School Of Information Technologies and Telecommunications, Department of Informatics

University Of Piraeus

DIPLOMA THESIS

Descriptive analysis, feature selection and forecasting of water quality features, using machine learning and artificial intelligence

Full Name*:* **MARKO PLAKU**

REGISTRATION NUMBER: P17107

SUPERVISOR PROFESSOR: DIMITRIS APOSTOLOU

Piraeus, 2021-2022

Keywords

Water Quality,

Water's Electrical Conductivity,

Feature Selection,

Pearson's Correlation,

Mutual Information,

ARIMA,

Forecasting,

Machine Learning,

Artificial Intelligence,

Multi-Layer Perceptron,

Recurrent Neural Network

Long Short-Term Memory

Abstract

The use of Machine Learning and Artificial Learning to give answers in real world problems , is getting more and more usual as they have proven decision making and predictive abilities, sometimes better than humans, which makes them an integral tool for operational tasks often found in the industrial sector. Industries processing and using the qualitative features of the water, whether it comes from biochemical processes or simply found in nature, are no exception. This paper is proposing the use of such tools to achieve the prediction of natural water quality features, focusing on the feature of Water's Electrical Conductivity. Descriptive analysis, feature selection techniques and various machine learning and Artificial Intelligence (A.I) models are applied and tested in order to find the best complementary features for Electrical Conductivity leading to the most accurate predictions, with the final purpose being the forecasting of future values. The structure of the thesis can be viewed in the fig.1, where the content of the thesis is categorized in three continuous phases. Explanatory, these phases are the Data Preprocessing phase, consisting of the data exploration and cleansing, the searching of outliers and the transformation of the dataset so it describes daily or hourly measurements. The second phase is the feature selection phase where methods such as the Pearson's correlation coefficient and the Mutual Information are implemented. Last but not least, the phase 3 where Machine Learning and Artificial Intelligence models are implemented with goal of predicting the water's electrical conductivity values.

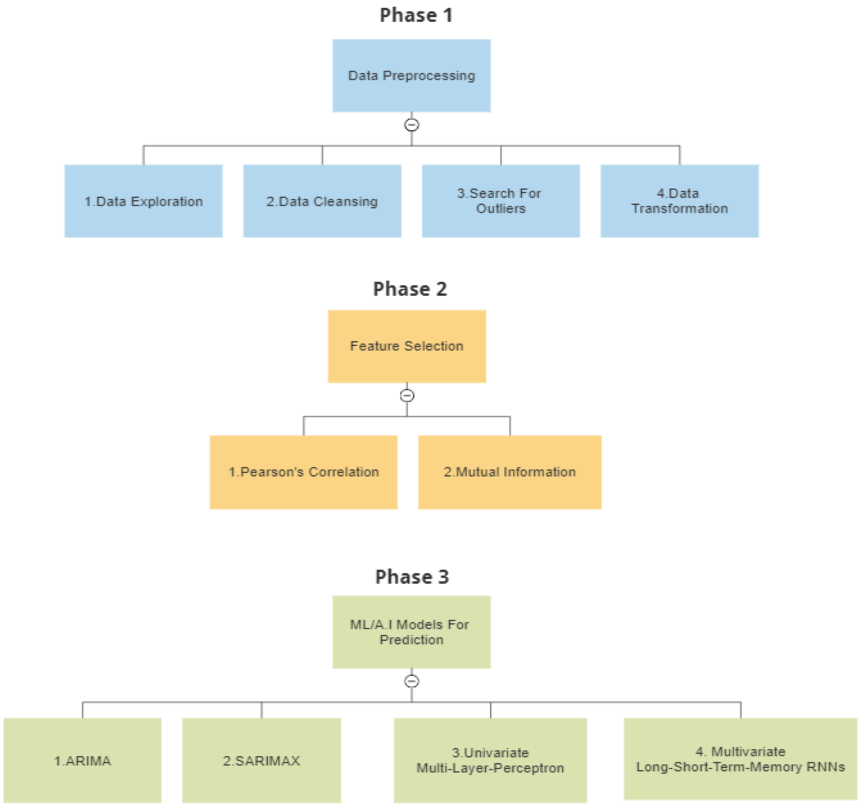


Figure 1*:Diagram describing the content structure of this thesis.*

1.Introduction

As previously mentioned, nowadays Machine Learning and A.I are constantly being applied at enterprise and industrial level, not only at a research and development level but also at production stages , as a logical resultant of their ability to give reliable answers to complex questions. They found their place at industries that process and researches the various Water Features, such as Water Treatment Plants. Extended research has already been done , examining the water quality features as they appear ,processed or not, at such Industries.

In this thesis, water data referring to characteristics found at natural water sources, are being examined , with emphasis given at the feature of Electrical Conductivity (EC). After the data acquisition from various sources, descriptive analysis as well as feature selection methods are applied in order for the most ideal dataset to be formed, containing the most correlated features with the water's electrical conductivity feature, aiming for the final goal which is the forecasting of the future values of water's EC values, in both shorter and longer time periods. Why the water's electrical conductivity feature? Because water's electrical conductivity is a very important water quality feature overall. That is because, water's electrical conductivity can indicate the water quality itself. More details about the importance of water's electrical conductivity will be given below.

When it comes to the data used for this thesis, different data sources are being used to collect both raw water features as well as weather data with the purpose of initially find as many features as possible. The Water quality data collected by a station that iscollecting data at Schellebelle Belgium and specifically from the Zeeschelde (De Schelde) river, with an hourly logging periodicity. The data acquired from the station are referring at a year's time span. The initial dataset , apart from the datetime log and the water's EC value itself, is consisted of thirteen other features. Afterwards, as the preprocessing phase of the raw data took place, the descriptive analysis on the multiple features is conducted, aiming to find the features that appear the highest correlation with the water's EC value. That is because the goal was not only to reduce and limit the initial dataset ,eliminating all the unnecessary or not correlated features, but also to identify and form the ideal dataset that will be used as an input , for the training phase of the machine learning and artificial intelligence models. Through this process of feature selection, the main methods that are implemented are Pearson's Correlations Coefficient and Mutual Information. The complementary use of both methods leaded to the final dataset formation , viz ,timeseries, containing the features that appear to most affect the water's electrical conductivity and ended up to be selected. Through this feature selection phase, data reduction was simultaneously happened as the complementary features are reduced from thirteen to only four.

After the feature selection phase, the analysis of water's electrical conductivity is conducted, with the goal of the further examination of the feature's characteristics , such as its seasonality. After that, ARIMA and SARIMAX models are trained, evaluated, and used for the forecast of the next day's mean water's EC value . Then, the water's EC values, as an individual feature and in a supervised formation, is inputted for the training of a Multi-Layer Perceptron (MLP) which after its evaluation , it is used for the prediction of the EC of the next 10 hours. Finally, a Multivariate Long Short-Term Memory (LSTM) Recurrent Neural Network is built , trained with the other four features from the final dataset as an input, and is able to predict the water's EC Value. The LSTM model ,at first, is designed to predict the next day's water's EC values based on the other features previous day data, but eventually designed to predict the water's EC values of the next hour, based on the time-corresponding values of the other four features. After evaluation from various statistical methods, such as the coefficient of determination (R^2) and the Mean Squared Error (MSE), the model was used for the forecast of real water's EC values for the next 10 hours. The data for the other four features were created by individual , high performance univariate MLPs for each feature.

2. Review of machine learning and artificial intelligence application for water resources problems

As prementioned, extensive research and applications of machine learning and artificial intelligence (A.I) models, referring to water sources and water quality data have already been conducted. The outcomes from these studies not only have given important insights but also have proven the ability of machine learning and A.I to aid to predict future values and eventually solve real world problems. This breakthrough use of machine learning and artificial intelligence models is definitively a game changer in the approach that is used to solve stochastic and deterministic problems as well. Water quality problems are not an exception, and therefor the applications of machine learning and A.I for solving the water quality and environmental problems, are becoming more and more popular. In this chapter, bibliographic description of similar work ,where machine learning and artificial intelligence models are used to solve problems referring to water quality and environmental issues will be referenced.

2.1 Big Data used for analysis of water resources analysis.

It is as fact that for water managers/engineers, big data is showing big promise in many waters related applications such as planning optimum water systems, detecting ecosystem changes through big remote sensing and geographical information system, forecasting/predicting/detecting natural and manmade calamities.

Big Data analysis and modelling can reach even farmers in poor, rural areas of developing countries through cellphones, providing access to weather forecasting and market information to make better decisions and thereby helping improve livelihoods as well as local water and food security. Big data is creating a new generation of decision support data management. A key to deriving value from big data is the use of analytics *[1]: Sirisha Adamala :An Overview of Big Data Applications in Water Resources Engineering(2017).*

2.2 Machine learning and Artificial Intelligence applications for Environmental Science

Innovations in Machine Learning and Data Analytics can possibly affect numerous aspects of Environmental Science (ES. When Machine Learning and Deep Learning are combined it is possible to unleash the supremacy of data analytics. These advancements also aid in bridging the gap between the theoretical backgrounds on ES to practical implementation. Machine Learning and Deep Learning are the sub fields of artificial intelligence which deals with training the models to learn from data without being explicitly programmed.

These techniques show high prospective for process optimization, information-centric decision making and scientific discovery. Scientific developments like these will assist ES to make real time autonomous decisions by extracting useful insights from huge data.*[2]: Tharsanee Raman Maganathan , Soundariya Ramasamy Senthilkumar , Vishnupriya Balakrishnan: Machine Learning and Data Analytics for Environmental Science: A Review, Prospects and Challenges(2020)*

2.3 Deep Learning applications in hydrology

The volume, variety and velocity of water-related data are increasing due to large-scale sensor networks and increased attention to topics such as disaster response, water resources management, and climate change. Combined with the growing availability of computational resources and popularity of deep learning, these data are transformed into actionable and practical knowledge, revolutionizing the water industry.

Deep learning approaches used in the water industry for generation, prediction, enhancement, and classification tasks, and serves as a guide for how to utilize available deep learning methods for future water resources challenges. Runoff prediction and flood forecasting are major tasks in rainfall-runoff modeling. Toward this end, many researchers have applied cutting-edge deep learning architectures to the runoff prediction and flood forecasting tasks. Since rainfall and runoff are both time-series data, the common networks for the streamflow prediction and flood forecast are RNN, LSTM *[3]: Muhammed Sit,Bekir Z. Demiray,Zhongrun Xiang;Gregory J Ewing,Yusuf Sermet,Ibrahim Demir: A comprehensive review of deep learning applications in hydrology and water resources(2020)*.

3.Data Acquisition and Preprocessing

3.1 Data acquisition

As previously mentioned, the data used for this project are referring to water quality features. However, since the process of feature selection will take place, it is ideal to collect as many features as possible, with the aim not only to have a bigger selection of features to choose from and eventually use the most appropriate ones, but also to have diversity between the dataset, as different features can provide different insights and therefore , features that at first glance seemed to have no correlation with the water's electrical conductivity, after research, to be found to provide significant information. This would make the analysis process of feature selection more surprising and may lead to unexpected but positive conclusions. The initial dataset that is used, is consisted of features from two sources. Firstly, from features directly associated with water quality such as its chemical characteristics . Secondly, from weather data which accompany the chemical features. Both sources are of course referring to the same geographical area and are in a format of continues time logs, at the level of minutes, making the dataset describing timeseries, and the analysis that will follow to be timeseries analysis. It is important to refer that the time-period that the data , both chemical and weather, collected from, studied and the whole project refers to, is the time period beginning from 03/04/2021 up until to 03/04/2022. Daily measurements starting from the first day up until and the very last day are logged, making the time period that the timeseries are describing, a whole consecutive year.

At first, data referring to chemical features of water, indicating its quality were used. This data derived from an online website where the water's qualitative features measurement, from Belgium stations, are public. The website is called waterinfo.be[1]. The website belongs to the Flemish Government and its goal is to inform about floods and droughts in order to minimize the water damage .The website offers a variety of stations available where water measurements are logged, throughout Belgium. Each station is able to log and different water features. The station where the most chemical features are logged, and for which the interest is about, is chosen. This station is called the station of Schellebelle. The station tracks and logs the qualitative features of water from and the Zeeschelde (De Schelde) river, which passes by the station. Every measurement of each chemical feature is logged at the rate of a minute. Each feature can be downloaded separately in a csv format file.

These chemical qualitative features that are measured are:

- **Water's PH**: As it is known in chemistry, pH is a measure that indicates the concentration of hydrogen ions in a water-based solution, measure that characterizes the acidity or alkalinity of a liquid. As it is known, distilled water's PH, which is neutral, is 7.

[1] waterinfo.be

- **Water's Temperature (Celsius metric system):** The temperature of the water measured at the time studied. The measurements were conducted using the Celsius metric system.
- **Water's Turbidity(NTU):** Turbidity is the measure of relative clarity of a liquid. Turbidity makes water cloudy or opaque .The higher the intensity of scattered light, the higher the turbidity. Material that causes water to be turbid include clay, silt, very tiny inorganic and organic matter, algae, dissolved colored organic compounds, and plankton between other microscopic organisms. Excessive turbidity, or cloudiness, in drinking water, represent a health concern.
- **Water's Dissolved Oxygen(mg/L):** Dissolved oxygen is the amount of oxygen that is present in water at any given time . If the levels of Dissolved Oxygen are 1 (mg/L),the waters are considered hypoxic and usually devoid of life.
- **Water's Salinity(psu):** Water's Salinity is the concentrations of salts in water. High levels of salinity are considered harmful to many plants and animals.
- **Water's Electrical Conductivity(μS/cm):** Electrical Conductivity(EC) is the feature that is set to be examined and forecasted. Electrical Conductivity of water indicates its ability to conduct an electrical current. Salts or other chemicals that dissolve in water can break down into positively and negatively charged ions. The electrical conductivity (EC) of water depends on the concentrations of free ions. Electrical conductivity it is also a strong indicator of water's purity. It is a fact that the purer the water the lower is its conductivity.
- **Water's Chlorophyll(mg/L):** Water Chlorophyll comes from algae that live in the water (also called phytoplankton). Algal growth is often linked to nutrient enrichment in freshwaters. Excessive algal growth is accompanied with the growth and multiplication of respiring bacteria which may use up existing dissolved oxygen in the river water.

Next, it was crucial to find a complementary data features , apart from water characteristics .Weather features data were ideal. The weather data that are used, derived from the closest weather station from Schellebelle station were the water features are measured. The weather station found was the Melle's weathers station .Melle weathers station and Schellebelle are approximately 10 kilometers (km) apart from each other. The fact that they have a very small distance, it is a strong , trustworthy factor that the weather data collected can indeed affect the water quality as it is measured nearby. The actual data were acquired by using a Python's API called Meteostat [2]. Given the coordinates of the Scellebelle , the API can find the closest weather station, to the given coordinated, that has available weather data. With given longitude : 51.0115409 and given latitude 3.917, Scellebelle's coordinates , it is returned that Melle weather station is the closest station available. The data acquired are logs based on the datetime(timeseries) , at the time-level of also a minute. The weather features that were measured and possible were:

[2] Meteostat API

- **Average Water Temperature, Minimum and Maximum Water Temperature(Celsius metric system):** The average, minimum and maximum atmospheric temperature measured using the Celsius Metric System.

- **PRCP(mm):** PRCP is the Total Precipitation indicating the total amount of liquid water equivalent of presumably all precipitation. It is at the combination of rain and liquid water equivalent of the known snow.

- **Wind Direction (Degrees):** Wind direction is the direction from which it the wind originates.

- **Wind Speed (km/h):** The wind's speed measured in km/h.
- **Air Pressure (hPa):** Air pressure is the weight of air molecules pressing down on the Earth. The highest pressure is at sea level where the density of the air molecules is the greatest.

After the data from both sources are acquired , the initial dataset that is formed , includes all fourteen features, both the water quality and the weather ones from Shcellebelle , Belgium. The data ,as they are measurements (logs) that differentiate by their unique timestamp, they create a timeseries dataset. These timestamps disclose the exact date (day, month, year) as well as the time (hour, minute).The dataset is ideal for timeseries analysis and also it can be used for the training of models with the goal of making predictions for the future.

3.2 Data preprocessing

The data preprocessing phase can be separated in four different phases. These are, data cleansing , the handle of missing values, the search for outliers and finally the transformation of the dataset to describe all the features with the date as the indication (primary key) and also the preprocessing in order to make the dataset describe hourly or daily measurements according to the design of each model.

Initially, after the data were collected from the data sources described, basic cleansing and preprocessing techniques were applied. Some of the registrations, derived from the chemical measurements, were logged using a comma (,). The comma was replaced by the dot (.). Afterwards, all values were declared as floats. Some columns are dropped and renamed too. Preprocess of the datetime column also takes place with the goal of isolating only the datetimes's part that contains its date and time. By doing these, it was achieved the uniformity of the dataset, which is crucial for the actual processing and analysis that will be followed.

Secondly, by exploring the dataset, it was found that empty values (Nan) existed. As the dataset was describing measurements taken in the timespan of minutes, it was rational to fill those values with the value existing right before. There were no big time periods where the measurement of any feature were not available, so no actual imputation techniques, such as interpolation, was necessary.

In addition, as it is known the presence of outliers can negatively affect the whole analysis, so it is crucial to find and remove them. At the fig. 2, by the scatter plot it is noticeable that no extraordinary or unusual values that they exceed the rational thresholds, were present for any feature, at least at a level that it may affect the outcome, so no measurements were taken. Some sudden fluctuations are considered to be normal.

