 doi:10.20944/preprints202206.0428.v1

Machine learning-based algorithms for weather forecasting.

Ismaila Oshodi
Department of Computing
Bournemouth University
Bournemouth England
S5423662@bournemouth.ac.uk

Abstract
Weather forecast has a big impact on the global economy, accurate and timely weather forecast is required by all, it affects many aspects of human livelihood and lifestyle, it also plays a critical role in decision making for severe weather management and for primary and secondary sectors like agriculture, transportation, tourism, and industry as they rely on good weather conditions for production and operations.
The erratic and uncertain complex nature of the weather makes traditional weather forecasting tedious and a challenging task, traditional weather forecast involves applying technology and scientific knowledge on numerical weather prediction (NWP), and weather radar to solve complex mathematical equations to obtain forecasts based on current weather conditions.
These traditional processes utilize expensive, complex physical and computational power to produce forecasts, which can be inaccurate and have various catastrophic impacts on society.
In this research, a machine learning-based weather forecasting model was proposed, the model was implemented using 4 classifier algorithms which include Random Forest classifier, Decision Tree Algorithm, Gaussian Naïve Bayes model, Gradient Boosting Classifier, these algorithms were trained using a publicly available dataset from Kaggle for the city of Seattle for the period 2012 to 2015.
The model's performance was evaluated; the Gaussian Naive Bayes algorithm proved to be the best performing algorithm with a predictive accuracy of 84.153 %.

Keywords: Machine learning, weather forecast

1. Introduction

Weather forecasting is the prediction of weather and conditions of the atmosphere for a specific time, weather conditions include rain, snow, temperature, fog, wind etc, they are various techniques for the prediction of weather which includes persistence forecast, climatological forecast, synoptic forecasting, statistical forecasting, computational forecast etc (Murugan et al. 2021). Due to the erratic, uncertain complex nature of the weather, forecasting has been a tedious, challenging task that involves expensive, complex, and computational processes. (Fang et al. 2021).

Weather forecast has a big impact on the global economy, these impacts can be categorized into four general categories, they include low-impact, moderate-impact, high-impact, and extreme-impact (McCarthy 2007), humans depend on weather forecast because it affects many aspects of our livelihood and lifestyle, precise weather forecast plays a critical role in decision making for severe weather management and for primary and secondary sectors like agriculture, transportation, tourism, and industry because they rely on good weather conditions for production and operations. (Singh et al 2019).

Weather forecast is made by collecting quantitative, atmospheric data about the past and current state of the atmosphere (Das et al. 2014) known as determinates, they include temperature, pressure, the humidity of the air, precipitations, wind, and other meteorological elements, these abiotic determinates are inputted into a model base, early success in forecasting was in terms of Numeric weather predictions (Haupt et al, 2018).

Numerical weather prediction (NWP) such as Weather Research and Forecasting (WRF) model uses mathematical equations or models to compute current weather observations collected from the weather station to produce weather predictions either for short term weather forecasts or long-term climate change (Hewage, et al. 201).

These current weather prediction models depend on the complex physical model and require large computational power to run, whose forecasts are often inaccurate (Jakaria et al. 2020), more recently with the advancement of technology and the availability of metrological big data, researchers had begun utilizing data-driven approaches in metrology, data-driven approaches include using machine learning methods and other technological advance methods which have achieved considerable success in weather forecast (Wang B. et al. 2019).

machine-learning integrates data from multiple sources together to produce a prediction as output, it identifies patterns in a set of data and then finds a relationship between those patterns, using a training dataset it learns and adapts to test or validate data without human interventions (Gagne et al. 2017), it has been implemented successfully in other fields like finance, agriculture and cancer diagnosis, Machine learning approach has also been applied to weather forecasting with remarkable success.

Machine learning approach to weather forecast uses quantitative metrological data to build models and tries to improve the performance of the model by learning from the dataset,

In this study, a machine learning model was proposed and implemented, this project aims to use a publicly available dataset from Kaggle to develop a machine learning model for the prediction of weather conditions, the weather conditions predicted in this project include drizzle, rain, sun, snow, and fog.

