SEP769
CYBER-PHYSICAL SYSTEMS & DEEP LEARNING

Water Quality Detection

Prepared for:

- Dr. Hamidreza Mahyar

Prepared by:

- Nour Ziena
- Catherine Chan
- Noor-adien Al-Shrah
- Hamid Sourghali

Contents

Abstract

All living organisms on earth rely heavily on water as one of their primary sources of nutrition. To ensure the stability and protection of the ecosystem, treated wastewater discharge quality must be monitored. Laboratories require a great deal of time and resources to collect and analyze water samples. Given access to data and machine learning algorithm models, this report aims to widely apply this computational power to distinguish between safe and unsafe water quality in a more efficient timeline.

1. Introduction

Environmental and public health is directly impacted by water quality. In addition to drinking, water is used in agriculture and industry. To protect public health, access to safe drinking water is a fundamental human right. If drinking water at an unsafe level, containing contaminants, it can potentially cost serious health issues such as gastrointestinal sickness, and chronic diseases such as cancer. There are different water contaminants that will affect the quality of water, such as industry and agriculture, human and animal waste, and even natural sources that will have an impact on the water quality. This is an important health and development issue on an international, regional, and local level. To assure humans have their essential needs and maintain good health. The reduction in adverse health effects and health care costs associated with investing in water supply and sanitation has been shown to outweigh their costs in some regions. It is time-consuming and sometimes expensive to assess water quality using conventional laboratory techniques. Water quality can be predicted within a short timeframe using the algorithms proposed in this paper. There are significant consequences of poor water quality for both aquatic life and the ecosystem as a whole. We conducted this systematic review in order to identify the effectiveness of applying machine learning methodologies to estimate water quality parameters.

1.1 Dataset Description

The dataset selected for this project is multivariate and comprised of 3276 instances of different water bodies, with 10 total attributes and 1 class. The 10 are used to predict Potability, and the class represents the safety of consumption, 0 for unsafe and 1 for safe. A split of 80/20 was decided by the team

for the limited dataset. A total of 80% of the data is used for training each methodology, and the remaining 20% is used for testing. The large testing dataset will allow for in-depth comparisons as is the goal. With respect to the dataset, that is approximately 2620 elements for training and approximately 1105 for testing.

Attribute	**Description**
pH value	When assessing the acid-base balance of water, PH is a key parameter. Additionally, it indicates whether water is acidic or alkaline. There is a maximum permissible pH range of 6.5 to 8.5 recommended by WHO. According to the current investigation, the range was between 6.52 and 6.83, which is within the WHO standard range.
Hardness	Calcium and magnesium salts are responsible for hardness. Water travels through geologic deposits where these salts dissolve. A raw water's hardness is determined by how long it has been in contact with hardness-producing material. According to the original definition of hardness, calcium and magnesium precipitate soap in water when they react together

Solids (Total dissolved solids - TDS)	Potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates, and more are all soluble in water, including some organic minerals or salts. Water appearance was diluted by these minerals, resulting in an unwanted taste and color. In order to use water effectively, this parameter is essential. TDS values above 20 indicate heavily mineralized water. To ensure that drinking water is safe to consume, there is a maximum limit of 1000 mg/l and a desirable limit of 500 mg/l.
Chloramines	Public water systems use chlorine and chloramine for disinfection. In treating drinking water with chlorine and ammonia, chloramines are most commonly formed. Generally, drinking water should contain no more than 4 milligrams of chlorine per liter (4 mg/L or 4 parts per million (ppm)).
Sulfate	Natural substances such as sulfates can be found in minerals, soil, and rocks. There are many sources of contaminants in the environment, including the air, groundwater, plants, and food. The chemical industry is the main commercial user of sulfate. Approximately 2,700 milligrams of sulfate are present in seawater. Some geographic areas have much higher concentrations (1000 mg/L) than others (usually 3 to 30 mg/L in freshwater supplies).
Conductivity	In contrast to the conductivity of electrical current, pure water is an excellent insulator. Water's electrical conductivity increases with increasing ion concentration. Water's electrical conductivity is primarily determined by its dissolved solids content. Electric conductivity (EC) is actually the measure of how efficiently a solution transmits current through its ionic process. WHO recommends that the EC value should not exceed 400 μS/cm
Organic_carbon	Biologically decomposing organic matter (NOM) and synthetic materials make up the total organic carbon (TOC) in source waters. The total organic carbon content (TOC) in pure water measures the total amount of carbon incorporated into organic compounds. In treated / drinking water, TOC levels are set at 2 mg/L, and in source water that is treated, TOC levels are set at 4 mg/L, according to the US EPA.
Trihalomethanes	In water that has been treated with chlorine, THMs can be found as chemicals. It is important to consider that THM concentrations in drinking water depend on several factors, including organic material in the water, the amount of chlorine that must be added to treat the water, and the temperature of the water being treated. It is considered safe to drink water with a THM content of up to 80 parts per million.
Turbidity	There is a direct relationship between the amount of solid matter present in a suspended state and the degree of turbidity in water. Using this test, waste discharge is analyzed for colloidal matter as well as its light-emitting properties. In comparison to the WHO-recommended turbidity value of 5.00 NTU, Wondo Genet Campus had a mean turbidity of 0.98 NTU.

