Water Quality Detection

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

Supervisor:
Dr. Hamidreza Mahyar

Project outline

- Introduction.
- Dataset description.
- Task Description.
- Data Visualization.
- Model building.
- Conclusion

Problem statement

- 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.

Introduction

- Environmental and public health is directly impacted by water quality. In addition to drinking, water is used in agriculture and industry.
- 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.
- 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 project.

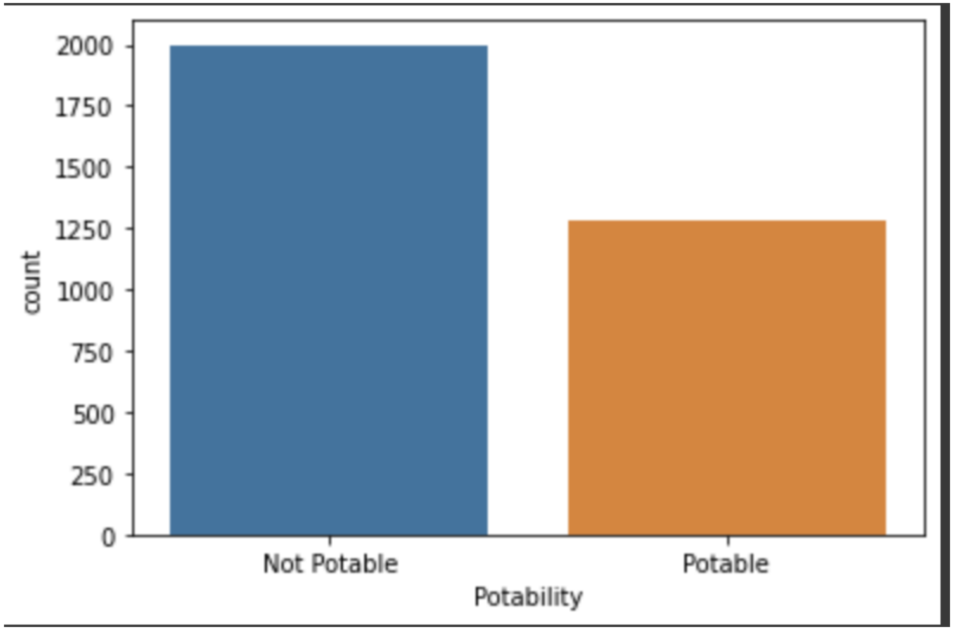
Dataset description

- The dataset contain input variables based on physicochemical tests

Dataset description
pH value
Hardness
Solids (Total dissolved solids – TDS)
Chloramines
Sulfate
Conductivity
Organic carbon
Trihalomethanes
Turbidity
Potability (Class)

