1. Introduction

In the time of growing urbanization and climate change, it is vital to have resilient environment in smart cities with secure water availability (Ebrahim Banihabib and Mousavi-Mirkalaei, 2019). There is noticeable discrepancy between water supply and water consumption in areas with expanding populations (Kesornsit and Sirisathitkul, 2022). Prediction of urban water consumption can help to minimize the imbalance between supply and demand and enhance the performance of water distribution systems and in recent times, scientific interest is growing for the development of prediction models for water consumption (Cutore et al., 2008). Also, With the accurate prediction of water supply and demand in the future, we can implement water resources planning in an effective and scientific way (Liu et al., 2022).

However, the estimation and prediction of water demand involves many uncertainties and parameters such as socioeconomic factors (population, income, education, household size , etc.) and environmental factors (precipitation, temperature, wind, evapotranspiration) (Pourmousavi et al., 2022). Along with need of considerable amount of data, the constant change in city with rapid urbanization, population growth, modernized construction and water resource pollution, prediction of the water supply and demand becomes more challenging (Liu et al., 2022).

Nowadays, machine learning methods to perform prediction have emerged as an attractive, rapid and reliable computing tool (Kesornsit and Sirisathitkul, 2022). Generally, for time series forecasting, both linear and nonlinear methods are used. Auto Regressive Integrating Moving Average (ARIMA) and Auto Regressive moving average (ARMA) are the most commonly used linear models and artificial neural network (ANN) is extensively used nonlinear method (Ebrahim Banihabib and Mousavi-Mirkalaei, 2019). Random forest algorithm has also been widely utilized in water demand prediction and good results have been obtained (Liu et al., 2022).

2. Aim and Objectives

Study of demand pattern and factors influencing the demand provides the groundwork for better understanding of the water distribution system which helps in advancing the network to the digitalization era. Extended water consumption timeseries data is key to start studying and analyzing water demand pattern. With the limited timeseries, it is tricky to carryout wholesome analysis. Thus, this project has focused on the specific objectives that was possible to explore with limited timeseries data. The objectives of the project were (a) to develop a machine learning predictive data driven model for summer months; (b) to understand the consumption behavior of the given municipality during summer months and to do so, following questions were answered:

- Is there relationship between temperature, rainfall, time of the day and water consumption?
- Are there monthly variations within the summer months?
- Are there any correlations between water usage and public holidays?

3. Methods

3.1 Study Area

Danderyd municipality, is a small municipality in the north of stockholm with a total area of 32.67 sq. km with a total population of 32,803. It is a small but affluent municipality.

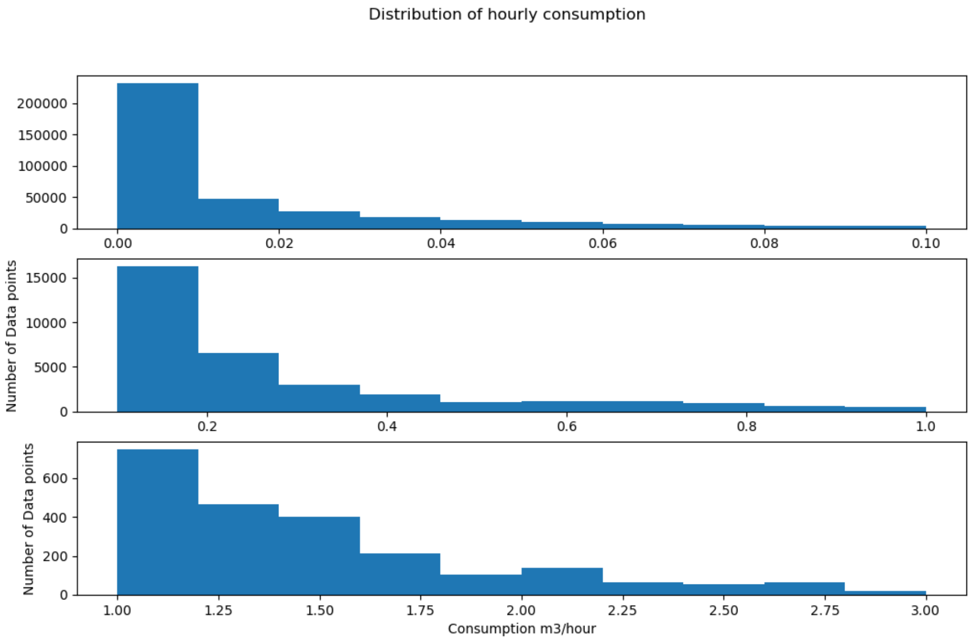
3.2 Assembling the data

The data file contained an hourly data of consumption from 2020-05-01 to 2020-09-15 recorded by 132 meters. The time of recording was approximately hourly as the times on which the meters had recorded the data was random. Hence there was a need to resample the data and even out the timestamps. The time series dataset was checked for any missing values and null values. Also, the consumption calculation for all the meter data was carried out in the beginning. For every meter reading, volume of water consumed was calculated to generate hourly consumption data.

