The International Journal of Engineering and Science (IJES)
|| Volume || 11 || Issue || 2 || Series I || Pages || PP 13-16 || 2022 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

Machine Learning Techniques for Atmospheric data using WEKA

P. Venkata Ramana Moorthy [1], B. Sarojamma [2] and S. Venkatramana Reddy [1*]
[1]*Department of Physics, S.V. University, Tirupati-517502, A.P., India.*
[2] *Department of Statistics, S.V. University, Tirupati-517502, A.P., India.*

----------ABSTRACT----------
Wind speed plays a vital role for rainfall and wind energy is produced with low cost using wind turbines. Wind speed is related with temperature, if temperature is high wind speed is also high and when temperature is low wind speed decreases with high moisture in air. In this paper we are fitted different regression models like Linear Regression, K-nearest neighbours, Reptree and Support Vector Regression using WEKA software. Variables taken for interpretation is wind speed as dependent variable and day wise date, temperature, visibility as independent variables from 1.1.2017 to 1.1.2019. Various measures of accuracy used are Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, Root Relative Squared Error.

Keywords: *Wind speed, regression models, MAE, RMSE, RAE, RRSE.*

Date of Submission: 05-02-2022 Date of Acceptance: 18-02-2022

I. INTRODUCTION

There are various atmospheric variables and some of them are Wind speed, Temperature, High Temperature, Temperature at different points, visibility, Rainfall, Humidity, wind gust, Precipitation etc. Wind speed plays a vital role for production of Electricity using wind Turbine with low costs. Wind speed is increased and decreased with temperature because of moisture in air.

Tayeb Brahimi et al[1] wrote an article "Prediction of Wind Speed Distribution Using Artificial Neural Network: The Case of Saudi Arabia". They explains prediction of the aerodynamic loads and performance of Wind Turbines using Artificial Neural Networks(ANN) methods to forecast and interpret. The variables taken for the training period is the time of the day, the year, Latitude and Longitude, air temperature, wind direction, Humidity and Pressure. For learning points of 60% of the training set, 40% of testing. They evaluated correlation Coefficient and Root Mean Square Error. WEKA software is used for numerical validation with data from Meteorological stations to models.

Somaieh Ayalvary et al [2] discussed in the paper "Select the most relevant input parameters using WEKA for models Forecast for Solar radiation based Artificial Neural networks". The variables taken for predictive accuracy is Monthly air Temperature, the average Minimum Temperature, average Maximum Temperature, Maximum windspeed, Maximum daily Rainfall, Wind, Rain and Latitude for using Neural Network models like ANN-1, ANN-2, ANN-3. They conclude that ANN-2 is the best model for prediction.

Xin Wang et al[3] in their paper "Mine fire prediction based on WEKA Data Mining", collected data on Temperature, Residual gas, Wind speed,O_2 concentration and Dangerous degree according to none, weak, medium and strong expressed as 1, 2, 3 and 4 respectively. They used three algorithms such as SUM, BP Neural Network and J48 Decision Tree. Ferreira et al[4] explained "Short Term Forecast of Wind speed through Mathematical Models". Kidmo et al[5] provided "Statistical Analysis of Wind speed distribution based on six Weibull Methods for Wind power evaluation in Garova.

II. METHODOLOGY

In this paper, we used Regression Machine Learning Method for interpreting Wind speed by taking time i.e. according to date, Visibility and Temperature. The popular Machine Learning Algorithms [6] are

- Linear Regression
- K-nearest neighbours
- Decision Trees- Reptree
- Support Vector Machine

Linear Regression

It gives the linear relationship of one dependent variable with several independent variables. If Wind speed (y) is dependent variable and time(x_1), Visibility(x_2) and temperature(x_3) variables to fit linear Regression.

$y = \beta_0 x_1 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Where y = wind speed

x_1 = time

x_2 = visibility

x_3 = temperature

β_0, β_1, β_2 are constants.

K-Nearest Neighbours

KNN algorithm is simple and easy to implement Machine Learning Algorithm used for both Regression and Classification problems. It is a simple algorithm that stores all variables and it predicts the target by using measures like Euclidean distance function, Mambattam distance function and Minkowski distance function for continuous variables and in case of categorical variables Hamming Distance function is used.

Decision Tree Regression

For Decision Tree Regression, we use Reptree algorithm. Reptree is the best and fast Decision Tree Learners. It is produced for Classifications or RandomTrees based on C4.5 Algorithms. It improves the model by removing the decisions of the tree that are not important in classification.

Support Vector Regression

To minimize error by maximizes the margin of Linear model by keeping that error is tolerated. General Linear Regression is

$y = \beta x + b$

Minimize $\frac{1}{2}||\beta||^2 + c\sum_{i=1}^{N}(\varepsilon_i + \varepsilon_i^*)$

Subject to constraints $y_i - \beta x_i - b \leq \varsigma + \varepsilon_i$

$$\beta x_i + b - y_i \leq \varsigma + \varepsilon_i^*$$

$$\varepsilon_i \, , \varepsilon_i^* \geq 0$$

Linear Support Vector Regression is

$$Y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\,(x_i \, , x) + b$$

For estimation of above four models we are using WEKA software.