Class	Description
Potability	The number 1 indicates potable(Safe for human consumption) water and the number 0 indicates unpotable water.

1.2 Task Description

The label column in the Water Quality data set clearly indicates that it is a classification. As part of this project, the team plans to create a classification model that distinguishes between two categories. Water with Class = 0 is unsafe for consumption, while water with Class = 1 is safe. This problem does not require regression since we are not interpolating predictions.

1.3 Methodology

This project aims to apply various machine-learning algorithms to the given dataset in a Python environment to assess the Quality of Water. To begin, the python environment will be used because of its simplicity, readability, access to powerful data analysis & visualization libraries, and consistent widespread use in the AI and ML industry. The following libraries will be utilized:

- SciKit Learn
- Pandas
- Numpy
- Matplotlib

2. Data Visualization

To understand the data we need to implement some data analysis and visualize the data to check for outliers and missing values.Also, to see the relationship between the features and the output variable.

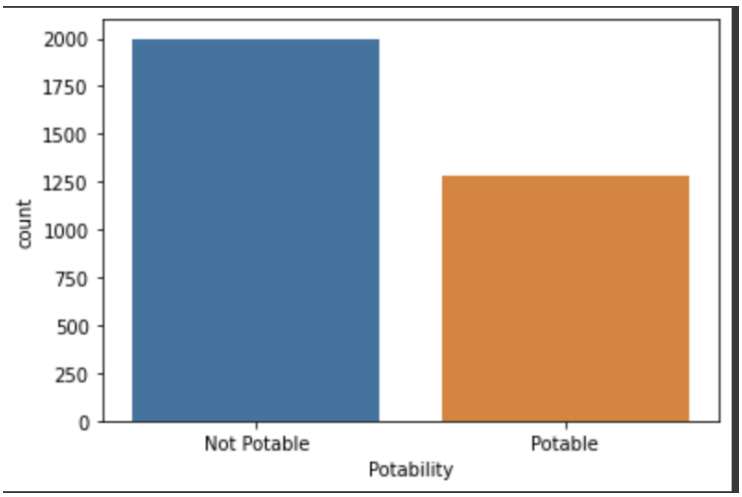


Fig. 1 (potable and non-potable count)

From the figure above we can observe the difference between the potable and non-potable values we need processing techniques such as imputation to balance the values so we can implement the classification.

