TECHNISCHE HOCHSCHULE INGOLSTADT
ALGORITHMS FOR AI 3

LAB 1: WORKING WITH HISTOGRAMS

FERNANDO FINCATTI

INGOLSTADT
MARCH 2023

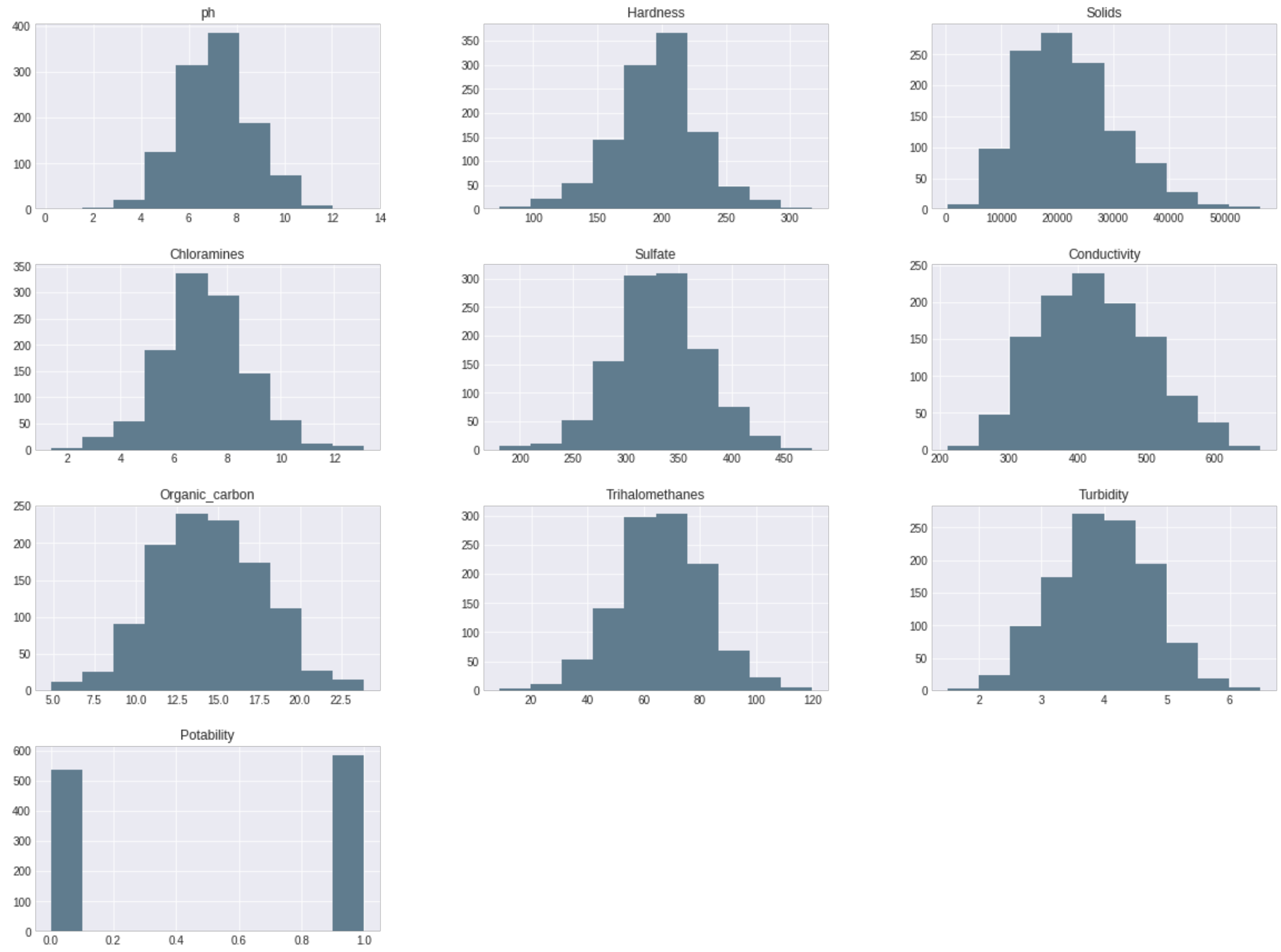
1. INTRODUCTION

The aim of this report is to assess the performance of a Support Vector Machine (SVM) classifier for a decision-making application related to predicting the potability of drinking water. The dataset used for this study is sourced from Kaggle and is known as "Drinking_Water_Potability." It consists of over 1000 samples, each described by nine attributes, including the target variable "Potability" that denotes whether the water is potable (Potability = 1) or not potable (Potability = 0).

