Research Project

PREDICTION OF THE AMOUNT OF RAINWATER TO BE RECYCLED WITH MACHINE LEARNING ALGORITHMS

Berk OZKAN

İZMİR
September - 2023

CONTENTS

ABSTRACT

In a world characterized by growing water scarcity, increasing population, and climate change, the effective utilization of rainwater can play a key role in fulfilling water demands while minimizing the ecological footprint of water supply systems. Estimating recycled rainwater amounts as a crucial component of sustainable water management practices is important.

This project aims to estimate the daily amount of rainwater per square meter and propose a prediction model for the problem. In this manner, machine learning algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine) are used, and the performance of the prediction model is tested on real-life data. The results show that machine learning algorithms could accurately estimate the daily amount of rainwater per square meter. Thanks to its high estimation success, the developed prediction model is used in a private company.

Keywords: Water Recycled, Rainwater Prediction, Machine Learning

1. INTRODUCTION

Water is a fundamental source of life, a fact that has been known to humanity for centuries. However, in today's world, the value of water resources has surged due to a host of pressing challenges. Climate change, for instance, has brought about unprecedented unpredictability in precipitation patterns, disrupting the traditional rhythms of water availability, and leading to more frequent and severe droughts and floods in various regions. Simultaneously, rapid global population growth, with its escalating demands for water, exerts significant stress on our already overburdened water resources. The world's population is expanding at an alarming rate, primarily driven by urbanization, which intensifies the demand for water in both residential and industrial sectors. As urbanization continues to draw more people into cities, the need for water for drinking, and industrial processes escalates, underscoring the urgent necessity for effective water resource management. Furthermore, the relentless expansion of industrial activities compounds the pressure on our invaluable water resources. Industries heavily rely on water for manufacturing, cooling, and various processes, often resulting in substantial water withdrawals and pollution. The discharge of pollutants into water bodies not only compromises the quality of available water but also poses a threat to the health of aquatic ecosystems. In response to these challenges, the preservation, effective management, and sustainable utilization of existing water resources have risen to the forefront of global priorities. Water resource preservation is no longer a matter of choice but an essential imperative to ensure the well-being of human societies and the health of ecosystems dependent on these vital resources. As we confront an increasingly uncertain future regarding water availability, the responsible stewardship of our water resources remains paramount.

As the importance of preserving and effectively managing water resources continues to grow, Machine Learning has gained considerable popularity for its potential for providing significant benefits to address this challenge. Machine learning offers diverse and extensive application areas in various sectors, including biology, transportation, healthcare, finance, manufacturing, and many others. For example, in the transportation sector, the development of autonomous vehicles is being driven by machine learning. In the healthcare sector, it is instrumental in disease forecasting and drug discovery. Additionally, in the finance sector, it is utilized for market analysis and the detection of fraudulent activities. With its various applications, machine learning enhances the efficiency of numerous industries by improving data analysis and predictive capabilities, while also providing valuable insights into future trends.

Predicting how much water can be recycled in advance is crucial, as it empowers us to make informed decisions and prepare for efficient resource utilization. On the other hand, the significance of water recycling should not be underestimated, as it presents a sustainable solution in response to the growing demand. In this context, machine learning algorithms can demonstrate a highly effective approach. These algorithms provide powerful tools for forecasting future water resource levels, and water demand by analyzing large amounts of data.

The aim of this project is to predict the daily rainwater per square meter. In this regard, machine learning algorithms (k-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine) are employed, and the performance of the prediction model is tested on real-life data. The prediction results show that the prediction model can be used for estimating daily rainwater per square meter. Thus, using the developed model, companies can predict how much rainwater they can recycle and make their plans based on this information.

The following sections of this project have been structured as follows: The second section presents a literature review. In the third section, the machine learning algorithms employed in this project are given. Section 4 presents the application and prediction results. An overall assessment of the study is included in the conclusion section, which also offers some suggestions for future research.

2. LITERATURE REVIEW

This section presents prediction certain studies on rainwater recycling using machine learning methods, arranged in chronological order.

Gan et al. [1] have explored the use of backpropagation neural networks for rainfall prediction. They conducted experiments with a proposed model using a dataset spanning from 1970 to 2000, consisting of 16 meteorological parameters. The model was implemented using the MATLAB neural network platform, achieving a remarkable accuracy of 100% for Backpropagation Network prediction, while regression prediction yielded a 67% accuracy rate.