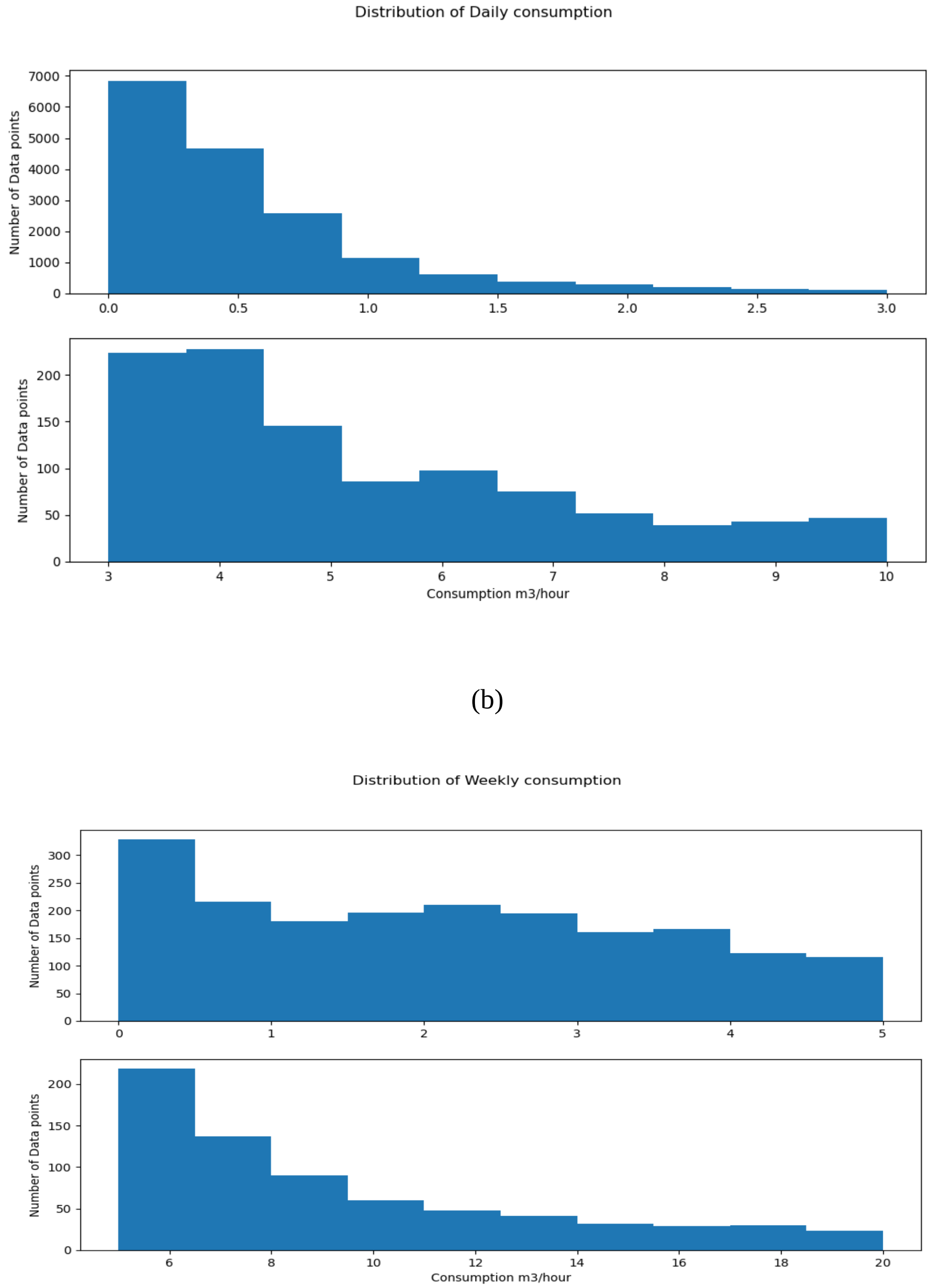
Further, precipitation and temperature data were taken from SMHI. Since, we did not have the municipality gauge station, we used the stations that were approximately was 8 kms away from given municipality. The assembling of data was done for two parts of this project i.e., one for modelling and another for hypothesis testing.

- For Modelling:

The data was resampled by Hourly, Daily, and weekly classes. The resample function of pandas was used to do the same. To observe the variability in the data histograms were constructed as in Figure 1.



