# Sea water pollution prediction

Mia Dudić, Karolina Škunca and Ana Zahtila

*Abstract*— **Based on the article on Temporal variations analyses and predictive modeling of microbiological seawater quality [1], where there was a case study on the effect of Escherichia coli on sea water pollution, a new case is made. A predictive model was made using previously collected data of different variables that could cause the pollution. In the before mentioned paper, SVR algorithm was used to make the model which did not give good results in our case. Instead, binary classification, logistic regression and SVM were used and produced a model that gave accuracy between 70% and 90%.**

## I. INTRODUCTION

The main premise of this project is to predict sea water quality pollution based on a set of data obtained from three locations in Rijeka - Pećine, 3 Maj and Kantrida. It is based on the article on Temporal variations analyses and predictive modeling of microbiological seawater quality [1], where there was a case study on the effect of Escherichia coli on sea water pollution and making a predictive model using previously collected data of different variables that could cause the pollution. The data set is divided into two parts: set A and set B. Set A contains data from the meteorological station in Rijeka and has a total of 966 measurements. Set B consists of Solcast site data and has 666 tests. To develop a reliable prediction model, it is extremely important that the data is accurate because, as such, they represent the foundation on which the model is built. Solar radiation and precipitation are the main parameters that affect the quality of water from an ecological point of view, and this is confirmed by the fact that the highest concentrations of bacteria were recorded in the early morning due to lack of solar radiation during the night. Also, rainfall was the second main parameter that affected the spread of bacteria. Simply, a larger amount of rain equals a larger number of bacteria. However, when we calculated the influence of the parameters, the results did not meet expectations. Namely, it turned out that the amount of precipitation does not have such a big impact at all. The research led to the conclusion that, in fact, the greatest impact on the quality of sea water has wastewater, about which we have no data at all. Because we cannot determine with 100% certainty what the most influential parameters are, we decided to use almost all the data. Although tidying up the database involves omitting those parameters whose null value occurs too often, for those parameters that we consider important (such as air temperature) we simply did not do so. Instead of omitting a perhaps important parameter, we decided to use linear interpolation to construct new data points within a certain range based on known values. The analysed parameters in seawater are faecal and non-faecal indicators, where the most familiar one is the bacterium Escherichia coli, also taken as the main indicator of water pollution. The defined tolerance threshold is 300 CFU/mL, that is, if the number of bacteria in the water is above that threshold, the water is treated as contaminated. The aim of the project was to develop a model for predicting the quality of the sea in Rijeka based on the tests performed. Machine learning algorithms that were used predict the value of the target variable based on several input data. Algorithms that were used were: Linear Regression, Support Vector Regression, Gaussian Process Regression, Classification and Regression Tree and Bayesian Ridge. From the previously mentioned methods, there were not good results with any algorithm on any data set. Prediction models were made, and actual and predicted values were graphically displayed, however, no algorithm resulted in an accuracy percentage greater than 5%. Given the obtained results, it was decided to make a binary classification for the bacterium and use classification algorithms. With the help of logistic regression and Support Vector Machine algorithms, models with an accuracy of 80% and 90% were obtained.

## II. METHODOLOGY

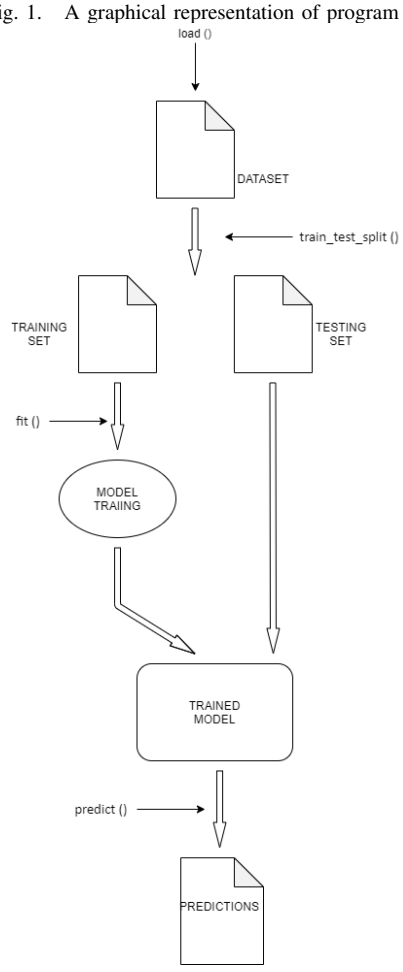
The aim of the project was to make a model for predicting sea water quality. As Escherichia coli is a continuous variable, it was decided to use regression algorithms to achieve the solution. The idea of regression algorithms is to predict the targeting of a variable (in this case bacteria) based on independent input variables which, in this case, are air temperature, sea temperature, solar radiation, precipitation, wind and others. Regression works by looking for the best matching line that will predict the target variable as accurately as possible. A connection is made between the independent variables and the target dependent variable and it should be linear in nature. After many unsuccessful attempts to realize a suitable model with the help of regression algorithms, it was decided to use classification. This was achieved by making a binary classification on the values of the bacterium Escherichia coli. If the values exceeded 300CFU/mL, it was defined as 1, and if the value was below the mentioned limit, it was defined as 0. Only after the classification was performed, solutions of higher accuracy were obtained.