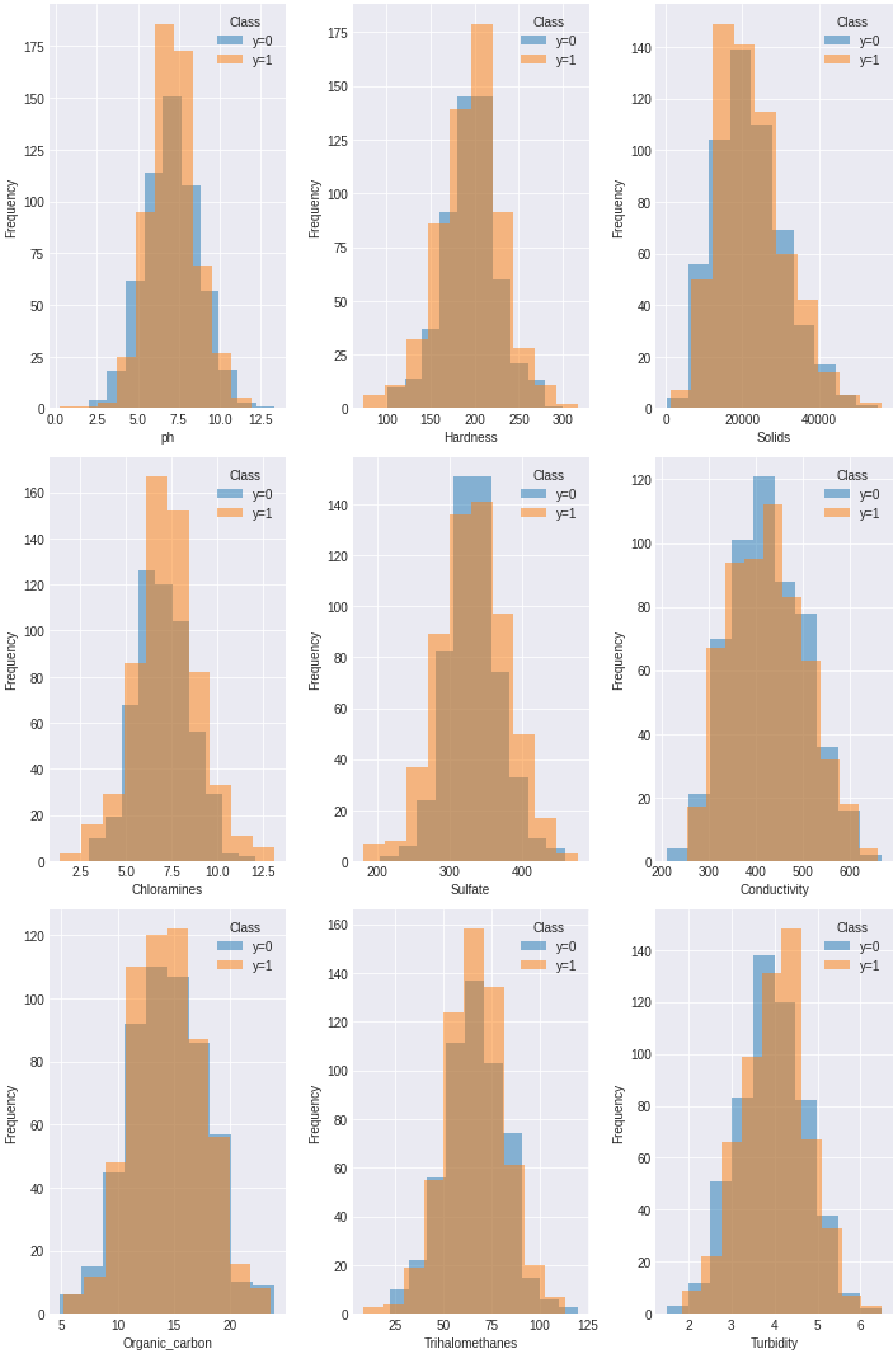
2. EXPLORATORY ANALYSIS

Histograms of all attributes, including the target variable "Y" (potability), provide a visual representation of data distribution and help identify patterns and outliers. Examining the histogram of "Y" allows assessment of class distribution, ensuring a balanced dataset for accurate classification. This analysis sets the foundation for further exploration and model development.