(a)



(b)

(c)

Figure 1 Histograms for Variability in consumption variations for different ranges; (a) Hourly Variation (b) Daily Variation (c) Weekly Variation

From the above figure we can see that histogram for Hourly, Daily and weekly measures of consumption. For the Hourly, three histograms were constructed with range [0,0.1], [0.1,1] and [1,3]. The following ranges were seen fit to capture the variability. Similarly for the daily values ranges from [0,3] and [3,10] were chosen, and [0,5] and [6,20] for the weekly values. With the help of the above classes were assigned to each data point. Classes here are defined as a range of consumption obtained from the histogram, Table 1 The range of consumption for each class (0 to 6) with its time specifies the range of consumption for each class (0 to 6) with its time.

Table 1 The range of consumption for each class (0 to 6) with its time

Time Interval	Classes							
	0	1	2	3	4	5	6	7
Hourly(m^3/h)	0 – 0.01	0.01 - 0.0196	0.0196 – 0.0296	0.0296 – 0.05	0.05 – 0.191	0.191 – 0.3	>=0.3	
Daily(m^3/d)	0 – 0.289	0.289 – 0.602	0.602 – 0.888	0.888- 1.487	1.487 – 4.39	4.39 - 5.79	>=5.79	
Weekly(m^3/w)	0 – 0.49	0.49 – 0.98	0.98 - 3	3 - 4	4 - 5	5 – 6.6	6.6 – 9.7	>= 9.7

- For hypothesis testing

The data sets were resampled to get hourly consumption. Then for all four summer months (May, June, July, and August), separate data frames were generated. These data frames were again grouped to get cumulative hourly consumption for all given meters for all days in a particular month. The outliers for these individual datasets were checked by plotting scatter plots and outliers were removed if there were any. Furthermore, to check the correlation between the temperature, precipitation, hour of the day and month, a separate data frame was generated. This data frame had daily resolution of all datasets and had 138 rows for each parameter. Similarly, three public holiday dates (May 26 Ascension Day, June 06, National day, June 25 Midsummer Day) were selected and data frames were formed to check the change in the monthly consumption pattern due to holiday.

3.3 Selection of input variables for the model

For modelling, following variables were selected

Table 2 Variables used for three different Models (Hourly, Daily and Weekly)

Variable	Resolution	Remarks
Mean Temperature	Daily and Weekly	Data grouped by Mean for weakly resolution
Precipitation	Daily and weekly	Data grouped by Sum for weekly resolution
Day of the month	Daily	-
Water Consumption	Daily, weekly, and hourly	Data grouped by Sum of consumption
Hour of the day	Hour	-

3.4 Modelling

Three models were chosen namely Random Forest, Boosting, Logistic Regression. All three models were imported from the sklearn package of python. Apart from classification Regression models were also used that have a discharge value as an output given the instances that are provided in the model. For calculating the class, the consumption column was removed before hand as class includes consumption. Vice versa was done to calculate the consumption. The Mean square error is used for the consumption model as the measuring parameter to select the model.

- Random Forest

It is a type of ensemble model used for various machine learning algorithms. It can be used for both classification and regression. It basically creates a multitude of decision trees and selects the class that is chosen by most trees. In the model for Random Forest depths from 1 to 10 were iterated and the mean error for each was calculated. For regression it takes the mean of all the values.

- Boosting

Boosting is an ensemble model. It splits the data into various subsets and creates models for the same. It then combines all the models, and the outcome is a combination of all the models learnt from the subsets.

- Logistic Regression

The following model models the probability of each outcome given a set of variables. And returns the value that has a high probability.

The following models were chosen citing many papers, such as (Herrera et al., 2010), (Shuang and Zhao, 2021). Both these papers have all the models mentioned above except for Logistic regression. Logistic Regression was chosen instead of Support vector machines (SVM) in this model as SVM take a very long time to train the model. Hence to reduce the computation time the later was chosen. Time was also included in the models as an integer namely Month, Year, Day, Hour, Week. Week number was only included in the weekly dataset, and Day was included in all the dataset. Hour was also only included in the hourly dataset. The Daily and the weekly models also included precipitation and temperature data.