As the initial goal was to predict the next day's values. So, the dataset at this point was preprocessed in a way that it would have daily values for each water chemical feature, not values per minute. This was achieved by taking the mean values for each day, simply group them by the date and take the average value per day for each feature. This resulted at a smaller dataset. It is worth mentioning ,that the research approach changed afterwards , as the goal altered to the prediction of the next hours values. So, a similar process was conducted, resulting in a dataset containing hourly measurements. The mean value per hour, for each feature, were calculated forming a dataset containing the requires hourly logs. So two datasets, one that contains daily and ones that contains hourly measurements are formed.

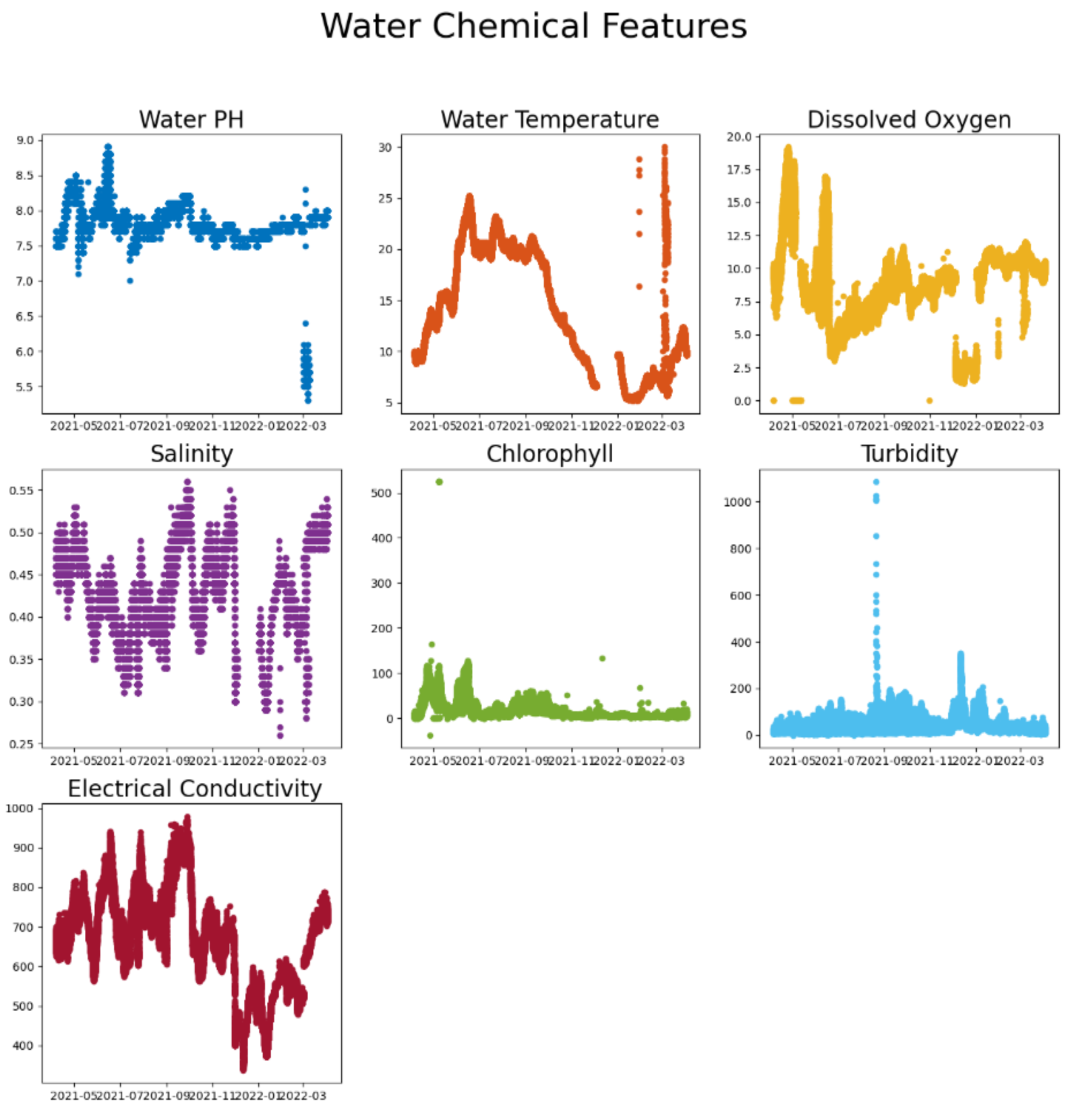


Figure 2:*Scatter Plot of Chemical Features.*

When it comes to weather data , same methods are also used. The missing values are also filled with the previous registration's value. Weather data are also measured per minute and are also referring to the exact same time-period as the water's chemical data, which is from 03/04/2021 until 03/04/2022. All features were converted to floats as well. Below , at fig.3,you can view the values of each weather feature at the field of time. As prementioned, as the approach of the time-period that will be predicted, two datasets for the weather dataset ended u to be created. One containing daily measurements and on containing hourly measurements.

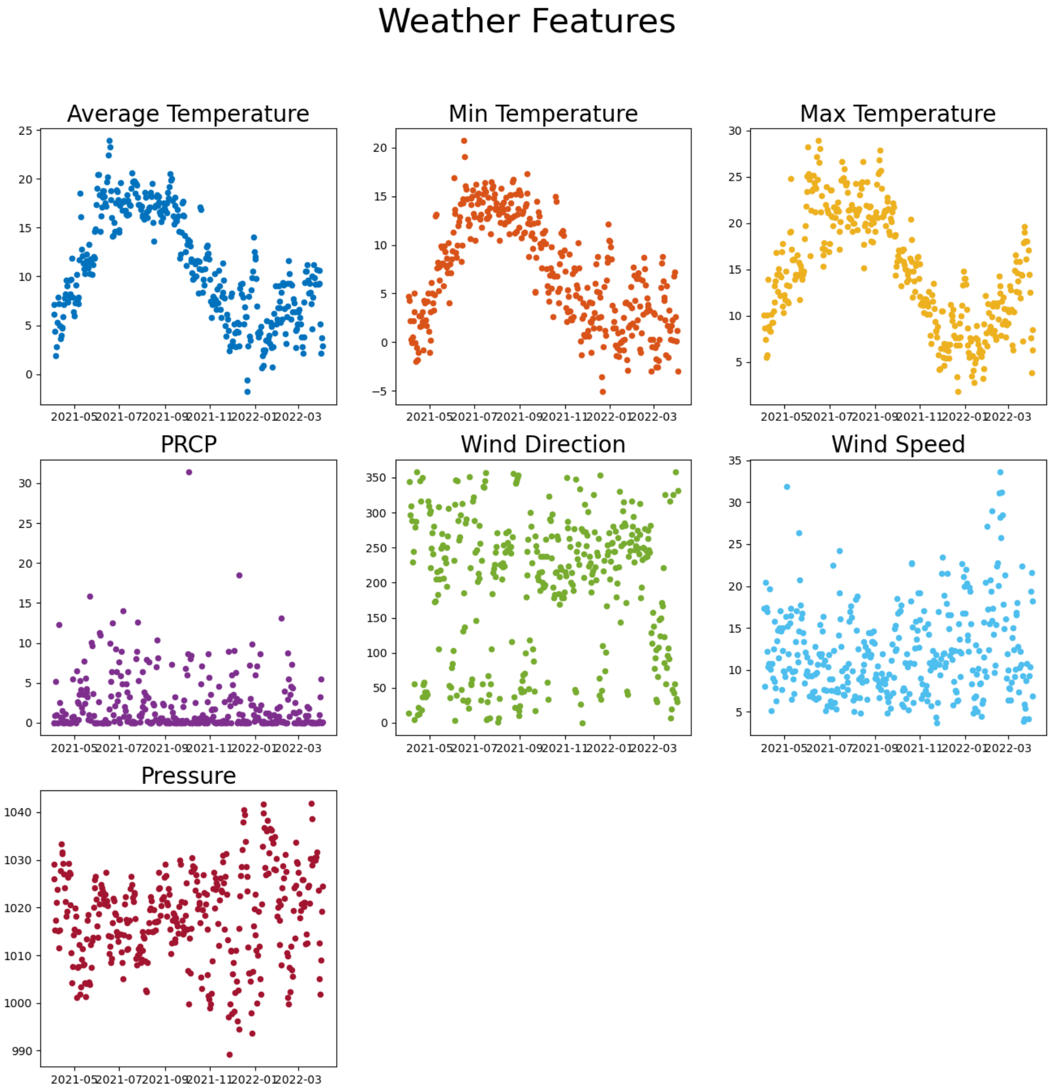


Figure 3:Weather Features Scatter plots.

Finally, the two datasets, the one that contains the water's chemical features and the one that contains the weather features were merged using the datetime column as the connecting link. This is done for both the datasets describing daily and hourly measurements. In both cases, this resulted in datasets containing timeseries, from both water's chemical features and weather features, fourteen features in total, ready to be furthermore processed.

At the *Table 1* below you can view basic descriptive analysis of the features such as the total number of registrations (Count), the feature’s mean value, feature’s standard deviation and feature’s min and max values as well as the feature’s variance.

Table 1: Descriptive analysis of the dataset’s features

	PH	Water-temperature	Turbidity	Dissolved Oxygen	Conductivity	Salinity	Chlorofyl	Aver.Temp.	Min Temp.	Max Temp.	PRCP	Wind Direction	Wind Speed	Pressure
Count	366.0	366.0	366.0	366.0	366.0	366.0	366.0	366.0	366.0	366.0	366.0	366.	366.0	366.0
Mean	7.76964	13.509840	43.858	8.3028	668.5535	0.4245	16.9474	10.97	7.10136	14.855	2.1081	193.819	12.22	1018.21
Standard Deviaton	0.3489	5.524325	24.374	3.038463	120.637	0.0518	17.60	5.506	5.49	6.1949	3.4931	100.444	5.35	9.67163
Min value	5.68298	5.218750	13.17	0.0	381.7152	0.29	3.419	-1.800	-5.1000	1.800	0.0	0.0	3.700	989.30
Variance	0.1217	30.518167	594.09	9.2322	14553.3	0.0026	309.951	30.326	30.2169	38.3771	12.201	10089.0	28.64	93.54
Max Value	8.6295	24.72	214.18	16.80972	901.343	0.527	98.162	23.90	20.7000	28.90	31.4	358.0	33.60	1041

4. Feature Selection

The feature selection phase is a crucial stage where the features, accompanying the water’s EC, for the upcoming predictive models training are selected. The reduction of dataset’s dimensions is an especially significant step with the goal of achieving better results from the machine learning and artificial intelligence models. As previously mentioned, the use case of this thesis is to predict the water’s electrical conductivity future values. So, its logical that the analysis and the feature that the final dataset would be consisted of, will strongly associate with the feature of water’s EC. Extended research has been done covering the matter of feature selection and reduction. Renown methods as PCA ,Anova or Chi-Squared test are usually used to achieve the desired dimensionality reduction.

However, in this thesis the proposed suggestion is the use of Pearson's Correlation as well as the method of Mutual Information , with the aim to conclude about the best features that can complement the water's electrical conductivity value, with the intention of training machine learning and A.I models. Both methods will be implemented and will be taken in considerations in order to aid in the decision making of the features that will form the final dataset. At first, the Pearson's correlations coefficient will be implemented ,examined and commented. Next, the Mutual Information method will also be implemented , examined, and commented. When both methods have been examined, their results and outcomes will be the determine the features that will compose the dataset that will be used afterwards.

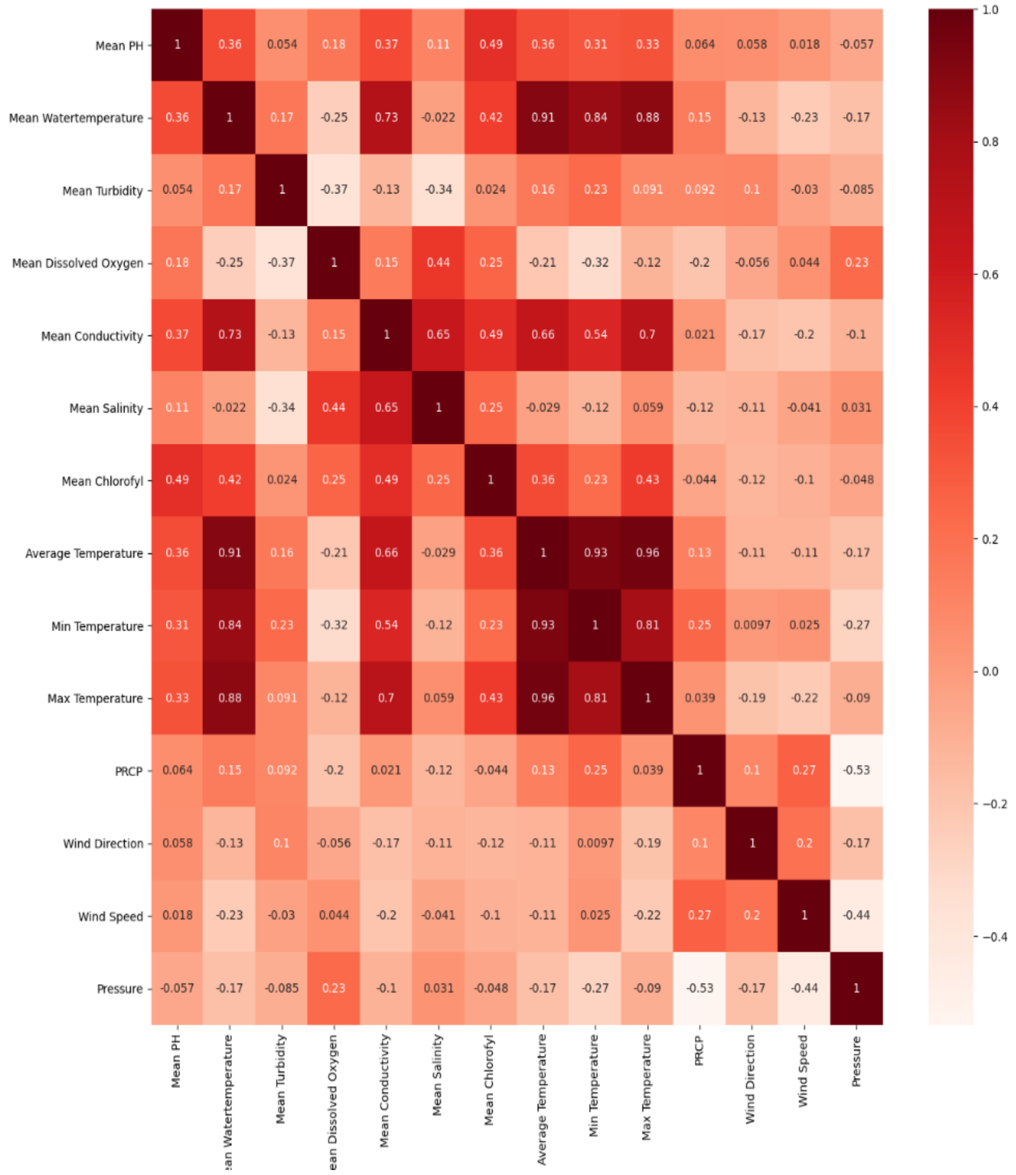
4.1 Pearson's Correlation Coefficient

Pearson's formula is a simple yet crucial formula that it is widely used to describe and perceive the correlation between the various features. A correlation coefficient of one(1) means that for every positive increase in one variable, there is also a positive increase in the other. That means that the two variables are highly positively correlated. So, the desirable correlation between two features is a number as close to one(1) as possible. From the other hand, if the correlations of the two features is minus one (-1), that means that those features are behaving counter to each other. That also showcases negative correlations, but the correlations still exist. Correlations close to minus one (-1) also are desirable as they indicate strong negative correlation. Correlations that are close to zero describe the absence of any correlation at all. Pearson's correlation coefficients, symbolized as r, as it was found at 1880s by Karl Pearson, is being calculated using the below formula, as show at fig.4.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

***Figure 4**:Pearson's Correlation Coefficient calculation formula.*

The first approach is to display the r between all features with the help of a heatmap, displayed at the fig.5, where deep red indicates high positive correlations and white indicates high negative correlations. At this point the dataset contains mean values per each feature, as it is in a format that describes daily measurements. Both the x and the y axis displaya each feature. So, each (x,y) point, which in the heatmap is displayed as a square of a specific color and a specific value, represents the correlation coefficient between the feature in the x axis and the feature in the y axis. It is expected that the correlations coefficient between the same features,where both x and y axis describe the same feature, are equal to one (1), as the correlations between a feature and the feature itself is the highest it can be as their relationship is absolute. The factor of color is added with the aim of a better visual understanding of the correlation. As the total relations that can be examined, therefore the number of squares that are being displayed in the heatmap, are in the class of the number of features that exist in the dataset multiplied with itself. As both x and y axis display the same number of features that exists in the dataset, that can be also described as x^2 or y^2 accordingly. In the dataset at this point , fourteen (14) features exist so the total number of correlations that will be calculated are a hundred-ninety-six (196) correlations coefficients. That means, that a better visual representation will come in clutch for a better and maybe an easier understanding. This is the reason that a heatmap representations is ideal for that kind of display.