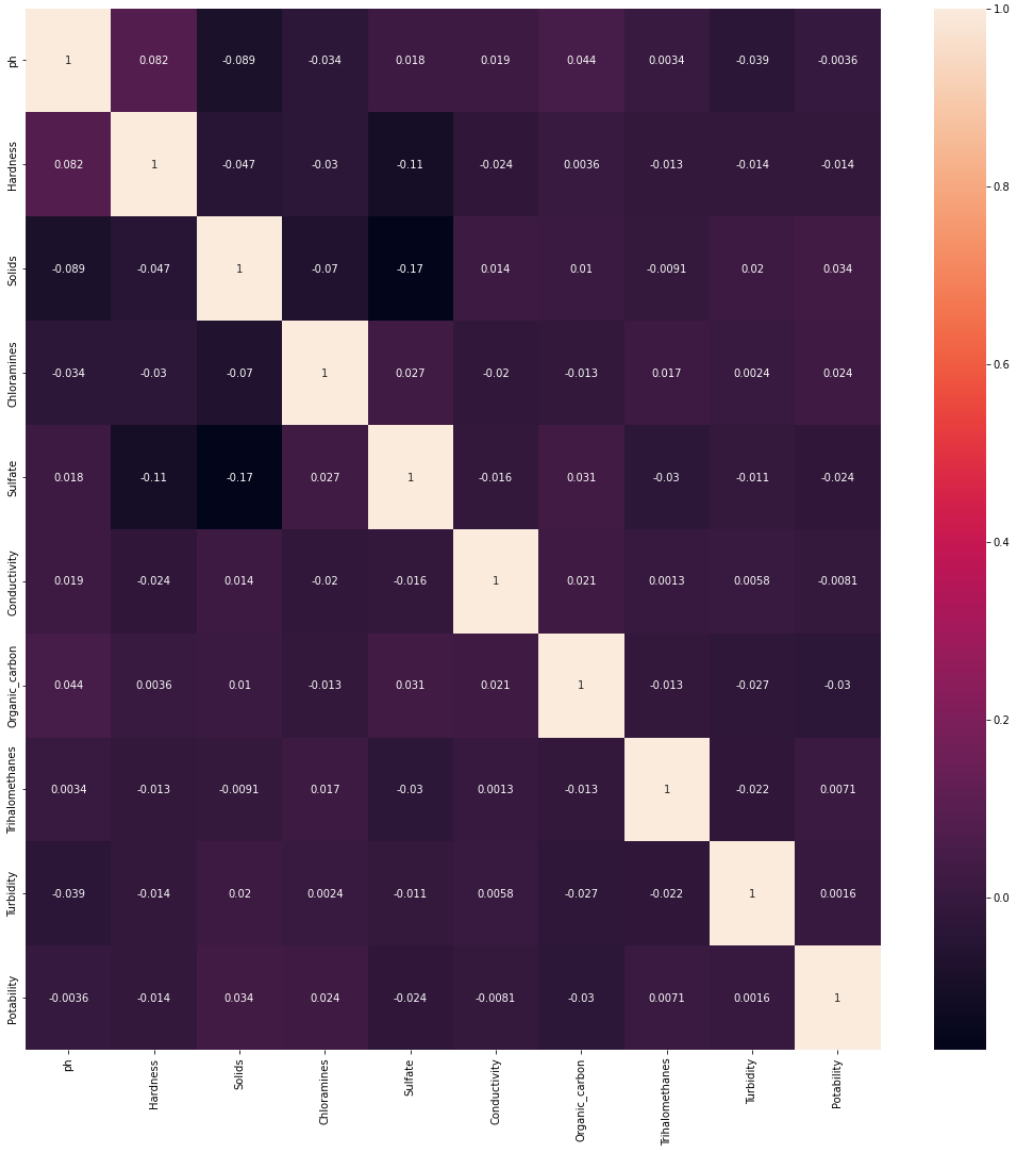


Fig. 2 (heat map)

Figure above shows the heat map for the attribute correlation. There is a less correlation between the features.so we need to implement more analysis to understand the data and investigate the missing values and correlate the variable's.

