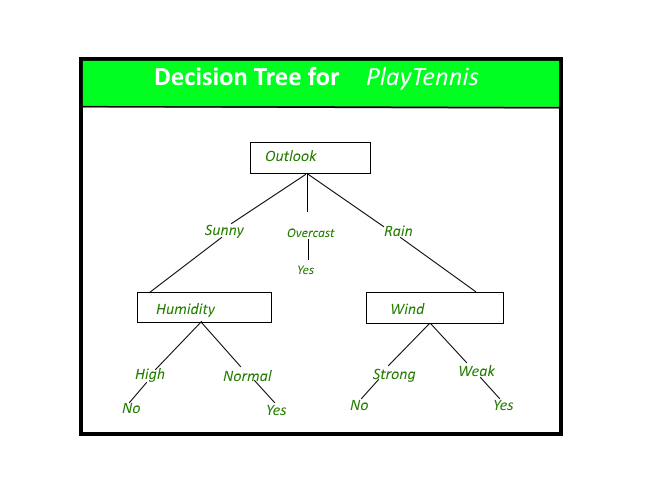
Logistic regression is the most basic model which just looks at data points and tries to fit a logistic function to classify the data points. As this is meant to be a baseline, no hyperparameter tuning was used.



**Figure 2:** Graphic of a linear regression model

## DECISION TREE

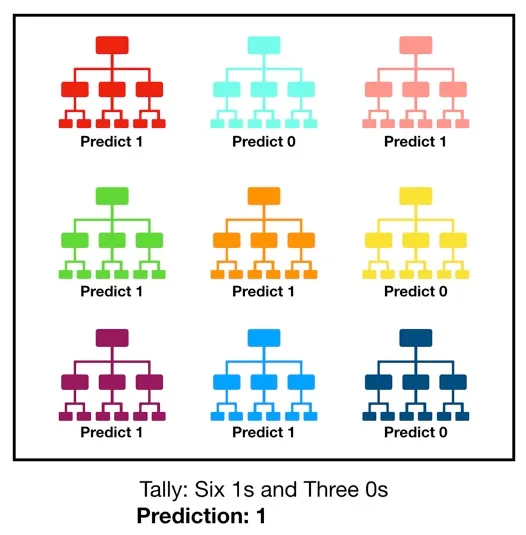
The decision tree model is useful for classifying a wide range of data points. It proposes a basic if statement and based on the result it classifies the data further and repeats until there are no more nodes to run through.



**Figure 3:** Graphic of a decision tree

## RANDOM FOREST

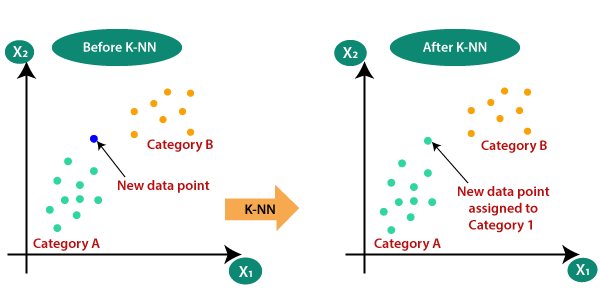
A Random forest model is essentially a more advanced version of the decision tree model. It creates many decision tree models that each make a classification. In the end, the random forest model looks at which classification has the most amount of decision trees classified as and that is the prediction.



**Figure 4:** Representation of a random forest model

***K-Nearest Neighbors(KNN)***

The KNN is a classification model that stores all previous data and classifies the new data based on how close that point is to other data points.