	Mean PH	an Watertemperature	Mean Turbidity	ean Dissolved Oxygen	Mean Conductivity	Mean Salinity	Mean Chlorofyl	Average Temperature	Min Temperature	Max Temperature	PRCP	Wind Direction	Wind Speed	Pressure
Mean PH	1	0.36	0.054	0.18	0.37	0.11	0.49	0.36	0.31	0.33	0.064	0.058	0.018	-0.057
Mean Watertemperature	0.36	1	0.17	-0.25	0.73	-0.022	0.42	0.91	0.84	0.88	0.15	-0.13	-0.23	-0.17
Mean Turbidity	0.054	0.17	1	-0.37	-0.13	-0.34	0.024	0.16	0.23	0.091	0.092	0.1	-0.03	-0.085
Mean Dissolved Oxygen	0.18	-0.25	-0.37	1	0.15	0.44	0.25	-0.21	-0.32	-0.12	-0.2	-0.056	0.044	0.23
Mean Conductivity	0.37	0.73	-0.13	0.15	1	0.65	0.49	0.66	0.54	0.7	0.021	-0.17	-0.2	-0.1
Mean Salinity	0.11	-0.022	-0.34	0.44	0.65	1	0.25	-0.029	-0.12	0.059	-0.12	-0.11	-0.041	0.031
Mean Chlorofyl	0.49	0.42	0.024	0.25	0.49	0.25	1	0.36	0.23	0.43	-0.044	-0.12	-0.1	-0.048
Average Temperature	0.36	0.91	0.16	-0.21	0.66	-0.029	0.36	1	0.93	0.96	0.13	-0.11	-0.11	-0.17
Min Temperature	0.31	0.84	0.23	-0.32	0.54	-0.12	0.23	0.93	1	0.81	0.25	0.0097	0.025	-0.27
Max Temperature	0.33	0.88	0.091	-0.12	0.7	0.059	0.43	0.96	0.81	1	0.039	-0.19	-0.22	-0.09
PRCP	0.064	0.15	0.092	-0.2	0.021	-0.12	-0.044	0.13	0.25	0.039	1	0.1	0.27	-0.53
Wind Direction	0.058	-0.13	0.1	-0.056	-0.17	-0.11	-0.12	-0.11	0.0097	-0.19	0.1	1	0.2	-0.17
Wind Speed	0.018	-0.23	-0.03	0.044	-0.2	-0.041	-0.1	-0.11	0.025	-0.22	0.27	0.2	1	-0.44
Pressure	-0.057	-0.17	-0.085	0.23	-0.1	0.031	-0.048	-0.17	-0.27	-0.09	-0.53	-0.17	-0.44	1

Figure 5:Heatmap displaying Pearson's Correlation Coefficients between all features.

However, as this use case revolves around the water's electrical conductivity, it is needed to isolate only the correlations referring to the EC. In the below table the Pearson's correlation between the water's EC and the rest of the features of the dataset is being displayed. Previously, all the correlations between the features of the dataset are being calculated. Now it is needed to isolate and further examine only the correlations coefficients between the water's electrical conductivity and the other thirteen(13) features of the dataset. The *Table 2* describes exactly that. At the first column every and each one of the rests of the features appears and in the second column, the correlations between them and the water's electrical conductivity is appearing in a descending order. The column that contains the Pearson's correlations coefficient is named as r between Electrical Conductivity (EC) where r is the symbol of Pearson's correlations coefficient.

Table 2: Pearson's Correlation Coefficients between water's EC and the rest of the features

	r between Electrical Conductivity (EC)
Water Temperature	0.734060
Max Temperature	0.701891
Average Temperature	0.655604
Salinity	0.645287
Min Temperature	0.542095
Chlorophyll	0.493334
Mean PH	0.369380
Mean Dissolved Oxygen	0.149492
PRCP	0.021385
Pressure	-0.104593
Turbidity	-0.125157
Direction	-0.175498
Wind Speed	-0.202309

The correlations can also be displayed with a bar plot so a better visual understanding will be feasible. In the case of the histogram, the x axis will contain the rest of the features and the y axis contains the calculated correlations between them and the water's electrical conductivity. The bar plot , displayed at the fig.6 ,displays the relationship in a descending order as well.

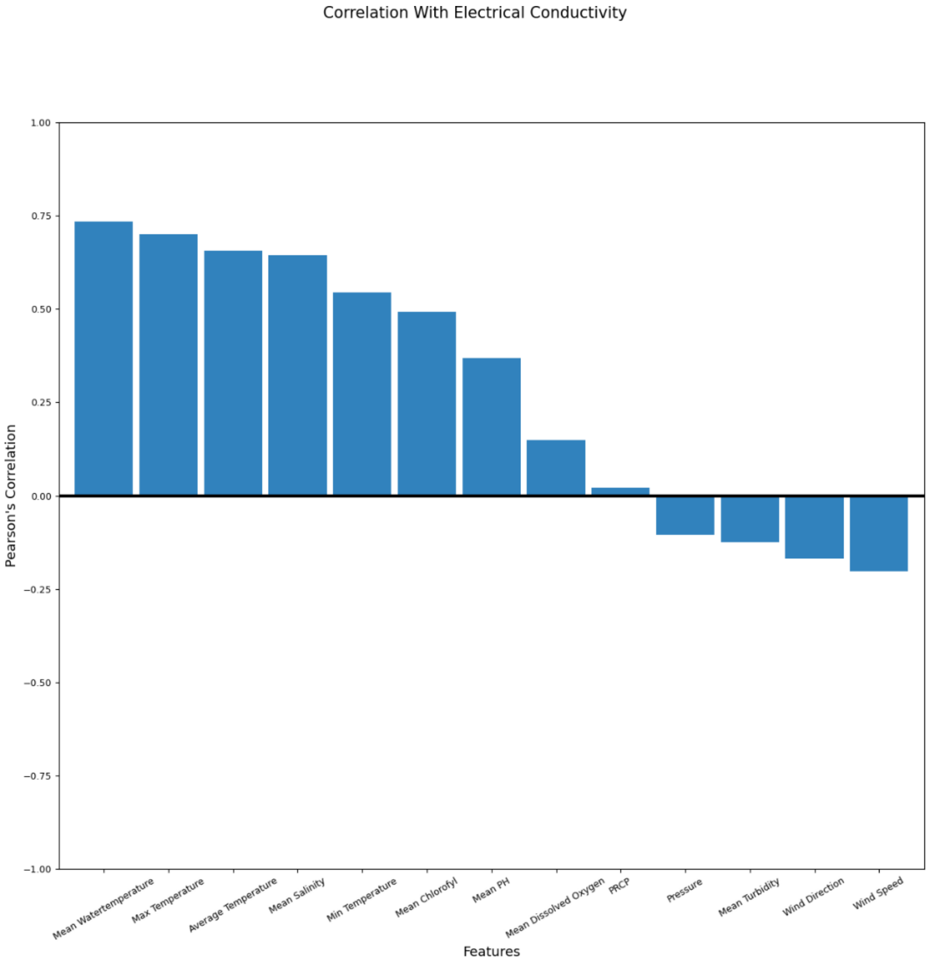


Figure 6*:Peatson's Correlation Coefficients between water's electrical conductivity and the rest of the features*

It is clear the top four features that appear the highest positive correlation with the EC ,are the water temperature, the atmospheric temperature (assuming that the average temperature , the minimum and the maximum temperature are referring to the same feature so only the average will be kept),the salinity and the chlorophyll. Negative correlations are not selected as there are not correlations negative enough that will describe a significant relationship. The rest of the correlations are close to zero and provides no important insight.

Their correlation with water's EC , displayed with a heatmap is shown below, at fig.7. The heatmap follows the same logic as previously described and contained only the top highest positive correlations found between the four features that previously referred and the water's electrical conductivity.

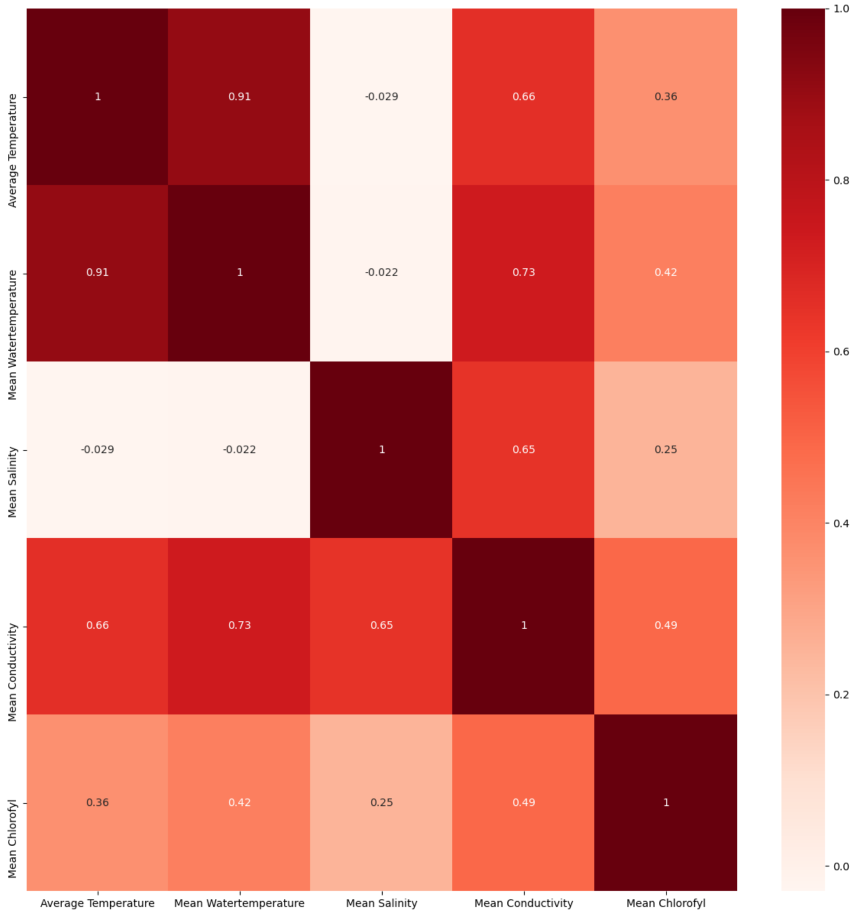


__Figure 7__:Heatmap displaying Pearson's Correlation Coefficient between water's electrical conductivity and the top four most correlated features.

According to the correlations , it will be interesting to display the actual measurements in the field of time, with line plots , fig.8 ,of the four most correlated features next to the water's EC ,so similar trends or patterns can be visually discovered .

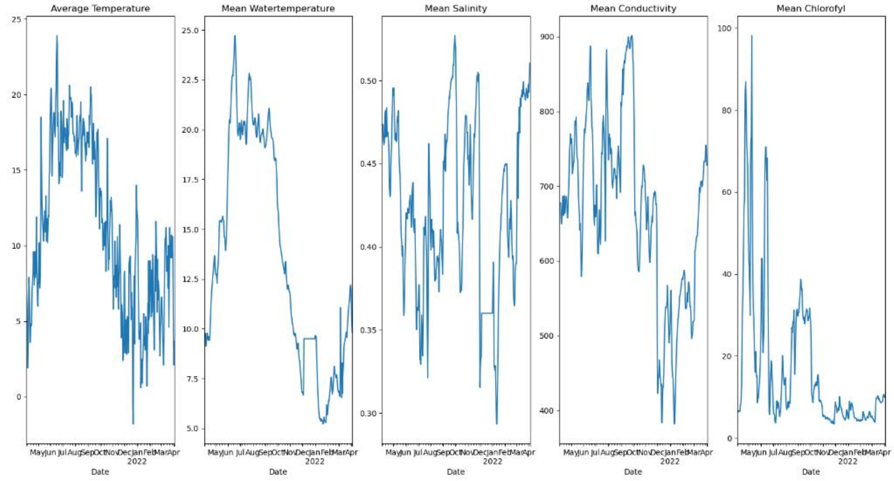


Figure 8:Line Plots of the Measurements from water's electrical conductivity and the top four most correlated features.

It is noticeable that some features to have similar seasonality patterns and follow the same trends. However, these assumptions will further be examined at the time series analysis that will be conducted afterwards.

4.1.1 Rejecting the null hypothesis to assure the correctness of Pearson's correlation coefficients

Even if the Pearson's Correlations have been calculated, the statistical theory suggests that the outcome must be supported with the corresponding p-value calculation. Explanatory, the p-value must be calculated in order to reject the null hypothesis that the correlations between the features is due to 'luck' and randomness, and they are not reflecting the truth. So , assuming that the null hypothesis that is about to be busted, suggests that the correlation is due to random causes meaning that there is no actual relationship between the two measured phenomena. The smaller the p-value the stronger the evidence against the null hypothesis, so a p-value close to zero (0) is wanted, so it can confirm the integrity of the calculated correlations coefficients.

In the *Table 3* , the Person’s correlation coefficients and the according p-value appears for each feature from the four features that are will furtherly be examined and shows appear to be the most interesting ones.

Table 3: Pearson’s Correlation Coefficients between water’s EC and the rest of the features accompanied with the corresponding P-Value

	Pearson’s Correlation	P-value
Average Temperature	0.656	0.0
Water Temperature	0.734	0.0
Salinity	0.645	0.0
Chlorophyll	0.493	0.0

As the p- values have been calculated and the result is very close to zero, as only one decimal digit was kept for the reason of display, that means that the were no random or lucky causes that resulted in the calculated correlations. The null hypothesis has successfully been rejected and the calculations ,as well as their accompanied importance of them, can be trusted and will be taken in serious consideration for the feature selection. This step is a perquisite in order to safely continue and take the correlations coefficients as a significant insight that describes the true relations among the features of the dataset.

The rejection of null hypothesis, explanatory the rejection of the possibility that the correlations that calculated, are a random outcome, makes the process of feature selection more reliable.

4.2 Mutual Information

After the calculations of correlations coefficients, the method of Mutual Information is implemented with the aim to provide more useful insights , that will help with the final features selection.

In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" (Entropy=Expected value of the surprise) obtained about one variable by observing the other variable. The concept of mutual information is intimately linked to that of entropy of a random variable, a fundamental notion in information theory that quantifies the expected "amount of information" held in a random variable.

Mutual Information is a method that reveals if there is any entropy between the two features being examined. It does not provide the information of positive or negative relationship between the features. Mutual Information, as it takes values between zero(0) and one (1), showcases the amount of entropy between two features. The closer the mutual information is to one , the higher the entropy between the features is. The Mutual Information formula is calculated by the mutual information algorithm as given by Shannon and Weaver in 1949.

$$I(X;Y)=\sum_{x,y}P_{XY}(x,y)\log P_{XY}(x,y)P_{X}(x)P_{Y}(y)=E_{P_{XY}}\log P_{XY}P_{X}P_{Y}.$$

Below, at fig.9, the Mutual Information score is calculated , between the water’s electrical conductivity and the rest of the features. The Mutual Information score is appearing as MI Score. The score is being displayed in descending order, firstly , with a histogram and secondly in a table format. The scores, have been calculated using both Mutual Information Regression algorithm from Scikit-Learn as well as the K-best algorithm using the Mutual Information method, searching for top 5 highest scores. In the x axis, of the histogram appearing below, all thirteen of the features , appear. In the y-axis, the calculated Mutual Information (entropy) between them and the water’s electrical conductivity appears. The order of appearance is descending. The *Table 4* follows, in which the actual MI calculations appear.

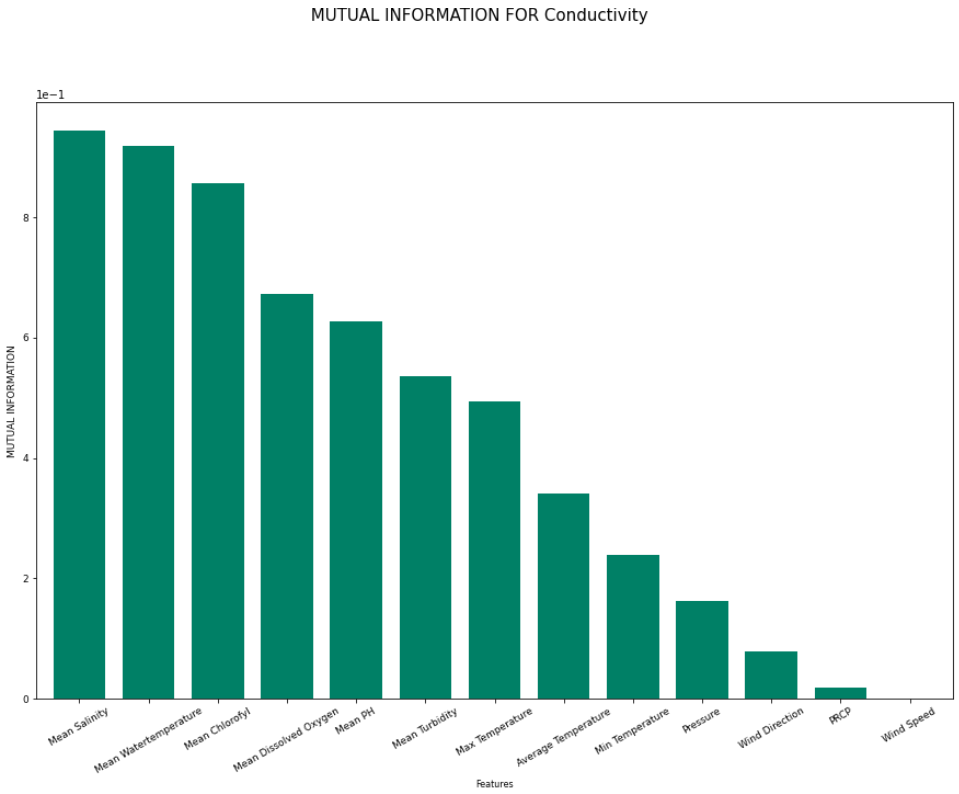


Figure 9:*Mutual Information Score between water's electrical conductivity and the rest of the features.*

Table 4: Mutual Information(MI) Scores between EC and the rest of the features

	MI Score in relation with EC
Salinity	0.943512
Watertemperature	0.920074
Chlorophyll	0.857600
Dissolved Oxygen	0.672092
PH	0.626645
Turbidity	0.536175
Temperature	0.484868
Average Temperature	0.332888
Min Temperature	0.240405
Pressure	0.171107
Wind Direction	0.109219
Wind Speed	0.017266
PRCP	0.003371

As it is noticeable, the outcome of the score is quite different from what the correlation coefficients were indicating. However, this is totally expected, as they are describing different things. Entropy ,as previously mentioned, does not describe a positive or negative relationship, it describes only that there is an actual relationship. For example, the feature of Salinity appears the highest mutual information score, outcome that is more than expected, as their close relationship, has already been studied and proven. Furthermore, both methods will be taken in considerations, with the goal of assisting in the final decision making of the dataset. Features that appear high mutual information but close to zero correlations are not ideal candidates for the dataset. It is more desirable ,features that contain both big amounts of entropy as well as a descent amount of positive or negative correlation with EC.