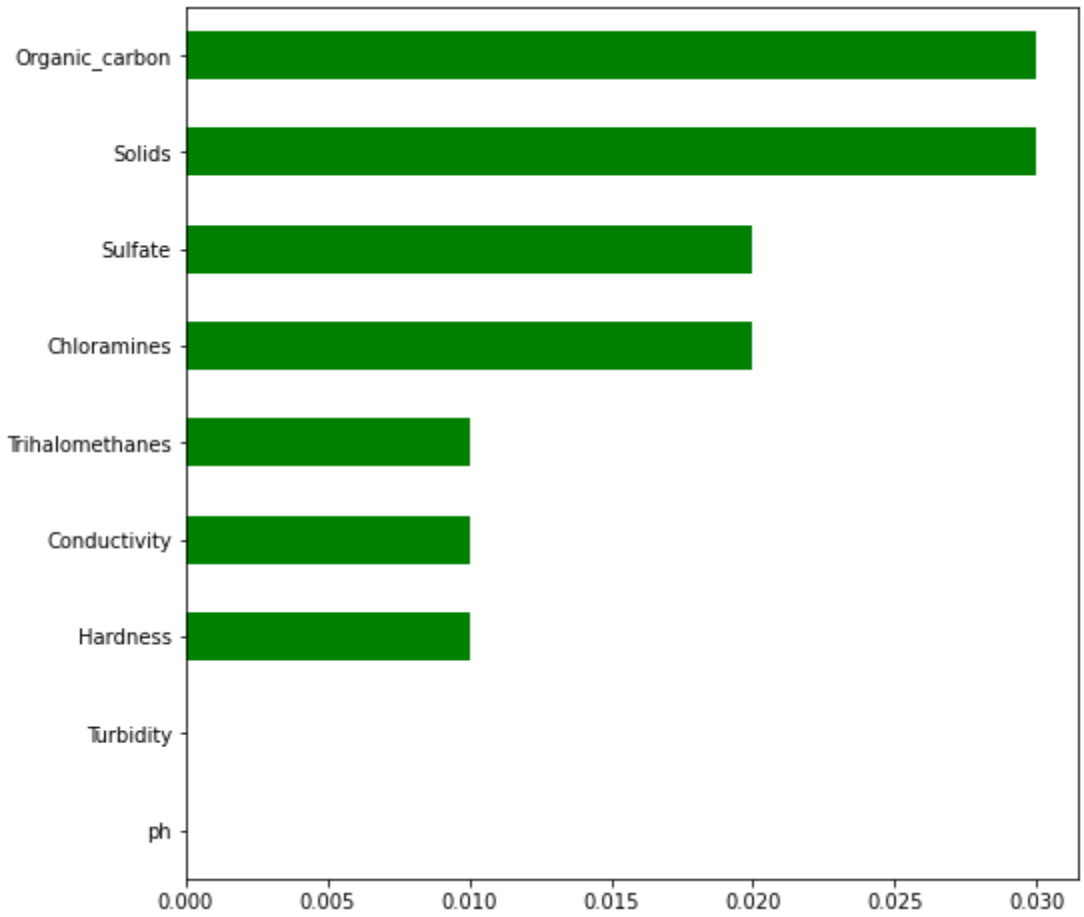


Fig. 3 (Potability count)

The figure above shows the Potability for each feature, and we can observe that (Ph and Turbidity) are not potable.