### A. Regression

When predicting the target variable Escherichia coli, machine learning algorithms were used to perform regression. Out of all the algorithms that were tested, only the results obtained by the SVR algorithm and linear regression will

be discussed in this paper. The algorithms used yielded poor solutions for each model and as such need not to be explained. Firstly, the task starts by loading a file with the appropriate set of data to use on which certain algorithms are then implemented. Loading is done using the *pandas* library. After successfully loading the file, the data is divided into input independent variables and one target variable. Using the *train_test_split()* function, the data is divided in an 80:20 ratio for learning and testing. For each algorithm we created an instance of the object on which the *fit()* function was used for training the model, and then the predicted values of the target variable are realized with the *predict()* function. Using the *matplotlib.pyplot* library, the obtained results are graphically displayed and they are shown in table format for easier readability. Program flow is represented on Figure 1.

Fig. 1. A graphical representation of program flow




The number of actual values within the test data set that exceeded 300 CFU/mL was calculated and also the number of predicted values exceeding the pollution limit. The results are presented in percentages. Also, the difference between actual and projected values was calculated, as described in section 4.1.

### B. Classification

The classification algorithms used are logistic regression and SVM. The steps for uploading the file, sharing the data, training, and predicting the target variable are the same as for regression. In classification, a confusion matrix is calculated that shows us how the algorithm predicted the values. Namely, the algorithm predicts based on the input variables whether the corresponding target class will be 0 or 1. If it is 0, the value of Escherichia coli is below the defined pollution threshold, and if the predicted class 1 exceeds the limit of 300CFU/mL. More is described in section 4.2.

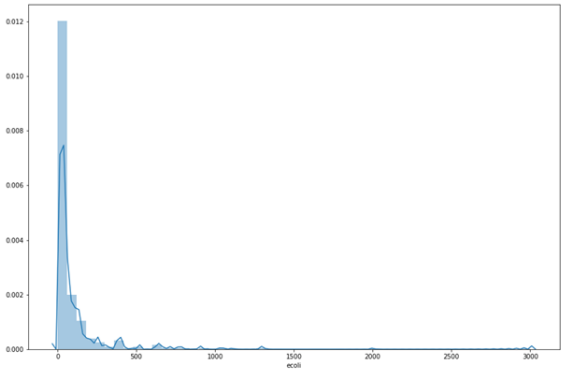
## III. CASE STUDY

The obtained data set was divided into two parts: set A and set B. Set A, for input parameters, contains air temperature, water temperature, salinity, pH, turbidity, solar radiation, precipitation, precipitation 24h, precipitation 48h, precipitation 72h, direction and wind strength. Escherichia coli was defined as the output parameter and prediction variable. Set B contains: air temperature, daily albedo, azimuth, cloud transparency, dew point temperature, dhi (Diffuse Horizontal Irradiance, W/mˆ2), days (Diffuse Normal Irradiance, W/m ˆ2), ghi (Global Horizontal Irradiance, W/m ˆ2), ebh (Direct Beam Horizontal Irradiance, W/m ˆ2), Gni (Global Normal Irradiance, W/m ˆ2), Gti Fixed Tilt - (Global Tilted Irradiation), Gti Trackig, sedimentation rate , relative humidity, snow depth, surface pressure, wind direction at 10m, wind speed at 10m, zenith. The output parameter that was defined is Escherichia coli.

### A. Distribution of Escherichia coli

A graphical representation of the distribution of the target variable Escherichia coli was made on both data sets as seen on figures 2 and 3.

Fig. 2. Graphical representation of the distribution of the target variable Escherichia coli on set A.



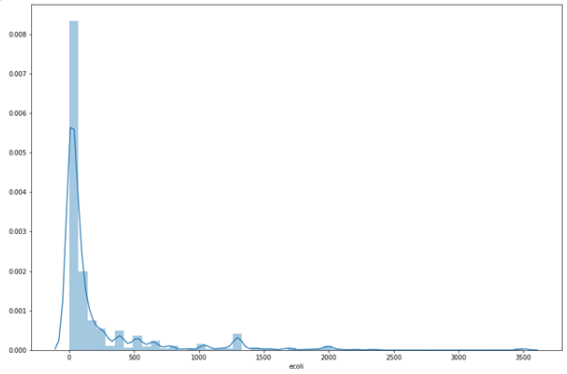

On set A of the data, it can be concluded that the bacterial value predominantly assumes values between 0 and 1000 with a small proportion of those ranging from 1000 to 3500 which can be seen on figure 2.