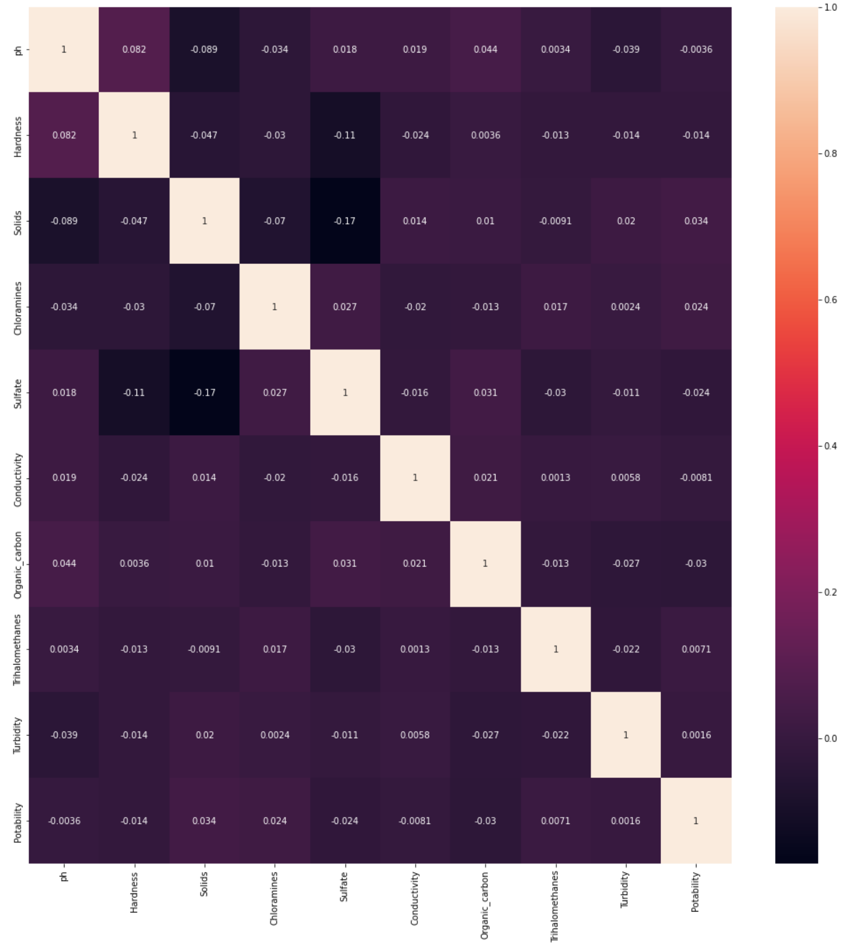
Data visualization

- The figure shows the difference between the potable and non-potable values.



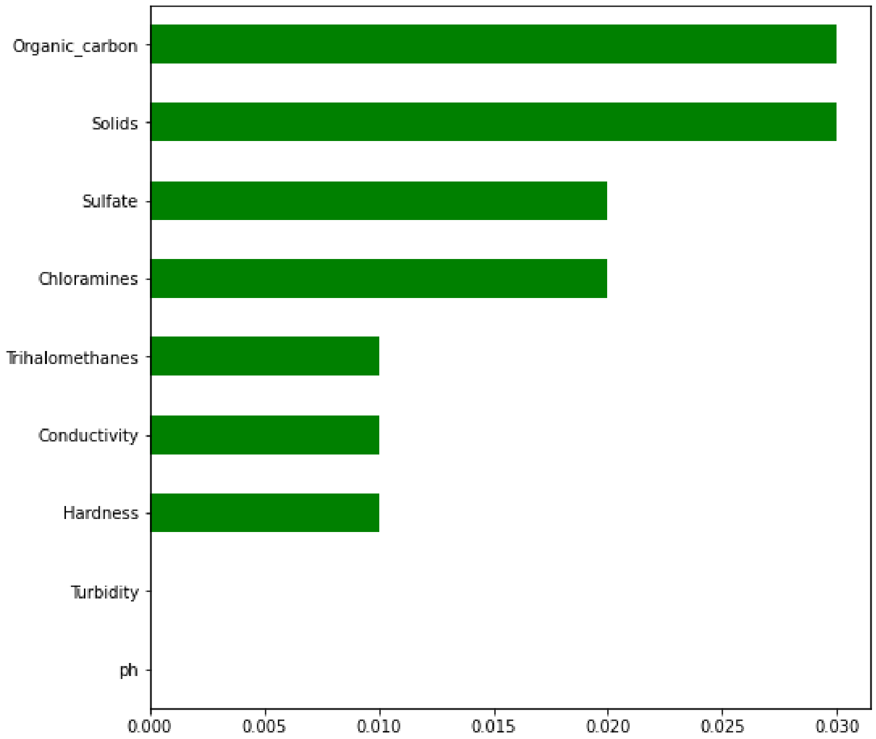
Data visualization

- The figure shows the heat map for the attribute correlation.
- there is a less correlation between the features.



