International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**8**, pp.**25-28**, August **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

An Investigation into the Applications of Machine Learning Algorithms on Wind Speed Prediction

Nitesh Kothari[1*]

[1]Greenwood High International School, Bengaluru, India

**Corresponding Author: niteshkothari2006@gmail.com*

Received: 25/Jun/2024; **Accepted**: 27/Jul/2024; **Published**: 31/Aug/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i8.2528

Abstract: In recent times, wind energy is a highly demanded source of renewable energy today. Consequently, global demand for wind energy has increased and thus the construction of wind turbines. However, wind turbines are often met with unfavorable conditions such as highly erratic and variable wind speeds or even storms. This has further consequences such as greatly reducing the efficiency of wind turbines and leaving its body damaged which is economically unfavorable. Particularly, wind speed prediction is a steady variable to consider while looking for viable options to increase the power generation from wind turbines. This paper aims to assess the performance of various machine learning models in time-series wind speed prediction. My hypothesis is that among the machine learning models tested, Random Forest Regression will outperform the others in predicting wind speed. After training and testing the data, I found out that Random Forest Regression had the best performance with a mean squared error of 5.64 and mean absolute error of 1.81. It also had the highest coefficient of determination of 0.68 and supported my hypothesis. Thus, these results show how machine learning models are reasonable tools for wind speed prediction as well as that Random Forest Regression can be used for real-time wind speed prediction after some hyper parameter tuning. This has major implementations as the model can be used to increase the efficiency of wind turbines, improve their safety and help in maintenance planning.

Keywords: Machine Learning, Wind Speed, Artificial Intelligence, Wind Energy, Sustainability, Renewable Energy

1. Introduction

In recent years, wind energy produced by turbines has emerged as one of the most significant sources of renewable energy. Notably, it is one of the fastest growing and advancing net-zero carbon sources of energy [1]. With global warming at its peak and fossil fuel reserves depleting at an extremely fast rate, there is worldwide need to switch to renewable sources of energy and meet the rising power demands. Wind energy, generated by wind turbines, has major implications to solve this problem, as wind is an abundant and inexhaustible resource.

Accurately predicting wind speed is a reliable approach to optimizing power generation in wind turbines. Forecasting wind speed enables model predictive control for wind turbines [2] and is crucial for the management of electrical grids [3]. It is also vital in reducing the expenditure to run electrical power systems [4]. However, the random and unsteady nature of wind speeds can create uncertainties. Therefore, advanced technologies have to be implemented to minimize uncertainties and accurately predict wind speed.

Machine Learning is one major technology that has been implemented for time-series wind speed prediction. Machine Learning is defined as the branch of Artificial Intelligence that imitates the human's learning behavior and uses algorithms to attempt to predict data. This branch has massive ascendancy as it has the ability to form sustainable algorithms from a large set of data, a key factor in the case of wind speed prediction. The different machine learning models tested in this paper include Linear Regression, K – Nearest Neighbors, Decision Trees, XGBoost and Random Forest Regression.

This is categorized as a supervised learning problem, where input data is provided to the model, which is then trained to predict wind speed, a quantitative variable. The input data in this case would be the past wind speeds recorded which come under the training data. Specifically, this is categorized as a time-series forecasting problem, where models predict future data based on previously inputted data. Time is the independent variable in this case, so the regression models are made to look at trends over time and predict the future data.

The dataset used in this research paper came from the Kaggle and was named "Wind Turbine Scada Dataset". The dataset comprised of 5 quantitative variables that the sensors detected every 10 minutes. This included the Date/Time, Wind Speed, Wind Direction, Active Power, and Theoretical Power Curve. The entire dataset was recorded in the form of a CSV file.

There were a total of 50,530 samples recorded. This is a healthy number of observations and will suffice for all the machine learning models used in this paper.

This research paper aims to investigate the performance of different machine learning models in the context of wind speed prediction. My hypothesis is that Random Forest Regression will outperform the other machine learning models tested, including Linear Regression, K-Nearest Neighbors, Decision Trees, and XGBoost, in terms of accuracy and predictive power. This is because Random Forest Regression can handle non-linear data with precision because of its decision tree structure. After testing all the regression models, the results attribute that Random Forest Regression had the highest R2 Score and the lowest errors of all models as well. Furthermore, this can indicate that after some hyper parameter tuning, the model can be used to predict real-time wind speeds

Rest of the paper is organized as follows, Section 1 contains the introduction of to wind speed prediction and machine learning, Section 2 contain the related work of ensemble methods used in the past to predict wind speed, Section 3 outlines the methodology and procedures, Section 4 presents the results and discussion, and Section 5 provides a conclusion of the research and suggests future directions.