On set B, the value of Escherichia coli, again, takes the highest values in the range from 0 to 1000 which is shown on figure 3.

### B. Influence of parameters

After correcting the null values and dropping the deviating values, the influence of the parameters was calculated. This

Fig. 3. Graphical representation of the distribution of the target variable Escherichia coli on set B.




was done using the Random Forest algorithm. In the first set of data, the greatest impact on sea quality has salinity with utility of 14%, and the least impact is precipitation with 4% or less. For set B data, relative humidity, and sediment waters with 10% or more utility have the greatest impact. Albedo has the least impact with less than 2%.

### C. Correlation of parameters

In order to better understand the data and their interrelationships, a correlation of parameters for both data sets was made. Correlation represents the interconnection of defined parameters and according to their relationship, the value of one variable can be overlooked based on the knowledge of the value of another variable. For set A, dark blue indices on figure 4. indicate a strong correlation between the variables, while red shows a weak correlation. Thus, it can be seen from the picture that precipitation in any time period has a strong influence on precipitation in another time period, which actually means a clear sequential connection through the seventy-two-hour period. Also, turbidity is quite related to rainfall, as is water temperature with salinity and air temperature. It can be seen that the sun and wind affect the air temperature and the sea temperature. Rainfall has the weakest correlation with air temperature, sea temperature and salinity. Also, there is a weak correlation between turbidity and temperature and salinity.

For set B, the following was obtained: Again, blue shows a strong correlation, while red shows a weak one. All judge radiation indices (Dhi, Dni, Ebh, Ghi, and Gti) have strong correlations with each other. Also, sun radiation indices have a strong correlation with air temperature. The weakest relationship was observed between relative humidity and air temperature, and sun's radiation. This is represented on figure 5.

### D. Model evaluation

The *score ()* function was used to evaluate the model, which calculates the accuracy of the prediction model. The mean absolute error value is then calculated using the *mean_absolute_error ()* function of the *sclearn* library, *mean_squared_error ()* and root mean square error as the root of the *mean_squared_error ()* function.

Fig. 4. Matrix representation of correlation between variables on set A.

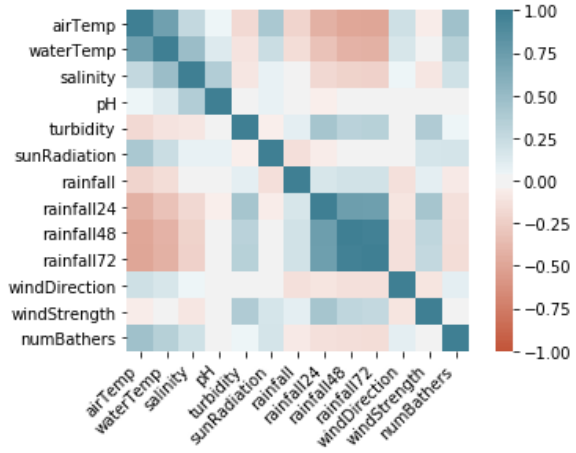
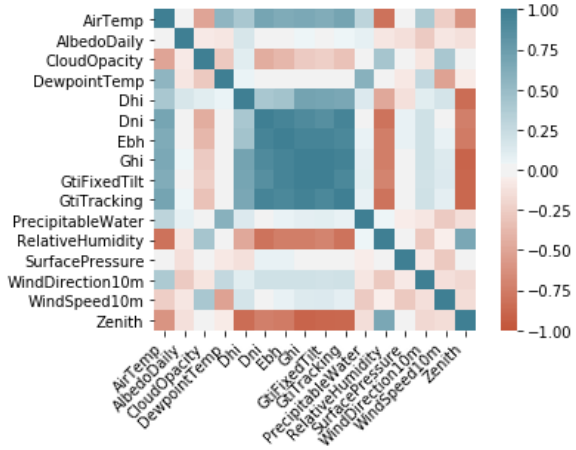



Fig. 5. Matrix representation of correlation between variables on set B.




1) Mean absolute error (MAE) - MAE measures the average magnitude of the errors and a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight [2]

Fig. 6. Mean absolute error formula. [2]

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

2) Root mean squared error (RMSE) - RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation [2]

Both MAE and RMSE express the average model prediction error. They range from 0 to infinity and are negatively oriented which means that lower values indicate better re-

Fig. 7. Root mean squared error. [2]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

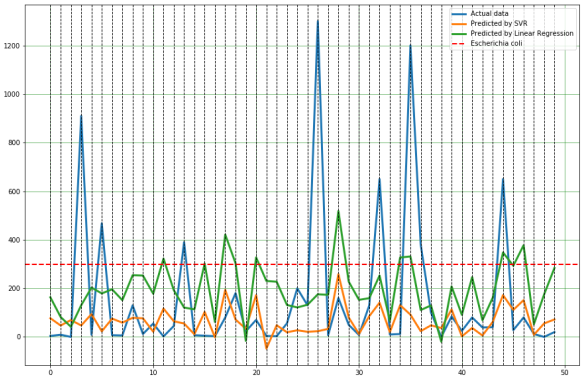
sults. In RMSE, errors are squared before their average is calculated, which means that RMSE should be more useful when large errors are particularly undesirable. [2]

## IV. RESULTS

### A. Results achieved by regression

Using the flow SVR algorithm and linear regression on given data sets, the following results were obtained.