The data was randomly split into two categories a test and a train. The chosen percentage of the split was thirty percent, test being the lower. Three such data sets were made, and each were used to make the models and calculate the error. The Final error was calculated as the arithmetic mean of the three Mean errors. As this type of error does not consider the distance between the predicted and the real value of the class a different type of error was used to see how much is the mean difference in all predicted values and the real values. For this the absolute difference of the same was taken and then the mean of that sum was calculated. The error was named Class Distance Error

The Mean error was calculated Table (123123)

$$Mean\ Error = \frac{\sum_{i=1}^{n}(y_{pi} \neq y_{ti})}{n}$$

$$Root\ Mean\ Square\ Error = \sqrt{\frac{\sum_{i=1}^{n}(y_{pi} - y_{ti})^2}{n}}$$

$$Final\ Error = \frac{\sum_{i=1}^{3} Mean\ Error_i}{3}$$

$$\text{Class Distance Error} = \frac{\sum_{i=1}^{n}\left|(y_{pi} - y_{ti})\right|}{n}$$

Where Y_p is the predicted class and Y_t is the test class, n is the number of data points.

Naïve Model

A naïve model was made which included all the classes 7 for hourly and daily and 8 for weekly. The test data was compared to the model and the above three errors were calculated.

3.5 Generating Plots

To fulfill the aim of the projects, multiple graphs were plotted. Four box plots were generated to see hourly consumption pattern for each summer months. Similarly, for selected three public holidays that lie in month of May and June consumption line graph was plotted over the monthly consumption graph of respective months. Temperature vs Consumption graph was also plotted to verify the relation between these two parameters.

4. Results

4.1 Relationship between temperature, rainfall, time of the day and water consumption

The correlation heatmap was generated from the dataset gave the results shown in Figure 2

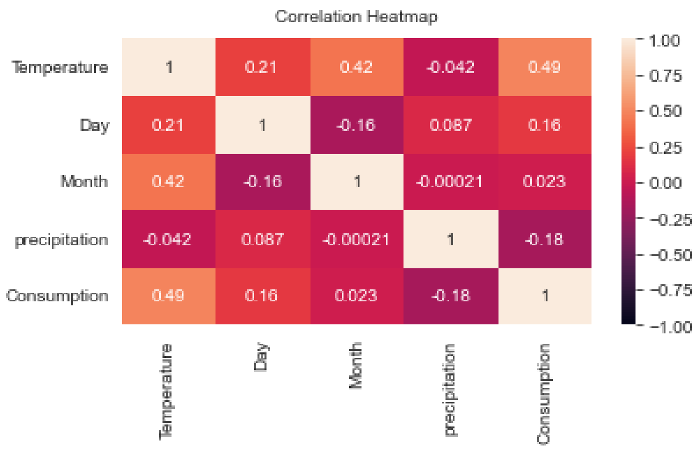


Figure 2 Correlation between temperature, rainfall, time of the day and water consumption

Further, to visualize the water consumption and temperature relationship, following graph show in Figure 3 was generated.

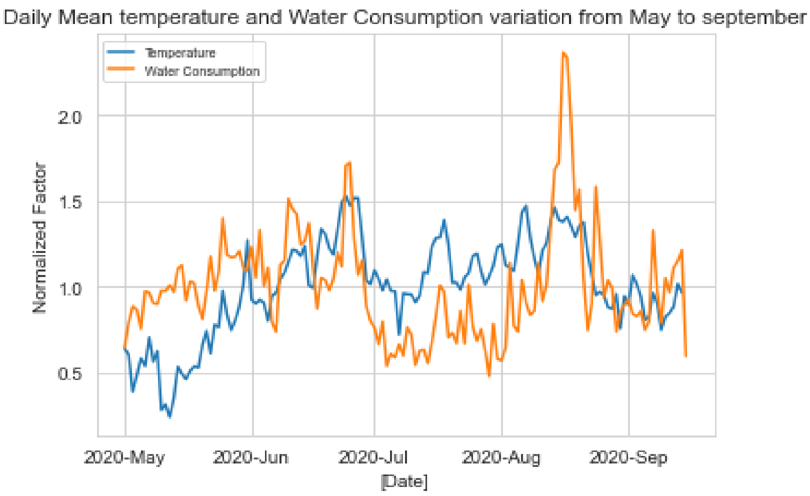
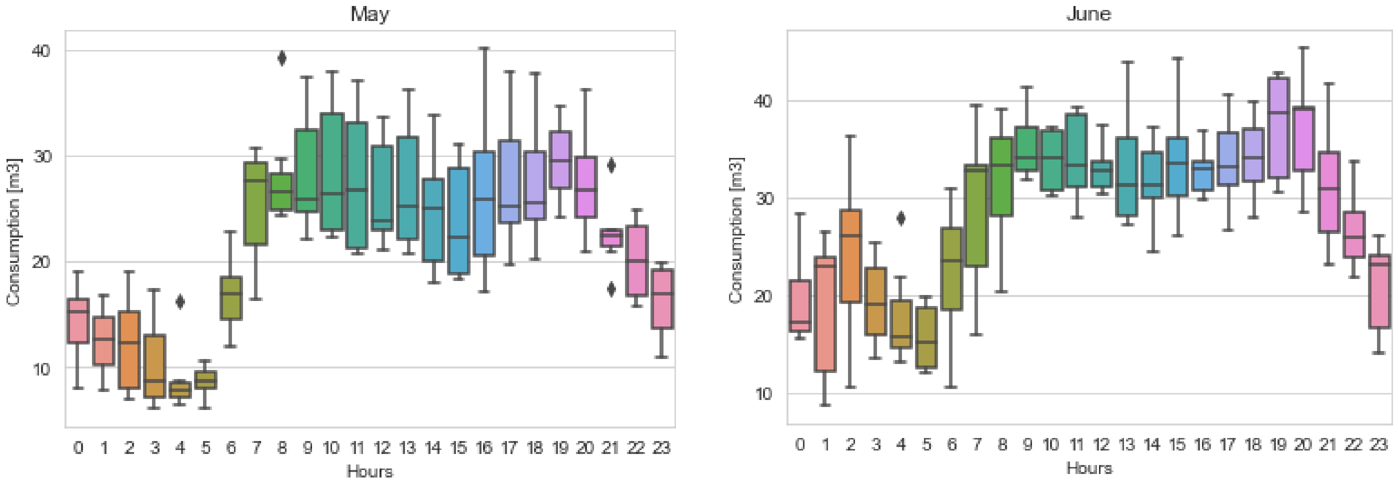