2. Related Work

Attempts in the past have been made to accurately predict the wind speed. Monfared et al. proposed the use of artificial intelligence to predict wind speed. He highlighted the application of fuzzy logic for predicting wind speed, noting that these methods proved to be more accurate than traditional statistical approaches. [5]. Researchers also delved deeper into the use of multilayer perceptron neural networks and the extreme learning models along with nearest neighbors approach to make a time-series prediction of the wind speed. His research touched upon the utilization of machine learning models to effectively predict wind speed, as it is a major factor in maximizing the power supply from wind turbines [6]. Research was also done to show how Artificial Neural Networks could be used in wind speed prediction [7]. Researchers also utilized Data Mining Methods for Wind Speed Prediction with Weka [8].

3. Procedure

3.1 Data Preprocessing and Formatting

Before importing the data, some unnecessary elements of the dataset had to be removed. Since this was specifically a time-series wind speed prediction problem, all elements in the dataset excluding the Date/Time and Wind Speed (m/s) were removed. The data was imported with 50,530 columns. Along with the data, some essential Python Libraries were imported. The essential libraries pandas, NumPy, scikit-learn were imported as well.

The Date/Time variable had to be formatted in the format days, months, years, hours and minutes. This would make a new column with the heading “date_formatted”. After trying to work with the “date_formatted” sequence, it was found that scikit-learn doesn’t work too well with datetime64 datatype. So, the date_formatted column was converted into an object using the toordinal () method which is from the pandas library.

3.2 Data Plotting

Next, plotting the data would give a good visualization of the data being analyzed. The matplotlib and seaborn libraries have to be used for this. Therefore, matplotlib library was imported as plt and seaborn was imported as sns. The seaborn lineplot was used to graph the data. The “date_formatted” column went on the X axis, as it is the independent variable. The wind speed was plotted on the Y axis, as it is the dependent variable.(Figure 1).

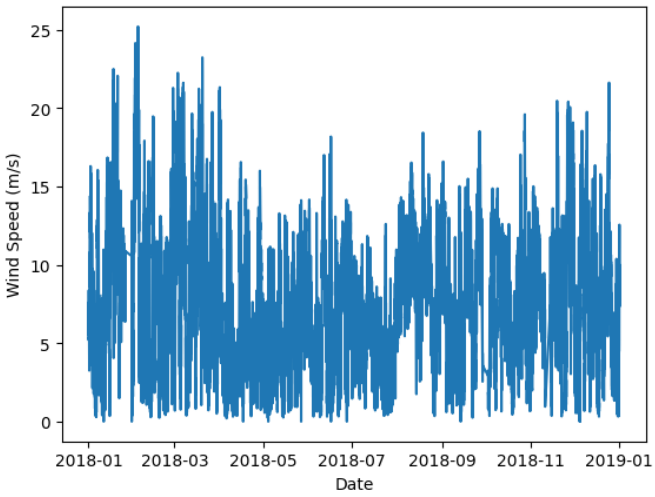

Figure 1. Visualization of the dataset.

3.3 Splitting the data

After this, the data was divided into 2 parts which is training and testing data. The training data consisted of 80% of the dataset, which contained both the time and the wind speed. This data was fitted into the model. The model was then tested on the testing data, which is 20% of the dataset. To do this, the train_test_split () function from the scikit-learn was used. This was ,then, used to initialize four variables that are X_train, X_test, Y_train and Y_test. The X variable was assigned to time and the Y variable was assigned to wind speed. As mentioned earlier, the test size was defined as 0.2 which is 20% of the dataset.

The regression metrics were then imported from the scikit-learn library. All models that were created in this research paper were evaluated on the basis of these three metrics. These include mean squared error, mean absolute error and R2 Score.