4.3 Final feature selection and the composure of the dataset

After the implementation and the examinations of both methods, Pearson's correlations coefficients and Mutual Information, a final decision can be made, referring to the final features that will best accompany the wanted to be forecasted feature, which is water's electrical conductivity and that they compose the dataset that will be used for the training and testing of the upcoming models. Taking both methods into considerations, the outcome is that top four features that will be selected are, atmospheric temperature , water temperature, salinity, and chlorophyll. These are the features that appeared either the highest correlations with water's electrical conductivity, either appeared large entropy with water's electrical conductivity, such as chlorophyll, followed of course by a descent amount of correlation coefficient, either appeared both , fact that indicates a very strong relation between the features such us salinity. In the case of features that may appeared high entropy but a mediocre amount of correlation , the reason that they are selected as a crucial feature, derives from the importance that both feature selections methods have and the fact that the methods can complement each other.

Explanatory, a feature that may appear large amount of entropy and it appears a medium amount of correlation coefficient, positive or negative, for example a r=0.5 with a feature, shows that even if positive correlation is not very high, but the feature appears to have a large entropy with the other feature, that may indicate that the correlation coefficient affects the features in a larger and a more significant scale, when their entropy is high enough. Both methods played their important role at the feature selection procedure.

All in all, both methods, Pearson’s correlation coefficient and Mutual Information, between the water’s electrical conductivity and the rest of the features, were taken in consideration and the insights that they provided leaded to the selection of atmospheric temperature , water temperature, salinity, and chlorophyll as the complementary features that will compose the dataset, accompanied by the water’s EC itself , in the form of timeseries as they describe measurements in time, with the goal of using them to try and predict the future values of water’s electrical conductivity.

5. Focus on water’s electrical conductivity (EC) and forecast it using ARIMA models.

Before the use of multivariate artificial intelligence multivariate models that will use the selected features ,as they were selected by the feature selection processes, to try to predict the water’s electrical conductivity, it will be quite interesting to focus more on the desirable feature itself. The goal at this point is trying to achieve the forecasting of future values of water’s electrical conductivity, using only past measurements for the feature itself. In statistics, such univariate models do exist and are significantly helpful to make predictions at this kind, using only past measurements of a timeseries and predict future ones. Such models are ARIMA (Autoregressive Integrated Moving Average), Linear Regression, ETS, and Prophet. At this thesis, ARIMA models are used , as they probably are the more commonly used and yet most efficient methods out there.

However , before the actual use of those statistical models, furthers research about the feature of water’s electrical conductivity must be conducted . Characteristics of the feature such as the seasonality patters or the trends it follows , are needed in order to determine the extra parameters that will compile the models. So, further search for trends and seasonality patterns of the water’s electrical conductivity feature will be followed. Afterwards, it will be attempted to predict the feature’s future values , by using only the feature’s past recorded values. This will be approached by the use statistical models such as ARIMA models, and variations of them.

To begin with, below, at fig. 10 , only the logged water’s electrical conductivity values will be displayed with a line plot. In the x axis, the dates are displayed, with the form of months. Jan 2022 is where the year is changed, so it easily observable that the timeseries refers to a year’s measurements. In the y axis the actual measurements of the water’s electrical conductivity are displayed , so the line-plot displays the measurements of water’s electrical conductivity in the span of the time.

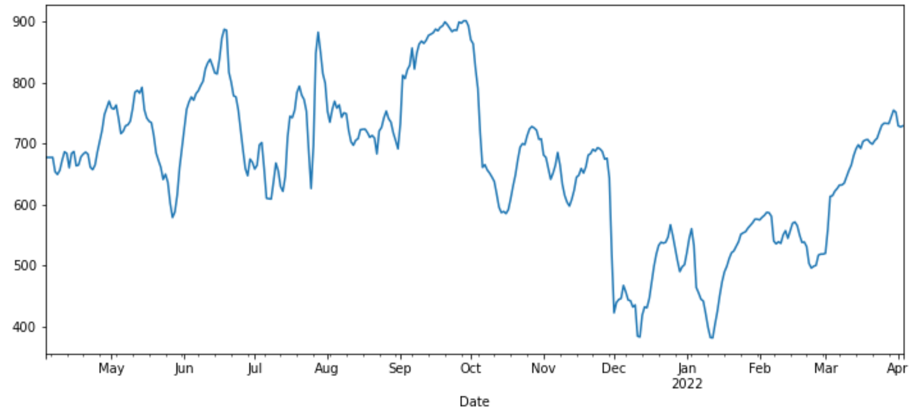


Figure 10:*Water's electrical conductivity values through the timespan of a year.*

5.1 Discovering the seasonality trends of water's electrical conductivity (EC)

Water's electrical conductivity, being a feature high correlated with features describing temperatures, either atmospheric or water, features with recognized seasonality trends, makes the water's EC a feature expected, with very high possibilities , to appear seasonality trends throughout time, similar to those that nature's temperatures appear. However, this hypothesis must be proven, and statistical methods and tests will be implemented to do so.

To begin with, by running a seasonal decompose algorithm, the seasonality, the trend, and the residual of the water's electrical conductivity, will appear. Below the plots at fig .11, describe all three of those characteristics in additions to the plot that displays the water's EC measurements throughout the year (the same plot that was displayed before at fig.10) . So, lets describe the subplot appearing to the left. The first plot of the subplot is the water's EC actual measurements. It is important to clarify, that the plot displays the values of mean electrical conductivity as the feature has been process in order to describe one measurement per day, for a year.

Following, the plot under that, it is the plot that describes the trend of the water's electrical conductivity measurements throughout the year. Below that, it is the seasonality of the feature, that it clearly appears periodic behavior and the last plot is the residual. The green plot at the right is the seasonality, as it has been enlarged and isolated to be displayed by its own, as it is the characteristic that will further be examined.

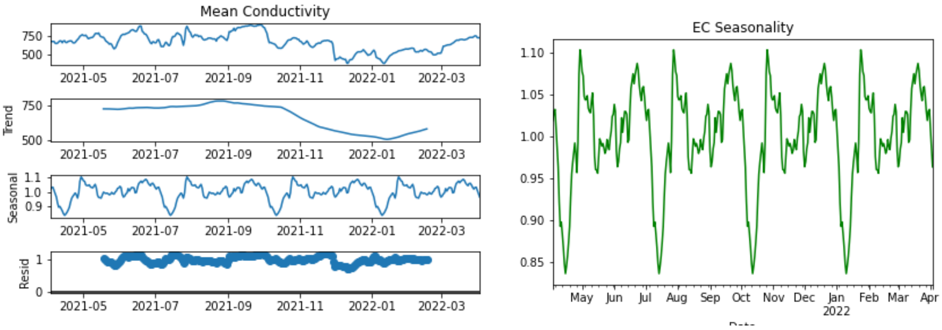
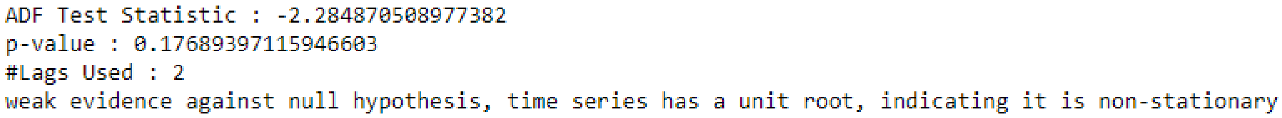


***Figure 11**:Seasonal Decompose of water's electrical conductivity.*

5.2 Run Augmented Dickey–Fuller test and make the feature stationary using moving averages

As the model have proven seasonality patters, the next step is to practically examine if it is stationary or not. Traditional ARIMA model, does not work with non-stationary data. To conclude if the feature is stationary or not, an Ad-fuller test will be run, with the null hypothesis that the feature is non-stationary. The test sets a p-value smaller than 0.05 in order for the null hypothesis to be rejected.

The results of the test are being displayed below at fig.12. The first ADF Test Statistic value is the critical value of the test. The p-value is the p- value calculated with the formula that MacKinnon recommended. Lags used is the actual number of lags.



```
ADF Test Statistic : -2.284870508977382
p-value : 0.17689397115946603
#Lags Used : 2
weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary
```

***Figure 12**:Augmented Dickey–Fuller test*

As the p -value is well over 0.05 so that means that the feature is indeed non-stationary. The feature must be in a stationary form, in order to be fitted at an traditional ARIMA model. There are plenty methods of making a timeseries stationary, viz remove its seasonality . Such are methods as log transforming of the data, taking the square root of the data, taking the cube root or proportional change. The moving averages, a renown soothing method will be used in this case. The results will be displayed below at fig.13. The blue line plot is the feature after the moving averages method has applied and the red line plot are the original feature's values.

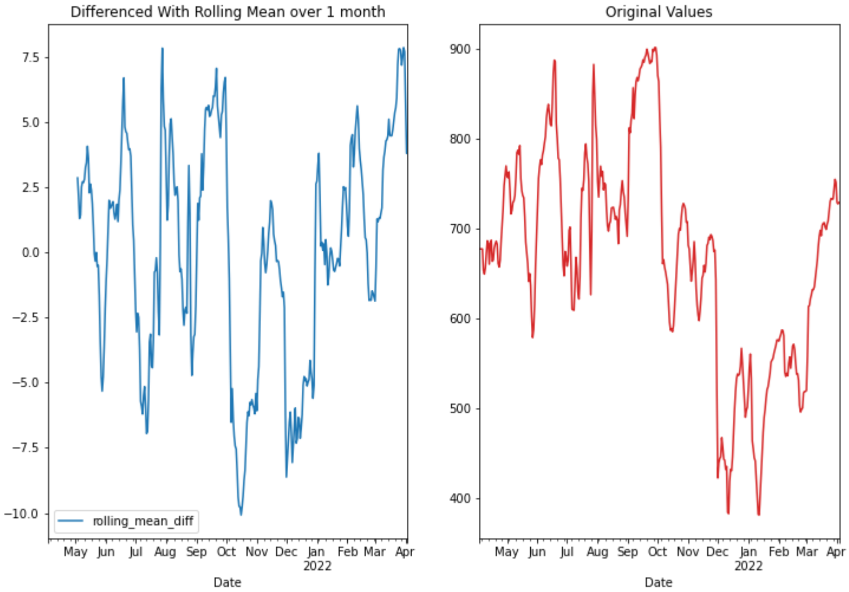
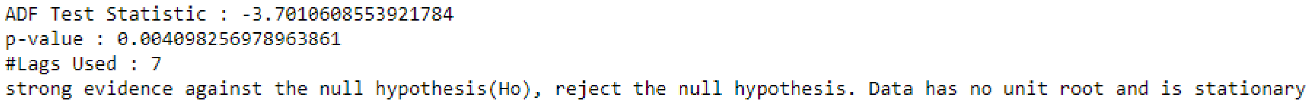


Figure 13*: Water's electrical conductivity values after and before the moving aveages application.*