Figure 3 Daily Mean Temperature and Water consumption Varriation from may to september

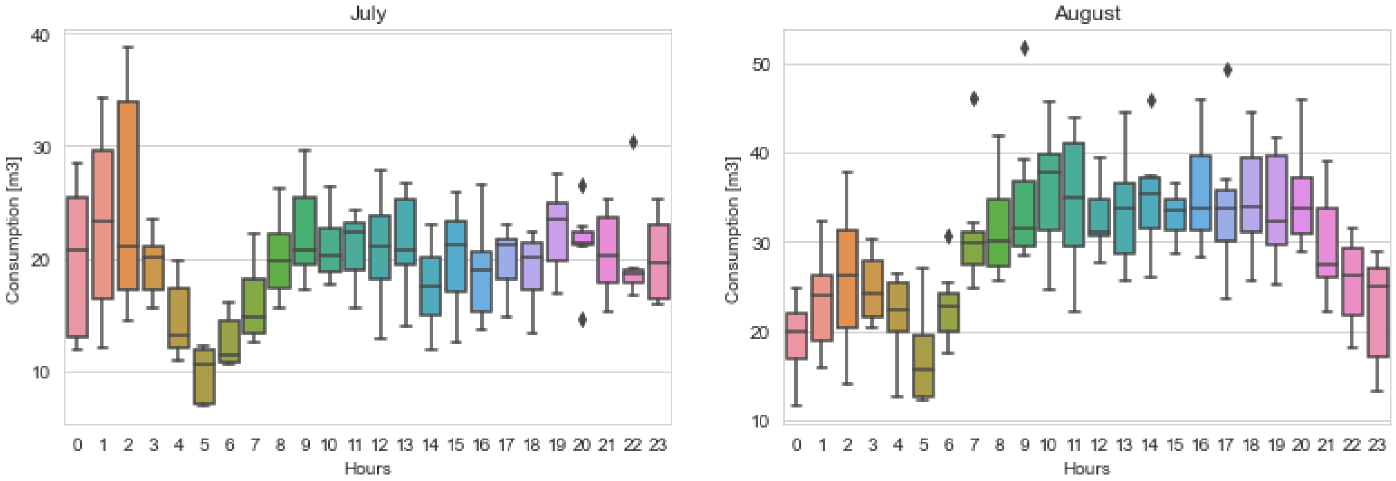
4.2 Monthly Consumption variation pattern within the summer months

The cumulative hourly consumption was plotted against the hours in a day for the entire month for all four months. The plots are shown in Figure 4 below:



(a)

(b)