Fig. 8. Graphical representation of results achieved by SVR and linear regression.




The graph on Fig. 8. shows the actual values of Escherichia coli marked in blue in the range of fifty days. The orange line shows the results obtained by the SVR algorithm, and the green line indicates the results achieved by linear regression in the same period. Red dashed indicates the threshold value of Escherichia coli of 300CFU / mL and also the limit of marine pollution. In this way, the actual values of the bacterium can be seen exactly for each day, and the adaptation of the models obtained by the regression algorithm.

The data of set A is divided in a ratio of 80:20 for learning and testing, which gives us the number of 193 data for testing, of which the figure shows 50 results for a more accurate definition. There are also 193 data of predicted values. For set B, there are 133 test samples and 133 corresponding predicted values.

Figure 18. and 19. (see Appendix) numerically shows the actual values, the values achieved by SVR and linear regression. Numerical representation of data allows for easier understanding of accurate values for bacteria and as such it is presented in Fig. 18. and Fig. 19..

On the whole data set, the actual value of the bacterium exceeded the limit of 300CFU/mL 167 times. In testing of the 193 samples, the predicted values exceeded the actual values 99 times, which means that in 51% of cases the predicted value was higher than the defined actual value. Figure 19. (see Appendix) shows the numerical difference between the values of the actual and predicted values for the SVR algorithm for the case of the first 50 test values.

When calculating the difference between the actual and the estimated value in percentages, the estimated value in relation to the actual value is looked at. There are two cases:

1) Estimated value less than actual value
   Calculated as (estimated value * 100) /actual value. In the specific case (Fig. 19., sample number 4): the actual value of the bacterium is 910, and the predicted value is 47. The error is in 863CFU/mL, i.e. in percentages it is 47 * 100/910 = 5%.
2) Estimated value greater than actual value
   It is calculated as 100 - (actual value * 100)/estimated value. In the specific case (Fig. 19., sample number 0): the actual value of the bacterium is 4 and the predicted value is 77. The error is in 73CFU/mL, i.e. in percentages it is 100 - (4 * 100) / 77 = 99%. Since the predicted value is higher than the actual one, the percentage is calculated as the opposite case.

Calculation of the exact number of times when the predicted value exceeded the amount of 300CFU/ml (and it amounts to 2% (all four times)), and was then divided into two cases:

1) When the actual value is greater than predicted – Such cases were recorded on 139th, 167th and 175th samples of 193 test values.
2) When the actual value is less than predicted – There was only one such case in the 141st sample.

Calculation of the exact number of times the predicted value is below the defined limit and this occurred in 98% of cases, a total of 189 measurements, and was then divided into two cases:

1) When the actual value is greater than predicted - There are a total of 91 such values.
2) When the actual value is less than predicted - There is a total of 98 such values.

In the actual values taken for testing, 38 of them exceed the defined contamination limit on a set of 193 samples. In the fig. 8 we can see that in the first fifty cases it happened seven times. This represents 19.689119171% of total test values.

The following table shows the accuracy of the model:

Fig. 9. Numerical representation of accuracy of SVR and linear regression models.

| | Support Vector Regression | Linear regression |
|---|---|---|
| Model accuracy | 0.12990416150753048 | 0.06445583050980475 |
| Mean absolute error | 209.59902906457543 | 249.14764967492712 |
| Mean squared error | 201519.9946945829 | 166855.61704691147 |
| Root mean squared error | 448.9097845832533 | 408.47964092095395 |

For a better graphical representation, the scatter function was used from the *matplotlib.pyplot* library for a more credible representation of the actual values and the values achieved by the algorithms. In the Figures 10. and 11., the real values of Escherichia coli are marked with blue dots, and the green line shows the adjustment of the model to the actual values.

Fig. 10. Graphical representation of scatter function using linear regression.