Histogram of all variables



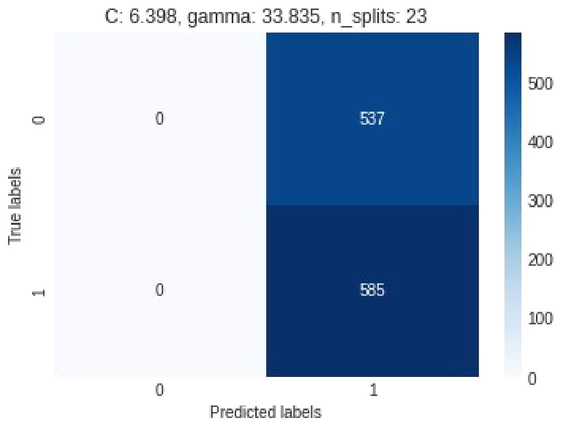
For a better understanding of the distribution of each variable for each different class, the following plot shows the distribution dividing by the Y value:

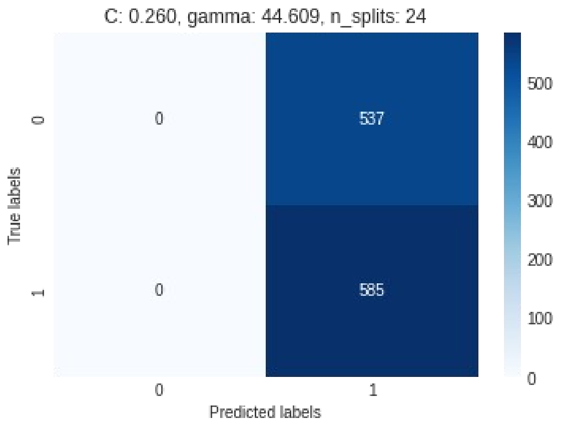

Class
y=0
y=1
Frequency
0
25
50
75
100
125
150
175
0.0
2.5
5.0
7.5
10.0
12.5
ph
Class
y=0
y=1
Frequency
0
25
50
75
100
125
150
175
100
150
200
250
300
Hardness
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
140
0
20000
40000
Solids
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
140
160
2.5
5.0
7.5
10.0
12.5
Chloramines
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
140
200
300
400
Sulfate
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
200
300
400
500
600
Conductivity
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
5
10
15
20
Organic_carbon
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
140
160
25
50
75
100
125
Trihalomethanes
Class
y=0
y=1
Frequency
0
20
40
60
80
100
120
140
2
3
4
5
6
Turbidity

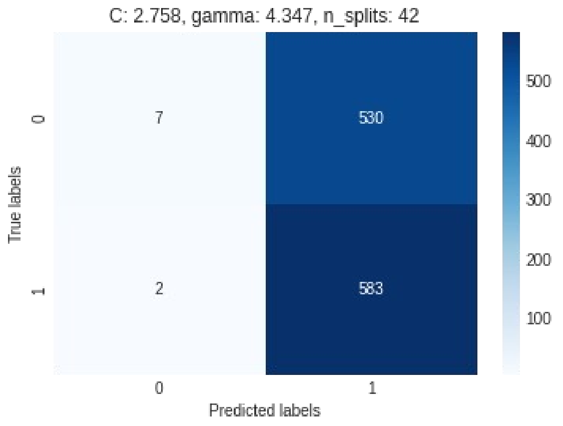
Once the histograms are similar for both classes, presenting significant overlap area, we can say that the classifier is unlikely to perform well. If the histograms of the attributes for the two classes are significantly different, it suggests that the attributes have different distributions for the two classes and are therefore informative for classification. On the other hand, if the histograms are similar, it suggests that the attributes are not strongly associated with the class labels and may not be useful for classification.