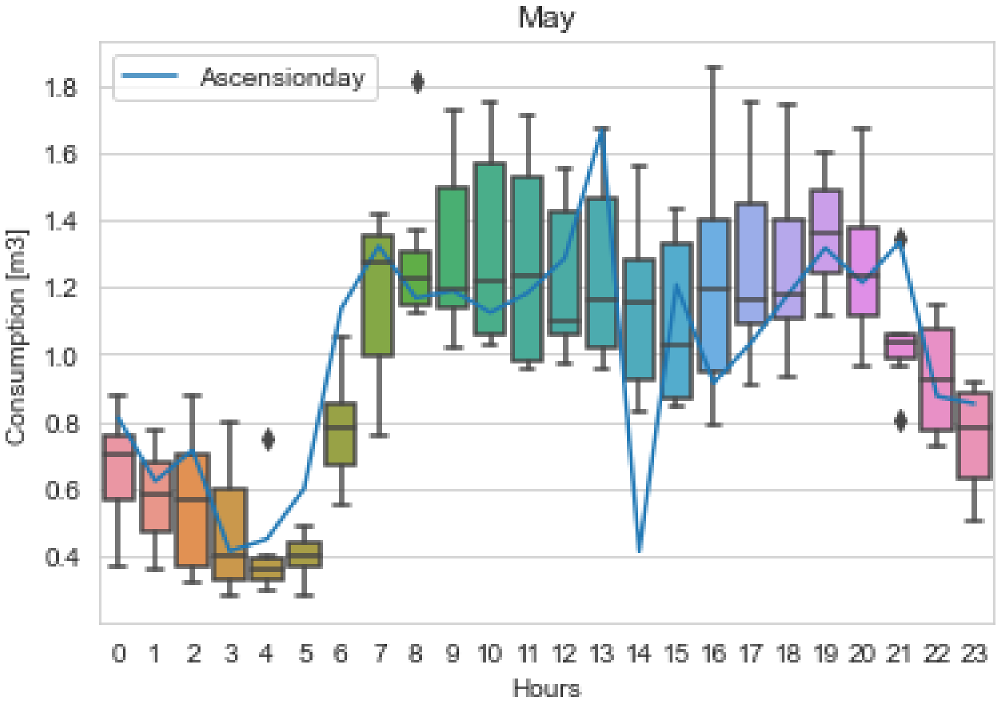
(c)

(d)

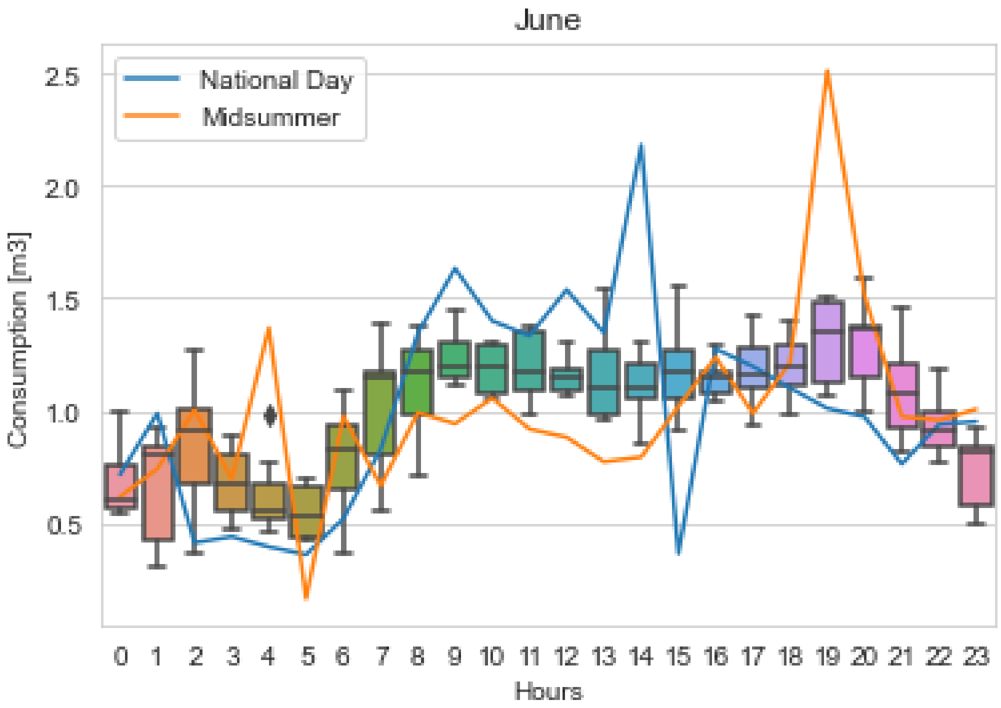
Figure 4 Cumulative hourly consumption variations for summer months; (a) May, (b) June, (c) July, (d) August

4.3 Correlations between water usage and public holidays

The change in pattern of water usage during the holidays in compared to entire month is shown in the Figure 5 below:



(a)



(b)

Figure 5 Change in pattern of water usage during the holidays in compared to entire month (a) May (b) June

4.4 Modelling

The total computational time taken to assemble the data and make the models was above one hour. The models were not tuned using any packages such as grid CV, but rather by catching the variability in the data and assigning appropriate classes. The errors for the classification models are reported in the following tables.