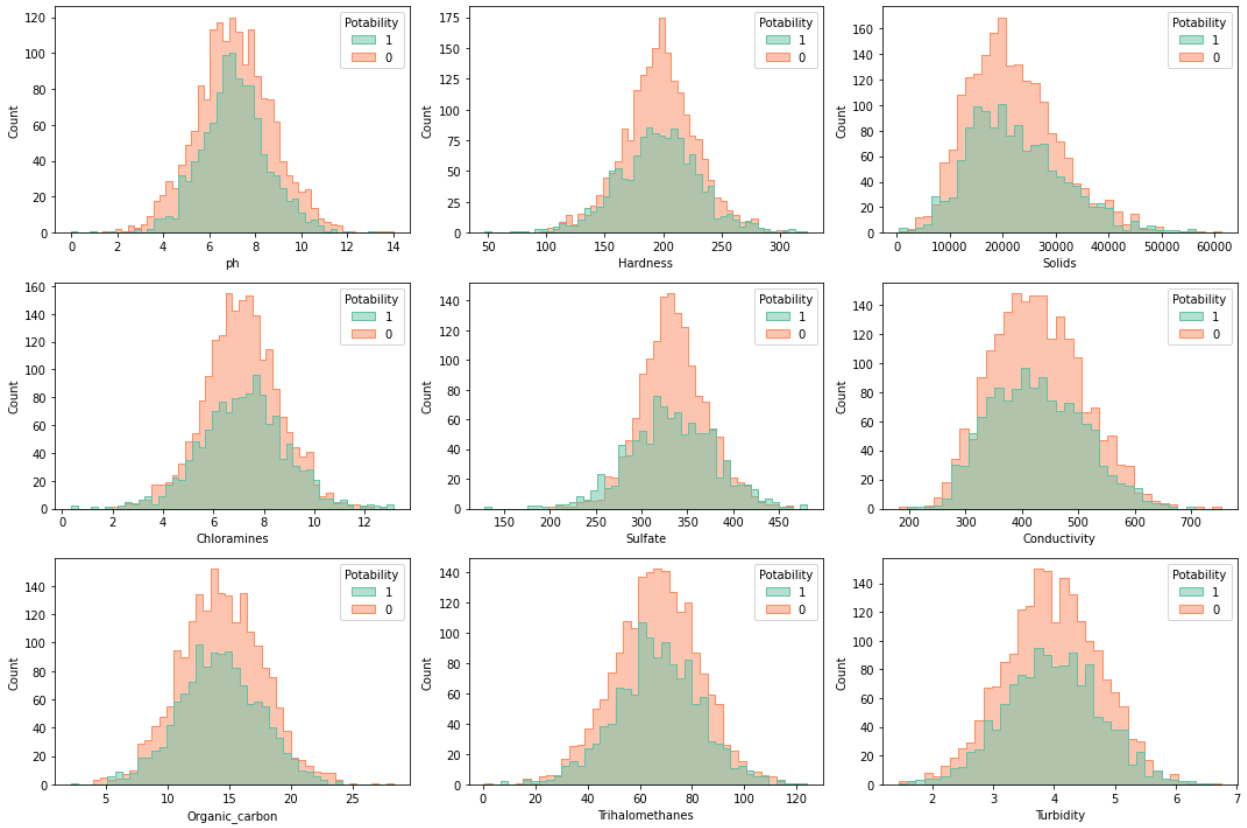


Fig. 4 (feature distribution)

The figure above shows the feature distribution for the Potability, we can observe that the distribution of non-potable is high compared to potable.

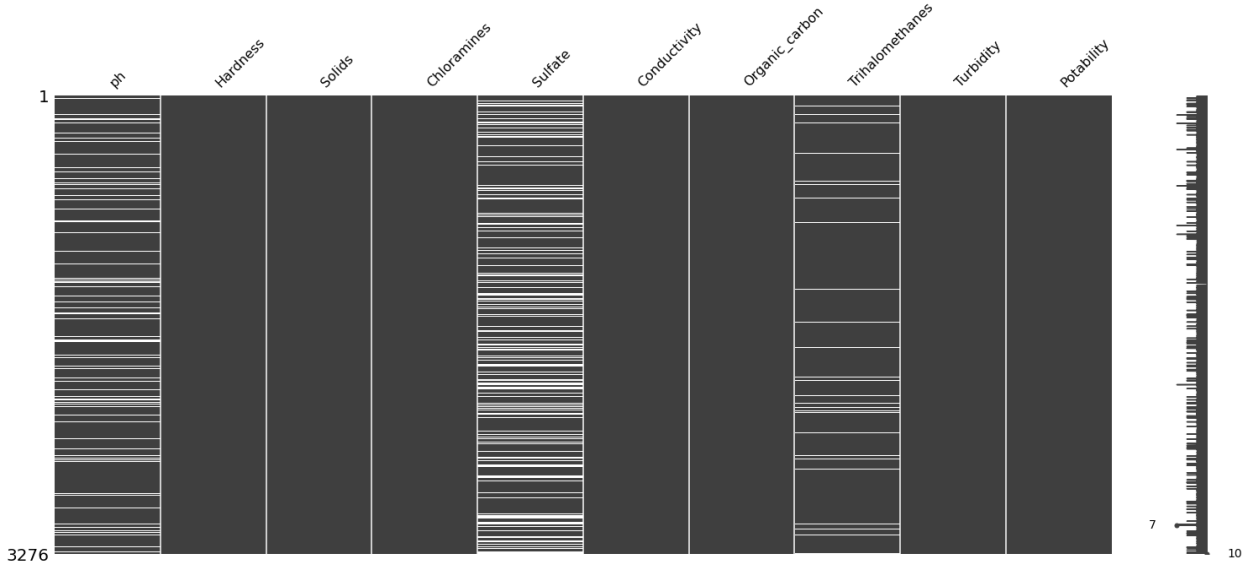
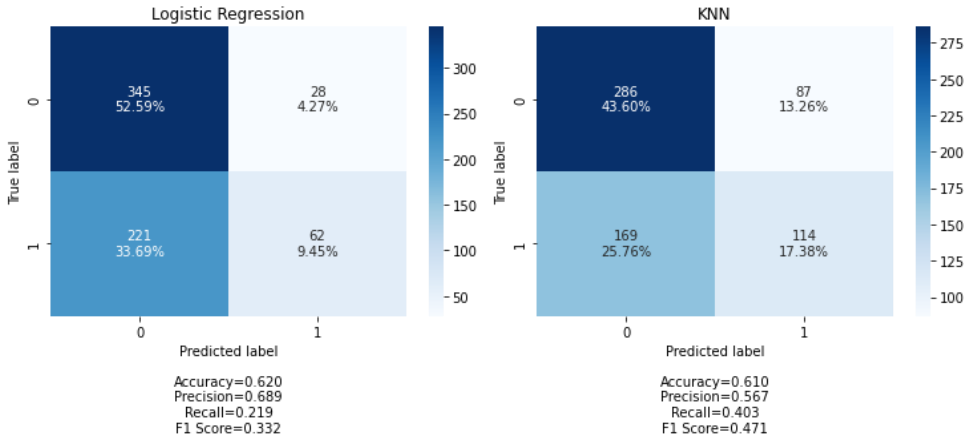