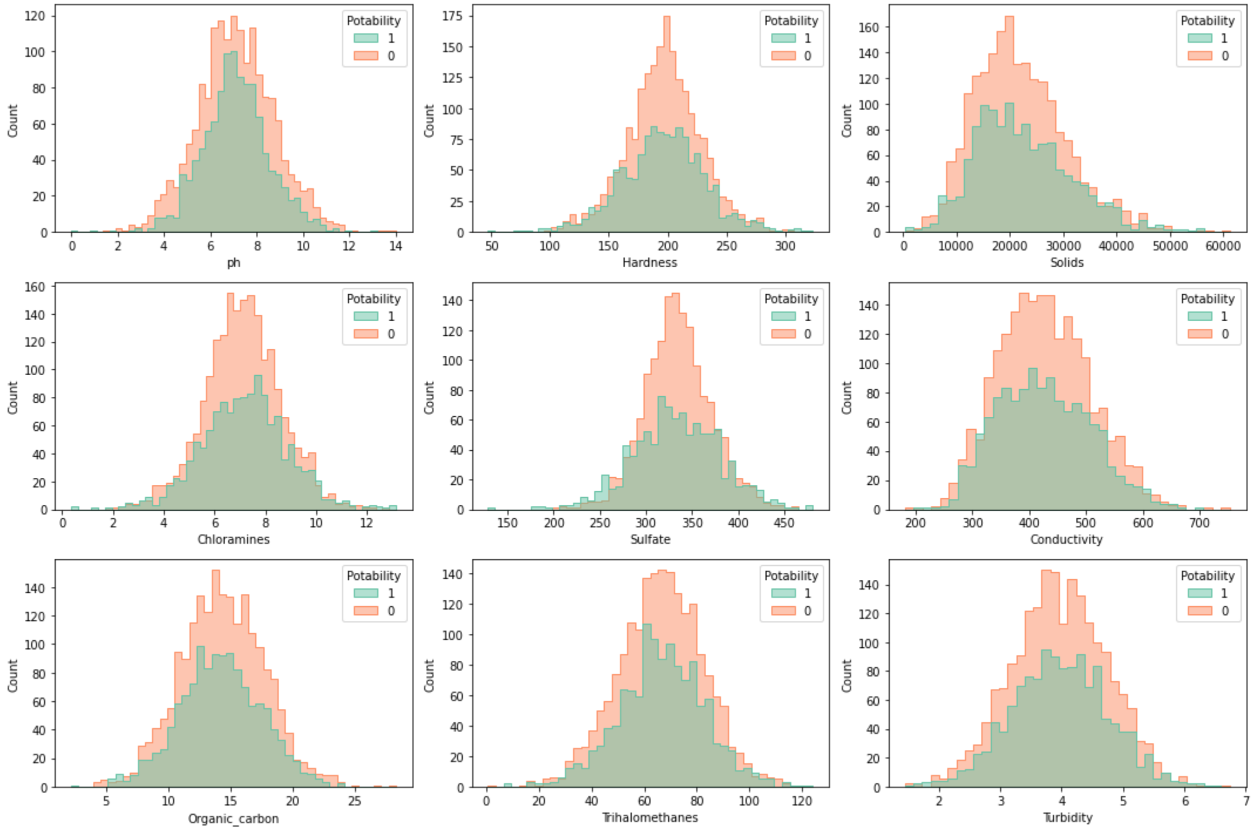
Data visualization

- The figure shows the potability for each attribute.
- PH and Turbidity are not potable.