The main aim of this research is to develop a predictive model for forecasting weather, the objectives include

1) Review related work on weather forecasting

2) Develop an efficient and effective model for weather forecasting.

3) Evaluate the performance of the models developed using accuracy and precision to determine which algorithm is best suited for the prediction of weather conditions.

The rest of the paper is organised as follows: chapter 2 describe related work on weather forecasting that used various machine learning algorithm and methods, chapter 3 describes the data and the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2022 doi:10.20944/preprints202206.0428.v1

methodology used to implement the artificial intelligence approach, and chapter 4 is evaluation and result in this chapter it contains how the approach result was evaluated, it also contains conclusions and future work, at the end of the paper references.

2.Related work

In today's data-driven world, artificial intelligence plays a key role in our society, researchers have successfully applied several algorithms and models to forecast weather conditions using various metrological features, attributes and data generated from different sources.

Murugan et al. 2021 proposed a hybrid NWP model for weather prediction model to improve the accuracy and efficiency of forecasting, they proposed using a hybrid C5.0 decision tree algorithm with k-means clustering algorithm for short-range prediction, the k-means clustering algorithm was used for grouping similar dataset together, the dataset was obtained from modern-Era Historical Analysis for research and applications (MERRA), They processed their data, and the selected attributes include temperature, pressure, humidity, wind speed, rainfall, snowfall, and snow depth, the model's performance was evaluated with mean absolute error and root mean square error and achieved a predictive accuracy of 90.18%.

The temperature of the next day at any hour in Nashville, Tennessee was predicted by jakaria et al. 2020 using different machine learning techniques including the Support vector Regressor, Ridge regression and multi-layer perceptron Regressor (MLPR), Random Forest Regressor (RFR) and extra- Tree Regressor (STR), after evaluation of their models using root mean squared error(RMSE), they were able to determine that that machine learning models can predict the weather accurately enough to compete with traditional models.

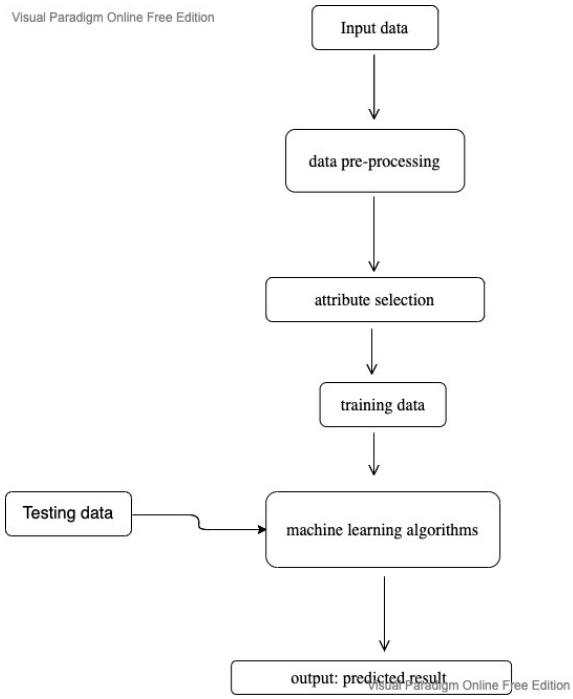
Singh et al 2019 proposed a new system for weather forecasting, In their proposed system, a low cost, reliable and efficient weather forecasting application was developed, it uses machine learning concept in python on Raspberry Pi board, and the application gets real-time data for humidity, temperature and pressure from its sensors to predict the possibility of rain on the present day, they used Delhi's weather data for the past 20 years which was split into 75% training and 25% testing, after applying the random forest classification algorithm it achieved an accuracy of 87.90%.

Petre 2009 used Pressure, clouds quantity, humidity, precipitation, and temperature as attributes for the proposed model, the model achieved an accuracy score of 83% when applied to the CART decision tree algorithm and had 16.67% incorrect classified instances though the model was trained with limited data of 48 instance for data recorded between 2002 and 2005, it produced a good prediction accuracy.