Now that the method is applied, seasonality must have been removed making the feature stationary. Augmented Dickey–Fuller will run once more , with the same null hypothesis. The results can be viewed at fig.14.



```
ADF Test Statistic : -3.7010608553921784
p-value : 0.004098256978963861
#Lags Used : 7
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

Figure 14: *Augmented Dickey–Fuller test after the moving averages was implemented.*

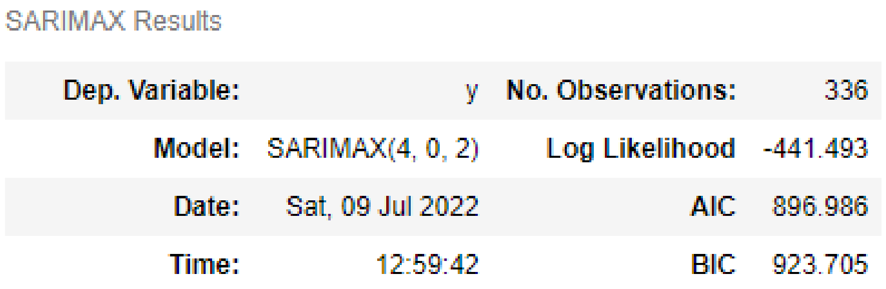
Indeed, the seasonality of the feature has been removed and it now stationary, therefore ready to be used as input at an ARIMA model.

5.3 Using Auto-ARIMA to select p,d,q parameters and train ,test and evaluate ARIMA Models

Now that the feature of electrical conductivity has no seasonality making it stationary, is ready to be used as an input to a conventional ARIMA model. ARIMA models are widely used to forecast timeseries . They do not require the existence of other features to make a prediction. The models train themselves by past values of a timeseries.

However, the traditional ARIMA model requires three parameters, to be set. These parameters are p, d and q where p is the number of autoregressive terms, d is the number of nonseasonal differences is the number of lagged forecast errors in the prediction equation. Statistical approaches of finding these parameters have extensively been proposed, such as the examination of autocorrelation function (ACF) and partial autocorrelation (PACF) plots. In this thesis, the auto Arima algorithm will be used to determine the best combination of these parameters. The *Table 5* below describes the output of the Auto-Arima algorithm. The Model is the model with its (p,d,q) proposed parameters defined. The Data and the Time are the timestamp that the algorithm executed. The No.Observations (Number of Observations) are the number of usable , non-null values contained in the dataset. As the moving averages method was applied to the dataset with a time window of 30, that means that the first 30 values were used to calculate the 31st one , so the 30 first values are null. Given that the initial timeseries measurements were 366 in total then 336 remain. The Log Likelihood is self-explanatory and finally, the AIC and BIC Scores are the scores the algorithm uses.

Table 5: Results of Auto-Arima after the Moving Averages Smoothing Method was applied



SARIMAX Results

Dep. Variable:	y	**No. Observations:**	336
Model:	SARIMAX(4, 0, 2)	**Log Likelihood**	-441.493
Date:	Sat, 09 Jul 2022	**AIC**	896.986
Time:	12:59:42	**BIC**	923.705

The fact that it the Auto-Arima model proposes a SARIMAX model will not be taken in consideration, yet ,as a simple ARIMA will be used. Only the proposed (p,d,q) parameters will be used. Specifically, the parameters best chosen by the Auto-Arima are p=4, d=0 as it is non seasonal and expected, and q=2.

Since the parameters have been chosen, now a ARIMA model , with the above parameters will be trained , tested, and evaluated.

The split between training and testing data wants the firsts 186 values from the total of 336 records (days) of measurements existing in the dataset, to be used for training and the rest of 150(last), to be used for testing. Considering that the dataset initially contained 366 measurements ,as it is referring to a whole year, and moving averages has been applied to it, with a window of 30 measurements which is the timespan of a month, it is no wonder that the data logs that initially were selected three-hundred-sixty-six(366) has now dropped down to three-hundred-thirty-six (336). The analogy of training and testing data, as well as the methods in use for the selection of the dataset can vary and tuned accordingly.

After the training and testing, the model is evaluated. The model performed poorly and totally failed to follow the actual original values whatsoever. Below ,at fig.15,the blue line expresses the original values of the stationary testing dataset , and the orange line expresses the values that the model predicted in the testing phase.

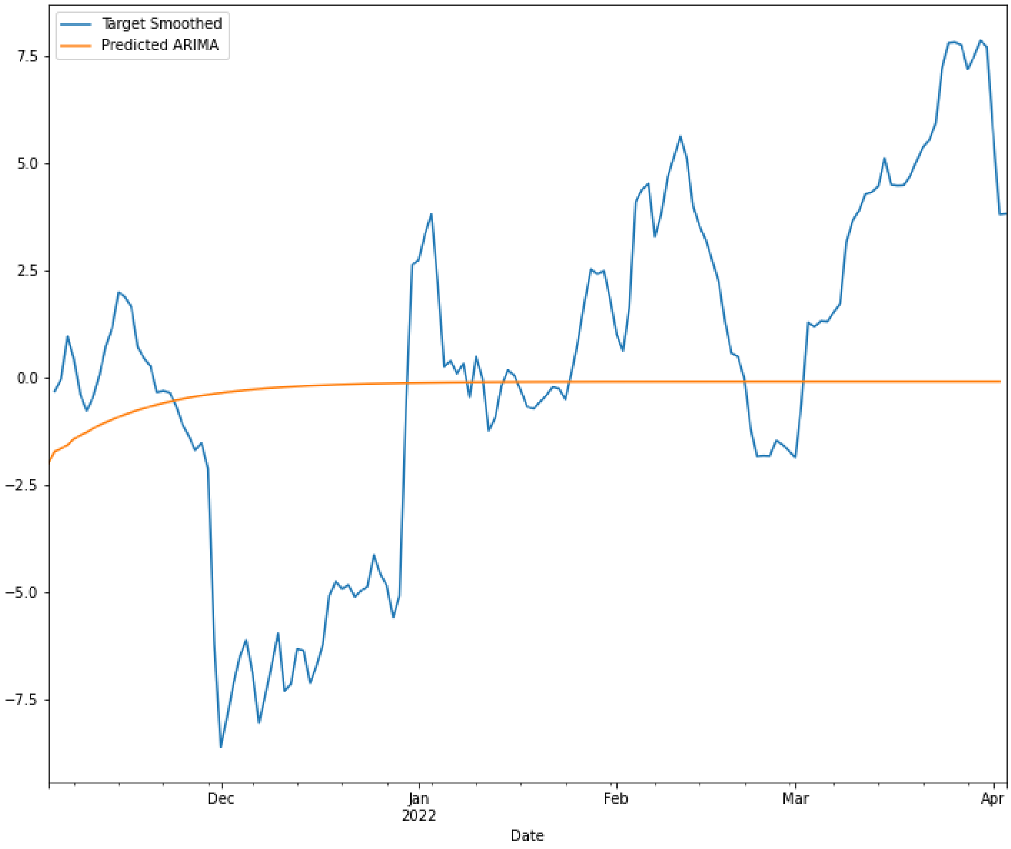
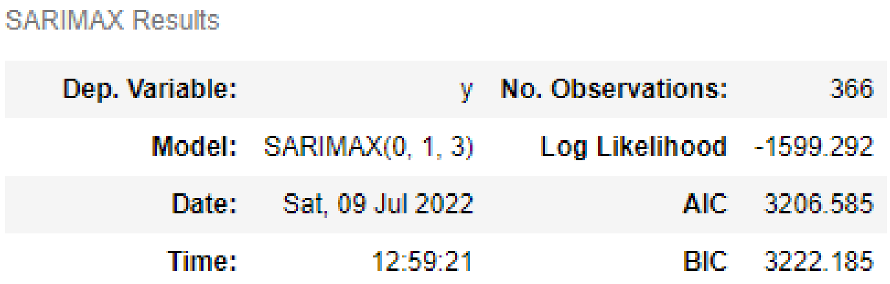


***Figure 15**: Testing-phase evaluation of the ARIMA (4,0,2) model.*

Since the model, ARIMA(4,0,2) performed so poorly, a different approach is taken. An approach where seasonality will not be smoothed away but it will be used as an asset, will take place. A non-traditional SARIMA model will be used , trained, and tested with the original non-stationary values of water's electrical conductivity.

Once again, the parameters will best be determined by the Auto- Arima algorithm. Below the *Table 6* describes the output of the Auto-Arima algorithm with the initial non-stationary dataset as an input. It is noticeable that the number of observations is the same as the original 366 values at the non-stationary dataset.

Table 6: Results of Auto-Arima with the non-stationary data



SARIMAX Results

Dep. Variable:	y	**No. Observations:**	366
Model:	SARIMAX(0, 1, 3)	**Log Likelihood**	-1599.292
Date:	Sat, 09 Jul 2022	**AIC**	3206.585
Time:	12:59:21	**BIC**	3222.185

A SARIMAX (0,1,3) model is proposed , and this will be applied subsequently. However, an additional parameter, the one of the seasonality, is required. This will be set as 90, the period of days that the feature may be appear seasonality patterns, as after trial and error this was performing the best. The splitting of the data to training and testing was the same as before with the difference that the testing dataset has now all the 216 timeseries available for training and the rest of 150 for testing.

The model can be evaluated for its performance at the testing phase. As it can viewed at the fig.16, , the blue line describes the original water's electrical conductivity measurements, at the testing phase and the orange line represents the water's electrical conductivity values as they are predicted by the mode. The model clearly caught the trends of the values, and it makes close predictions, but it tends to predict higher values that the true ones. The model of course can be tuned better , and with the existence of a larger training dataset better results may be possible. However , for this point the performance is acceptable and no further tuning will take place.

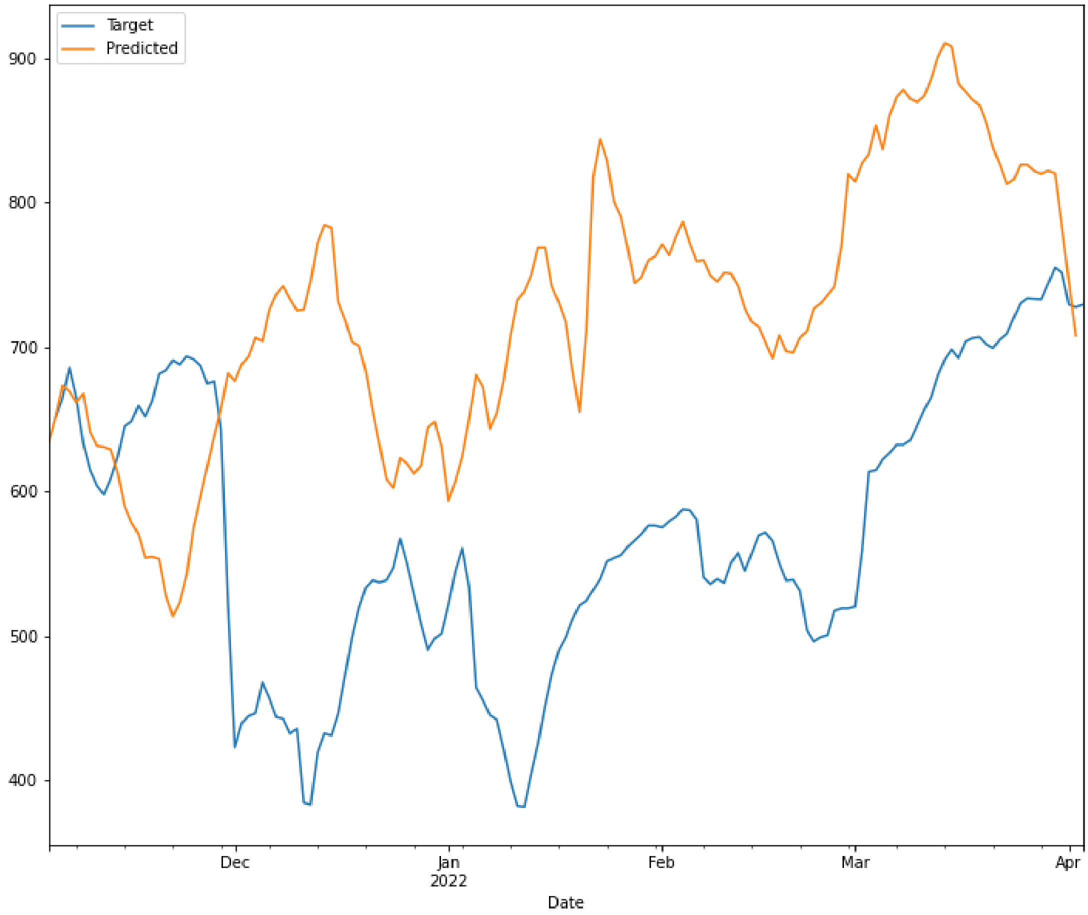


Figure 16: *Testing-phase evaluation of the SARIMAX(0,1,3,90) model.*

The model seems to have followed the actual trends much better from the simple ARIMA model that implemented previously. However , when its performance scores are calculated, the coefficient of determination (R^2) and Mean Squared Error (MSE), it is was found that the performance was far from ideal. As the R^2 is negative, -3.2, a value that is out of possible value range possible for R^2, it is clear that the model appears to follow the opposite trend, than the actual water's electrical conductivity value, something that is not clear from the plot that appears at fig.16. At Table 7 the actual scores are being displayed.

Table 7: R^2 and MSE for SARIMAX(0,1,3,90)

Model	R^2 Score	MSE Score
SARIMAX(0,1,3,90)	-3.2000	39243.490

As prementioned, the model can of course be tuned in a better way so it can be more reliable. Nevertheless, the already trained model will be used for this point .The model will be used to forecast future unknown values of water's electrical conductivity for the upcoming 5 months. The creation of the labels needed to describe the dated of the predictions, specifically the labels for the dates from 2022-04-04 until 2022-09-04 are made. At the following line plot,fig.17, the actual original values of the water's electrical conductivity, for the period from the April of 2021 up until the last existing value in the dataset which is at the 03/04/2022, are being displayed with a blue line. With the orange line, the values that the SARIMAX model forecasted for the future values, for the next five months, are being displayed. As noticed, the values do not seem extraordinary or that they deviate in a non-acceptable way.

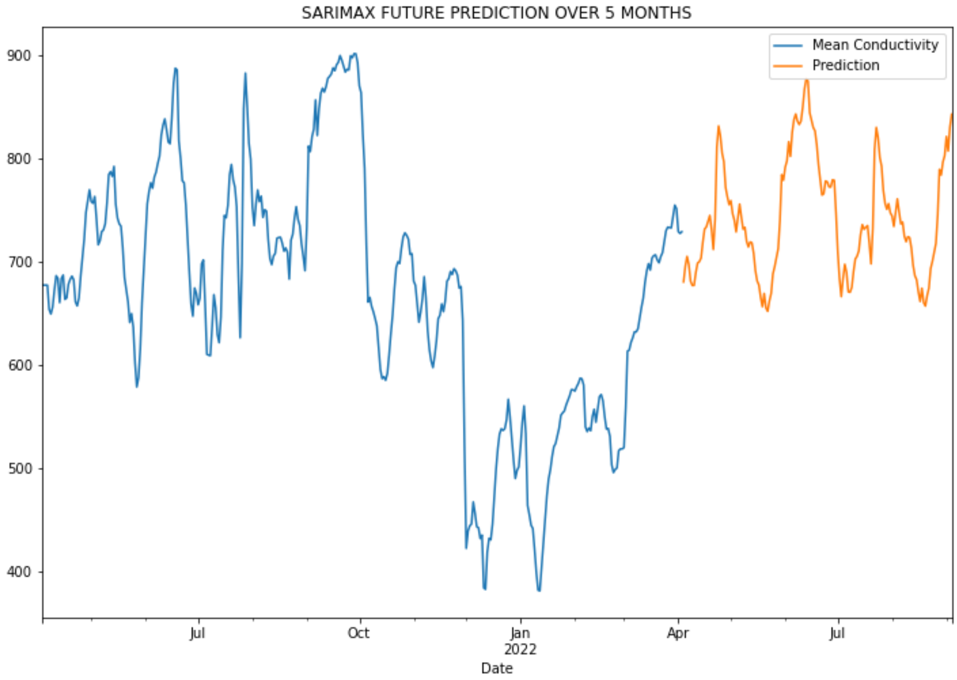


__Figure 17__:Forecasting of the water’s electrical conductivity values for the next five months.

6. Predict water’s electrical conductivity using artificial intelligence models

In the next step, the approach of using artificial intelligence models with the purpose to predict the future water’s conductivity values is made. At this part of the thesis, the goal is the prediction for a very shorter time-period and not predictions over lots of days or months. Specifically, the following approaches, aim to make predictions for the next hours and maximum for the next day. In order to do so, several approaches were made. Approaches such as a univariate Multi-Layer-Perceptron (MLP), as well as multivariate models that take advantage of the complementary features selected in the feature selection stage. The multivariate models that were used , were Long-Short-Term (LSTM) Memory Recurrent Neural Network models, with two different design approaches. All models were evaluated and used to forecast future water electrical conductivity values.

It is important to specify the fact that the dataset was preprocessed again so it can be modified and fit the purpose. The measurements, for all five features that were selected, including the water's electrical conductivity itself, at this phase the datasets describes hourly measurements, for the same time period as previously mentioned. Explanatory, the dataset contains timeseries of the features selected at the feature selection phase ,as well as the water's EC, hourly measurements , referring to same dates as previously, meaning it contains hourly logs for a whole year. This was possible, as the initial data were measurements per five minutes, for each day. The measurement per hour is the average value from the measurements taken each five minutes, referring to the same hour and date as previously explained.

At first, the prediction of water's EC is made by using a univariate Multi-Layer Perceptron(MLP). For this model implementation, the dataset contained only the water's electrical conductivity values, but transformed in a supervised dataset. Afterwards , in the multivariate approaches that are implemented, the complementary features selected at the feature selection phase are exploited , with the goal to predict the water's electrical conductivity values for the next hours. This approach was made using Long- Short-Term Memory (LSTM) Recurrent Neural Network (RNN)models. The first model was tuned so it would predict the water's electrical conductivity value based on the other four features values that were measured twenty-four hours (24h) ago. The second LSTM model is designed to predict the next water's electrical conductivity given the values of the other four features. However, as this approach does not leave any measurements for forecasting, the values of the other four features were also forecasted using individual Multi-Layer-Perceptron models for each feature.

6.1 Predict water's electrical conductivity using a univariate Multi-Layer-Perceptron

6.1.1 Epigrammatic Reference to Multi-Layer-Perceptron models

To begin with, it is known that the Multi-Layer-Perceptron neural network, is a simple yet very efficient feed forward model that can be used for both classification and prediction use cases, given that the use case refers to supervised learning and the right transformation of the input data has been done. Multi-Layer-Perceptron consists of three types of layers. Them will be, the input layer, the hidden layer, and the output layer. The input layer receives the input data that will be used .This , requires the ideal design and shape of the input, as well as the dataset itself, so they can best serve the design and the use case of the problem.

Afterwards , usually several hidden layers that are placed in between the input and output layer is where the computations of the model take place. As it is a feed-forward model, the data flows to the output layer. The output layer should be designed also to serve the type of use case that the model will be used to. Different use cases call for different designs of the output level . The main and most common use cases for Multi- Layer -Perceptron models ,are classifications ,predictions and pattern recognitions. The neurons in the Multi-Layer-Perceptron are trained with the back propagation learning algorithm. At Fig.18, a hypothetical MLP design is displayed. The number of the input neurons , the hidden layers as well as the number of neuros consisting them ,and also the number of output neurons can be tuned according to the use case scenario.

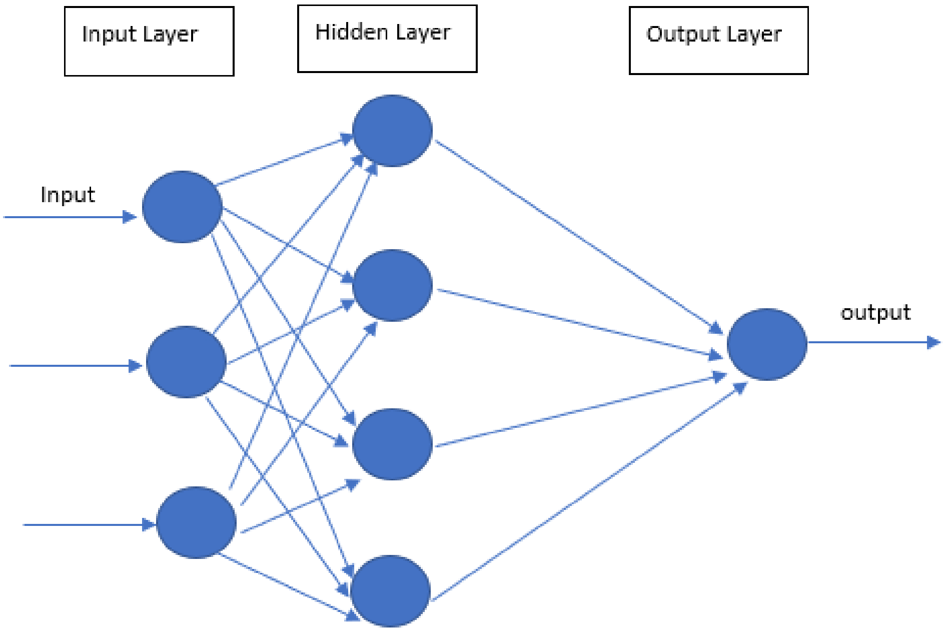


Figure 18:A hypothetical MLP Design

When it comes to the computations that takes place in the neurons from which the weights of each neuron are being adjusted as well as the way the back propagation algorithm, minimizes the loss function and readjusts the weights has already been extensively described by publicized scientific research some of which can be found in the referenced work section*[18]: A. Najah, A. El-Shafie, O. A. Karim & Amr H. El-Shafie : Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring (2014)*.,so the actual calculations and math that take place in the MLP will be not be further analyzed no special tuning or something that will differentiate them from the standard artificial intelligence models takes place.