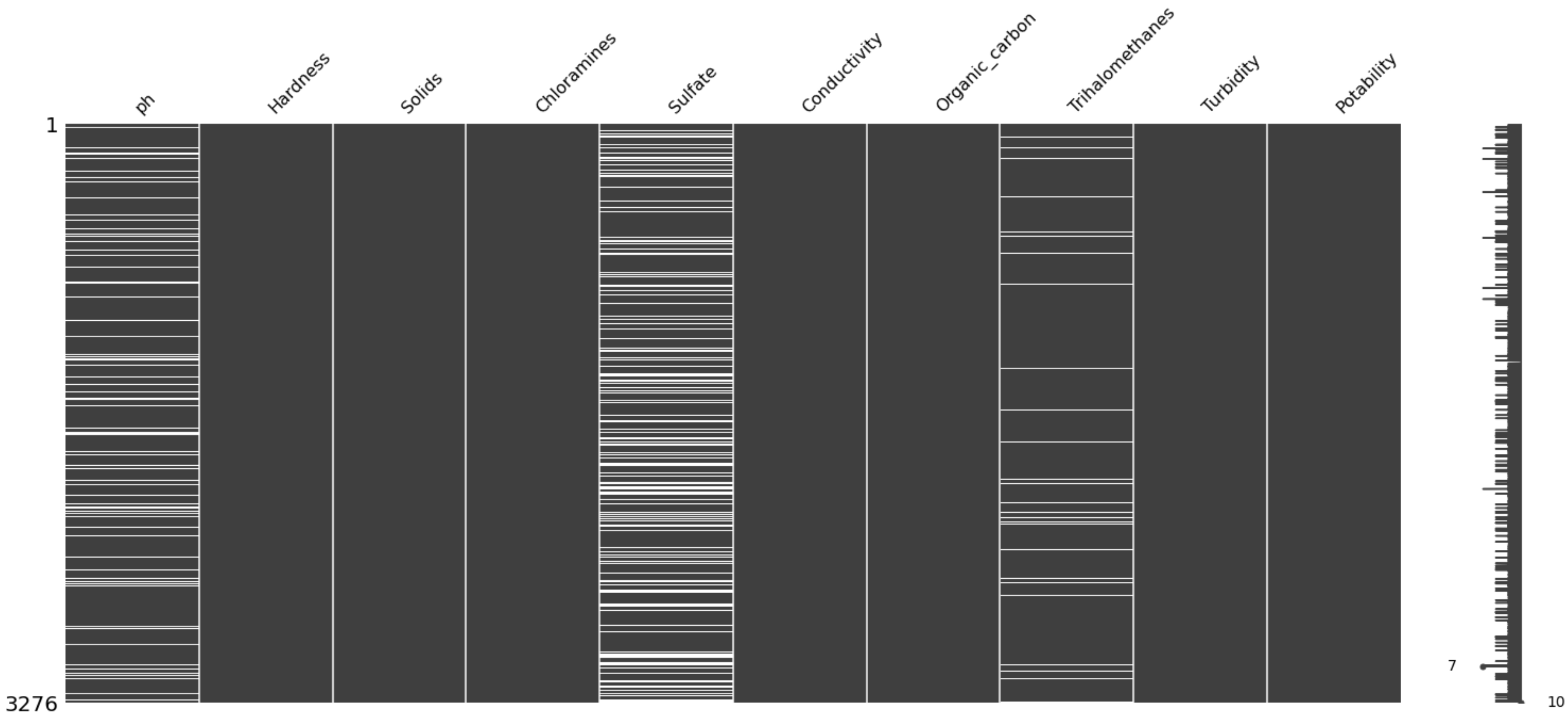


Data visualization

- The figure shows the feature distribution for the Potability.
- the distribution of non-potable is high compared to potable.