**Figure 5:** Graphic of a KNN model

***XGBoost***

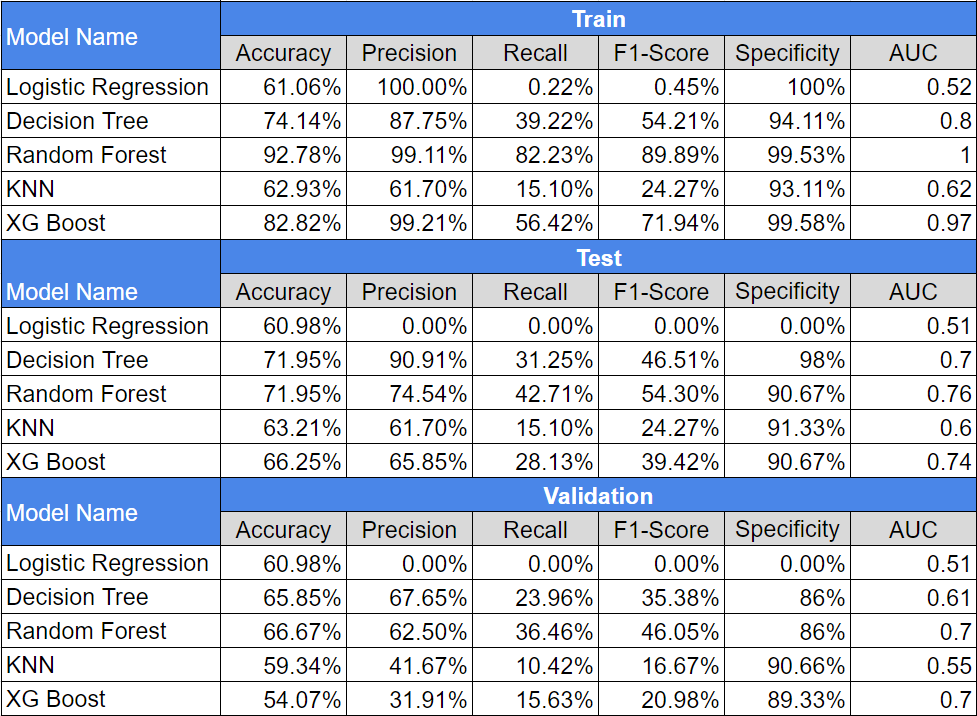
The XGBoost model is extremely similar to the random forest model in which many decision trees are used and the final prediction looks at all the predictions of the individual trees. The difference comes in which when the XGBoost builds a new decision tree it looks at what points were misclassified in the previous tree and tries to mainly classify those points correctly.

***Results Comparison***

The models will be compared by accuracy, precision, recall, F1-score, and the area under the curve. Accuracy is the most basic result of seeing what percentage of the values were predicted properly Precision measures how many of the predictions predicted as potable were correct Recall is the opposite of precision in that it measures what percentage of the potable values in the data were predicted correctly F1-score combines both precision and recall and finds its harmonic average. Specificity is the inverse of recall and tells us how many of the values predicted as unpotable were correct . AUC stands for area under the Receiver Operating Characteristic curve and measures the model’s ability to distinguish between potable and unpotable cases.

Since the model is meant to help predict whether water is potable. The worst possible prediction is when unpotable water is predicted as potable, causing people to drink unsafe water and possible health issues. Since Specificity is the measure of the chance that water predicted as unpotable is predicted correctly, it is preferred to have a higher specificity to prevent people from drinking unsafe water.

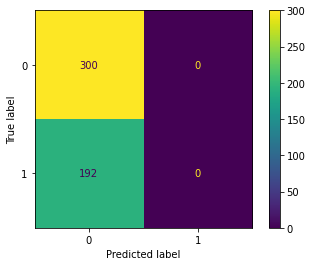
**RESULTS**

****

**Figure 6:** Datatable of training, validation and testing metrics for all models

## LOGISTIC REGRESSION

This model has an accuracy of just 60.98% but it also has a precision, recall, and F1-score of 0%, and an AUC very near 0.5, meaning the model predicted every datapoint the same. Taking a look at the confusion matrix, all 492 data points were predicted as unpotable.