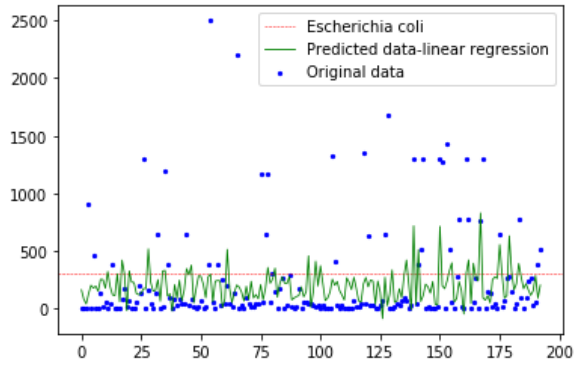
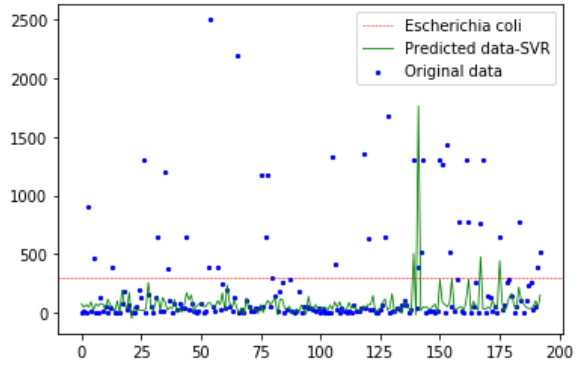



Fig. 11. Graphical representation of scatter function using SVR regression.




The graphs on Figures 10. and 11. show all 193 values taken for testing and their values and prevalence within the specified range. A red dashed line indicates a bacterial indicator value of 300CFU/mL. The x-axis contains 193 samples of the bacterium, while the y-axis contains the values that the bacterium takes for each sample ranging from 0 to 2500.

### *B. Results achieved by classification*

Support Vector machine and logistic regression were used for classification. Two data sets were also worked on: set A and set B.

*1) Set A:* Figure 12. shows the confusion matrix showing the result obtained by logistic regression. Of the 193 test values, for as many as 162, algorithms got a true positive result. In 25 cases the result turned out to be false positive. For two values results are false negative and for four values results are true negative.

For the SVM algorithm the results are shown on Figure 13. and are as follows: 164 values are true positive, for 29 values the result is false positive. In 0 cases the result is false negative and true positive.

The table Figure 14. shows the accuracy of the model.

*2) Set B:* The confusion matrix for SVM on set B of the data in Figure 15. shows a true positive result for 107 values, 26 false positive results, zero false negative and zero true negative results.

For the logistic regression performed on set B, shown on Figure 16., 104 the values are true positive, 20 false positive, 3 false negative and 6 true negative.

The table Fig. 17. shows the accuracy of the model.

Fig. 12. Confusion matrix results obtained by logistic regression on set A.

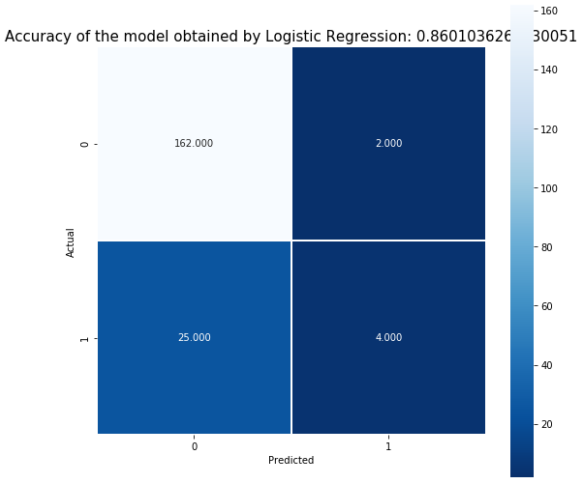
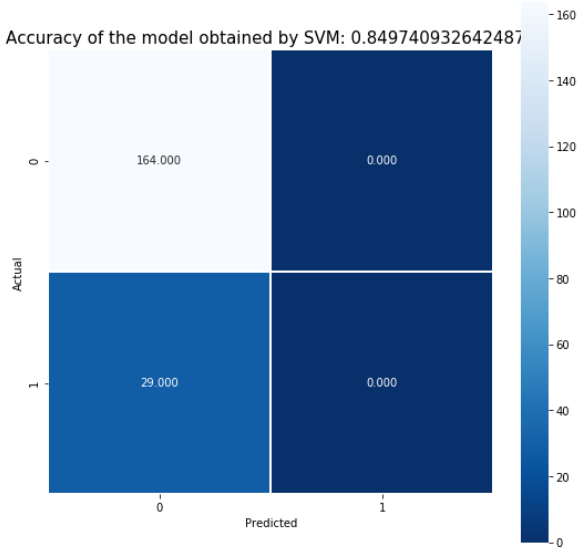



Fig. 13. Confusion matrix results obtained by SVM on set A.