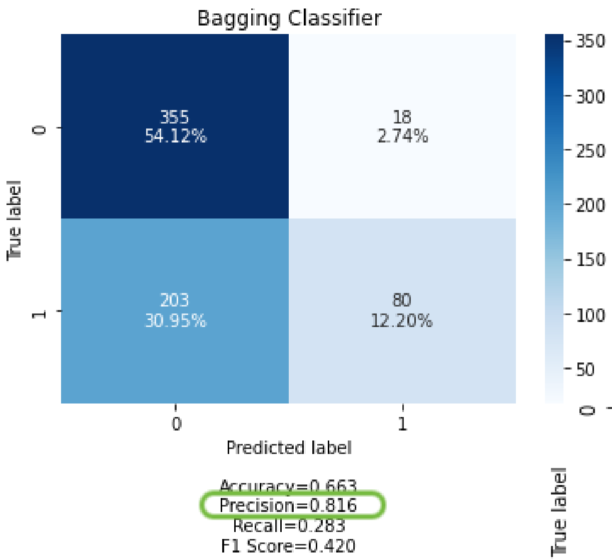
Data visualization

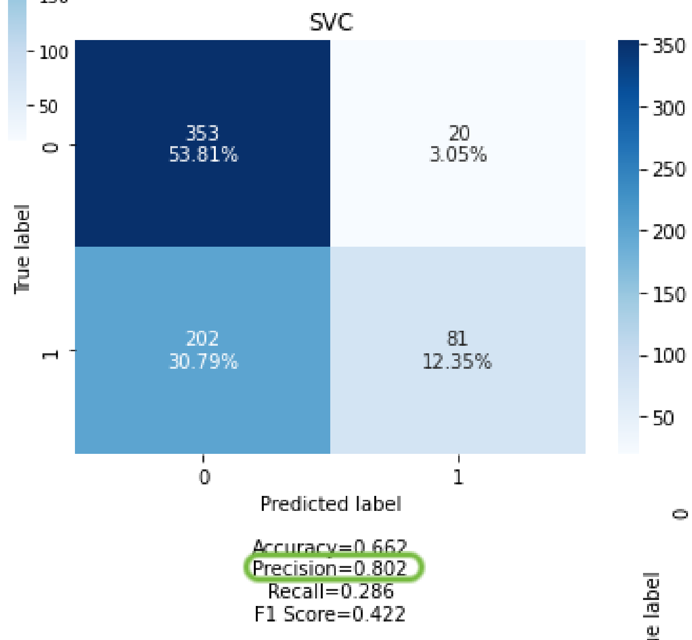


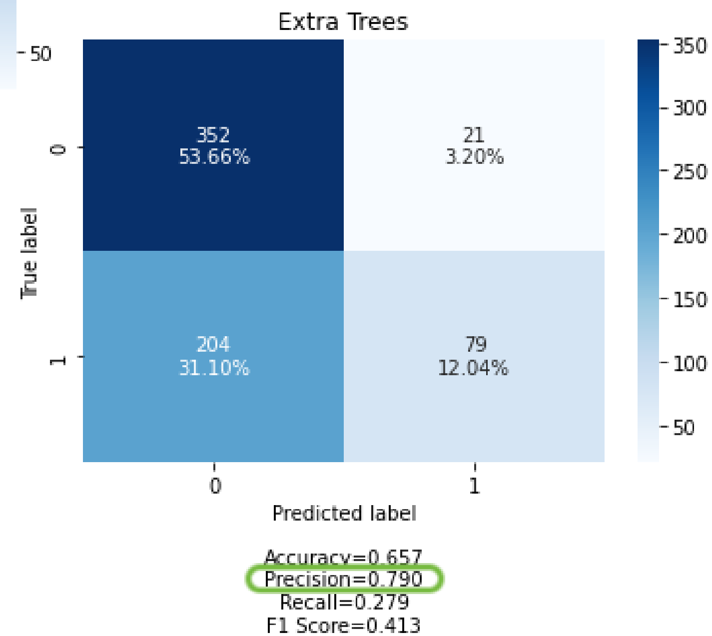
- The Figure shows the missing values count for each feature
- sulfate and ph. have a lot of missing values which can affect the classification and decrease the accuracy

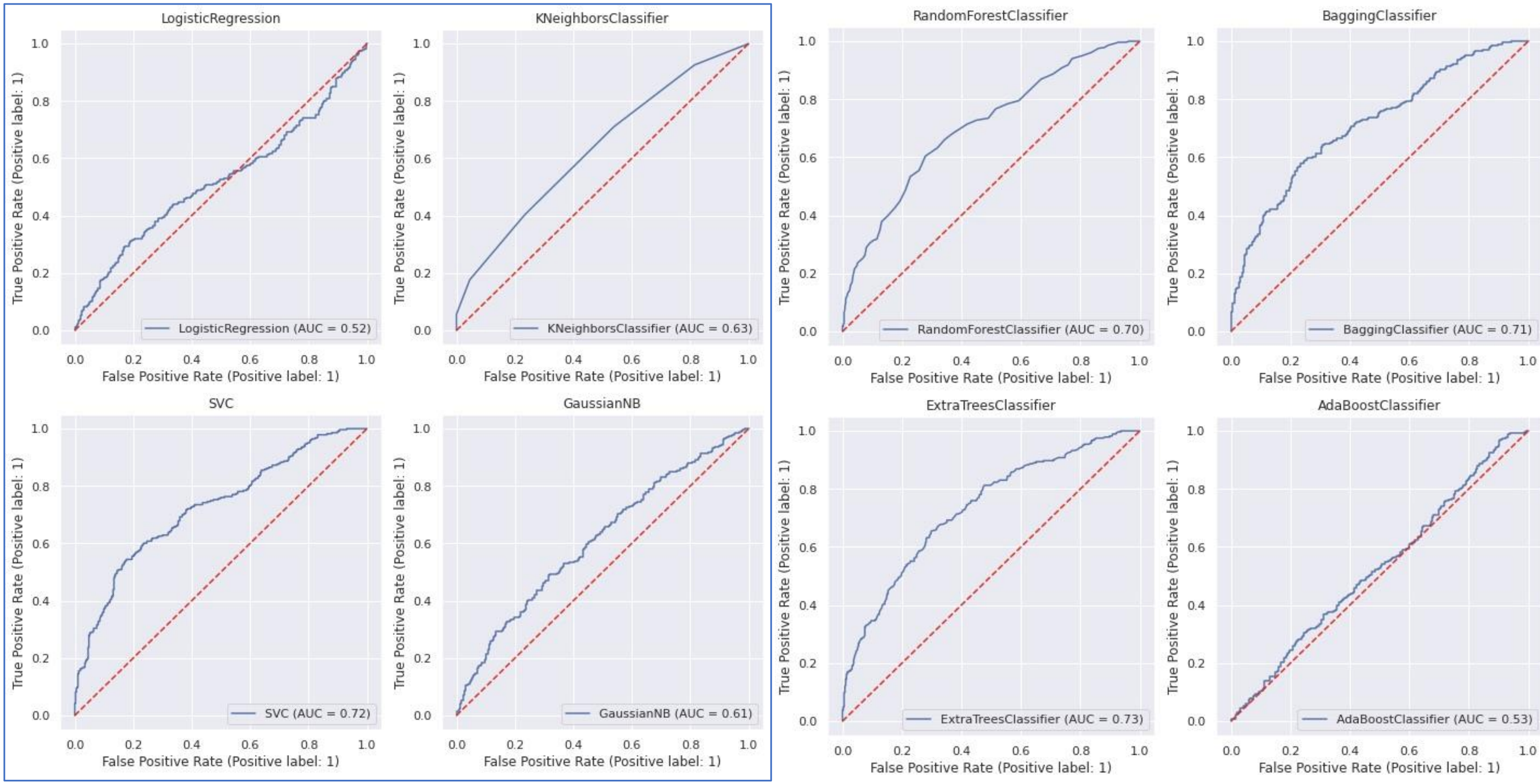
Machine learning model (Classification)