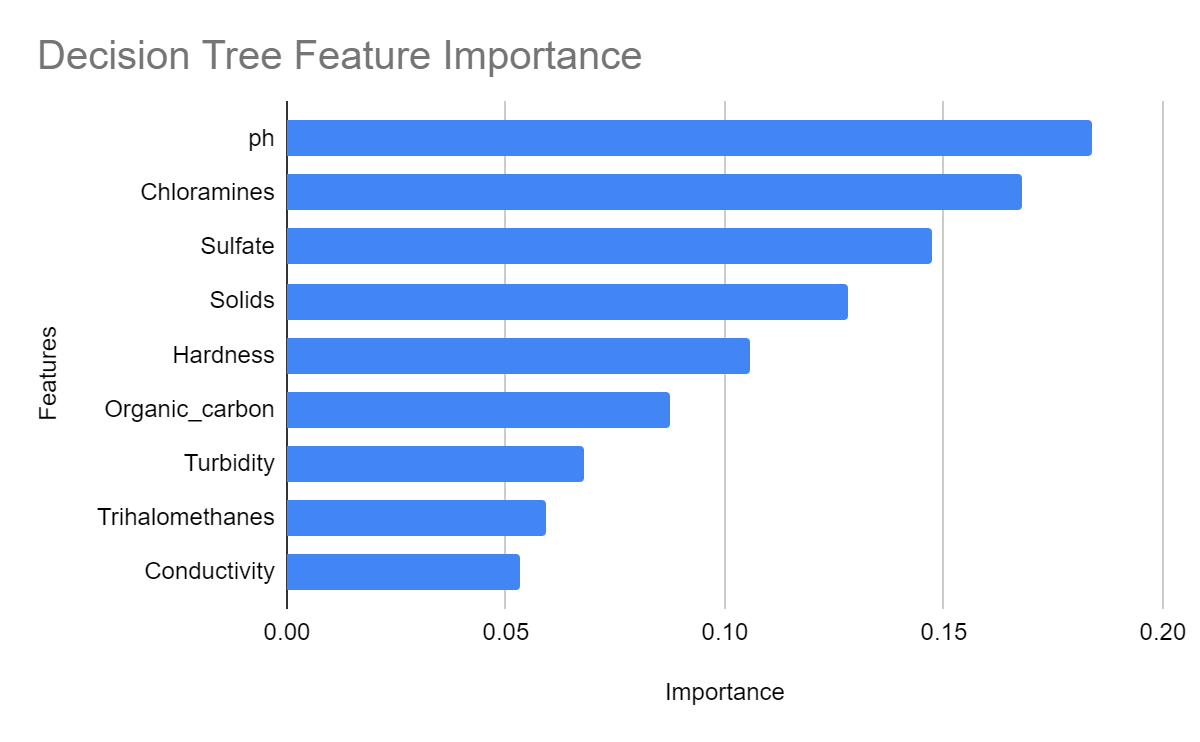
**Figure 6:** Confusion matric for the logistic regression model

## DECISION TREE

For this model, the optimal hyperparameters are:

* max\_depth=15
* max\_features=4
* min\_samples\_split=1e-14
* min\_impurity\_decrease=1e-14
* min\_samples\_leaf=0.001

The feature importance graph can help to analyze the importance of each feature in our decision tree classifier.



**Figure 7:** Feature importance for the decision tree model

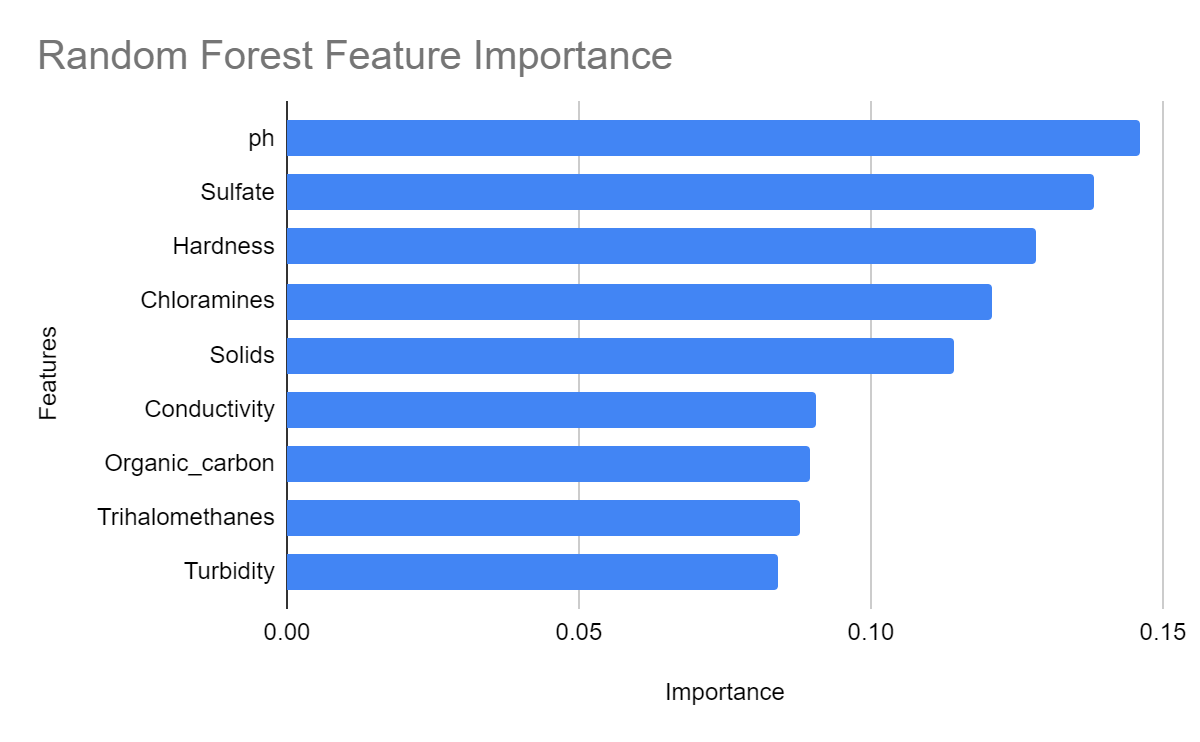
pH is observed to have the greatest importance, closely followed by chloramines. This makes sense as ph measures the acidity of the water and ph that is not within the WHO recommended range of 6.5-8.5 can begin to corrode the tissues inside our body. Chlorine is also an extremely toxic chemical that can cause scarring in the intestinal tracts if ingested, so it is extremely important the chloramine levels are not too high. There also seems to be a very gradual decline in the importance of each feature with Turbidity, Trihalomethanes, and conductivity being among the least important features.

