A&O 204 Final Project

Modeling of Water Consumption for Small Rural Communities via Decision Tree Regressor vs Random Forrest Regressor

Maria Soto

Introduction

Access to potable water is a universal concern that goes especially unaddressed in financially and technologically disadvantaged communities. In particular, small (< 11 residential units) communities with extensive agricultural activities must rely on local and usually contaminated sources of groundwater for domestic, irrigation, and livestock purposes [1]. In an attempt to treat nitrate (NO_3^2) and salinity contaminated water, engineers have developed a diversity of environmental, simple, and highly efficient water treatment systems at municipal scales. Nevertheless, there remains a lack of high resolution (hourly – seasonal variability) water consumption data modeling for precisely configuring and optimizing treatment systems whilst conserving technological and financial feasibility. Understanding water consumption trends becomes especially important for developing novel water treatment systems in remote areas as there is no in-person technical or engineer supervisory and support. As such, understanding and predicting community water consumption trends is critical for designing robust system autonomous operation and self-maintenance [2]. In addition to guiding the design of water treatment systems, forecasting the water consumption of small communities is also important for managing water storage within a reasonable capacity and monitoring groundwater replenishment and/or exhaustion to ensure communities will keep on having access to water.

In collaboration with the state of California, the Water Technology Research (WaTeR) Center of the University of California, Los Angeles (UCLA) is actively studying the water consumption trends of a small community located in Salinas Valley, CA. In this joint effort, the WaTeR Center developed a flexible reverse osmosis (FLERO) water treatment system with autonomous operation and a cyberinfrastructure for remote management and operation [2]. Nevertheless, the community water consumption has been observed to vary seasonally and to optimize the FLERO system operation, it is necessary to study and forecast the water consumption in Salinas.

The development of algorithms and statistical models with forecasting capabilities is a subset of artificial intelligence known as machine learning. The three main types of machine learning are supervised learning, unsupervised learning, and reinforcement learning. Under supervised learning, a model is trained on a labeled dataset and the input data is paired with corresponding output labels to develop predictions on unseen data. In contrast, unsupervised learning utilizes unlabeled data and the algorithm attempts to find patterns, correlations, or some structure within the data. Finally, reinforcement learning involves an agent that learns to make decision by receiving feedback in the form of rewards or “reinforcement” and alternatively, penalties for different solutions. Some of the most common machine learning models include linear regression, logistic regression, decision trees, random forest, and neural networks. Each type of model is designed to address specific tasks and types of datasets and thus, having a general understanding of the data, features, and desired outputs is necessary in selecting a model. In the case of small community water consumption, the data is expected to reflect a seasonal change in water demand that is impacted by the daily average weather (temperature), humidity, wind speed, and precipitation in the area. As such, a water forecasting model should consider various input features to arrive at a single output feature (water as the target value).

Accordingly, the objective of this project was to develop a model capable of forecasting the average daily water consumption (gallons) of small communities in Salinas Valley, CA by considering features such as ambient temperature (°F), precipitation (inches), windspeed (mph), humidity (%), and heating degree days (HDD). To accomplish this, (i) the raw meteorological and water consumption data was visualized, (ii) modeled with both decision tree regressor and random forest regressor, and (iii) optimized via hyperparameter tuning and feature importance ranking. Finally, the root mean squared error (MSE) and R^2 scores were compared to analyze the accuracy of the models and propose future work plans for improving the forecasting performance.

Data and Modeling

The objective of this project was approached in a four-step process that began with data acquisition, followed with data visualization, then the model development, and concluded with a discussion of the model limitations and required future work, **Figure 1**. The temporal water consumption of Salinas Valley was monitored via a wireless water meter (Spectrum 88DL, Metron-Farnier LLC, CO US) installed in the study community's main groundwater distribution point. The water consumption history was accessible through internal authorization to Water Scope and was initially visualized to demonstrate the water consumption from 2014 to present at different temperatures, **Figure 2**. The meteorological data was then obtained from the National Weather Service (NWS) online database. During the data acquisition and preprocessing, blocks of windspeed and/or humidity data (~50 points of data) would randomly be missing during the years 2014-2020 and as such, data from before 2020 was removed from the study and missing data from 2020 – present was filled in with manual meteorological online searches in NWS. In the second phase of the project, a brief statistical description (**Figure 3**) of five attributes and the target values (water consumption) was obtained and visualized with a histogram (**Figure 4**) to understand the general nature of the data before model training.