Chatterjee et al. [2] proposed a rainfall prediction model using Neural Networks based on data from the Meteorological Station in Dumdum, West Bengal, spanning from 1989 to

1995. They employed k-means clustering to organize the data and compared a Hybrid Neural Network with a Multilayer Perceptron Feed-Forward Network classifier. The Hybrid Neural Network demonstrated superior performance, achieving an accuracy rate of 89.54% with feature selection and 84.26% without feature selection, outperforming the Multilayer Perceptron Feed-Forward Network in both cases.

Grace and Suganya [3] have developed a rainfall prediction model based on multiple linear regression using meteorological data from India. In this research, a dataset comprising a total of 4,116 data points was employed, and this model achieved a high level of accuracy, with a success rate of 0.99%. This study has emphasized the significance of rainfall prediction in India and has indicated its contribution to obtaining reliable precipitation forecasts that can offer more effective planning opportunities, particularly in sectors such as agriculture.

Mohammed et al. [4] focused on two types of rainfall predictions, namely short-term and long-term predictions, emphasizing that short-term predictions yielded more accurate results. They collected rainfall data between 1901 and 2015 and analyzed it to understand precipitation patterns in various regions. The tuned Support Vector Machine algorithm was found to provide the best predictions, with a Mean Absolute Error of 4.35 and an R^2 value of 99%.

Uddin et al. [5], the Northwest Pacific Ocean region, has developed a modified cluster-based typhoon rainfall forecasting model with a lead time of 1-6 hours. Various machine learning and deep learning techniques were integrated into the model, and a grid-search cross-validation method was employed to optimize their parameters. The key finding of this study indicates that when combined with the grid-search algorithm, Support Vector Machines outperforms other algorithms in forecasting cluster-based hourly typhoon rainfall, resulting in an improvement in forecast efficiency by 45% to 90%.

Reddy et al. [6] combined the data pre-processing technique called Singular Spectrum Analysis with supervised learning models, namely Least-Squares Support Vector Regression and Random Forest, for rainfall prediction. A prediction model was developed by integrating Singular Spectrum Analysis with Least-Squares Support Vector Regression and Random Forest, and it was compared with traditional methods (Least-Squares Support Vector Regression and Random Forest). Monthly climate data was used in the study. Proposed model producing Root Mean Square of 71.6%.

3. MACHINE LEARNING ALGORITHMS

Machine learning is a subset of artificial intelligence that focuses on developing algorithms that enable computers to learn from data, recognize patterns, and make predictions or decisions without being explicitly programmed. Two fundamental concepts in machine learning are Supervised Learning and Unsupervised learning. In supervised learning, the algorithm is trained on a labeled dataset, where the input data is paired with corresponding output labels. The goal is to learn a mapping from input features to output labels so that the algorithm can make predictions or classify new, unseen data. Whereas Unsupervised learning, the algorithm deals with unlabeled data or data where the output is unknown or not provided during training. The primary goal of unsupervised learning is to discover patterns, structure, or relationships within the data without specific guidance. Another important distinction in machine learning is based on the type of target. If the goal is to predict a continuous or numeric output, then Regression algorithms are used. However, if the goal is to categorize data into discrete classes or labels, then the Classification algorithms are used. Since the aim of the Project is the predict the amount of the daily rainwater per square meter, regression algorithms are employed on the dataset. In the rest of this section, the machine learning algorithms used in the project are introduced.

3.1 k-Nearest Neighbors Algorithm

k-Nearest Neighbors (k-NN) algorithm [7] works by finding the ‘k’ nearest data points to a given point in a dataset and then makes predictions based on the average or weighted average of the target values of those nearest neighbors. It’s like asking your closest friends for advice you make a decision based on what your friends would do. k-NN can be applied to various real-world problems, from predicting house prices to weather forecasts. It’s a great introduction to the world of machine learning and data science, offering a simple yet powerful way to make predictions based on existing data.