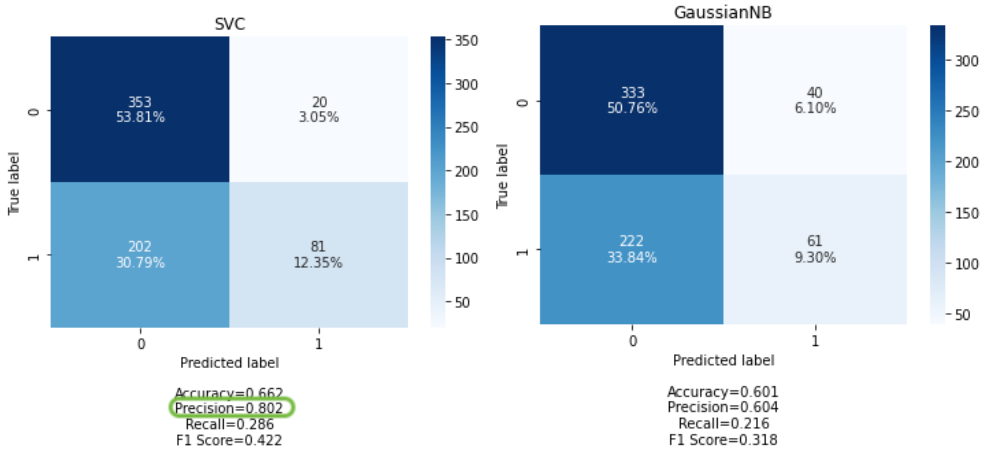
Fig. 5 (missing values count)

Figure above shows the missing values count for each feature. And we can observe that sulfate and ph. have a lot of missing values which can affect the classification and decrease the accuracy. A lot of techniques are used to handle missing values such as imputation or filling the values with mean or median it depends on the model accuracy which method is better.

3. Model Design

First, we used several machine learning techniques to examine the processed data set in order to create the model. The list of these models, along with their performance, is shown below. The proposed models were first chosen by looking at them with the fewest parameter modifications. But since several models specially ensemble ones, including Random Forest, Bagging Classifier, Gradient Boosting and Extra trees use the Decision Tree model as base estimator by default, in order to prevent them from being too similar to each other without significantly manipulating their main structure, the Bagging Classifier and Adaboost models are designed based on SVC in this project, which Finally, we can explore more variety and modes. As mentioned earlier in the data exploration section, the data set that we are facing has a nature of uncorrelated variables, which makes it difficult for any model to predict (The dataset is not linearly separable). In principle, this amount of non-dependence cannot be investigated with a linear model, although to ensure this, the Logistic Regression model has also been investigated to see how linear model performs on this dataset.