However, basic concepts as weight calculations will epigrammatically be mentioned below.

$$\mathrm{o(x) = G(b(2) + W(2)h(x))} \qquad (1)$$

$$\mathrm{h(x) = \Phi(x) = s(b(1) + W(1)x)} \qquad (2)$$

Where the bias vectors are b(1) and b(2) , the weight matrices are W(1) and W(2) and the activation functions are G and s. The set of parameters to learn is the set θ = {W(1), b(1), W(2), b(2)}.

$$u(\mathbf{x}) = \sum_{i=1}^{n} w_i x_i .$$

The input features are passed on to an input function *u*, which computes the weighted sum of the input features.

6.1.2 The implementation of the univariate Multi-Layer-Perceptron model used for water's electrical conductivity prediction.

In this thesis, the first approach of a neural network was made using a simple but yet renown type of neural network, the Multi-Layer-Perceptron. The use case wants the model to be univariate and to predict the water's electrical conductivity future values, for the next hours. However, the model uses only measurements from the water's EC itself and no other feature from the other four selected is yet to be included.

The dataset that will be the input to the MLP , contains only the feature of water's electrical conductivity, measured by hour for a year, but transformed appropriately to be the input of a the MLP neural network. The transformation of the unsupervised timeseries formation to a supervised dataset is crucial to happen appropriately so it can be inputted to a MLP neural network design.

In addition, before the actual shape transformation takes place, the dataset is standardized. It is known that the standardization of the dataset before it enters a neural network it is required for the achievement of better results, as it provides stability and increases calculation speed. In this case, the Min Max scaler was used for the data standardization. The Min Max scaler is a well-known standardization algorithm that is being commonly used for the standardization process. The values of dataset ,once being processed by the algorithm, will be converted so they are values only between the close space of 0 and 1 , including zero and one [0,1]. Below, at the fig.19, the formula that expresses the Min Max Scaler is being displayed. The scaler will then be used to inverse the transformation after the actual predictions will made, so the true values of the predictions will be possible.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 19:Min Max Scaler calculation formula.

As previously mentioned, the dataset that contains the hourly measurements of water's electrical conductivity, making it essentially an unsupervised timeseries, must be reshaped so it meets the supervised criteria that the MLP sets. Explanatory, it should be transformed in supervised dataset, that contains time steps describing the values of water's electrical conductivity. The steps that were used in this use case are fifteen (15). That means that each value of water's EC will be predicted based on its past fifteen values. The number fifteen was selected after trial-and-error experimentation. Considering that the initial hourly dataset contains 8665 records, the supervised converted dataset with a step of fifteen will contain 8665-15=8650 records, as the first fifteen will be used to form the first entry.

So, a dataset that contains the values of water's conductivity, is standardized and reformed so that each entry record contains values in the form of water's electrical conductivity (t-15) until the value of water's electrical conductivity(t), where t starts from 15 and ends to 8664 and increases by one each time, forming a supervised dataset that contains 8650 total entries.

As, the dataset at this point is standardized and supervised, it is time to split it in training and testing. Validation phase was not used in this case. Probably the most used dataset splitting technique was used ,the 70%-30% split. This technique indicates that the 70% of the dataset will be used for training and the rest of 30% will be used for testing. As the data sample that is possible is 8650 total entries, the training will contain 6055 entries and the rest 2595 entries will be used for testing. It is important to mention that the dataset containing the target value, which is the water's electrical conductivity value after t meaning the t+1 value for each record, will also contain 6055 values for training and 2595 for testing. However, it is also important to mention that the testing dataset will contain the value of the very last record used at the training dataset to complete its first fifteen steps. That means that the very last record of the dataset is not being used ,yet. It will be used ,at the forecasting phase as it is needed to have a value where the model has not been training or tested with, before. The last value of the initial dataset , along with the other fourteen last values, therefore the last fifteen values of the supervised dataset will be used for the upcoming forecast. The shape of the training dataset is (6055,15) as they are 6055 total entries and each entry is consisting of fifteen individual steps and also the shape of target dataset is (6055) or (6055,1). Similar, the shape of the testing dataset is (2595,15) and the target training dataset is (2595) or (2585,1).

Now that the dataset has been converted to supervised the model can be designed and compiled. A TensorFlow sequential model was used . The input layer will be consisted of fifteen neurons hosting the fifteen (15) steps that each record contains. The hidden layers are consisted of two (2) layers containing fifty (50)neurons each and finally the output layer contains one neuron that that will output the predicted value of water's electrical conductivity. It is noticeable that the Dropout factor with a ratio of 20% is added in order to prevent overfitting from happening. As for the activation function, Relu is chosen. For the loss function, the Mean Squared Error function is chosen. Last but not least , as optimizer the Adam optimizer is used. It is also important to declare the batch size which is one (1) as in each epoch only 1 batch of records, containing the record itself will enter the input layer. The epochs are ten (10). All the prementioned parameters mentioned , except from the output layer design, were tuned and changed several times until an acceptable performance were reached.

After the model was designed ,compiled and trained it is time to test it. After the testing phase the predictions, that the model made , are inverted from the standardization process, using the same Min Max scaler that it was used to standardize them ,so they can describe reality. After that, both the original and the predicted water's electrical conductivity values were put side by side, specified by the datetime and then the actual evaluations were made.

For the evaluation of the model, considering it is a regression model, measures such as Mean Squared Error, the Mean Absolute Error and most importantly the coefficient of determination ,the R squared (R^2)‚were calculated. The calculations of each formula are being displayed at the fig.20.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

$$\text{MAE} = \frac{|(y_i - y_p)|}{n}$$

y_i = actual value
y_p= predicted value
n = number of observations/rows

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

Where,
$\hat{y}\ -predicted\ value\ of\ y$
$\bar{y} - mean\ value\ of\ y$

***Figure 20**: Calculation of Mean Squared Error, Mean Absolute Error and R^2*

The calculations of the above formulas , which they express the model's performance evaluation , are displayed at *Table 8.* The most important measurement is the determination coefficient (R^2) which is almost ideal as it is very close to one (1). Being very high at 0.973, that makes the model trustworthy to continue with and forecast the future value of the next hour.

Table 8: MSE,MAE and R^2 of the univariate MLP model after testing.

Evaluation Method	**Score**
Mean Squared Error (MSE)	214.329
Mean Absolute Error (MAE)	[12.262]
Coefficient of Determination (R^2)	0.973

The performance of the model at the testing phase can be viewed by fig.21 where a plot containing both the original and the predicted by the model values are displayed. The green plot expresses the actual values of water's electrical conductivity, while the red plot describes the values of water's electrical conductivity that were predicted by the Multi-Layer-Perceptron model. As it is noticeable, the model did a great job ,as the lines are almost perfectly touched.

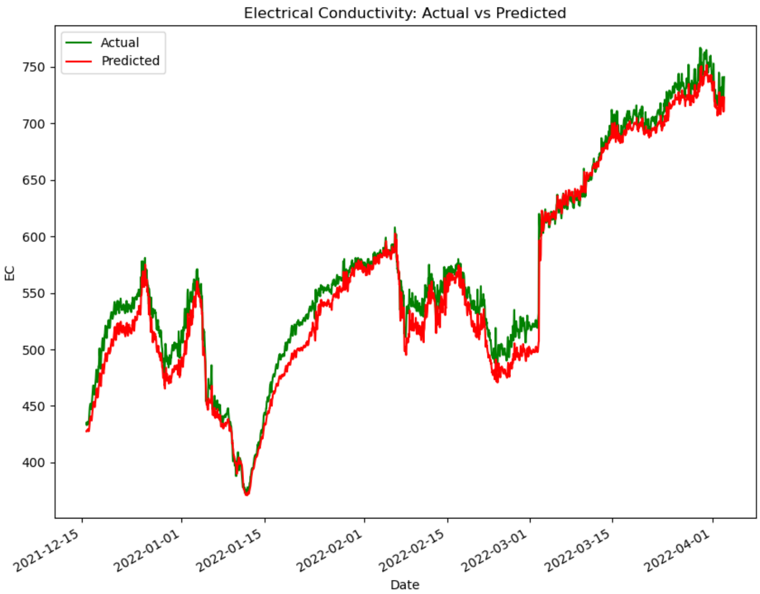


Figure 21: *Multi-Layer Perceptron Testing Phase Performance with line plots.*

After the testing and evaluation phases it is time to make the actual future prediction. As previously explained, the very last value of the initial dataset of the water's electrical conductivity measurements is not being used in the supervised testing dataset. This value, as well as the other fourteen last values of the supervised dataset, will be form one unknown step consisting of the desirable fifteen entries that will be used for the forecasting of the next water's electrical conductivity value. So, the step will also be standardized with the same Min Max scaler and will be inputted at the model so it can make one prediction. Now, it is important to clarify the ability of creating some sort of a loop where the forecasted value will be used for the next forecast and this will go on and on, in order to forecast more values. This technique will be used afterwards but considering that the more the input step is consisting of forecasted values the less likely is the predictions to be trustworthy, the prediction was limit only to one,the next hour's, as the possible real values can be form only one input step. After the prediction was made, the predicted value was also inversed from standardization so it can express the actual water's EC value.

The existing dataset contains measurements up to the specific datetime of 2022-04-03 00:00:00, so the model will predict the value of water's electrical conductivity at 2022-04-03 01:00:00. The water's EC value that was predicted is 727.74646.

At fig.22 the forecasted value ,it can be displayed as a blue dot, where the red line are the actual previous values of water's EC.

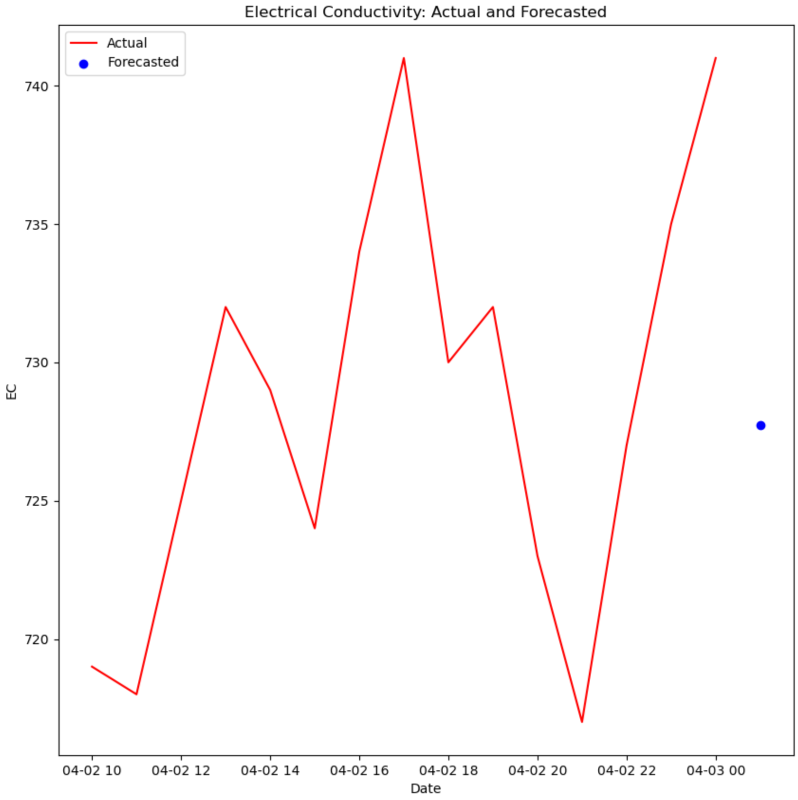


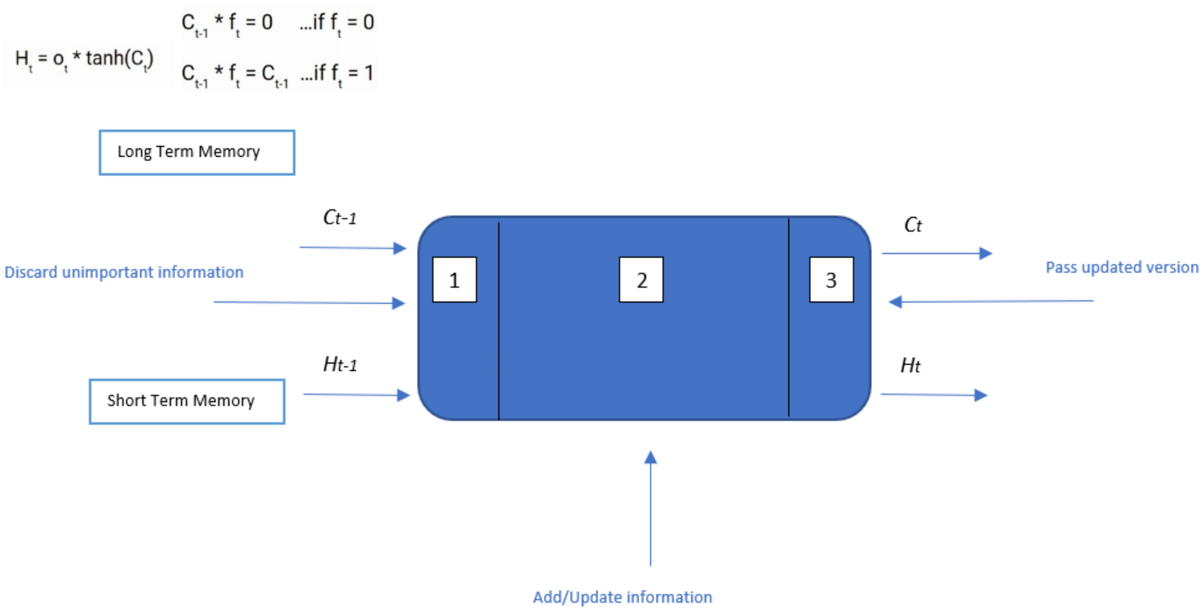
Figure 22:*Multi-Layer Perceptron's forecast for the next hour's value of water's EC.*

6.2 Predict water’s electrical conductivity using multivariate Long-Short-Term-Memory RNNs

6.2.1 Epigrammatic Reference to Long-Short-Term-Memory RNNs

Long Short-Term Memory networks (LSTMs) are Recurrent Neural Networks (RNNs), capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used for use cases such as prediction or image processing problems. They support supervised learning, so they tick all the boxes that the use case in this thesis suggests.

Long-Short-Term-Memory networks are explicitly designed to avoid the long-term dependency problem. Their asset is their ability to remember information for long periods of time and use it to make decisions. All this, is working of recurrently. All recurrent neural networks have the form of a chain of repeating modules of neural network. A basic LSTM’s architecture will consist of three parts. The first part chooses whether the information coming from the previous timestamp is worth to be remembered or not so it can be forgotten. In the second part, the cell tries to learn new information from the input. Last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp. At fig.23, a basic architecture of a LSTM RNN is displayed, among some of the formulas that are used for the calculations.



$$H_t = o_t * \tanh(C_t)$$

$$C_{t-1} * f_t = 0 \quad \text{...if } f_t = 0$$

$$C_{t-1} * f_t = C_{t-1} \quad \text{...if } f_t = 1$$

__Figure 23__:LSTM Architecture

6.2.2 Implementation of multivariate Long-Short-Term-Memory RNNs for the prediction of water's electrical conductivity

After the first approach with a MLP neural network, a second attempt with the use of LSTM models were made. This was needed to utilize the features that were selected at the feature selection phase. The LSTMs have a renowned reputation for their performance in prediction use cases , therefore that was a main reason that were selected for implementation. As it is already known, the other four features that were selected from the feature selection methods, the atmospheric temperature, the water temperature the water's salinity and the water's chlorophyll, provide significant insights for the behavior of the water's electrical conductivity value. All for features will be used as an asset so they can predict the water's EC value.

By using this design, the LSTM models that will be used are characterized as multivariate. That means that they will be designed to predict the value of water's EC by using the values from the other four features that were selected in the feature selection stage. This is the reason that the feature selection was conducted in the first place. In this thesis , two different design approaches of LSTM models are made.

The first one suggests that the previous day's values of the other four features will predict the water's electrical conductivity next days, value at a specific hour. The second approach that took place, was more straight forward and tradition , as the model was designed so using only past values of the other features will predict the water's electrical conductivity for the exact next hour from when the last step used was measured. Both approaches have their pros and cons. They will be extensively analyzed below.