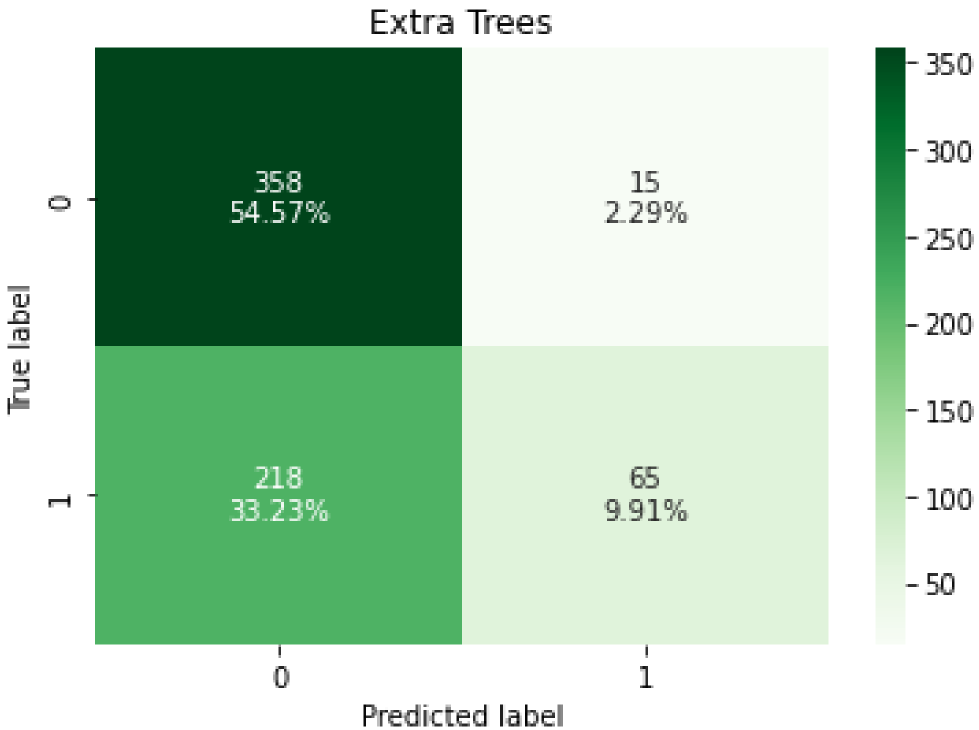
```
LogisticRegression 0.6209923664122138
KNeighborsClassifier 0.6194656488549618
SVC 0.6748091603053435
GaussianNB 0.6320610687022901
RandomForestClassifier 0.666030534351145
BaggingClassifier 0.667175572519084
ExtraTreesClassifier 0.665648854961832
AdaBoostClassifier 0.6202290076335878
GradientBoostingClassifier 0.6358778625954199
```


Bagging Classifier
355
54.12%
18
2.74%
203
30.95%
80
12.20%
True label
Predicted label
0
1
350
300
250
200
150
100
50
Accuracy=0.663
Precision=0.816
Recall=0.283
F1 Score=0.420