3.2 Decision Tree Algorithm

Decision Tree (DT) algorithm [8] are powerful tools for classification and regression tasks. They’re easy to understand and interpret, making them a great choice for solving various real-world problems, including medical diagnosis, credit scoring, and more. They provide a structured and intuitive way to make decisions based on data, which is why they’re a key concept in machine learning. Decision Trees are a visual representation of a decision-making process. They work by breaking down a complex problem into a series of simpler decisions based on the input features, much like a flowchart. Each decision is represented as a node, and the branches leading to other nodes represent different possible outcomes.

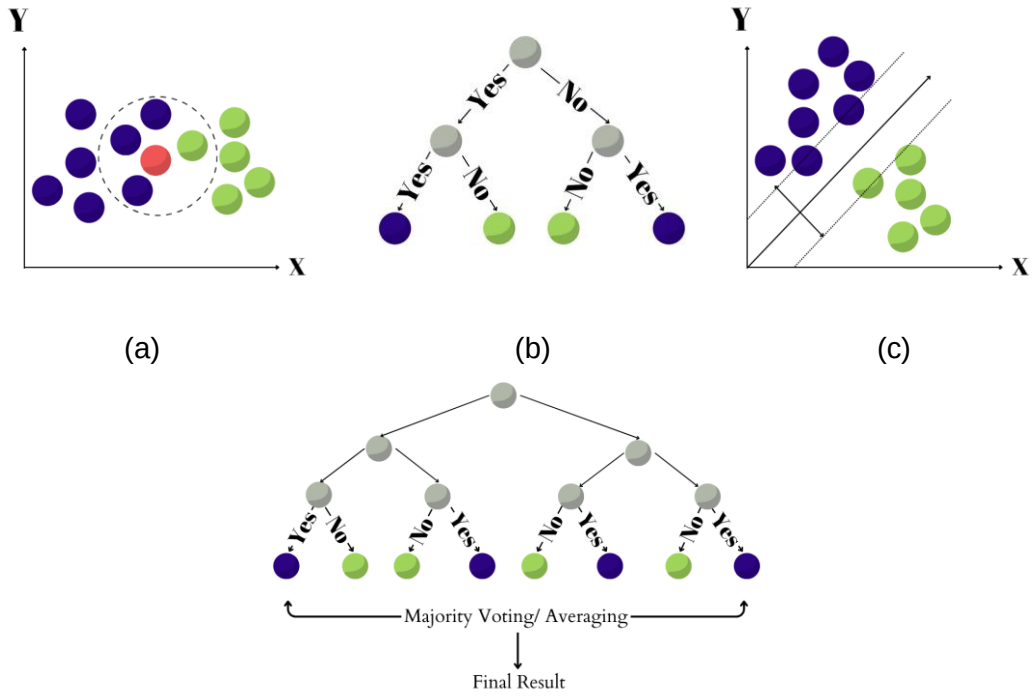
3.3 Random Forest

Random Forest (RF) [9] is an ensemble learning method, which means it combines the power of multiple decision trees to make more accurate predictions. RF builds a ‘forest’ of many decision trees, each trained on a random subset of the data and using a random subset of the features. These individual trees are like a group of experts, each offering their opinion. When you want to make a prediction, the Random Forest combines the predictions from all these trees, either through majority voting or averaging.

3.4 Support Vector Machine Algorithm

Support Vector Machine (SVM) [10] is a powerful algorithm used for classification and regression tasks. At its core, SVM is all about finding the best possible decision boundary that separates different classes in your data. SVM’s objective is to find the hyperplane that maximizes the margin between these classes. SVMs have a wide range of applications, including text classification, image recognition, and biological data analysis. They are a fundamental concept in machine learning that helps solve challenging problems by finding the optimal boundary to distinguish between different classes in your data.

Figure 1 illustrates the conceptual drawings of the algorithms used.



(a) (b) (c)

(d)

Figure 1. Representative illustrations of machine learning algorithms: a) k-NN, b) DT, c) SVM, and d) RF, respectively.

4. APPLICATION

In this section, the prediction results of the applied algorithms on real-life data are provided. To execute the algorithms, Python 3.9.7 programming language and associated packages were used. Additionally, Microsoft Excel (Microsoft 365) was employed as the database due to the convenience it offers in terms of data control and operations that can be applied.