## RANDOM FOREST

For this model, the optimal hyperparameters are:

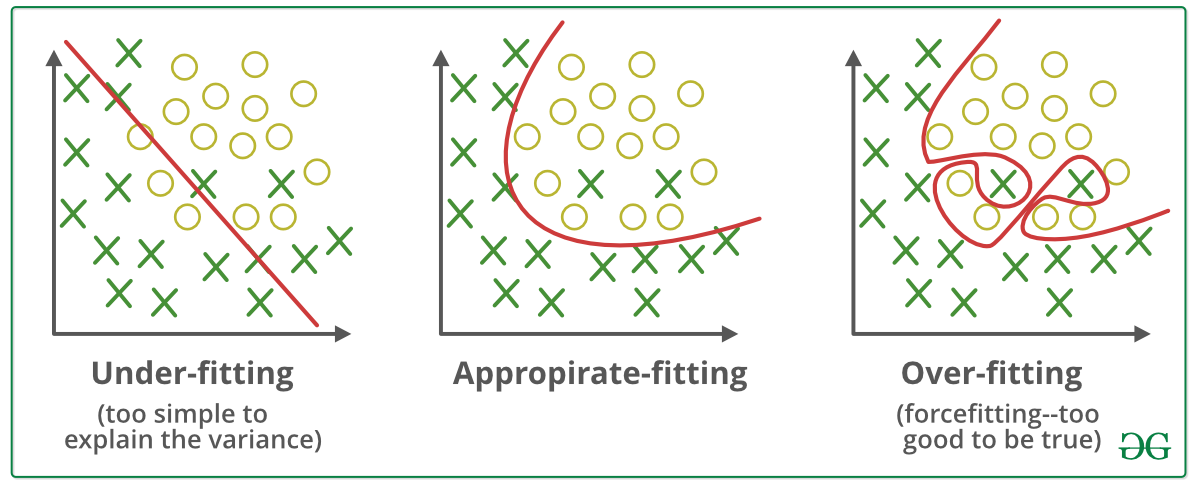
* criterion='gini'
* max\_depth=17
* max\_features=0.78
* max\_leaf\_nodes=174
* n\_estimators=124

The random forest model is the most accurate of all the models we trained, with an accuracy of 66.67%.



**Figure 8:** Feature importance for the random forest model

Looking at the feature importance graph for this model, all 9 features have much closer amounts of importance than the decision tree in which the difference between the most important and least important features is substantially smaller. The maximum importance does not even reach 0.15 compared to the decision tree, where the maximum importance was 0.183, and the minimum value for the random forest is 0.03 higher than that of the decision tree. This is likely because random forest uses more than 1 decision tree; therefore, there is much less bias, and each feature is considered much more equally. Ph is still the most important feature in this model, but interestingly sulfates and hardness are much more important, going from 3rd and 5th to 2nd and 3rd in importance. The model also has much higher precision than recall. From [figure 6] the training metrics of the random forest model are near perfect, with every value in the high 80s or 90s and an AUC score of 1. Whereas in the validation metrics, they are substantially lower with the accuracy and precision in the low 70s, recall taking the biggest hit falling by 45%, f1 score dropping by over 40%, and the AUC being 0.26 lower. Such large differences between training and validation metrics tell us our model overfitted the dataset. Overfitting means the model is trained to fit the training dataset too closely, so it is likely unable to predict newer data points with the same accuracy.



**Figure 9:** Graphic of different types of fitting

In order to reduce overfitting, the most general approach for any machine learning model is to increase the amount of training data. Increasing training data will allow for more variations, making it much harder for the model to train itself to fit the training data specifically. Specifically for the random forest model, the model's maximum depth can be lowered, which lowers the maximum number of decision layers in each decision tree, decreasing the model’s overall complexity and making it able to predict varying types of data more accurately.