6.2.2.1 Implementation of a multivariate LSTM for the prediction of the next day's water's EC

As mentioned, the first approach of the implementation of the LSTM RNN is by using a design that will allow the prediction of the water's electrical conductivity value, based on the values of the other four complementary features, as they were measured twenty-four (24) hours ago. This approach is quite interesting as it provides practical benefits. It can be more insightful , suggesting that the values of the previous days are more likely to have already been measured so the water's electrical conductivity prediction can be predicted for the next day.

The prediction of the next day's water's electrical conductivity value may provide more actual value to the prediction purpose in some cases. In addition, when it comes to the implementation itself, by following this approach, there are data left 'to the tank' ,twenty-four records to be exact, that can be used for the forecasting phase.

At this point of the thesis, hourly data are being used , meaning that the dataset contains measurements per hour , for a year, therefore contains timeseries, from the features selected from the feature selection, as well as the corresponding measurements of water's electrical conductivity that will be used to form the target dataset. Firstly, the dataset is split to training and testing. Specifically, the first six thousand (6000) records from the other four, complementary features, starting from the very first measurement, are used for the training, accompanied with the target set containing the values of the water's electrical conductivity. It is rational that the firsts twenty-four (24) values from the target dataset are skipped. The target dataset contains also 6000 records meaning that values selected were from the 24th record up the 6024th record. The rest from the measurements of water's EC ,from the 60025th record until the very last measurement in the dataset are used for the target set of the testing dataset, meaning that it is consisted of a total of 2641 (8665-6024) records.

When it comes to the testing set , its values contain the measurements from the 6001st record up until the records that will cover the same number of records contained in the target set, 2641 records ,so the testing dataset contained the measurements of the other four features, starting from the 6001st until the (6001+ (8665-6024))=8642 record. Therefore, the testing dataset also contains 2641 records. Given that the actual measurements start from 0 point, that leaves the twenty-four (24) last records to be exploited for the forecasting.

After the splitting of the dataset in training and testing, the datasets were also standardized for the same reasons as previously explained. However, this time the Standard Scaler was used for this process. This time ,the values contained in the datasets, are not in the closed space of [-1.1] but are also preprocessed to aid the calculation processes (standardized) that take place in the model while maintaining the individuality of each measurement. The formula for the Standard Scaler uses is displayed at the fig.24.

$$z = \frac{x - \mu}{\sigma}$$

Standardization

Figure 24:*Standar Scaler Calculation Formula*

Once the standardization of the training dataset took place, the next step is to reshape the data appropriately so they can be inputted at the LSTM. At this approach, the step of the input will be fourteen (14).However, each step will include the measurements of all four selected by the feature selection, features. In more detail, each step will be consisted of measurements that will contain the measurements of atmospheric temperature, water's temperature , water's salinity, and the water's chlorophyll, for the past fourteen timeseries. This pattern will go on and on until the limit of the first 6000 records will be reached. Now, this data design, will lead to the formation of 6000-14=5986 inputs with the two-dimensional shape of (14,4),referring to the fourteen measurements of the four features. The final shape of the input is three-dimensional as it is (5986,14,4).The step of fourteen was also selected by trial and error.

This input design, is necessary , as the LSTM model that is set to be design will be multivariate. After the transformation of the training dataset, the design of the model takes place. The model chose, is a sequential LSTM model from TensorFlow. The input shape is set to be in the shape of (14,4). Important is define the return sequences parameter equal to True. Two hidden layers of 50 neurons each, was also added, as well as the Dropout factor was set to remove randomly the 20% of existing neurons in the hidden layers, preventing overfitting. The output layer consists of one neuron that will predict the water's electrical conductivity value. As an optimizer, once again Adam is chosen, and the loss function is the Mean Squared Error. After the compilation, the model trained, with 250 epochs and a batch size of 16. This combination of epochs and batch sizes as well as the other parameters of the models were tuned to achieve better performance.

After the training phase, the testing phase took place. The model made predictions and as the original values of those that the model tried to predicted exists, the model can be evaluated. As it can be easily viewed at the fig.25, where the testing phase is plotted by line plots. The values that the model predicted, the blue line, is not quite following the same behavior as the original values, the red plot. There are some points, especially through the end of the training phase that they touch but overall, the performance looks poor.

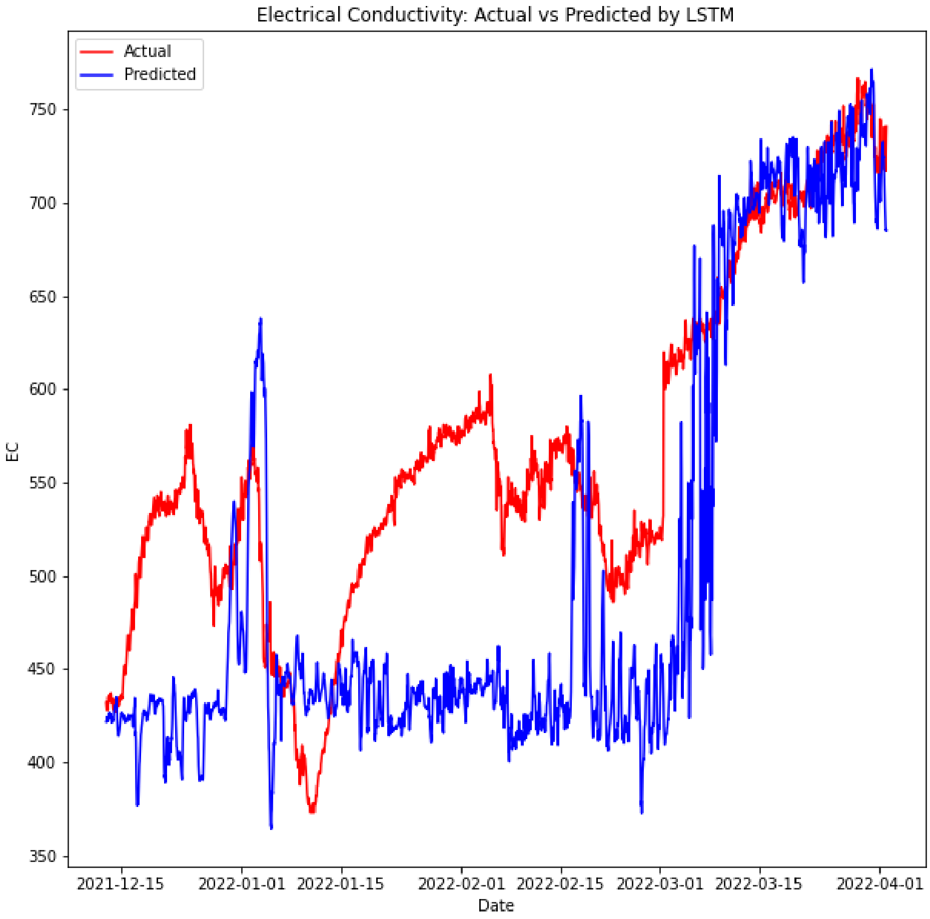


***Figure 25**:LSTM that predicts the next day's water's EC , at the testing phase.*

The inefficiency of the model can easily be views at *Table 9*. Judging by the coefficient of determination which is quite low at 0.0021 and the mean squared error that is high at 8390.180 , the evaluation makes the model unreliable and not usable.

Table 9: MSE and R^2 of the multivariate LSTM model ,that predicts the next day's water's EC value, after testing.

Evaluation Method	**Score**
Mean Squared Error (MSE)	8390.180
Coefficient of Determination (R^2)	0.0021

Despite the poor performance, the stage of the actual forecasting took place. For the forecast , as previously mentioned there are twenty-four unusable measurements of the other four features. These 24 measurements, with the hop of fourteen can make in total ten input steps, meaning ten inputs at the model. This will produce 10 forecasts in total. At the fig.26 , the forecasts are displayed. The ten forecasts made are displayed with the green plot. The ten forecast describes the water's electrical conductivity values as predicted from the model, for the time window that begins 24 hours after the last step. The last forecast is referring to the water's EC value for the next day's first hour which is 00:00 a.m. The red plot describes the actual last values of water's EC and the blue line describes the corresponding predicted values by the model.

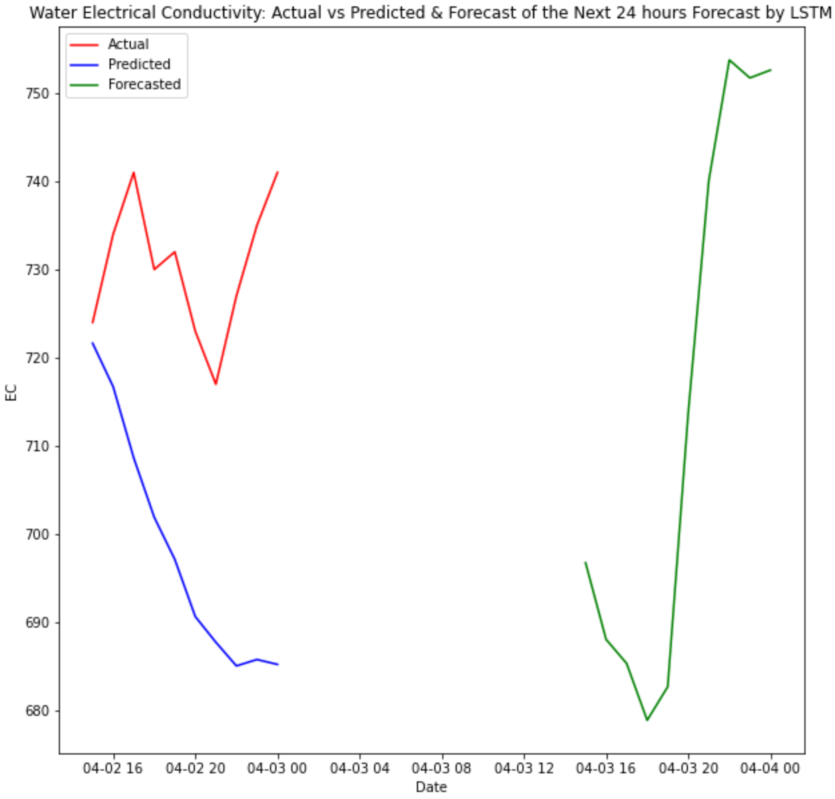


Figure 26:Forecast of the water's EC value after 24 hours.

6.2.2.2 Implementation of a multivariate LSTM for the prediction of the next hour's water's EC

As the first approach did not perform se well a more stable and conventional approach is made. A LSTM model is implemented again, but the design now wants the value of water's electrical conductivity, to be predicted by the hour, and also the prediction to be defined by the fourteen values of the other four features, as they were measured exactly fourteen hours prior from the hour that is about to be predicted .Meaning that, every fourteen (14) hours of the other four features will predict the next hours value of water's electrical conductivity. This ,approach meant to be safer and more stable, aiming for higher performance evaluation scores but it also comes with some drawbacks. The main drawback being that all the measurements that exist in the dataset are exploited , leaving no margin for forecasts. However, this issue will be overcome, using individual univariate high performance MLPs for each of the other features ,that predict multiple future values providing data for LSTM forecasts.

The main difference in this approach to the previous LSTM approach is the shaping of the data entering the model. Now each fourteen values of the other features will predict the corresponding next value of water's EC, for the next hour, so the shaping of the dataset is required to be different. The splitting to training and testing is similar. Specifically, the first six-thousand (6000) timeseries of the dataset are used for training as before, and the remaining two-thousand-sixty-five-thousand-sixty-five (2665) timeseries are used for testing. The target set for both training and testing will contain the corresponding timeseries of the water's electrical conductivity. However, as the step is also consisting of fourteen (14) measurements, the first 14 measurements will form the first input step, so in total 5986 input steps of shape (14,4) exist. The same goes for the target at training as its shape is (5986,1).

After the training and testing splitting, as well as the reshaping, the model was compiled and trained. The architecture of the model is exactly the same as the previous LSTM approach. Meaning that the input shape is also set to be in the shape of (14,4). The return sequences parameter is also equal to True. There are also two hidden layers of 50 neurons each as well as the Dropout factor is also set the same as before. The output layer consists of one neuron that will predict the water's electrical conductivity value. As an optimizer, once again Adam is chosen, and the loss function chosen is also the Mean Squared Error. The difference between the two LSTM models ,comes after the compilation, as this model trained, with 100 epochs and a batch size of 32. The same parameters were also tested with the previous model. The though behind keeping both models architecturally the same, was that it would be interesting the test them while they have the same design. However, the parameters of epochs and batch sizes best fitted this model , with this data shape, and therefore the were selected. They were also defined after trial and error.

After the training of the model the testing phase took place. At testing phase the model performed quite well, as it can be visually viewed by the fig.27 where the plot of the testing is displayed. The red line describes the actual water's EC values, and the blue line describes the water's EC values as they are predicted by the LSTM model. It seems that the model follows almost the same behavior of the original values. In this case the almost is wanted as it does not imply overfitting.

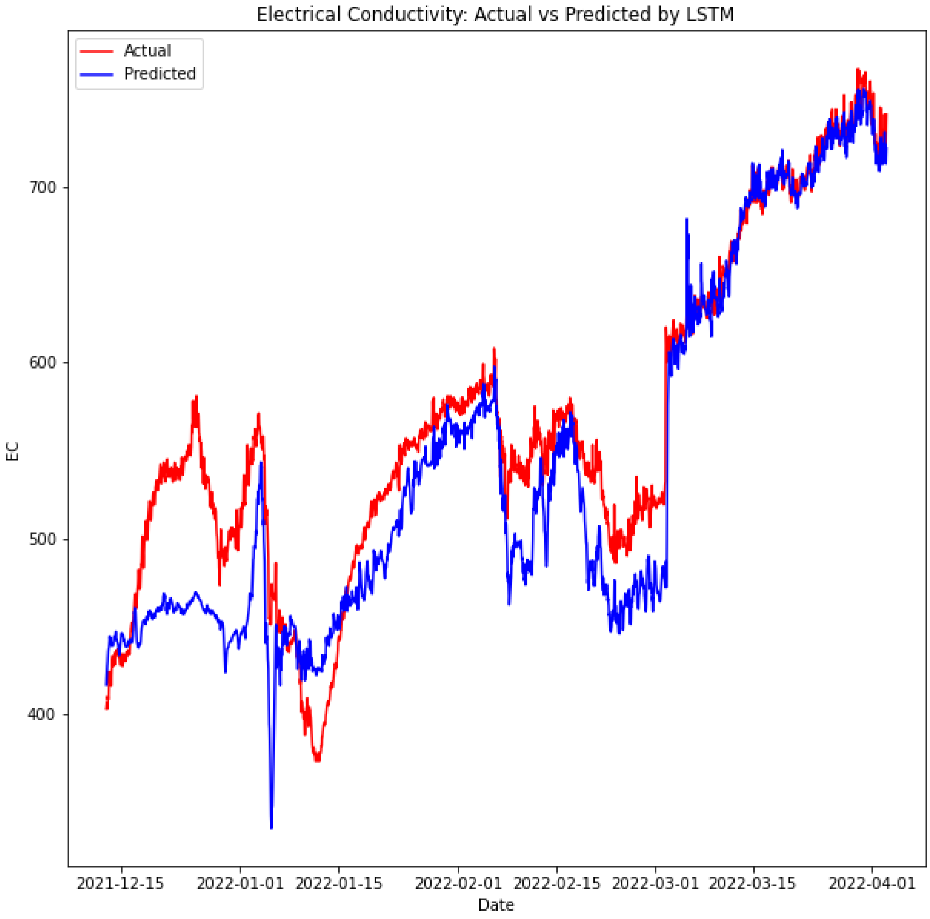


Figure 27:*LSTM that predicts the next hour's water's EC, at testing.*

At Table 10, the mean squared error (MSE) and the coefficient of determination (R2) are shown. The MSE is quite low at 933.01 and the coefficient of determination is very high at 0.8 906 indicating the almost excellent performance of the LSTM RNN model. This model is reliable and trustworthy and it will be used for forecasts.

Table 10: MSE and R^2 of the multivariate LSTM model ,that predicts the next hour's water's EC value, after testing.

Evaluation Method	Score
Mean Squared Error (MSE)	933.01
Coefficient of Determination (R^2)	0.8906

As previously mentioned, there is a difficulty with this approach as all the existing timeseries are used for training and testing, therefore no data are left to make forecasts. However, using individual ,high performance , univariate Multi-Layer-Perceptron models for each of the other four features, solved this issue. The architecture of the models as well as the data input shape are exactly the same as the MLP created for water's EC. The coefficient of determination (R^2),the Mean Absolute Error(MAE) and the Mean Squared Error(MSE) for every model implemented can be views at *Table 11,* proving the high-performance statement.

Table 11: R^2 ,MSE and MAE of the univariate MLP models ,for each of the four complementary features.

	R^2 Score	MSE Score	MAE Score
MLP For Atmospheric Temperature	0.93	1.15	0.75
MLP For Water Temperature	0.90	0.43	0.47
MLP For Salinity	0.92	0.00027	0.92
MLP For Chlorophyll	0.84	0.67	0.71

After the training of these MLP models , the forecasts for each value is made. The standardization, and predictions are conducted in a loop, so each forecasted value will be used to forecast the next one. Inside the loop the last fifteen values are kept ,as the MLP already displayed for water's EC ,had a step of fifteen (15) , then the dataset is scaled ,and it is used to make the next prediction, the scaling is inverted and appended to the bottom of the existing dataset while the first measurement is excluded, keeping the number of available measurements always at fifteen (15). This is repeated for fourteen (14) times ,so 14 forecasts are made per feature. The algorithm design as described is displayed below with the form of pseudocode.