SVC
353
53.81%
20
3.05%
202
30.79%
81
12.35%
True label
Predicted label
0
1
350
300
250
200
150
100
50
Accuracy=0.662
Precision=0.802
Recall=0.286
F1 Score=0.422


Extra Trees
352
53.66%
21
3.20%
204
31.10%
79
12.04%
True label
Predicted label
0
1
350
300
250
200
150
100
50
Accuracy=0.657
Precision=0.790
Recall=0.279
F1 Score=0.413


LogisticRegression
LogisticRegression (AUC = 0.52)
KNeighborsClassifier
KNeighborsClassifier (AUC = 0.63)
RandomForestClassifier
RandomForestClassifier (AUC = 0.70)
BaggingClassifier
BaggingClassifier (AUC = 0.71)
SVC
SVC (AUC = 0.72)
GaussianNB
GaussianNB (AUC = 0.61)
ExtraTreesClassifier
ExtraTreesClassifier (AUC = 0.73)
AdaBoostClassifier
AdaBoostClassifier (AUC = 0.53)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)

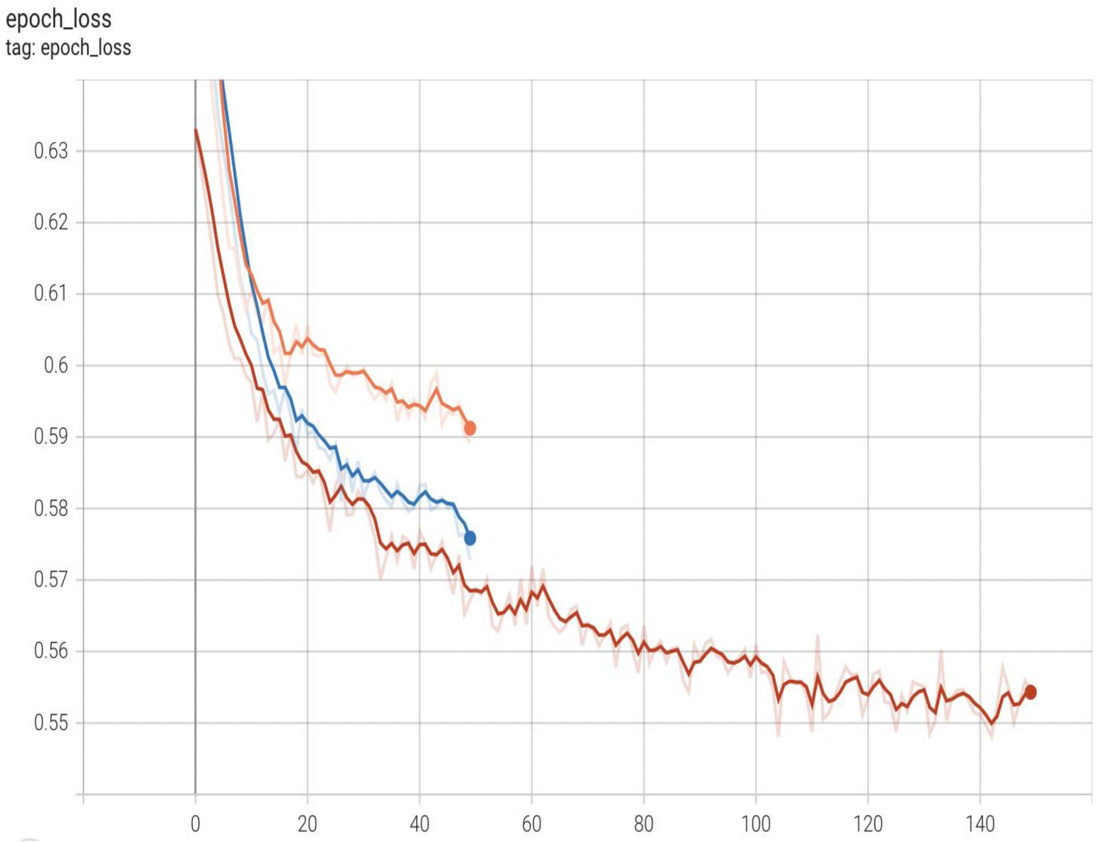


Accuracy=0.645
Precision=0.812
Recall=0.230
F1 Score=0.358

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

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




Thanks