**CONCLUSION**

Although the models' accuracy is not exceptional, a lot can still be learned from the results. We can infer that more accurate models will use every feature in the dataset as the feature importances of our random forest model were much more compact than that of the decision tree. This information tells us that although it may seem like certain features are much more important than others, once we get more accurate models, all 9 features will be equally as important in determining the potability. The largest reason for our model’s low accuracy is our relatively small dataset; with a larger dataset, the model could be drastically improved. This also raises questions like whether more features would allow us to create more accurate models. I also wonder if injecting the median values into nulls was the best option and if we should have removed those data points entirely to avoid creating outliers. In the future, I would like to try out different models like SVC, ANN, Soft Sensor, and more.

# References

Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie; Water quality prediction using machine learning methods. Water Quality Research Journal 1 February 2018;53(1):3–13.doi:<https://doi.org/10.2166/wqrj.2018.025>

Florida Department of Health. “Chemicals in Drinking Water Florida Department of Health, Bureau Of ...” *Total Trihalomethanes (TTHMs)*, May 2017, www.floridahealth.gov/environmental-health/drinking-water/\_documents/thms-fs-2017.pdf.

GeeksforGeeks. “A Decision Tree for the Concept PlayTennis.” *Decision Tree*, GeeksforGeeks, 28 Mar. 2023, https://www.geeksforgeeks.org/decision-tree/ .

Geeksforgeeks. “Overfitting Examples.” *ML | Underfitting and Overfitting*, Geeksforgeeks, 20 Feb. 2023, https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/ . Accessed 7 Apr. 2023.

“Hardness of Water Completed.” *Hardness of Water | U.S. Geological Survey*, U.S. Department of the Interior, 11 June 2018, https://www.usgs.gov/special-topics/water-science-school/science/hardness-water#:~:text=Measures%20of%20water%20hardness&text=General%20guidelines%20for%20classification%20of,mg%2FL%20as%20very%20hard.

Javatpoint. “Visualization of a KNN Model.” *K-Nearest Neighbor(KNN) Algorithm for Machine Learning*, Javatpoint, 2021, https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning . Accessed 7 Apr. 2023.

Jones, Sarah. “Conductivity in Drinking Water - Water Library: Acorn Water.” *H2OLabcheck*, 2 Jan. 2020, www.h2olabcheck.com/blog/view/conductivity#:~:text=The%20limit%20for%20drinking%20water,centimetre%20(%C2%B5S%2Fcm).

Kadiwal, Aditya. “Water Quality.” *Kaggle*, 2021, https://www.kaggle.com/datasets/adityakadiwal/water-potability. Accessed 7 Apr. 2023.

Kouadri, S., Elbeltagi, A., Islam, A.R.M.T. et al. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl WaterSci11,190(2021). https://doi.org/10.1007/s13201-021-01528-9

Mengyuan Zhu, Jiawei Wang, Xiao Yang, Yu Zhang, Linyu Zhang, Hongqiang Ren, Bing Wu, Lin Ye, A review of the application of machine learning in water quality evaluation, Eco-Environment & Health, Volume 1, Issue 2, 2022, Pages 107-116, ISSN 2772-9850, <https://doi.org/10.1016/j.eehl.2022.06.001>. (<https://www.sciencedirect.com/science/article/pii/S2772985022000163>)

Moore, D.R.J. “British Columbia Approved Water Quality Guidelines: Aquatic Life ...” *Ambient Water Quality Criteria for Organic Carbon in British Columbia*, Apr. 1998, www2.gov.bc.ca/assets/gov/environment/air-land-water/water/waterquality/water-quality-guidelines/approved-wqgs/wqg\_summary\_aquaticlife\_wildlife\_agri.pdf.

“Sulfate in Drinking Water.” *EPA*, 6 Mar. 2012, archive.epa.gov/water/archive/web/html/sulfate.html.

“Waterborne Disease in the United States.” *Centers for Disease Control and Prevention*, U.S. Department of Health & Human Services, 4 Jan. 2023, https://www.cdc.gov/healthywater/surveillance/burden/index.html#:~:text=About%207.2%20million%20Americans%20get,from%20diseases%20spread%20through%20water.

Wiki, Artificial Intelligence. “Logistic Regression Example.” *Logistic Regression*, 2020, https://machine-learning.paperspace.com/wiki/logistic-regression . Accessed 7 Apr. 2023.

Yiu, Tony. “Visualization of a Random Forest Model Making a Prediction.” *Understanding Random Forest*, Towards Data Science, 12 June 2019, https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree. Accessed 7 Apr. 2023.