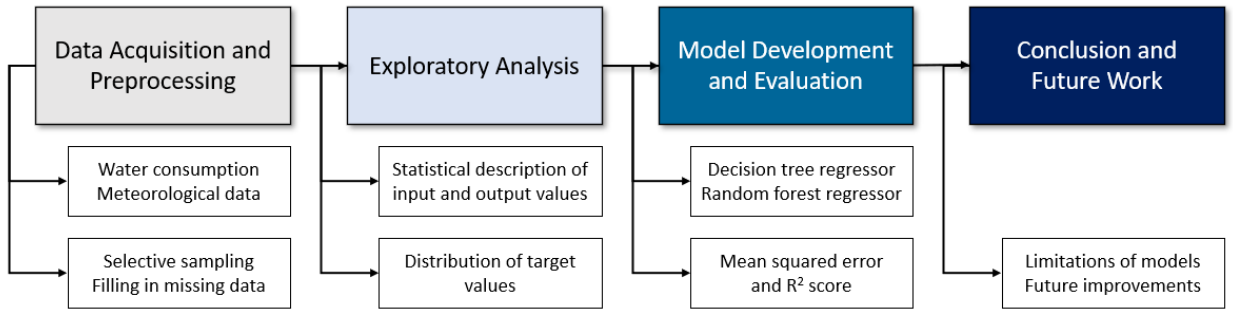

Figure 1. Workflow of the four project phases.

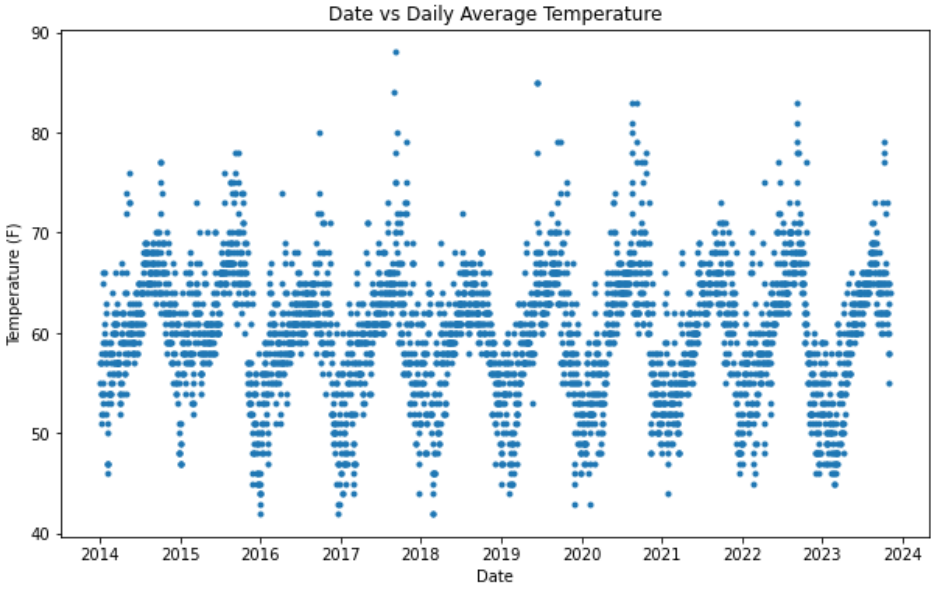

Figure 2. Initial data visualization of water consumption from 2014 – present at different temperatures.

```
df.describe()
```

	Water	DBTemp	Humidity	WindSpeed	HeatingDegreeDays	Precipitation
count	1316.000000	1316.000000	1316.000000	1316.000000	1316.000000	1316.000000
mean	821.004488	59.646657	70.119301	7.823100	6.130699	0.024164
std	165.968782	6.862512	9.775825	2.447206	5.718991	0.132020
min	376.354600	43.000000	20.000000	1.400000	0.000000	0.000000
25%	710.644475	54.000000	65.000000	6.200000	0.000000	0.000000
50%	803.968450	60.000000	72.000000	7.600000	5.000000	0.000000
75%	908.566300	65.000000	77.000000	9.100000	11.000000	0.000000
max	2310.534000	83.000000	94.000000	25.400000	22.000000	2.550000

Figure 3. Input and output value display.