Table 3 Errors by Logistic Regression model and Boosting Model (class)

Time Interval	Models	
	Logistic Regression	Boosting
Hourly	0.4288	0.4288
Daily	0.6316	0.6304
Weekly	0.7066	0.6992

Table 4 Errors by Random Forest Model (class)

Time Interval	Random Forest (Depth)								
	1	2	3	4	5	6	7	8	9
Hourly	0.4288	0.4288	0.4288	0.4288	0.4288	0.4288	0.4288	0.4288	0.4288
Dailly	0.6361	0.6285	0.6288	0.6278	0.6265	0.6273	0.6272	0.6266	0.6282
Weekly	0.7149	0.7060	0.7014	0.6987	0.6978	0.6978	0.6964	0.6964	0.6964

For all the three models we have a similar behavior for hourly time interval. This can be attributed to the model sensitivity of the classes that have been assigned. The classes were not able to capture the variability in the data as was expected. As in the case of hourly assigning more classes would certainly help. The models showed high error for both Daily and weekly consumption. This could be attributed to the availability of data points. Very less data points were available for the weekly model and hence the fit was not that good. The error here can be interpreted as 42% of all values precited as classed in hourly model of Random Forest were wrong. The same analysis can be used for the other two which give 63% and 71% wrong predicted classes.

Table 5 Class Distance Errors

Time Interval	Models		
	Logistic Regression	Boosting	Random Forest
Hourly	1.250	1.25	1.25
Daily	1.379	1.3287	1.3
Weekly	1.9493	1.9775	2.040

The class Distance error gave a better picture of the models. It can be seen in table 6 that for hourly intervals the absolute mean difference in the predicted classes and real classes is between one and two. Hence we can say that the model is good, but can be better if we have a larger training data and the classes more sensitive towards the variability. For Daily the model is also predicting the classes well, but as can be seen from the error in tables 3 & 4 most values are wrongly predicted. The models are not able to predict the weekly classes. They are having high values for both Class Distance Error and the Mean Errors.

Table 6 Root Mean Square Errors by Models (Consumption)

Time Interval	Boosting	Polynomial	Random Forest
Hourly (m^3/hour)	0.2	0.197	0.204
Daily (m^3/day)	2.5	2.323	2.55
Weekly (m^3/week)	10.72	10.317	10.72

In table 5 for the Polynomial and the Random forest the errors were averaged for all the degrees and depths. It can be seen in table 5 that the errors are very high for hourly we have a mean square error of around 0.2. Which can be interpreted as a deviation of 200ml per hour by the model for each of the 132 different meters. Which is a very high value for a small municipality. The Daily and the weekly errors are also very high.

Naïve Model

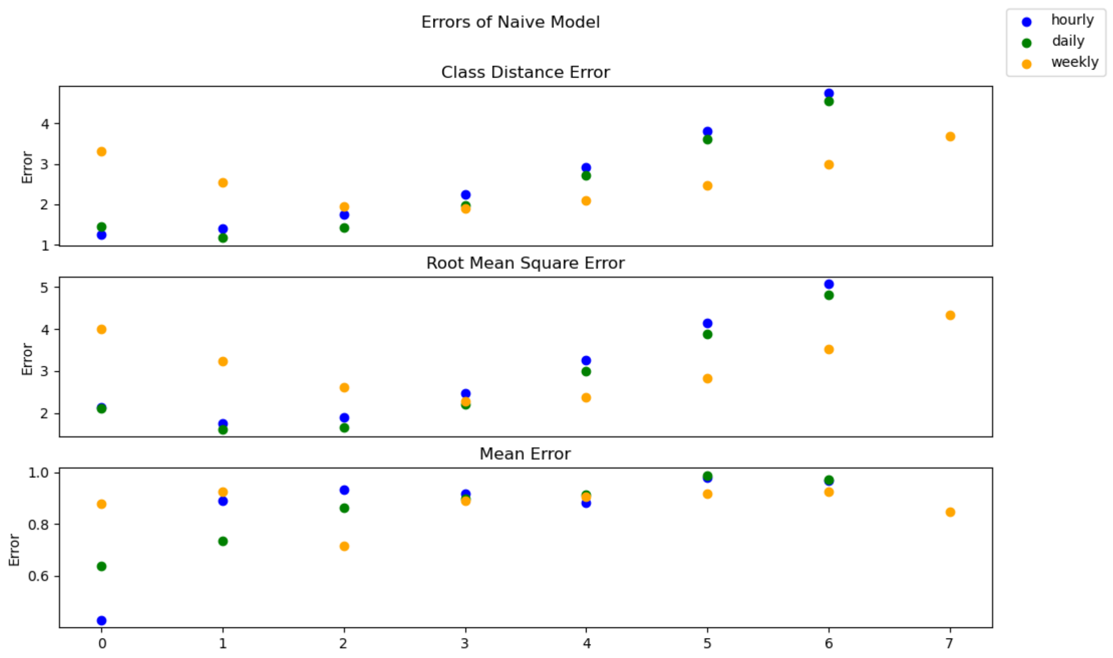


Figure 7 Naïve models of every class, hourly daily and weekly.