```
i=0
While i<14:
        Keep the last fifteen measurements
        Normalize them using MinMax Scaler
        Reshape to (1,15) to fit the model
        Predict with the individual MLP per feature
        Inverse scaling
        Append the forecasted value to the existing dataset, at the end
        i=i+1
```

After the creation of the forecasted values per each feature are made , a dataset that contains the last fourteen measurements from the existing dataset and the fourteen new forecasted values are appended, with respect to the time factor. This happed, so as the step of the LSTM is fourteen, the total inputs that can be entered the model are not just one (1) consisted only of the forecasted values but each step , until the last one, is consisted of existing values a well. Fourteen steps are formed to be exact that provide fourteen forecasts. It is worth mentioning ,that the less the true values that each step contains the less accurate the next forecast will be. As the dataset is created as described, the dataset then is standardized , and the fourteen (14) future forecast are made and inverted from the standardization to express reality. The forecasts are displayed at the fig.28where the actual forecasts can be viewed in a plot, as they are represented by the green line. The red line shows the real water's EC values for the previous hours and the blue line shows the corresponding predicted values at the testing phase.

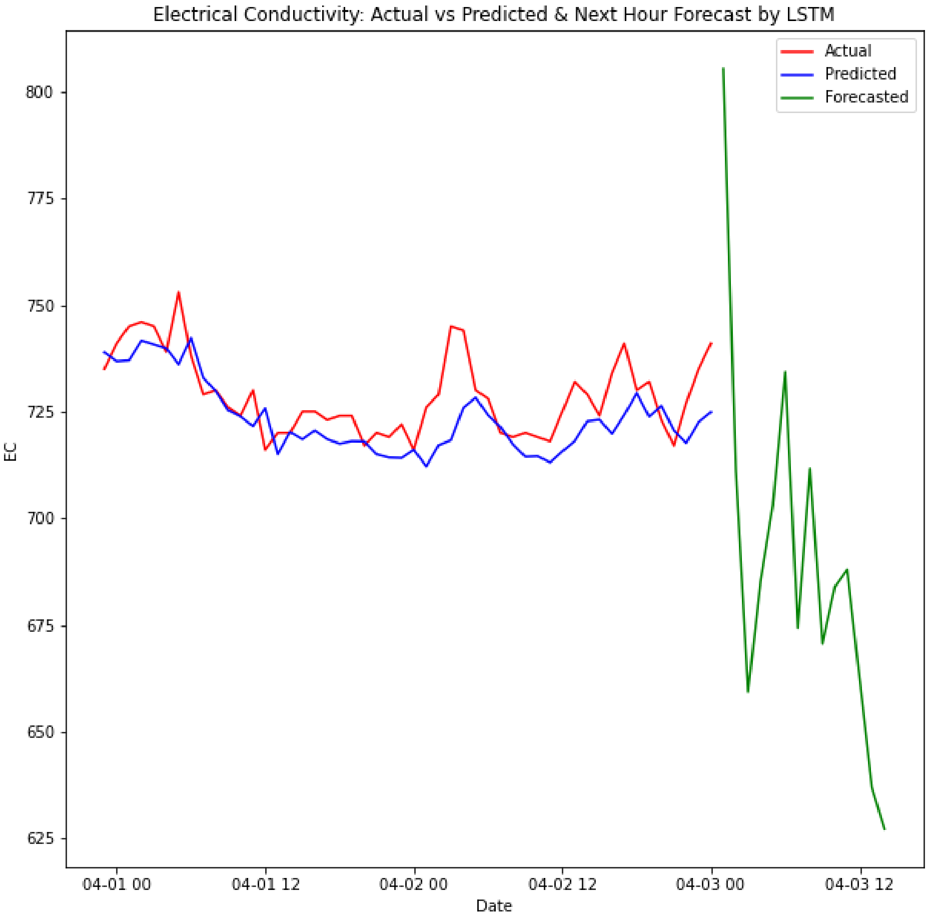


Figure 28:*Water's EC value for the next 14 as they are forecasted by the LSTM RNN model.*

After the implementations of both machine learning and deep learning (A.I) models, several results can be concluded. Results not only about the performance and the evaluation of the models that were implemented and tested, but also about the usability and the individuality of those models. Water's electric conductivity is a very important and interesting feature to conduct research on and the ability to predict its feature values is very useful. This thesis proposed several methods , some more traditional than others, that can be used to forecast water's EC feature values. Not only that but the research on the complementary features that can affect the various changes of the water's EC values, can also be proven very helpful.

To begin with, at the feature selection phase, several complementary features, both weather and water chemical features, were studied with the goal of finding if they actually affect the water's electrical conductivity value. By calculating both the Pearson's correlation coefficient and the Mutual Information score , among the water's EC and the other thirteen features, the dimensionality reduction was achieved and the final decision about the dataset that will be used was made , considering of course both methods. This decision suggested the ideal complementary features for water's electrical conductivity are: atmospheric temperature , water temperature, salinity, and chlorophyll.

Afterwards, as the feature selection phase was completed, the actual models can be implemented. At first, more traditional univariate forecasting models were implemented. A simple ARIMA and a variation of it, the SARIMAX , were designed ,trained and tested. The better results were given by the SARIMAX model. From the research of the water's electrical conductivity, it is obvious that the feature presents strong seasonality trends, therefore no wonder that the SARIMAX model performed better.

As the machine learning statistical models were implemented, more unconventional and state-of-the-art approaches were taken. This were the implementations of deep learning neural networks. A Univariate Multi-Layer-Perceptron that was able to predict the water's electrical conductivity EC and individual univariate MLPs for the other features were implemented ,trained and tested. This model performed excellent. Finally, multivariate LSTM RNN models that actually uses the other four features selected at the feature selection phase were implemented, in two variations. The first one suggested the predictions of the water's EC value by the values of the complementary features as they were measured 24 hours ago. The second version uses the measurements of the other four features to predict the water's EC value for exactly the next hour. The second version that predicts the values in a shorter time-period performed much better.

In the Table 12, the coefficient of determinations (R2) and the Mean Squared Error (MSE) for each model that was used to make future predictions about the water's electrical conductivity values are being displayed. As it is easily viewable the worst scores are given by the SARIMAX(0,1,3,90) machine learning model and the best scores are given by the univariate Multi-Layer-Perceptron.

Table 12: R^2 and MSE per model that predicted the water's electrical conductivity value implemented and used in this thesis

Model	R^2 Score	MSE Score
SARIMAX(0,1,3,90)	-3.2000	39243.490
Univariate MLP	0.9730	214.329
Multivariate LSTM (for next day's prediction)	0.0021	8390.180
Multivariate LSTM (for next hour's prediction)	0.8906	933.010

At the fig.29 the coefficient of determination per model is displayed in a descending order, with the form of a bar plot. The last bar has the highest R^2 score , indicating that it was the one that performed best. This score belongs to the univariate MLP. However, close is also the multivariate LSTM that predicts the water's electrical conductivity value for the next hour.

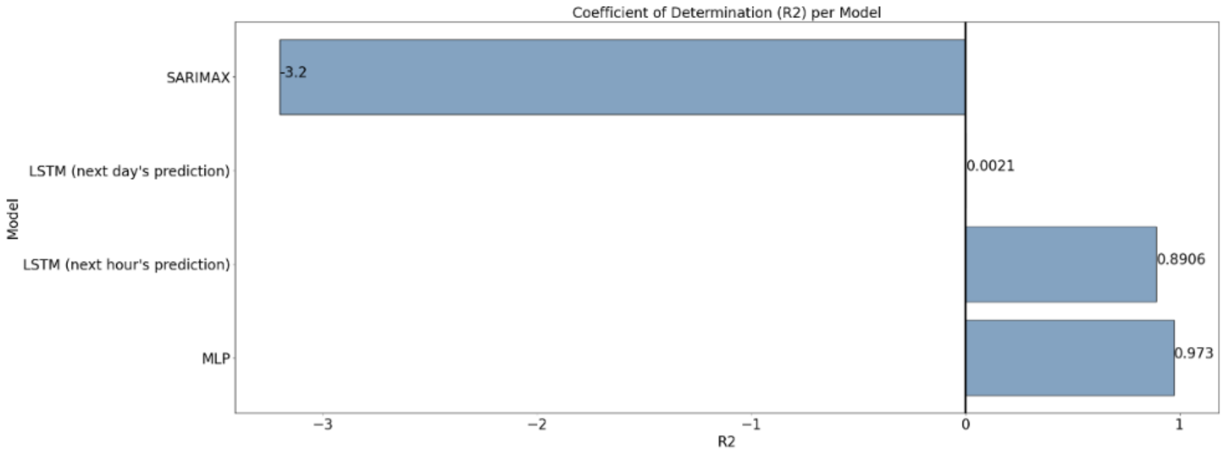


Figure 29:*Coefficient of determination (R^2) per model.*

At fig.30, the Mean Squared Error (MSE) per model that is used to make forecasts is displayed in the form of a bar plot in an ascending order. The first bar, that appeared the lowest MSE score, is representing the univariate MLP model, proving that it is the one that performed the best. In the second place with a small difference is the multivariate LSTM RNN that predicts the water's electrical conductivity value for the next hour.

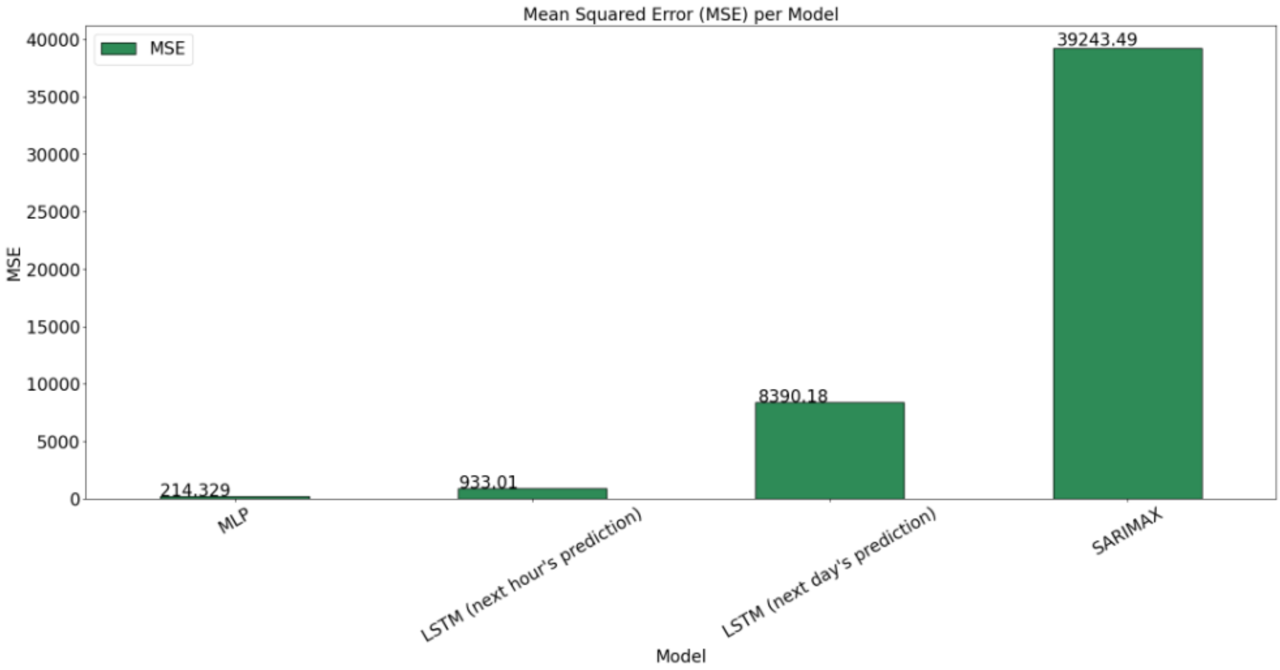


Figure 30: Mean Squared Error (MSE) per model.

In conclusion, judging only by the performance scores, the ideal model if the ability of using a univariate model is given, is hands down the univariate Multi-Layer-Perceptron. When it comes to the multivariate models, the LSTM model, designed to predict the next hour's water's electrical conductivity value, is the winner. However, it gives the limited ability of only short time predictions. The use case will be the one factor that will select the best model. For this thesis, as the use case defined the features that were selected at the feature selection phase to be used in as the complementary features that will predict the water's EC value, the LSTM model , able to predict the next hour's water's electrical conductivity value, is the ideal model.

In addition, it is important to mention that the univariate MLP design came in handy when data for the forecasting phase of the LSTM , needed to be generated.

In this thesis, two feature selection methods, Pearson's correlation coefficient and Mutual Information, proposed and implemented. In addition, two types of forecasting methods using machine learning, ARIMA and SARIMAX were also proposed , implemented, and evaluated. Furthermore, two Deep Learning (A.I) neural networks, the univariate Multi-Layer-Perceptron, and the multivariate Long-Short-Term-Memory RNN in two variations, were also proposed , trained ,tested and evaluated. Almost all methods were also used to make future forecasts of the water's electrical conductivity future value. However, it comes with no doubt that not only the existing models can be further tuned to perform better, but also the whole data input approach can be differentiated.

Moreover, the feature selection phase can be further enriched in order to give even more insights about the water's electrical conductivity from the co-existing features. Feature selection and reduction methods such as PCA , decisions trees, ANOVA and Chi-squared methods can be implemented. Such methods maybe can provide more useful information about the water quality features and how they affect each other while simultaneously can achieve the desirable feature reduction. It would be really interesting to further investigate feature selection methods at the future. When it comes to the dataset itself, it would be ideal to acquire a larger dataset from measurements about water quality features. It is known that it is the time of the big data , this will not be too hard. More available timeseries, describing measurement in the long past, can undoubtedly provide a better view about the behavior of water quality features, such us the water's electrical conductivity ,throughout the time. It is a goal to further research about more data and possibly even more features that can assist with the prediction of the water's electrical conductivity value.

Finally, as previously mentioned the phase of the model prediction can be also tuned in multiple ways. Not only the models already implemented in this thesis , can be tuned so they will perform better, but also different models can also be implemented and evaluated in the predictions of water's electrical conductivity. Future work of this thesis can be the experimentation with different approaches and parameterization of the existing models used but also the tryout of different artificial intelligent models used for prediction such us GRNN (Generalized Regression Neural Network) or even machine learning methods such us SVM (Support Vector Machines) can be tested.

9. References and Bibliography

[1]: Sirisha Adamala :An Overview of Big Data Applications in Water Resources Engineering(2017).

[2]: Tharsanee Raman Maganathan , Soundariya Ramasamy Senthilkumar , Vishnupriya Balakrishnan: Machine Learning and Data Analytics for Environmental Science: A Review, Prospects and Challenges(2020).

[3]: Muhammed Sit,Bekir Z. Demiray,Zhongrun Xiang;Gregory J Ewing,Yusuf Sermet,Ibrahim Demir: A comprehensive review of deep learning applications in hydrology and water resources(2020).

[4]: Yeonjung Lee&Sun-Yong Ha&Hae-Kyung Park&Myung-Soo Han&Kyung-Hoon Shin: Identification of key factors influencing primary productivity in two river-type reservoirs by using principal component regression analysis (2014).

[5]: Hong Guo, Kwanho Jeong, Jiyeon Lim, Jeongwon Jo, Young Mo Kim, Jong-pyo Park,Joon Ha Kim, Kyung Hwa Cho: Prediction of effluent concentration in a wastewater treatmentplant using machine learning models(2014).

[6]: Amir Hamzeh Haghiabi, Ali Heider Nasrolahi and Abbas Parsaie: Water quality prediction using machine learning methods(2018).

[7]: Luis Arismendy, Carlos Cárdenas, Diego Gómez, Aymer Maturana, Ricardo Mejía2and Christian G. Quintero M.: Intelligent System for the Predictive Analysis ofan Industrial Wastewater Treatment Process(2020).

[8]: Liya Fu,You-Gan Wang: Statistical Tools for Analyzing Water Quality Data(2012).

[9]: Jonathan Keck, Juneseok Lee: Embracing Analytics in the Water Industry(2020).

[10]: Abhilasha Sharma ,Himanshu Shekhar: A predictive analytics framework for Sustainable Water Governance(2021).

[11]: Nirav Raval , Manish Kumar : An Overview of Big Data Analytics: A State-of-the-Art Platform for Water Resources Management(2020).

[12]: Ravesa Akhter,Shabir Ahmad Sofi: Precision agriculture using IoT data analytics and machine learning(2021).

[13]: Maryam Rahbaralam, David Modesto, Jaume Cardús, Amir Abdollahi, Fernando M Cucchietti: Predictive Analytics for Water Asset Management: Machine Learning and Survival Analysis(2020).

[14]: S. C. Hillmer,G. C. Tiao: n ARIMA-Model-Based Approach to Seasonal Adjustment(2012)

[15]: J. Contreras; R. Espinola; F.J. Nogales; A.J. Conejo: ARIMA models to predict next-day electricity prices(2003)

[16]: Niematallah Elamin, Mototsugu Fukushige: Modeling and forecasting hourly electricity demand by SARIMAX with interactions (2018).

[17]: Mohammad Ali Ghorbani, Hojat Ahmad Zadeh, Mohammad Isazadeh ,Ozlem Terzi: A comparative study of artificial neural network (MLP, RBF) and support vector machine models for river flow prediction(2016).

[18]: A. Najah, A. El-Shafie, O. A. Karim & Amr H. El-Shafie : Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring (2014).

[19]: Jacob Benesty, Jingdong Chen, Yiteng Huang: On the Importance of the Pearson Correlation Coefficient in Noise Reduction(2008).

[20]: Benedikt Gierlichs, Lejla Batina, Pim Tuyls & Bart Preneel : Mutual Information Analysis (2008).

[21]: Yuanyuan Wang, Jian Zhou, Kejia Chen, Yunyun Wang, Linfeng Liu: Water quality prediction method based on LSTM neural network(2018).

[22]: Qiangqiang Ye, Xueqin Yang, Chaobo Chen, Jingcheng Wang: River Water Quality Parameters Prediction Method Based on LSTM-RNN Model (2019).

[23]: Yurong Yang, Qingyu Xiong, Chao Wu, Qinghong Zou, Yang Yu, Hualing Yi & Min Gao: A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism(2021).

[24]:Md. J. B. Alam, M. R. Islam, Z. Muyen, M. Mamun & S. Islam : Water quality parameters along rivers(2007).

[25]: Mauno Vihinen : How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis(2012).