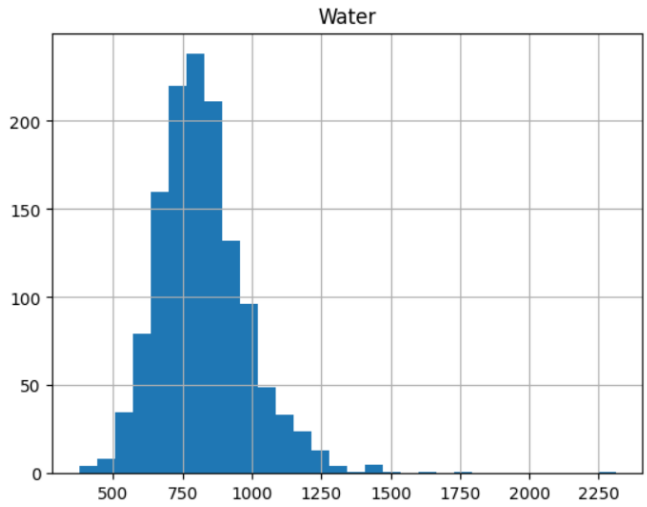

Figure 4. Distribution of target values.

When selecting the type of machine learning model, it was important to note that the project dataset was not expected to follow a linear relationship between features and water consumption, there was a potential usage for both numerical and categorical (i.e., weekday) data, and an expected difference in feature importance that would need to be investigated. As such, decision tree regressor was selected first. Nevertheless, to obtain more practice with model development and to compare two very popular and helpful techniques, random forest regressor (RFR) was also used. In contrast to decision trees, RFR mitigates overfitting by combining predictions from multiple decision trees and results in a more generalizable model. Moreover, RFR was expected to be less sensitive to the various outliers observed in this project's dataset. After each model was built with default hyperparameters, a grid search was executed to tune the hyperparameters and subsequentially increase the accuracy of the model during testing. Finally, the MSE and R^2 score of the original and tuned models were compared.

Results and Discussion

The decision tree regressor model was developed by training the model with 80% of the data and splitting 20% for testing. Based on the testing data, the model predictions scored a MSE of 28633.83 and a R^2 of -1.22 which were pretty unperforming scores relative to the data set, **Figure 5**. While an R^2 value of one reflects a perfect fit between the testing target values and the predicted output values of the decision tree regressor model, a negative R^2 score reflects a very inaccurate model. After this disappointing model performance, the tuned hyperparameter model was found to produce a slightly better MSE value (23531.01) and R^2 value (-0.073). Nevertheless, the R^2 was still negative and subsequently, this model was determined to be extremely inaccurate.

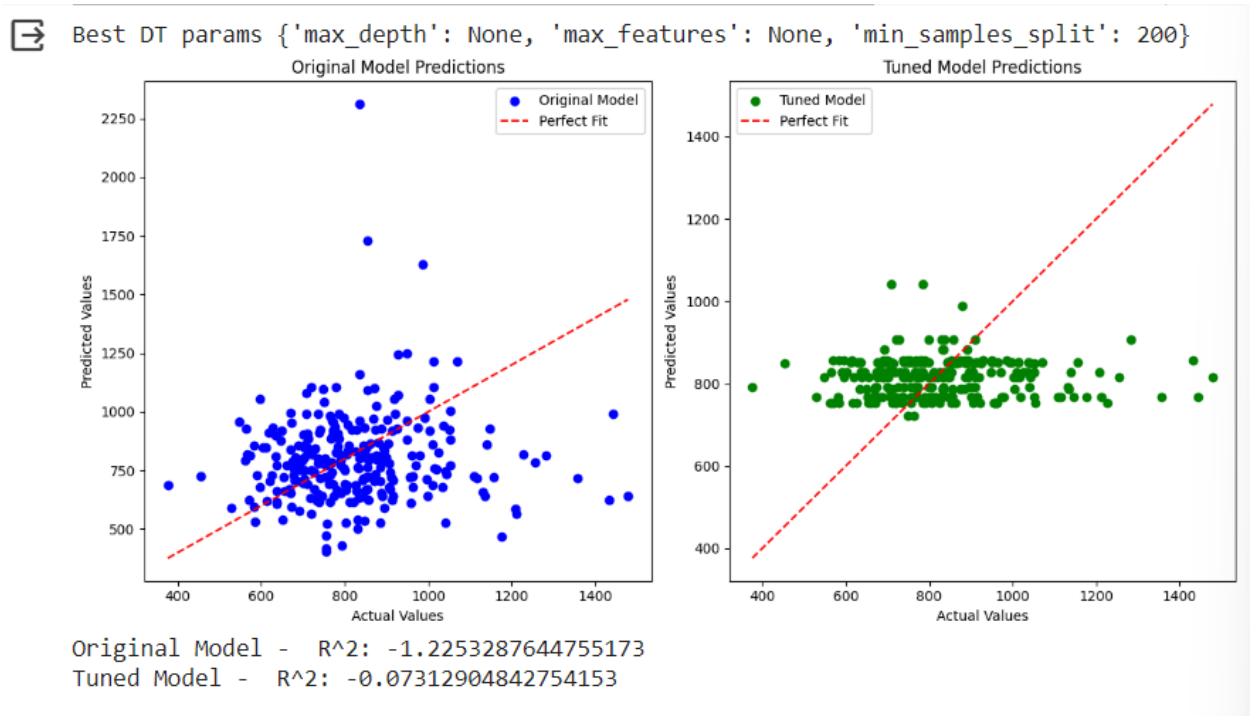

Figure 5. Final comparison between the original decision tree regressor model and the tuned model with optimized hyperparameters.

