Rainfall Prediction using Decision Tree: A Case Study of CST, Phuentsholing

Manoj Chhetri[1], Lily Gurung[2,]

[1]Information Technology Department College of Science and Technology, Royal University of Bhutan
[2]Civil Engineering Department, College of Science and Technology, Royal University of Bhutan
Email: manoj_chhetri.cst@rub.edu.bt[1],lilygurung.cst@rub.edu.bt[2]

Abstract
In this study we perform hourly rainfall prediction. Climatic data is chaotic in nature and performing regression analysis for short time periods, using limited data recorded by the weather station does not yield good results. Hence, in this study we consider rainfall prediction as a binary classification problem and classify rainfall events into two classes: rainy (positive class) or non-rainy (negative class). Using the independent climatic parameters of the current hour the rainfall status of the next hour is predicted. The dataset used was collected from CST weather station and contains records of 8 weather parameters recorded hourly. We want to study the usability of this data collected by CST weather station for predictive tasks. Since, there is no baseline prediction result on this dataset, we used logistic regression as the baseline model. The accuracy score of logistic regression was 73%. Decision tree which is the focus of this study to perform binary rainfall classification is a popular supervised machine learning algorithm, which forms a flowchart like structure where each internal node represents a feature. The optimization of parameters was conducted through grid search and we used k-fold validation with k value of five and we achieved an accuracy score of 79 percentage.

Key words: *Timeseries, Decision tree, Rainfall, Machine learning, Logistic regression*

1. INTRODUCTION

Rainfall prediction is an important field of research as it has significant implications for agriculture, water resource management, and disaster preparedness. Decision trees are a popular machine learning technique that has been widely used for rainfall prediction due to their ability to handle complex data patterns and provide interpretable results.

Phuntsholing experiences a subtropical climate with distinct seasons. The summers, from June to August, are hot and humid with temperatures ranging from 25 to 35 degrees Celsius. Monsoon rains are common during this time, with heavy rainfall contributing to the lush greenery of the region. The monsoon season brings about abundant precipitation, which can sometimes result in flash floods and landslides in the surrounding areas.

The author of this paper has previously worked on monthly prediction of rainfall data using the Simtokha dataset (Manoj et al.,2020) and acknowledge that the climatic data is very chaotic and the parameters that we have is not sufficient enough to perform an hourly regression study (finding exact amount of rainfall in an hour). Hence, we approach rainfall prediction as a binary classification problem, where we aim to classify rainfall events into two classes: rainy (positive class) or non-rainy (negative class). We employ decision tree algorithms, which are popular and interpretable machine learning techniques, to develop predictive models for rainfall prediction. We focus on a specific case study in the CST (College of Science and Technology) as a weather station is situated in the college.

In this research paper, we present a case study on rainfall prediction using decision trees in the context of CST, Phuntsholing. We aim to investigate the effectiveness of decision tree models in predicting rainfall in the CST region based on historical weather data.

2. LITERATURE REVIEW

Singhal and Tiwari (2019) conducted a comprehensive review of various decision tree algorithms, such as C4.5, CART, Random Forest, and Gradient Boosting, used for rainfall prediction. The authors discussed the methodologies employed in different studies, including data preprocessing techniques, feature selection, and model evaluation methods. They also highlighted the advantages and limitations of decision tree algorithms for rainfall prediction. The review concluded that decision tree algorithms are effective in capturing complex relationships in rainfall data, but their performance may vary depending on the dataset and algorithm used.

Geetha and Nasira (2014) used decision tree to predict weather phenomena like fog, cyclones, rainfall, and thunderstorms. using the open-source data mining tool Rapidminer. They trained the decision tree on data from 2013 and used 2014s data for testing. They got an accuracy score of 80.67 percentage.

Basha et al. (2020) compared performance of ARIMA, artificial network, support vector machine and logistic regression to predict rainfall. They proposed a method of rainfall prediction using a combination of a neural network and an auto-encoder. Their proposed model gave better RMSE and MSE scores compared to other standalone models.

Abhishek and Reddy (2020) conducted a comprehensive review of decision tree-based rainfall prediction models specifically for agricultural applications. The authors discussed the use of decision tree algorithms, such as ID3, C4.5, and CART, in predicting rainfall for agricultural decision-making. They highlighted the importance of input features, such as temperature, humidity, wind speed, and solar radiation. The review concluded that decision tree-based models are useful for agricultural applications, providing accurate rainfall predictions that can aid farmers in making informed decisions.