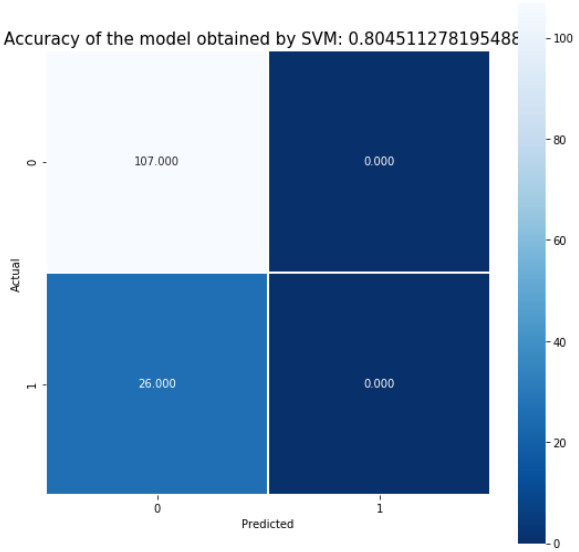

## V. DISCUSSION

Based on the regression algorithms used (five of them), it can be concluded that the regression did not give good results. Even after omitting the deviation of the Escherichia coli values and performing linear interpolation for the missing values, the algorithms resulted in almost no accuracy. There was no overlap between actual and predicted values at all, while the values were approximately close in only 5% of cases. Using the methods for evaluating the results (score function, MAE and RMSE), it is noticeable that the results are poor. Namely, for the SVR algorithm the accuracy of the model is 0.12990416150753048% and for the linear regression 0.06445583050980475. Since the accuracy of the models does not exceed even 15%, it can be concluded that the algorithms used are not the best solution in this case. The MAE for SVR is 209.59902906457543, and for linear regres-

Fig. 14. Numerical representation of accuracy of SVM and logistic regression models on set A.

| | Support Vector Machine | Logistic regression |
|---|---|---|
| Model accuracy | 0.8497409326424871 | 0.8601036269430051 |
| Mean absolute error | 0.15025906735751296 | 0.13989637305699482 |
| Mean squared error | 0.15025906735751296 | 0.13989637305699482 |
| Root mean squared error | 0.3876326448552972 | 0.3740272357155222 |

Fig. 15. Confusion matrix results obtained by SVM on set B.




sion this value reaches 249.14764967492712. The RMSE for SVR is 448.9097845832533, and for linear regression 408.47964092095395. As stated earlier, these functions for evaluation of results and show better results if you take on smaller values. Also, SVR assumes lower values for MAE than linear regression but linear regression expects lower values for RMSE. Separately from stated, MSE and RMSE expect too high values for both algorithms for them to be used at all. By that, it can be concluded that the pair of SVR and linear regression are equally not good choices for regression.

After the elaborated binary classification for Escherichia coli and the implementation of classification algorithms, better results were obtained. In relation to the regression algorithms, the results were better for 70% - 80%. What can be concluded is that classification works better because of the prediction of only two classes, 0 or 1, while regression works on predicting a continuous value. In any case, it is easier to predict a single class than an exact value in the range of 0 - 3500. In SVM algorithm, in set A, for the 164 test values algorithm predicted that the target variable belongs to a class 0 and 29 times it predicted that belong to the class 0, although they were supposed to belong to class 1. For zero values it predicted to belong to class 1 and for zero values it predicted wrongly to belong to class 1. It gives us the accuracy of the model of 84%.For set B, the algorithm predicted that it would belong to class 0 for 107 values, that it would belong to class 1 for zero values, and incorrectly

Fig. 16. Confusion matrix results obtained by logistic regression on set B.

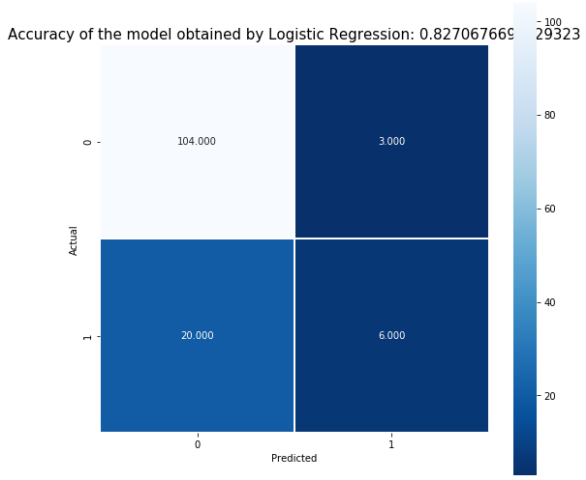



Fig. 17. Numerical representation of accuracy of SVM and logistic regression models on set B.

| | Support Vector Machine | Logistic regression |
|---|---|---|
| Model accuracy | 0.8045112781954887 | 0.8270676691729323 |
| Mean absolute error | 0.19548872180451127 | 0.17293233082706766 |
| Mean squared error | 0.19548872180451127 | 0.17293233082706766 |
| Root mean squared error | 0.4421410655034333 | 0.4158513326022506 |

predicted that it belonged to class 0 for 26 values. It resulted in an accuracy of 80%. Logistic regression for set A resulted in a 86% accuracy. For 162 values the algorithm predicted that the target variable would belong to class 0, for 25 values it was predicted to belong to class 0 even though it belong to class 1, for two values the algorithm predicted that they belong to class 1 in spite of belonging to class 0 and for four values it was predicted to belong to class 1 which is true. For Set B, 104 values were defined to belong to class 0, 20 values are false positive, meaning they actually belong to class 1, three values are predicted to belong to class 1 but indeed belong to class 0 and six values are predicted to belong to class 1 which is true. The model gives an accuracy of 82%.