Figure 7 shows the errors of the daily, hourly and weekly time intervals for the naïve models. There are in total 7 such models which predict only one class given any input. These models are used as a base for comparison to the other models reported above. If we see the class distance error, we can clearly see that the classes zero and one are behaving like a naïve models for all our models. This can be attributed to the high bias in the data set. With almost 36% of all values are zero, and almost 50% are below one. These two classes in the hourly time interval are overly represented. For the rest of values for RMSE and Mean Error, our models work fine as the error for the models(Random forest, Logistic Regression, Boosting) is below the errors of the naïve models for each class.

5. Discussion and Conclusion

We can see high correlation between the temperature and demand. There was not much of difference observed in consumption pattern for holidays but different holiday had different consumption behaviour. This may be due to limited resolution of data. Precipitation however showed the negative correlation.

It can be very tricky to predict the water demand of a particular area as many factors at play. In this model only some are included. Sociological factors, political factors (holidays) or other unknown factors are not included in this model. Hence this model is a very general one that includes only two meteorological factors namely temperature and precipitation. The model also includes time in terms of Month, day, hour, and year.

The regression models predicting values as an output for consumption had a very high error. But interestingly the models that were predicting classes had a relatively low error. By calculating the Class Distance Error we can see that the wrong predictions are wrong by one or two classes. The number 1.25 suggests that most values were wrong by only one class. But that is not the case for the weekly model. The weekly model had the least number of data points and hence could not generalize. It can also be seen by the Mean Error, which just sees if the predicted classes are wrong or right. Which means that it does not consider the magnitude of the incorrect answer. Hence if a model is off by two classes or by four classes is considered to be the same. We can see that for hourly and daily 42% and 61% of all predicted values of the classes are wrongly predicted. But The predictions are off by only one class. The error is slightly higher for the daily time interval and hence it is best to have more daily data points to train the model. Hourly has a more data points and it can be seen that the error is relatively small. The error for the weekly consumption is very high, and again this can be attributed to the limited data. But the bias in the weekly data in very less. Hence a little more data points would make a good weekly model using the same method above. There is bias in the daily model which is relatively less than that of the hourly model.

It can be concluded that it is possible to predict trends with a 60% accuracy for the hourly model given the same size of the data. Daily and weekly models cannot perform well with this training data size.

References

Cutore, P., Campisano, A., Kapelan, Z., Modica, C., Savic, D., 2008. Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. Urban Water Journal 5, 125–132. https://doi.org/10.1080/15730620701754434

Dahlke, H.E., Lyon, S.W., Jansson, P., Karlin, T., Rosqvist, G., 2014. Isotopic investigation of runoff generation in a glacierized catchment in northern Sweden: ISOTOPIC INVESTIGATION OF RUNOFF GENERATION IN A GLACIERIZED CATCHMENT. Hydrol. Process. 28, 1383–1398. https://doi.org/10.1002/hyp.9668

Ebrahim Banihabib, M., Mousavi-Mirkalaei, P., 2019. Extended linear and non-linear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region. Sustainable Cities and Society 48, 101585. https://doi.org/10.1016/j.scs.2019.101585

Kesornsit, W., Sirisathitkul, Y., 2022. Water consumption prediction based on machine learning methods and public data. Advances in Computational Design 7, 113–128. https://doi.org/10.12989/ACD.2022.7.2.113

Liu, X., Sang, X., Chang, J., Zheng, Y., Han, Y., 2022. Sensitivity analysis and prediction of water supply and demand in Shenzhen based on an ELRF algorithm and a self-adaptive regression coupling model. Water Supply 22, 278–293. https://doi.org/10.2166/ws.2021.272

Pourmousavi, M., Nasrollahi, H., Najafabadi, A.A., Kalhor, A., 2022. Evaluating the performance of feature selection techniques and machine learning algorithms on future residential water demand. Water Supply 22, 6833–6854. https://doi.org/10.2166/ws.2022.243

Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. J. Hydrol. 387, 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005

Shuang, Q., Zhao, R.T., 2021. Water Demand Prediction Using Machine Learning Methods: A Case Study of the Beijing–Tianjin–Hebei Region in China. Water 13, 310. https://doi.org/10.3390/w13030310