3. Methodology

In this section, I present the proposed model for weather forecasting using machine learning, it also contains a description of the dataset, its acquisition and pre-processing, and the analysis of the Algorithm used.

3.1 Proposed model



Section 3.2 Data acquisition and pre-processing

The dataset used for this analysis was acquired from an online data repository company called Kaggle and can be found here, the data was for the town of Seattle for the period 1st January 2012 to 31st December 2015, the dataset is licensed under CC BY-NC-SA 4.0, the file format was a comma-separated file (CSV) which contains 1461 rows and 6 columns, the columns which were identified as

- Date
- Precipitations
- Temp_max
- Temp_min
- Wind
- Weather

```
#peaking at the data head overview
data.head(20)
```

	date	precipitation	temp_max	temp_min	wind	weather
0	2012-01-01	0.0	12.8	5.0	4.7	drizzle
1	2012-01-02	10.9	10.6	2.8	4.5	rain
2	2012-01-03	0.8	11.7	7.2	2.3	rain
3	2012-01-04	20.3	12.2	5.6	4.7	rain
4	2012-01-05	1.3	8.9	2.8	6.1	rain
5	2012-01-06	2.5	4.4	2.2	2.2	rain
6	2012-01-07	0.0	7.2	2.8	2.3	rain
7	2012-01-08	0.0	10.0	2.8	2.0	sun
8	2012-01-09	4.3	9.4	5.0	3.4	rain
9	2012-01-10	1.0	6.1	0.6	3.4	rain
10	2012-01-11	0.0	6.1	-1.1	5.1	sun
11	2012-01-12	0.0	6.1	-1.7	1.9	sun
12	2012-01-13	0.0	5.0	-2.8	1.3	sun
13	2012-01-14	4.1	4.4	0.6	5.3	snow
14	2012-01-15	5.3	1.1	-3.3	3.2	snow
15	2012-01-16	2.5	1.7	-2.8	5.0	snow
16	2012-01-17	8.1	3.3	0.0	5.6	snow
17	2012-01-18	19.8	0.0	-2.8	5.0	snow
18	2012-01-19	15.2	-1.1	-2.8	1.6	snow
19	2012-01-20	13.5	7.2	-1.1	2.3	snow

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2022 doi:10.20944/preprints202206.0428.v1

Figure 3.2: data head overview

The dataset contains 5 weather classifications they are drizzle, rain, sun, snow, and fog.

To achieve accurate forecasting and ensure high performance of the algorithm, the data must be pre-processed.

Data pre-processing consists of transforming the acquired data into an understandable format, removing duplicate or null values, and dropping undesired attributes, for this analysis the data was pre-processing by removing the column date as it was deemed irrelevant, the dataset was complete and had no null value.

```
#checking for null values in the dataset
#to ensure data completeness.
data.isnull().sum()
```

```
date             0
precipitation    0
temp_max         0
temp_min         0
wind             0
weather          0
dtype: int64
```

Figure 3.3: checking data completeness

```
#removing the unwanted columns
data = data.drop('date',axis=1)
x = data.drop('weather',axis=1)
y = data['weather']
```

Figure 3.4: Removing the column date

Following the pre-processing of the dataset, to better understand the attributes and their relationship, a graph count of the attributes was generated.