3.4 Regression

Next, all machine learning algorithms were imported from scikit-learn. Linear regression was built using LinearRegression (), Random Forest was constructed using RandomForestRegressor (), K- Nearest Neighbours was made using KNeighboursRegressor(), XGBoost with

XGBRegressor () and Decision Trees with DecisionTreeRegressor(). All models were trained by using the fit() function and were trained using the X_train and Y_train data. After this, they were tested on the X_test data. Their prediction was stored in a variable and this data was compared to the Y_test variable by the three regression metrics. The results of the regression metrics were printed out using the print() function.

4. Results and Discussion

Table 1. Tabulation of all the Machine Learning models and the results of their regression metrics

Model	MSE	MAE	R2 Score
Linear Regression	17.6	3.40	0.003
Random Forest	5.64	1.81	0.68
KNN	5.81	1.84	0.67
XGBoost	8.15	2.17	0.55
Decision Trees	5.88	1.85	0.67

After testing all models, their performance was tested on three regression metrics and tabulated the results (Table 1). First, the Linear Regression model exhibited the poorest performance. It had the highest MSE and MAE and also a very low R2 Score of 0.003. This can be reasoned as Linear Regression plots the line of best fit through the data and this might be a poor prediction method as wind speed constantly fluctuates and the values can range from very high to very low.

Next, the XGBoost model had a coefficient of determination of 0.55. This shows that its performance is comparatively better than the other models. It has relatively MSE and MAE than Random Forest Regression, Decision Trees and K Nearest Neighbors. The difference in MSE and MAE can signify that model has some large errors and this accumulates in the MSE. Its R2 score of 0.55 indicates that it is much more suitable for wind speed prediction than Linear Regression which is the worst performing model out of the scope covered in this paper. This shows the model's ability to adjust to data ranging from high and low values. Its performance can increase with several hyper parameter tunings.

K-Nearest Neighbors and Decision Trees have almost identical performances. They had a negligible difference in all of their metrics. The K Nearest Neighbors model had optimum performance when the nearest neighbors was equal to a 100. This can be interpreted as grouping 100 variables and takes the average of them to make the next prediction. Decision Trees also had a solid prediction and had similar performance as K Nearest Neighbors. It can be inferred that its tree structure makes it suitable to be a good predictor for wind speed. Both the K-Nearest Neighbors and Decision Trees model had a higher R2 threshold than my hypothesis which was 0.60.

Random Forest Regression was the best performing algorithm. It has a very close but higher R2 score than K-Nearest Neighbors and Decision Trees. It had a slightly less MSE and MAE than the two models. This supported my hypothesis as Random Forest Regression outperformed all other models in all three metrics. Its difference in MSE and MAE also shows how some large errors increased the MSE. The results of this experiment suggests that Random Forest Regression is the most suitable when it comes to time-series wind speed prediction. Furthermore, this can indicate that after some hyper parameter tuning, the model can be used to predict real-time wind speeds. This can also indicate that Random Forest Regression is more promising when it comes to variable and dynamic datasets and can be implemented in other regions such as solar energy or financial markets.

5. Conclusion and Future Scope

This experiment was done to see which machine learning approach would best predict wind speed with time as the only independent variable. Wind speed prediction is a mechanism that has major implications in maximizing the power output in wind turbines. This, in turn, helps increase the efficiency of power generation from renewable sources of energy. The machine learning models were tasked with finding patterns and trends in the data, and attempted to predict the erratic wind speed. The coefficient of determination is a strong regression metric. The coefficient of determination of the best performing model had to be more than 0.60 in order for the hypothesis to be correct and showed good execution in predicting the wind speed.

The dataset was taken from Kaggle, which contained data collected by a wind turbine in Turkey. After some data pre-processing, the data was inputted into Google Colaboratory and split into training and testing data. After this, all machine learning models were trained. They were then used to predict the wind speed and their performance was evaluated through three regression metrics. These included mean squared error, mean absolute error and R2 Score. The R2 Score is also known as the coefficient of determination and is used to interpret how well the independent variable predicts the dependent variable. It is good to keep in mind that the higher the R2 Score, the better the model's prediction. The R2 Score can a maximum value of 1.

The major results of this study showed that Random Forest Regression was the best performing model and Linear Regression was the worst performing model for time-series wind speed prediction [Table 1]. Random Forest Regression had the highest coefficient of determination and also the least error. However, there was also a trend in the results where the MSE was greater than the MAE in all machine learning models that were tested. This shows two things. First, this speaks to the large errors the machine learning models make which make the MSE higher than the MAE. Second, this shows the uncertainty of wind speed as it signifies the variability of the data. The final result showed that the Random Forest Regression had a R2 Score of 0.68, which is more than 0.60, proving my hypothesis correct.