With a bit of hope, the RFR model was developed and a slight improvement of the R^2 score (-0.136) relative to the original decision tree model. Still, this score is substantially far from an acceptable R^2 value of one. The tuned hyperparameter RFR model then developed and found to produce a slightly better R^2 value of -0.058, **Figure 6**. Visually, plotting the actual output test value against the predicted output values of the RFR models demonstrated that there were at least a couple of outliers in the data and that tuning the hyperparameters only seemed to condense the distribution of the data into a horizontal pattern. Admittedly, this behavior was not easily interpretable. Then, the feature significance was investigated. Contrary to Zhou et al. (2021) findings, the calculation of feature significance revealed that there was a much stronger feature significance for wind speed (34.69%) than daily average temperature. Moreover, the daily average temperature was only the third feature with the highest significance of 18.08%, **Figure 7**. In Zhou et al. (2021)'s study, precipitation was found to have the second highest feature significance however, the dataset of this study scored precipitation with the lowest significance (0.04%).

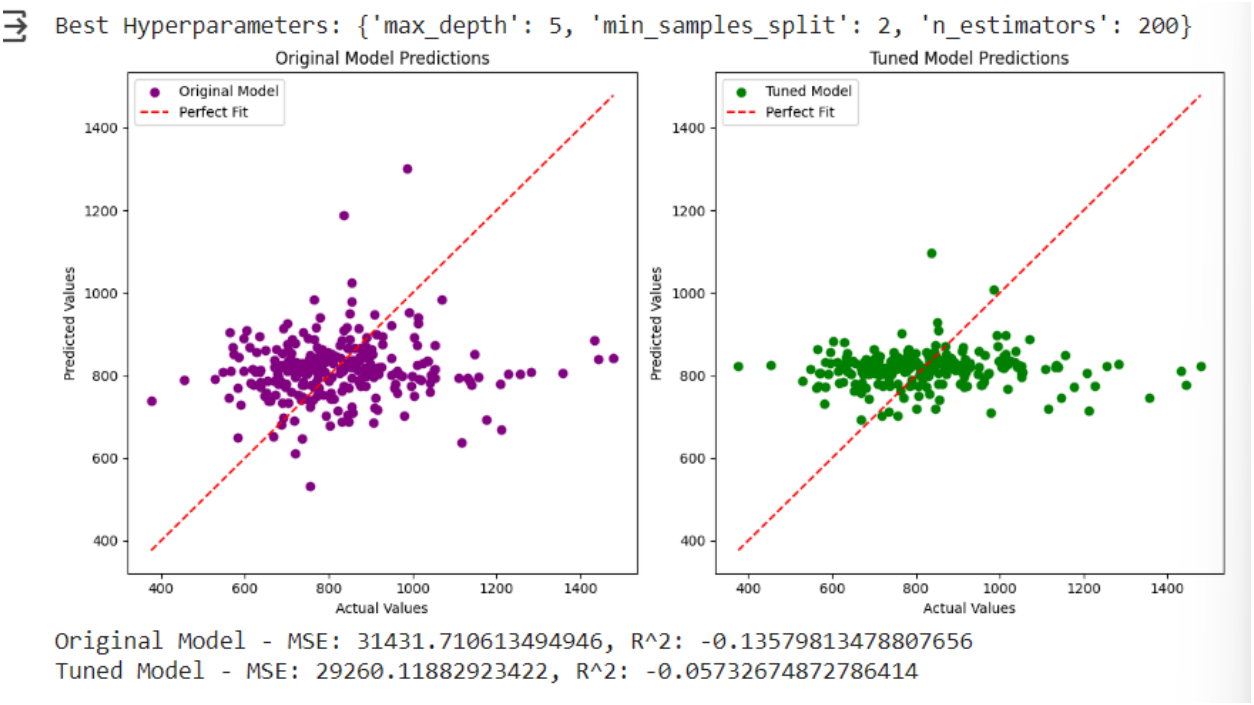

Figure 6. Final comparison between the original RFR model and the tuned model.