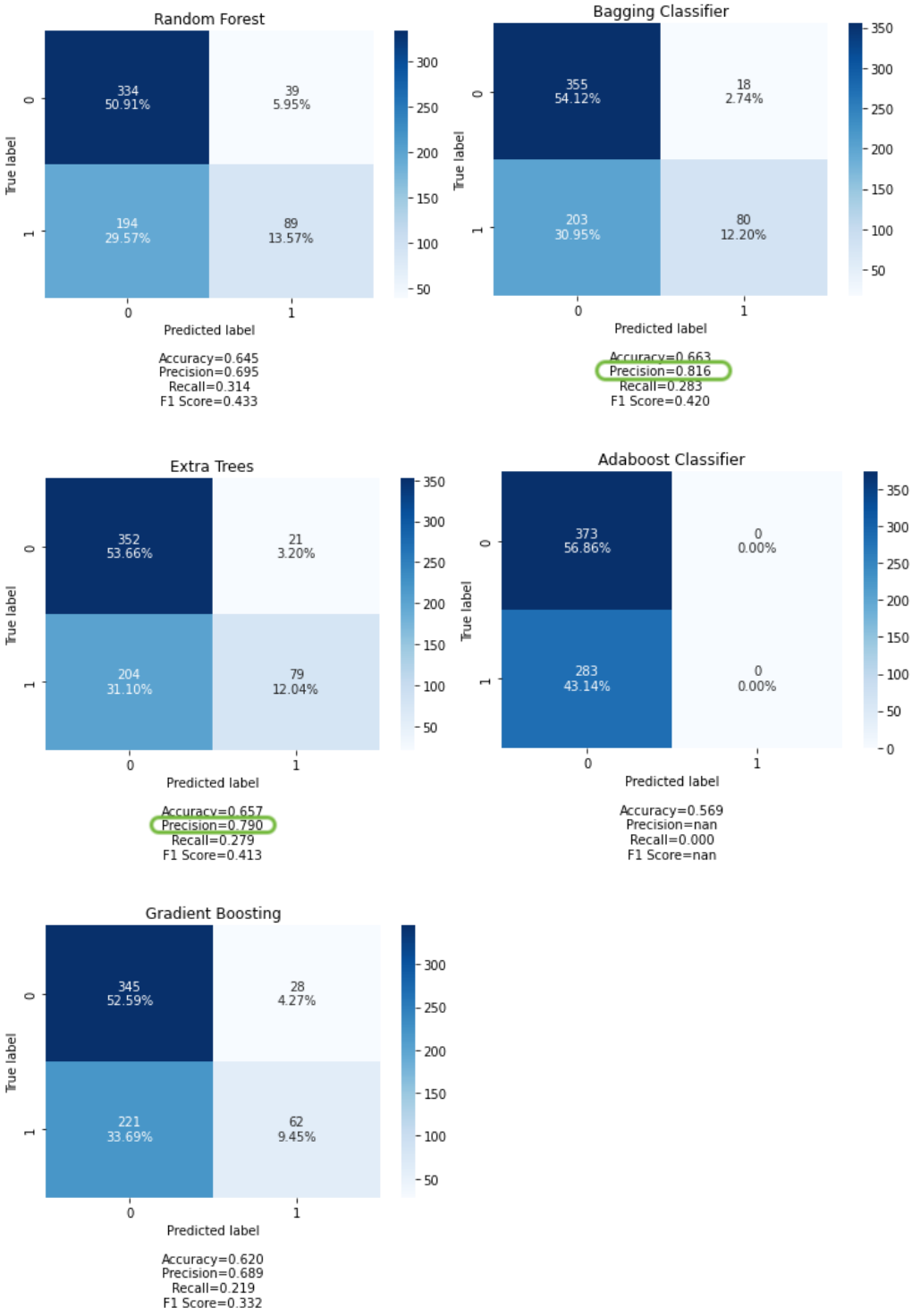


Fig. 6 (Predicted labels)

After the initial training phase, it is clear that the models Extra Trees, Bagging, and SVC have narrowly edged out the rest of the field. It should be noted that the underlying model for Bagging is SVC, which immediately slows down training speed since SVC employs an RBF kernel, and this problem eventually has a significant impact on the model's pace. Despite the significant improvement in the model's capacity to handle complicated datasets brought on by this kernel, the choice of this value must ultimately rely on the problem at hand.

The dataset's relationship to the health sector is a key consideration. The study should lead us to the conclusion that there is a high confidence that the test samples are safe to consume. Given this situation, Precision is more crucial than accuracy and recall in this case, and the three models listed above have the highest percentage of Precision among all. Extra Trees appears to be a superior alternative in the end, although having less precision than the other two models.

First, as mentioned, the SVC model has a low speed, and the Bagging model will naturally have the same defect because it uses SVC in our project. Secondly, in the topic of regularization and generalization, the Extra Trees model will be more capable when applied to flying data because it generalizes better than the other two models. Thirdly, according to the ROC curve and the area under that, it can be said that the Extra Trees model works more stably and in the process of improving and upgrading in the future, it will tend to the upper left of the graph in a more appropriate way. It should also be taken into account that tuning the Extra Trees model takes much less time. All these are convincing reasons that can be a wise justification for choosing this model over the other two.