Future questions could combine these models together to see if they work well together by amplifying their positive features and further improving their predictions. They can

also diversify the inputs and account for more variables such as wind speed, hub height etc.

Data Availability
All machine learning model results were empirical in nature and could vary slightly every time they are run.

There can be some possible limitations in the dataset. The dataset was recorded from a wind turbine in Turkey. This is a limitation as the data in different wind turbines in different location and hub heights can vary by shape a lot. This means that other machine learning models could have performed better if the general spread or shape of the graph was different. Another limitation to consider is the lack of hyper parameter tuning in most of the models. This could have allowed for even better performance of many of the models. Furthermore, if there was a variety of inputs, such as wind speed direction, the model could have predicted the wind speed better with more inputs as it would have helped find trends better and also consider more variables.

Conflict of Interest
I do not have any conflict of interest.

Funding Source
None.

Authors' Contributions
Nitesh Kothari was responsible for the literature review, analysis and evaluation of the entire paper. He also wrote the entire research paper.

Acknowledgements
I would like to thank Ms. Lalitha Rao in helping me proofread my paper.

References

[1] Olabi, A. G., et al. "Renewable Energy Systems: Comparisons, Challenges and Barriers, Sustainability Indicators, and the Contribution to UN Sustainable Development Goals." International Journal of Thermofluids, Vol.**20**, p.**100498, 2023.** DOI: https://doi.org/10.1016/j.ijft.2023.100498.

[2] Chen, Kuilin, and Jie Yu. "Short-Term Wind Speed Prediction Using an Unscented Kalman Filter Based State-Space Support Vector Regression Approach." Applied Energy, Vol.**113**, pp.**690–705, 2014.** https://doi.org/10.1016/j.apenergy.2013.08.025.

[3] Pfeifer, Sascha, and Hans‐Jürgen Schönfeldt. "The Response of Saltation to Wind Speed Fluctuations." Earth Surface Processes and Landforms, Vol.**37**, No.**10**, pp.**1056‐1064, 2012.** https://doi.org/10.1002/esp.3227

[4] Elyasichamazkoti, Farhad, and Abolhasan Khajehpoor. "Application of Machine Learning for Wind Energy from Design to Energy-Water Nexus: A Survey." Energy Nexus, Vol.**2**, pp.**100011, 2021.** https://doi.org/10.1016/j.nexus.2021.100011.

[5] Monfared, Mohammad, et al. "A New Strategy for Wind Speed Forecasting Using Artificial Intelligent Methods." Renewable Energy, Vol.**34**, No.**3**, pp.**845–848, 2009.** DOI.org, https://doi.org/10.1016/j.renene.2008.04.017.

[6] Ak, Ronay, et al. "Two Machine Learning Approaches for Short-Term Wind Speed Time-Series Prediction." IEEE Transactions on Neural Networks and Learning Systems, Vol.**27**, No.**8**, pp.**1734–1747, 2016.** https://doi.org/10.1109/TNNLS.2015.2418739.

[7] Elyasichamazkoti, Farhad, and Abolhasan Khajehpoor. "Application of Machine Learning for Wind Energy from Design to Energy-Water Nexus: A Survey." Energy Nexus, Vol.**2**, pp. **100011, 2021**. ScienceDirect, doi:10.1016/j.nexus.2021.100011.

[8] B. Hari Mallikarguna Reddy, S. Venkatramana Reddy, B. Sarojamma, "Data Mining Techniques for Estimation of Wind Speed Using Weka", International Journal of Computer Sciences and Engineering, Vol.**9**, Issue.**9**, pp.**48-51, 2021.**

AUTHORS PROFILE

Nitesh Kothari is a senior at Greenwood High International School, Bengaluru, India. He is an avid tech-pragmatist and is interesting in leveraging artificial intelligence to solve real world problems. From his past experiences, he has had an artificial intelligence internship at the NUS School of Computing and worked under Professor Sanka Rasnayaka to build a customer churn prediction model. He is also an AWS Certified Cloud Practitioner. He is also researching currently in the field of computational biology and is mentored by a UPenn Alum. He continuous to create problem-solving projects and apply them to the real world using entrepreneurship.