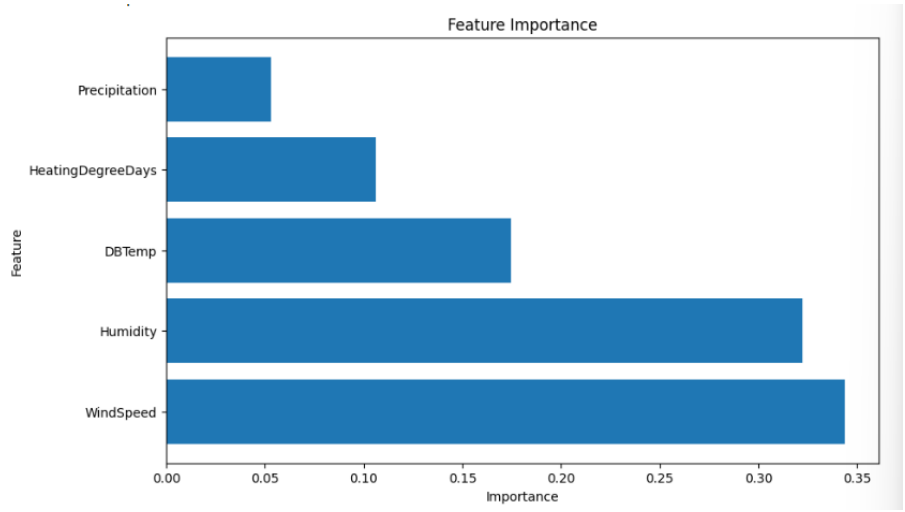

Figure 7. Comparison of determined feature importance.

Discussion with a research peer mentor identified a couple of potential reasons for the unacceptable model performance. First, it was determined that the exclusion of weekday and month categorical data might have substantially impacted the accuracy of the model as there is a clear trend in water consumption according to the day of the week (as a result of the community population working schedule) and month (as a result of the season and agricultural activities). Second, the size of the dataset was decreased from 115,167 data points to a mere 1,316 data points because of the missing blocks of humidity and/or windspeed values before 2020. Finally, there was very little (if any) data preprocessing. After plotting the four models, it was revealed that there were many outliers in the

dataset and the data was not initially scaled before model training. More time must be invested in preprocessing the data to generate a more interpretable MSE value for each model.

Conclusion

While there was a noticeable improvement in prediction accuracy from the decision tree regressor to the RFR model, the produced R^2 value was still unacceptable. Despite the hyperparameter tuning, the model accuracy was not really improved and the MSE calculated value remained uninterpretable. This model cannot be recommended for usage. Despite this, the project created for the perfect opportunity to contemplate, plan, and develop every aspect involved in machine learning model development. In other words, a variety of conceptual machine learning knowledge that typically goes unconsidered was more thoroughly understood through this project. For example, the importance of understanding one's dataset fundamentally was seen when it was revealed that windspeed contained a higher significance than daily average temperature. If the scientific understanding of how wind speed could have a higher impact on community water consumption was not understood from a environmental multimedia perspective, then one could have easily disregarded the higher feature significance score as some error in the calculations. However, by researching this phenomenon, one may now explore how this feature weight can be better taken into account during model development. Additionally, this project repeatedly encouraged the thorough understanding of how decision trees and RFR models can be limited by their tendency to overfit data or their sensitivity to outliers.

Still, more effort should be placed into developing an acceptable model. Specifically, categorical data for weekday and month should be included as a value of 1-7 and 1-12. Instead of just using daily average water consumption data points, the hourly water consumption data should be taken into account to increase the size of the dataset and potentially reduce the presence of outliers. Finally, data preprocessing should be performed to scale the data to interpretable metrics.

Much beyond the model performance, this project (and class) has given me the confidence to experiment with other datasets, explore new methods to optimize model accuracy, and derive new understandings of diverse datasets and models in different fields.

References

[1] K. R. Burow, B. T. Nolan, M. G. Rupert and N. M. Dubrovsky, "Nitrate in Groundwater of the United States," *Environmental Science & Technology,* vol. 44, pp. 4988-4997, 2010.

[2] Y. Zhou, B. M. Khan, J. Y. Choi and Y. Cohen, "Machine Learning Modeling of Water Use Patterns in Small Disadvantaged Communities," *water,* vol. 13, 2021.

[3] J. Y. Choi, T. Lee, A. B. Aleidan, A. Rahardianto, M. Glickfeld, M. E. Kennedy, Y. Chen, P. Haase, C. Chen and Y. Cohen, "On the feasibility of small communities wellhead RO

treatment for nitrate removal and salinity reduction," *Journal of Environmental Management,* vol. 250, 2019.