4.1 The Dataset

This study aims to predict monthly rainfall amounts based on daily meteorological data from the Menemen region of Izmir, Turkey. The dataset used includes daily measurements of meteorological variables representing the weather conditions in the Menemen region, including Vapor Pressure (VP), Minimum Daily Temperature (MDT), Mean Actual Pressure (MAP), Mean Relative Humidity (MRH), and Daily Total Precipitation (DTP). The dataset employed in this study comprises observed values spanning the years 2021 to 2023, encompassing a total of 860 rows, each representing daily values. However, due to privacy concerns, this specific dataset will not be disclosed.

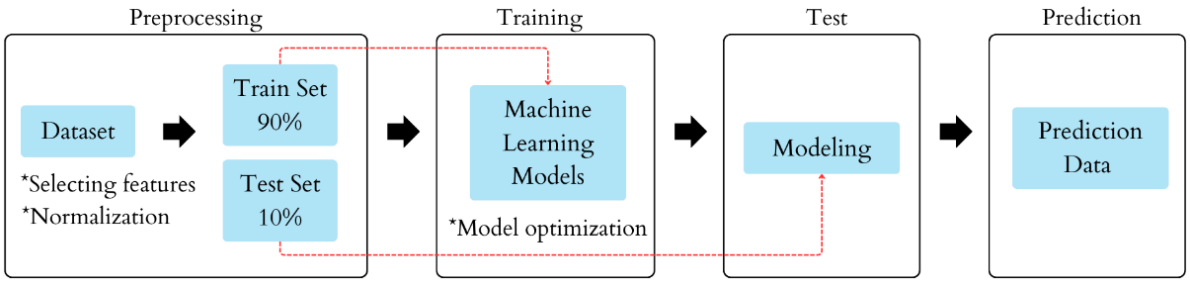


Figure 2. Flow chart to predict rainfall.

To assess the effectiveness of the algorithms, the dataset was partitioned into 90% for training and 10% for testing (Figure 2). Consequently, the initial 774 rows served as the training dataset, while the remaining 86 rows constituted the test dataset. Table 1 displays the first and last 5 rows of the dataset (Note that these values are random numbers not the actual values because of the privacy concerns). Scaling has also been done on the data to get better performance from the algorithms. For this reason, the variables in the dataset were brought to the same metric by taking values between 0 and 1.

Table 1. The sample dataset.

	VP	MDT	MAP	MRH	DTP
0	10.5	9.7	1018.5	83.7	0
1	8.7	8.8	1017.7	74.6	0
2	10	9.2	1020.3	70.5	0
3	10.1	9.5	1019.6	81.5	12.1
4	9.8	8.8	1018.4	81.3	0
…	…	…	…	…	…
856	15.4	17.2	1006.7	52.6	0
857	16.3	19.5	1005.3	55.4	0.5
858	15.9	18.4	1009.7	56	0
859	15.3	17.6	1008.1	57	0
860	17.6	18.3	1004.8	55.7	0.7

4.2 Fine-Tuning

Hyper-parameters of the machine algorithms have a big impact on the prediction performance. In this project, a comprehensive exploration of parameter configurations was conducted for algorithms. Specifically, the ‘n_neighbors’ parameter for the k-NN was systematically adjusted within a range of 3 to 7, while the ‘metric’ parameter was varied among ‘minkowski’, ‘manhattan’, and ‘euclidean’. The Decision Tree’s ‘max_depth’ parameter underwent a thorough examination, spanning from 3 to 20. Likewise, for the Random Forest, ‘max_depth’ was explored from 3 to 10, and ‘n_estimators’ ranged from 100 to 1000. The Support Vector Machine was assessed with various kernels, including ‘linear’, ‘poly’, ‘rbf’, and ‘sigmoid’. Specifically, we systematically examined SVR’s performance across a spectrum of degrees, ranging from 1 to 5.

As a result of these analyses, the k-NN algorithm was employed with a parameter setting of ‘n_neighbors’ equal to 7, signifying the utilization of seven nearest neighbors for each prediction. The ‘minkowski’ metric was chosen as the measure for distance calculations. For the Decision Tree, the ‘max_depth’ parameter was set to 4, restricting the depth of the constructed decision tree to a maximum of four nodes. In the case of the Random Forest, the ‘max_depth’ was specified as 3, regulating the maximum depth for individual trees, while the ‘n_estimators’ parameter defaulted to 100. Furthermore, the Support Vector Machine was formulated with a ‘poly’ kernel of degree 5.