When using classification algorithms, the model accuracy for the SVM algorithm resulted in 84% and 80% accuracy, respectively, for set A and set B data. In logistic regression, the accuracy percentage reaches 86% for set A and 82% for set B data. It can be noticed that the results with both algorithms are better for the B data set. The MAE for set A is 0.15025906735751296 for SVR and 0.13989637305699482 for logistic regression. Since the values are close to zero, this evaluation method confirms that the results for this model are appropriate. For set B data, the SVR has a value of 0.19548872180451127 and the logistic regression 0.17293233082706766. The RMSE for SVR for set A is 0.3876326448552972, and for logistic regression 0.3740272357155222. In the case of set B data, these values are slightly higher: 0.4421410655034333 for SVR and 0.4158513326022506 for logistic regression. In the case of classification, logistic regression proved to be a better

choice compared to the SVM algorithm. It can be concluded that the results are equitably acceptable in both cases, set A and set B for classification algorithms, even though, results for set A are slightly better. logistic regression gives us a little bit better results. Also, logistic regression gave somewhat greater results that Support Vector Machine. Comparing the results for MSE and RMSE, it can be noted that they are better when using classification algorithms than when using algorithms for regression. Although the data sharing ratio for learning and testing was changed, no higher accuracy results were obtained. It can be that the data is not of good quality and not completely accurate. There are unexplained deviations in the values of Escherichia coli as well as shortcomings in the measurements for temperature, turbidity, and salinity. Even after supplementing them using linear interpolation, no better solutions were obtained when using regression. The data sets consist of 966 and 666 measurements, which can certainly be a larger number because the algorithms will work better on a larger number of samples because there is more data to learn.

## VI. Conclusion

We have defined two sets of data that individually contain data from the Rijeka meteorological station and data from the Solcast site. Both data sets formed the basis for model development and were used in regression algorithms as well as in classification algorithms. The influence of the parameters was calculated in order to know as precisely as possible which input variable is the most influential, and whose usefulness is almost negligible. Also, a correlation of all input variables was made in order to be able to define the relationship between the parameters used. Each regression algorithm used to make a prediction model was compared to the model of actual bacterial values. Each model is graphically represented, defined on 50 measurements, and table representation of actual and anticipated results in a numerical format. Complete value overlaps were not the case with either algorithm. In actual values, there are deviations of bacteria that take values up to 3500 CFU/mL, and most often this value is 1300 CFU/mL. Although given the values of the independent variables there is no reason for these values to be so high. And if we follow the results obtained by the prediction model, in the case where the actual values bounced (exceeded the threshold of 300CFU / mL), the prediction model gave better results. In addition to the algorithms mentioned in this paper, the CART algorithm, Bayesian Ridge and Gaussian Process Regression were tested , but they gave poor results and their graphical presentation is not shown in this documentation because we already have an algorithm with poor results, it holds no value to list them all. Algorithms for predicting other bacteria (UBB / 37) were also tested, and the results were no better. Since Escherichia coli is kept as a defined indicator of marine pollution in the further development of the project, it was used as the only target variable. Also, other bacteria in the sea are affected by Escherichia coli, which means that if this bacterium is absent or its value is negligibly small, no other bacteria will be present. We conclude that it does not make sense to take other bacteria as input parameters when they are under the direct influence of Escherichia coli. After unsuccessful attempts to develop a prediction model using regression algorithms, a binary classification for the bacterium was made and classification algorithms were used. With Support Vector Machine and logistic regression, the obtained results hold greater accuracy and smaller square deviation which has shown that the classification is more than adequate in relation to the regression. Given that the goal of the project was to develop a model that will predict the quality of the sea, we believe that we have successfully completed the task. By regression, we presented the exact predicted continuous values for Escherichia coli relative to the actual values, and the classification gave even more accurate results by predicting which class the bacterial value for a particular sample would belong to.

## References


[1] S. M. D. L. S. J. Z. L. L. B. L. G. N. B. Darija Vukic Lusic, Lado Kranjcevic. Temporal variations analyses and predictive modeling of microbiological seawater quality.

[2] JJ. Mae and rmse — which metric is better?