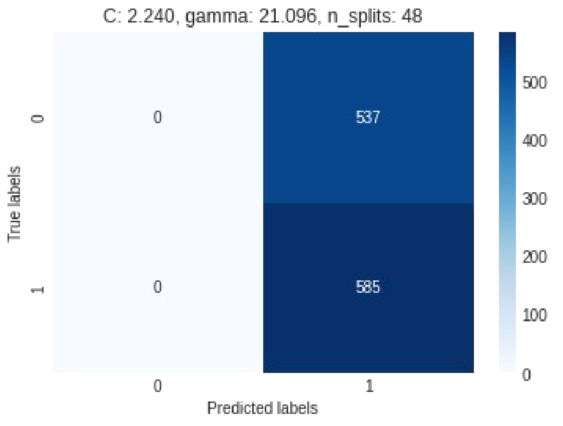
3. SVM

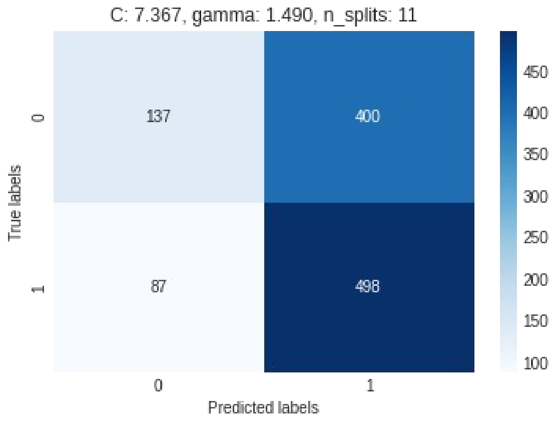
Multiple runs were conducted using SVM with varied C and gamma parameters and cross-validation without shuffling. Confusion matrices were displayed for each run to assess test data performance. The average accuracy of the test data after 5 runs was calculated, and the result was 53.12% providing an overall measure of the SVM classifier's effectiveness in predicting water potability. The results are shown bellow:





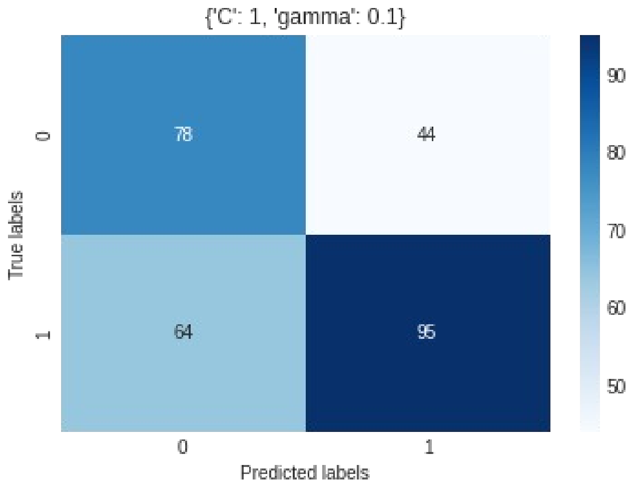


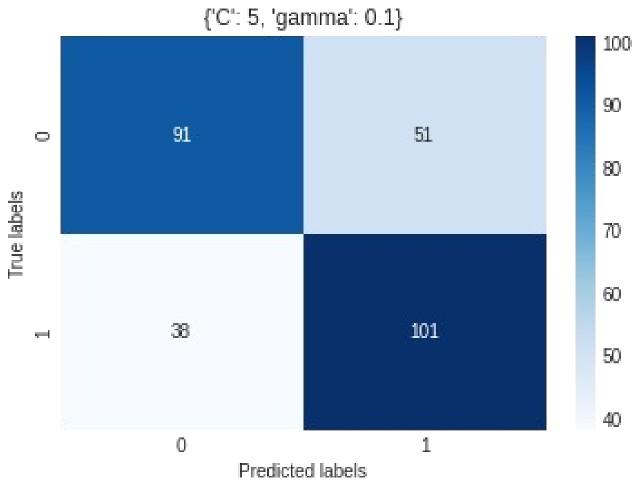


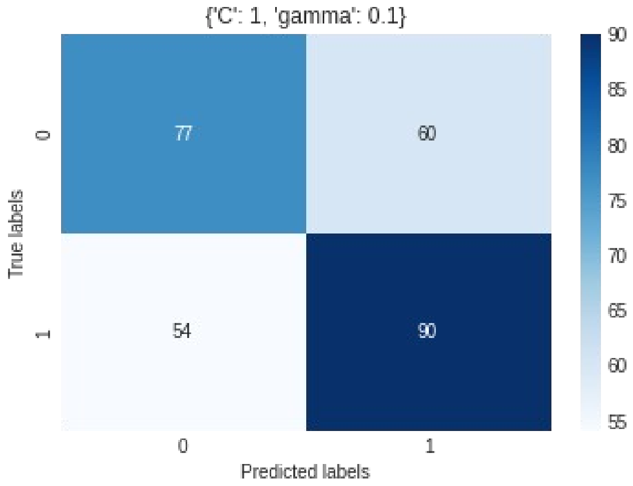


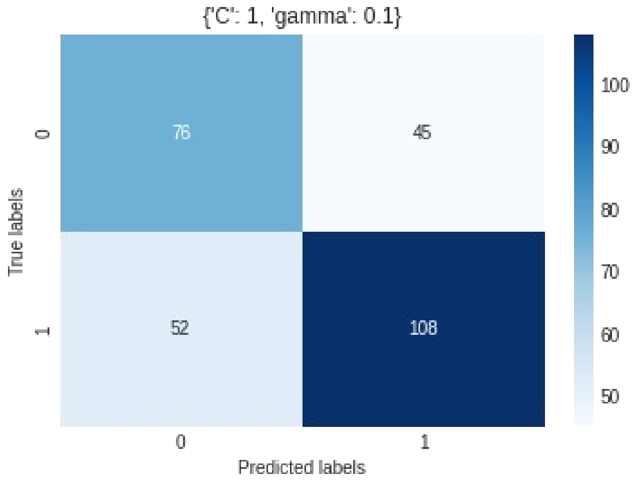
As we can see in the above results, changing the parameters of the model doesn’t has a huge effect on the algorithm performance. This can be due to a overfitting,

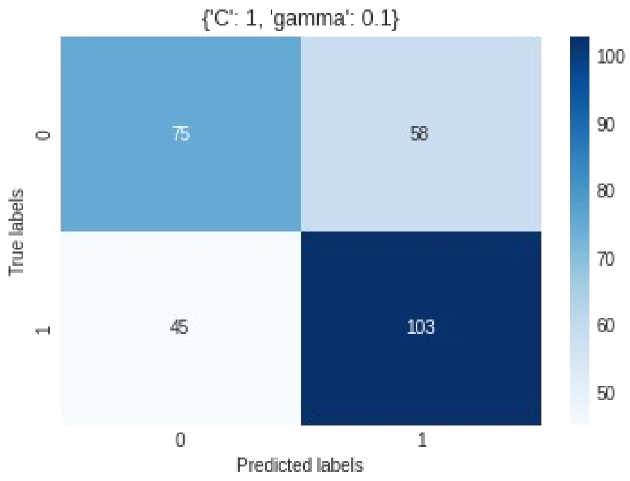
Five runs were performed using SVM with grid search. For each run, the data was reshuffled and was displayed the best hyperparameter results obtained, indicating the optimal values for C and gamma. The average accuracy was 62,2%. Additionally, we displayed the confusion matrix of the test data, providing an insight into the classifier's performance. Finally, we calculated the average accuracy of the test data across the 5 runs, serving as an overall measure of the SVM classifier's performance in predicting water potability.


{'C': 1, 'gamma': 0.1}
True labels
0
1
78
44
64
95
0
1
Predicted labels
90
80
70
60
50


{'C': 5, 'gamma': 0.1}
True labels
0
1
91
51
38
101
0
1
Predicted labels
100
90
80
70
60
50
40


{'C': 1, 'gamma': 0.1}
True labels
0
1
77
60
54
90
0
1
Predicted labels
90
85
80
75
70
65
60
55


{'C': 1, 'gamma': 0.1}
True labels
0
1
76
45
52
108
0
1
Predicted labels
100
90
80
70
60
50


{'C': 1, 'gamma': 0.1}
True labels
0
1
75
58
45
103
0
1
Predicted labels
100
90
80
70
60
50