Khamparia, Singh, and Singh(2018) proposed a rainfall prediction model using decision tree algorithms, namely C4.5 and CART (Rutkowski et al, 2014) along with ensemble learning approaches, such as Random Forest and Gradient Boosting. The authors compared the performance of these algorithms on a rainfall dataset and found that the ensemble learning approaches outperformed the standalone decision tree algorithms in terms of accuracy and prediction performance. The study concluded that the ensemble learning approaches, in combination with decision tree algorithms, can improve rainfall prediction accuracy.

Patil, D et al. (2017) compared the performance of decision tree algorithms, specifically CART and C5.0, with the Naïve Bayes classifier for rainfall prediction. The authors used meteorological data from a region in India and evaluated the accuracy and prediction performance of the models. The results showed that the decision tree algorithms outperformed the Naïve Bayes classifier in terms of accuracy and prediction performance. The study concluded that decision tree algorithms can be effective for rainfall prediction and can contribute to improved water resource management and agricultural planning.

3. METHODOLOGY

The following methodology was followed for the creation of the dataset.

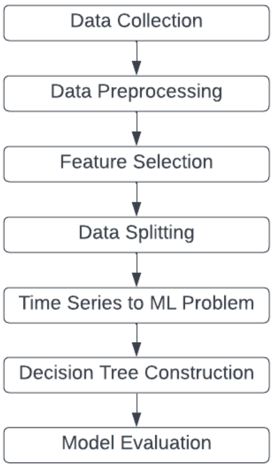
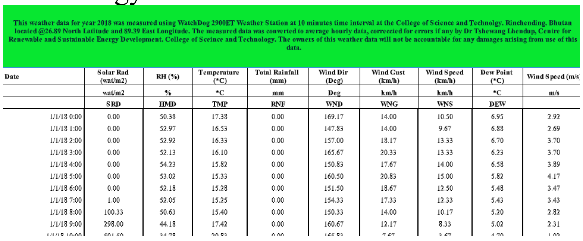


Fig. 1 *Methodology*

3.1. Data collection

The dataset was measured using WatchDog 2900ET Weather Station at 10 minutes time interval at the College of Science and Technology, Rinchending, Bhutan located at 26.89 North Latitude and 89.39 East Longitude. The measured data was converted to average hourly data and corrected for errors if any by Centre for Renewable and Sustainable Energy Development, College of Science and Technology.



This weather data for year 2018 was measured using WatchDog 2900ET Weather Station at 10 minutes time interval at the College of Science and Technolgy, Rinchending, Bhutan located @26.89 North Latitude and 89.39 East Longitude. The measured data was converted to average hourly data, correccted for errors if any by Dr Tshewang Lhendup, Centre for Renewable and Sustainable Energy Development, College of Sceince and Technology. The owners of this weather data will not be accountable for any damages arising from use of this data.

Date	Solar Rad (wat/m2)	RH (%)	Temperature (*C)	Total Rainfall (mm)	Wind Dir (Deg)	Wind Gust (km/h)	Wind Speed (km/h)	Dew Point (*C)	Wind Speed (m/s
	wat/m2	%	*C	mm	Deg	km/h	km/h	*C	m/s
	SRD	HMD	TMP	RNF	WND	WNG	WNS	DEW	
1/1/18 0:00	0.00	50.38	17.38	0.00	169.17	14.00	10.50	6.95	2.92
1/1/18 1:00	0.00	52.97	16.53	0.00	147.83	14.00	9.67	6.88	2.69
1/1/18 2:00	0.00	52.92	16.33	0.00	157.00	18.17	13.33	6.70	3.70
1/1/18 3:00	0.00	52.13	16.10	0.00	165.67	20.33	13.33	6.23	3.70
1/1/18 4:00	0.00	54.23	15.82	0.00	150.83	17.67	14.00	6.58	3.89
1/1/18 5:00	0.00	53.02	15.33	0.00	160.50	20.83	15.00	5.82	4.17
1/1/18 6:00	0.00	52.18	15.28	0.00	151.50	18.67	12.50	5.48	3.47
1/1/18 7:00	1.00	52.05	15.25	0.00	154.33	17.33	12.33	5.43	3.43
1/1/18 8:00	100.33	50.63	15.40	0.00	150.33	14.00	10.17	5.20	2.82
1/1/18 9:00	298.00	44.18	17.42	0.00	160.67	12.17	8.33	5.02	2.31
1/1/18 10:00	501.50	34.78	20.83	0.00	165.83	7.67	3.67	4.70	1.02

Figure 2 Screenshot of dataset

The dataset consists of eight parameters. They are as follow:

- Solar Radiation (wat/m2)
- Relative Humidity (%)
- Temperature (oC)
- Total Rainfall (mm)
- Wind Direction (Deg)
- Wind Gust (km/h)
- Wind Speed (km/h)
- Dew Point (oC)

3.2. Data Preprocessing

The collected dataset was preprocessed using the following data preprocessing pipeline:

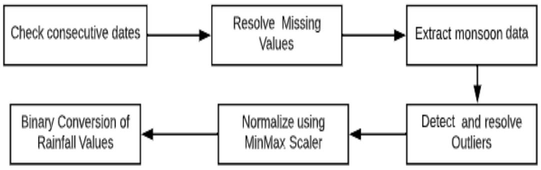

Check consecutive dates
Resolve Missing Values
Extract monsoon data
Detect and resolve Outliers
Normalize using MinMax Scaler
Binary Conversion of Rainfall Values

Figure 3 Data Preprocessing pipeline

The dataset was passed through basic dataset pipeline. The dataset was provided to us as four separate sheets sorted according to the year. The first step involved checking for consecutive dates and hours within each year's dataset to ensure that all records were present.

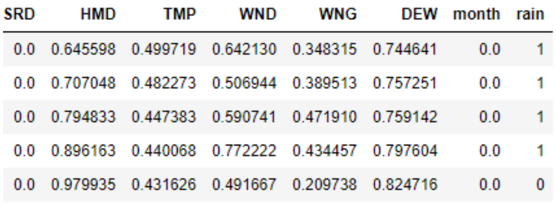
The data was checked for any missing values, and any rows with missing values were removed. The data contained record for every month of the year. Since rainfall normally occurs only during the monsoon season including all the records leads to massive class imbalance problem as there are almost no rainfall during other periods. For the purpose of our study only the monsoon data from July to September was extracted. To identify potential outliers, a basic box plot was utilized. Subsequently, any outliers detected were deleted. Mean imputation was not conducted as it can led to mean bias in the dataset (Bakker et al., 2014).

The data of different parameters are of varying ranges. This can lead machine learning models to learn that parameters with higher range have more impact on the rainfall. Therefore, all the values were normalized in the range of 0-1 using a simple min-max normalization function.

The rainfall values were converted to binary classes having hours with rainfall and hours without rainfall. In our study we want to predict whether it will rain in the next hour or not given the independent variables of the present hours.

3.3. Feature Selection

For feature selection a basic correlation test was performed. But since the target class is a binary value correlation results are difficult to interpret and hence p-value tests was conducted. But after experimental results the feature selection ended by removing the parameter having a lot of missing or noisy data. From the collected dataset with 8 parameters, we dropped wind speed and included an addition parameter of month.



SRD	HMD	TMP	WND	WNG	DEW	month	rain
0.0	0.645598	0.499719	0.642130	0.348315	0.744641	0.0	1
0.0	0.707048	0.482273	0.506944	0.389513	0.757251	0.0	1
0.0	0.794833	0.447383	0.590741	0.471910	0.759142	0.0	1
0.0	0.896163	0.440068	0.772222	0.434457	0.797604	0.0	1
0.0	0.979935	0.431626	0.491667	0.209738	0.824716	0.0	0

Figure 4 Normalized independent variables

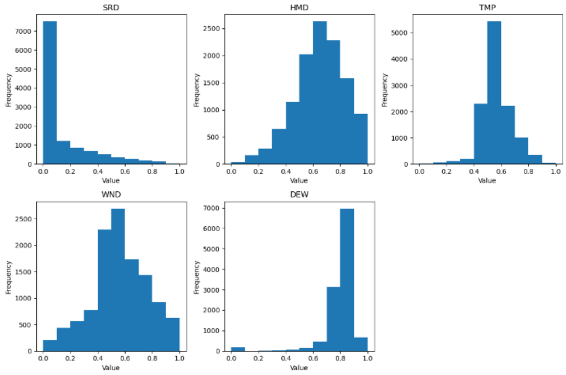


***Fig. 5** Histogram of continuous parameters*

3.4. Data Splitting

Machine learning usually involves converting the dataset into training and testing sets. In our research we perform an 80-20 split whereby 80% of the dataset were used for training and the remaining 20% were used for testing.