## APPENDIX

Fig. 18. Numerical representation of data and results obtained using SVR and linear regression.

| | Actual values | Predicted values using SVR | Predicted values using linear regression |
|---|---|---|---|
| 0 | 4 | 77 | 163 |
| 1 | 9 | 47 | 82 |
| 2 | 0 | 69 | 41 |
| 3 | 910 | 47 | 130 |
| 4 | 9 | 94 | 203 |
| 5 | 468 | 23 | 179 |
| 6 | 7 | 75 | 195 |
| 7 | 6 | 58 | 151 |
| 8 | 130 | 78 | 254 |
| 9 | 12 | 76 | 252 |
| 10 | 55 | 21 | 177 |
| 11 | 2 | 116 | 322 |
| 12 | 45 | 64 | 192 |
| 13 | 390 | 55 | 120 |
| 14 | 7 | 10 | 113 |
| 15 | 5 | 103 | 302 |
| 16 | 3 | -1 | 61 |
| 17 | 80 | 193 | 421 |
| 18 | 180 | 71 | 303 |
| 19 | 23 | 33 | -16 |
| 20 | 70 | 171 | 327 |
| 21 | 4 | -48 | 230 |
| 22 | 3 | 48 | 227 |
| 23 | 56 | 19 | 131 |
| 24 | 200 | 27 | 121 |
| 25 | 130 | 21 | 132 |
| 26 | 1300 | 23 | 175 |
| 27 | 7 | 32 | 173 |
| 28 | 160 | 258 | 518 |
| 29 | 50 | 80 | 227 |
| 30 | 8 | 12 | 152 |
| 31 | 130 | 85 | 159 |
| 32 | 650 | 141 | 252 |
| 33 | 10 | 19 | 66 |
| 34 | 12 | 130 | 327 |
| 35 | 1200 | 91 | 331 |
| 36 | 300 | 23 | 110 |
| 37 | 100 | 47 | 129 |
| 38 | 0 | 36 | -19 |
| 39 | 84 | 113 | 207 |
| 40 | 25 | 2 | 90 |
| 41 | 80 | 36 | 246 |
| 42 | 39 | 5 | 68 |
| 43 | 40 | 61 | 161 |
| 44 | 650 | 172 | 347 |
| 45 | 28 | 111 | 292 |
| 46 | 80 | 151 | 377 |
| 47 | 12 | 8 | 51 |
| 48 | 0 | 55 | 175 |
| 49 | 20 | 71 | 283 |

Fig. 19. Numerical representation of data and results obtained using SVR and the difference between the actual and estimated value.

| | Actual values | Predicted SVR values | Differences in values expressed in percentages |
|---|---|---|---|
| 0 | 4 | 77 | 95% |
| 1 | 9 | 47 | 81% |
| 2 | 0 | 69 | 100% |
| 3 | 910 | 47 | 5% |
| 4 | 9 | 94 | 91% |
| 5 | 468 | 23 | 4% |
| 6 | 7 | 75 | 91% |
| 7 | 6 | 58 | 90% |
| 8 | 130 | 78 | 60% |
| 9 | 12 | 76 | 85% |
| 10 | 55 | 21 | 39% |
| 11 | 2 | 116 | 99% |
| 12 | 45 | 64 | 31% |
| 13 | 390 | 55 | 14% |
| 14 | 7 | 10 | 37% |
| 15 | 5 | 103 | 96% |
| 16 | 3 | -1 | -49% |
| 17 | 80 | 193 | 59% |
| 18 | 180 | 71 | 39% |
| 19 | 23 | 33 | 32% |
| 20 | 70 | 171 | 60% |
| 21 | 4 | -48 | -1200% |
| 22 | 3 | 48 | 94% |
| 23 | 56 | 19 | 33% |
| 24 | 200 | 27 | 13% |
| 25 | 130 | 21 | 16% |
| 26 | 1300 | 23 | 1% |
| 27 | 7 | 32 | 79% |
| 28 | 160 | 258 | 39% |
| 29 | 50 | 80 | 38% |
| 30 | 8 | 12 | 34% |
| 31 | 130 | 85 | 65% |
| 32 | 650 | 141 | 21% |
| 33 | 10 | 19 | 50% |
| 34 | 12 | 130 | 91% |
| 35 | 1200 | 91 | 7% |
| 36 | 300 | 23 | 6% |
| 37 | 100 | 47 | 47% |
| 38 | 0 | 36 | 100% |
| 39 | 84 | 113 | 26% |
| 40 | 25 | 2 | 9% |
| 41 | 80 | 36 | 46% |
| 42 | 39 | 5 | 15% |
| 43 | 40 | 61 | 36% |
| 44 | 650 | 172 | 26% |
| 45 | 28 | 111 | 75% |
| 46 | 80 | 151 | 48% |
| 47 | 12 | 8 | 72% |
| 48 | 0 | 55 | 100% |
| 49 | 20 | 71 | 73% |