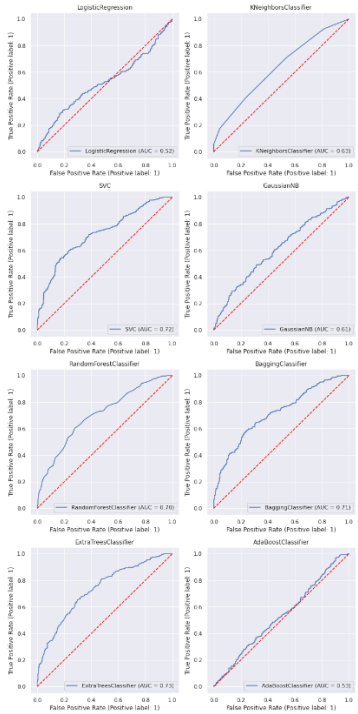
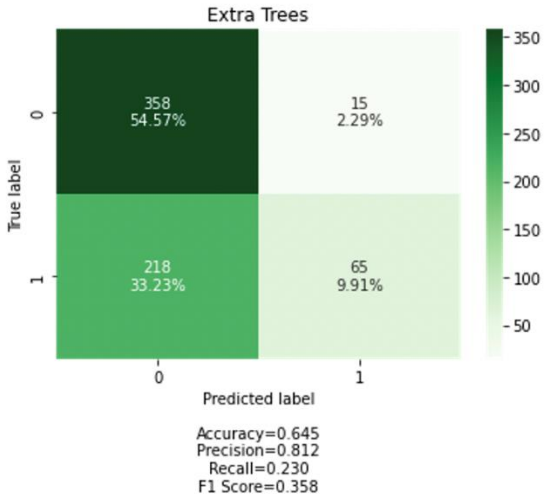


Fig. 7 (ROC Curve)

After performing the tuning, we see an increase of more than two percent in the performance of the model in precision, which will certainly come at the cost of reducing the recall and accuracy.



Parameters	value
Estimators	260
Max F	'Auto'
Max D	None
MSS	2
MSL	2

Fig. 8 (Predicted labels)

In the deep learning part, we tried to start by selecting a hidden layer with twelve neurons and then start making changes through grid search. In the case of this dataset, adding any hidden layer would lead to overfit, so it can be said that this single layer is the most appropriate choice for this model. After applying the tuning and obtaining the appropriate value of each of the parameters mentioned in the table below, our model was able to reach sixty-seven percent accuracy on the test set, which is not a significant improvement compared to the classical learning machine. Although, as can be seen in the Tensor Flow graph, with the iteration of more epochs we see a greater decrease in the loss function, but this issue practically has no effect on the performance of the model on the test set, and the result is the same as with fifty epochs.

Neural network model

Parameter	value
Batch	10
Epoch	50
Optimization	RMSp
Learning Rate	0.001
Opt. Momentum	0.2
Initializer	Glorot Uniform
Drop Out	0.1
W. Constraint	4.0
Neurons	15

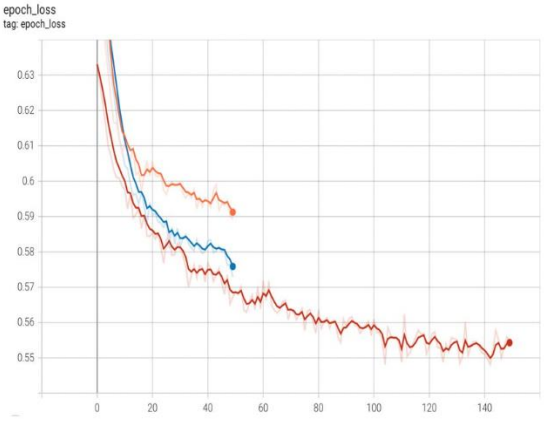


Fig. 9 (Model accuracy)

4. Conclusion

In Conclusion, for this particular dataset, classic machine learning models are more recommended because the dataset is not large enough for deep learning models to be trained well, as well as the performance of deep learning is not such that it can be the priority.