3.5. Time Series to Machine Learning Problem Conversion

The recorded dataset is a timeseries dataset and all the records are ordered chronologically, with a timestamp associated with each observation. Machine learning models like decision tree requires the dataset to consist of a set of independent variables(y) and a dependent variable(x). Using the variables in set 'y' the decision trees calculate 'x'. Since we want to predict the status of rainfall in the next hour using the parameters which are currently available to us, we convert the timeseries to a machine learning problem by shifting the binary rainfall records by a timestamp of one. So, the climatic parameters available becomes independent variables in set 'y' and the status of rainfall becomes the dependent variable 'x'.

3.6. Decision Tree Construction

The decision tree (Myles et al.,2014) was created using the sklearn library. The training set was used to create the decision tree using the CART algorithm which uses Gini index as a metric to evaluate the split of a feature node in the decision tree. The spits are made such that the objective is to minimize the Gini index value.

3.7. Model Evaluation

The evaluation metric used in our study is accuracy. In order to improve the generalization of our results we used k-fold (Anguita et al.,2012) validation with a k value of 5. With a k value of 5, it means that the dataset is divided into 5 equally sized "folds" or partitions. The model is trained on 4 of the folds (i.e., 80% of the dataset) and evaluated on the remaining 1-fold (i.e., 20% of the dataset). This process is repeated 5 times, with each fold used as the evaluation set once, while the remaining folds are used for training. The k value of 5 was chosen as we wanted the training and test set to be divided into splits of 80% and 20% respectively.

4. RESULTS AND CONCLUSION

We conducted multiple experiments with different maximum depth values of the decision tree. We achieved the highest accuracy of approximately 79% at the maximum depth value of four (4). Since there was no baseline score on the dataset, we used logistic regression as the baseline model for comparison. The accuracy score of logistic regression was 73%. Decision tree was able to outperform logistic regression but not by a huge margin. The authors conclude further improvement in accuracy score can be achieved using decision tree if the size of the dataset is increased by incorporating more parameters but it is out of the scope of this work as one of the main aims of the project is to study the usability of data collected by weather station. Since, Decision tree also have an issue of data overfitting and furthermore they are very sensitive to small changes in the training set. In the future we will basic MLP and other advanced deep learning models and compare its results with the result from the decision tree.

5. ACKNOWLEDGEMENT

The authors would like to thank Mr.Gom Dorji, electrical department for providing us the dataset and the authors would further like to acknowledge the help and guidance provided by the research officer of the College of Science and Technology.

REFERENCES

Abhishek, A., & Reddy, V.V. (2020). Decision tree-based rainfall prediction models for agricultural applications: *A review. Computers and Electronics in Agriculture*, 169, 105184. doi:10.1016/j.compag.2019.105184

Alam, A., Vennila, S., & Basha, A.A. (2020). Decision tree-based rainfall prediction using meteorological data: A case study of Tamil Nadu, India. *Journal of Water and Land Development,* 44(1), 33-42. doi:10.24425/jwld.2020.131649

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012, April). The'K'in K-fold Cross Validation. *In ESANN* (pp. 441-446).

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychological methods,* 19(3), 409.

Basha, C. Z., Bhavana, N., Bhavya, P., & Sowmya, V. (2020, July). Rainfall prediction using machine learning & deep learning techniques. *In 2020 international conference on electronics and sustainable communication systems (ICESC) (pp. 92-97). IEEE*

Chhetri, M., Kumar, S., Pratim Roy, P., & Kim, B. G. (2020). Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan. *Remote sensing*, 12(19), 3174.

Geetha, A., & Nasira, G. M. (2014, December). Data mining for meteorological applications: Decision trees for modeling rainfall prediction. *In 2014 IEEE international conference on computational intelligence and computing research* (pp. 1-4). IEEE.

Khamparia, A., Singh, U., & Singh, R. (2018). Rainfall prediction using decision tree and ensemble learning approaches. *In 2018 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) (pp. 529-534). IEEE.* doi:10.1109/ICIoTCT.2018.8479573

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: *A Journal of the Chemometrics Society*, 18(6), 275-285.

Newbold, P. (1983). ARIMA model building and the time series analysis approach to forecasting. *Journal of forecasting*, 2(1), 23-35.

Patil, D., & Padole, P. M. (2017). Rainfall prediction using decision tree and Naïve Bayes classifier. *In 2017 3rd International Conference for Convergence in Technology*

(I2CT) (pp. 1543-1548). IEEE. doi:10.1109/I2CT.2017.8226250

Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences,* 266, 1-15.

Singhal, A., & Tiwari, S. (2019). Rainfall prediction using decision tree algorithms: Rainfall prediction using decision tree algorithms: *A review. International Journal of Environmental Science and Technology,* 16(3), 1329-1342. doi:10.1007/s13762-018-1971-6