4.3 Prediction Performance of the Algorithms

To evaluate the predictive performance of the algorithm, Mean Absolute Error (MAE) (Equation (1)) Coefficient of Determination (R^2) (Equation (2)) metrics are considered.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\bar{y}_i - y_i| \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (2)$$

where, $\bar{y}_i$ is the i-th predicted value; y_i is the actual value of the ith sample, and $\hat{y}$ is the mean of the y values; n represents the total number of samples.

Table 2. Test data error values of machine learning algorithms.

Error Metrics / Models	MAE	R^2
k-NN	1.06	0.81
DT	1.14	0.82
RF	1.49	0.66
SVM	1.56	0.43

The error metrics of the algorithm are given in Table 2. The k-NN algorithm achieved a MAE of approximately 1.06 and an R^2 value of approximately 81%. These results indicate that the k-NN performed reasonably well in predicting the target variable, with an R^2 value of 81% suggesting that approximately 81% of the variance in the data was explained by the algorithm. Similarly, the DT algorithm demonstrated promising results with an MAE of approximately 1.14 and an R^2 value of approximately 82%. These figures suggest that the DT algorithm was effective in capturing the underlying patterns in the data. On the other hand, the RF algorithm exhibited a higher MAE of approximately 1.49 and a lower R^2 value of approximately 66%. These outcomes suggest that the RF algorithm had challenges in providing accurate predictions compared to the k-NN and Decision Tree algorithms. Lastly, the SVM algorithm exhibited less favorable results with a considerably higher MAE of 1.56 and a lower R^2 value of 43%. These outcomes suggest that the SVM algorithm struggled to provide accurate predictions, and its ability to explain the data's variance was limited. Based on the results, the DT algorithm showed the best predictive performance in this project, while the SVM algorithm appeared to lag behind in terms of predictive accuracy.

5. CONCLUSION

Rainwater recycling plays a crucial role in terms of environmental sustainability and the preservation of water resources. This study is aimed at estimating the amount of rainwater to be recycled using machine learning algorithms. In this context, firstly, a literature review was conducted. Subsequently, the prediction performance of the k-Nearest Neighbors, Decision Tree, Random Forest, and Support Vector Machine algorithms for daily total precipitation was examined based on the real-life dataset. The results were subjected to statistical analysis using error metrics, including MAE and R^2. The R^2 values for the algorithms on the test data 81% for the k-Nearest Neighbors, 82% for the Decision Tree, 66% for the Random Forest, and 43% for the Support Vector Machine. The results indicate that the Decision Tree algorithm outperforms the other algorithms used in terms of prediction performance.

This study makes a valuable contribution to the literature in terms of preparing for and implementing preventive measures in regions likely to face water resource challenges, using machine learning algorithms within the scope of sustainable water resource management. Additionally, thanks to its high estimation success, the developed prediction model is used in a private company.

ACKNOWLEDGMENTS

I would like to thank Asist. Prof. Onur UGURLU for his valuable suggestions throughout the project.

REFERENCES

1. Gan, X., Chen, L., Yang, D., & Liu, G. (2011, September). The research of rainfall prediction models based on matlab neural network. In 2011 IEEE International Conference on Cloud Computing and Intelligence Systems (pp. 45-48).
2. Chatterjee, S., Datta, B., Sen, S., Dey, N., & Debnath, N. C. (2018, January). Rainfall prediction using hybrid neural network approach. In 2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom) (pp. 67-72). IEEE.

3. Grace, R. K., & Suganya, B. (2020, March). Machine learning based rainfall prediction. In 2020 6th International conference on advanced computing and communication systems (ICACCS) (pp. 227-229). IEEE.
4. Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. International Journal of Scientific and Technology Research, 9(01), 3236-3240.
5. Uddin, M. J., Li, Y., Sattar, M. A., Liu, M., & Yang, N. (2022). An Improved Cluster-Wise Typhoon Rainfall Forecasting Model Based on Machine Learning and Deep Learning Models Over the Northwestern Pacific Ocean. Journal of Geophysical Research: Atmospheres, 127(14), e2022JD036603.
6. Reddy, P. C. S., Yadala, S., & Goddumarri, S. N. (2022). Development of rainfall forecasting model using machine learning with singular spectrum analysis. IIUM Engineering Journal, 23(1), 172-186.
7. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.
8. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. Wadsworth International Group.
9. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.