Measures of Accuracy

Different measures used for accuracy of above models i.e., Linear Regression, K-Nearest neighbours, Reptree and Support Vector regression are as follows

(i) **Mean Absolute Error:** Average of absolute error, Error is difference between prediction value and true value.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where MAE = Mean Absolute Error

y_i = prediction value

x_i = true value

n = total number of data points

(ii) **Root Mean Square Error:** Root Mean Square Error is also called as Root Mean Square Deviation. Positive Square root of mean to squared Errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$$

Where RMSE = Root Mean Square Error

y_i= prediction value

x_i = true value

n = total number of data points

(iii) **Relative Absolute Error:** Ratio of Sum of errors with sum of true values multiple with 100%

$$RAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{\sum_{i=1}^{n}|x_i|}\,100\%$$

Where RAE = Relative Absolute Error

y_i = prediction value
x_i = true values
n = number of observations

(iv) **Root Relative Squared Error :** The formula for Root Relative Squared Error is

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Where $\bar{x} = \frac{\sum x_i}{n}$

RRSE = Root Relative Squared error
y_i = prediction value
x_i = true value
n = number of observations

III. EMPIRICAL INVESTIGATIONS

By taking Wind speed as dependent variable and Time day wise, Temperature and Visibility as independent variables, we perform K – nearest neighbours, Reptree, Linear Regression and Support Vector Regression using WEKA software as follows. The data taken for analysis is from 1.1.2017 to 1.1.2019[7].

K-Nearest Neighbours: By taking test models, 4-fold classification, we obtained summary as

Correlation Coefficient	0.1931
Mean Absolute error	2.1923
RMSE	2.786
RAE	121.7697%
RRSE	124.5794%

Reptree: Under Decision tree of Regression, popular model is Reptree, 4-fold cross validation as test mode summary as follows

Correlation Coefficient	0.0249
Mean Absolute error	1.8002
RMSE	2.2363
RAE	99.9875%
RRSE	100.02%

Linear Regression: The fitted linear model by taking Wind speed as dependent and Temperature, Visibility and time daywise is as follows
Wind speed = 0.04 date + 0.001 temperature + 0.0001 Visibility + 0.2801
Summary for linear regression is as follows

Correlation Coefficient	-0.0923
Mean Absolute error	1.8018
RMSE	2.2373
RAE	100.0756%
RRSE	100.0432%

Support Vector Regression: 4 fold cross validation as test mode and variables, we are taken as Daywise time, Temperature and Visibility as independent and Windspeed as dependent. The equation with normalized weights is

Y_t = 0.032(1|1|2017)-0.1063(2|1|2017)+.....-0.006(12|03|2018)
+0.1145(1|1|2019)+0.0299 Temperature + 0.1993 Visibility +
0.0968.

Summary of the model is as follows

Correlation Coefficient	0.3906
Mean Absolute error	1.
RMSE	2.
RAE	90.8
RRSE	92.033

IV. SUMMARY AND CONCLUSIONS

By taking daywise time, temperature, Visibility as independent variables and Wind speed as dependent variable, We perform various Regression models such as K-Nearest Neighbours, Reptree, Linear Regression and Support Vector Regression by taking test mode as 4-fold Cross Validation using WEKA software. The best model among these four models is choosed by accuracy measure Root Mean Square error criterion. Below table explains the model and their corresponding RMSE values.

Model	RMSE
K Nearest Neighbours	2.786
Reptree	2.2363
Linear Regression	2.2373
Support Vector Regression	2.0581

Best model among four Regression models is Support Vector Regression model with least Root Mean Square error value.

REFERENCES

[1]. Tayeb Brahimi, Fatina Alhebshi, Heba Alnabilsi, Ahmed Bensenouci, Mumu Rahman, Procedia Computer Science **163**,41–48 (2019).
[2]. Somaieh Ayalvary, Zohreh Jahani and Morteza Babazadeh, Advances in Computer Science: an International Journal, **4(6)**, 38-44 (2015).
[3]. Xin Wang, Jian Hao, Jun Chen, Weijia Cheng, IOP Conf. Series: Earth and Environmental Science **384,** 012164 (2019).
[4]. M. Ferreira, Santos, A., Lucio, P., Preprints **2018**, 2018070501.
[5]. D.K. Kidmo, R. Danwe, S.Y. Doka, and N. Djongyang, Revue des Energies Renouvelables, **18(1),** 105 – 125 (2015).
[6]. H. Simon, Neural Networks:A Comprehensive foundation(Prentice Hall of India, New Delhi) 2008.
[7]. WWW.VisualCrossing.Com

P. Venkata Ramana Moorthy, et. al. "Machine Learning Techniques for Atmospheric data using WEKA." *The International Journal of Engineering and Science (IJES)*, 11(2), (2022): pp. 13-16.