3.2 Analysis

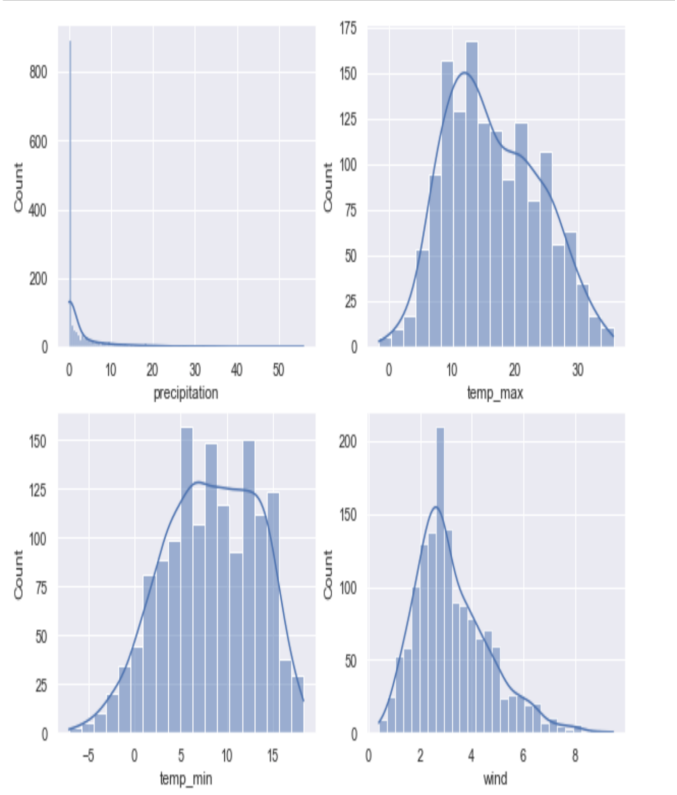


Figure 3.5: a graph count of the attributes

For the analysis of our data, the data was split into tests and Training dataset, using 25% for testing, four machine learning classification algorithms were implemented, they are:

- Random. Forest classifier
- Gradient Boosting Classifier
- Gaussian Naïve Bayes model
- Decision Tree Algorithm

- Random Forest Classifier

A random forest classifier is a collection of tree-structured classifiers whose results are compounded into one result; it is an ensemble machine learning algorithm which can be implemented for both classification and regression tasks and is made up of a set of classifiers known as a decision tree (Tin Kam Ho 1998) (Breiman Leo 2001), random forest classifier is known to produce accurate predictions, provides flexibility and reduced the risk of overfitting.

- Gradient Boosting Classifier

Gradient boosting classifier is a group of machine learning algorithms that combine many weak learning models together especially decision trees to create a strong predictive model, Gradient Boosting classifiers can be used for regression and classification tasks of machine learning, they are effective at classifying complex datasets and prediction accuracy is improved through developing multiple models in sequence, each additional model aimed at correcting the mistakes of the previous one. (Zhang, Y. et al, 2015)

- Gaussian Naive Bayes model

Gaussian Naive Bayes model is based on Bayes theorem, it assumes that a particular feature is independent of the value of any other feature, they can be trained very efficiently and are highly scalable, and the likelihood of the features is assumed to be -

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

the variance is independent of y and x.

- Decision Tree algorithm

Decision tree algorithm. Belongs to the family of supervised machine learning, where the data is continuously split according to a certain parameter and represented by a tree structure. It is used to solve classification and regression tasks and is among the most popular machine learning algorithm. (Xindong et al. 2018)

These algorithms were selected for the task because they were utilized by other weather forecasting models and have shown to have high predictive performance capabilities, they

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2022 doi:10.20944/preprints202206.0428.v1

were Implemented with python programming language using the Jupiter notebook, and various libraries were imported for the analysis including pandas, ski-learn, TensorFlow, and Matplot library.

4 Evaluation and result

A variety of results was obtained from the models trained with 75% of the data and tested with 25% of the data, in this section, we evaluate our models developed using various metrics, the metrics used for evaluation include:

- Accuracy
- Precision
- Recall
- Fi-score

Accuracy: This is the total proportion of observations that have been correctly predicted mathematically, accuracy is defined as:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Where TP= True positive, TN= True Negative, FP= False-positive and FN = false negative.

Precision: This is the percentage of the positive instance predicted that was correct, mathematically it is defined as

$$\frac{TP}{TP + FP}$$

Where TP= True positive, FP= False Positive

Recall: This is the percentage of the positive instance out of the total actual positive, mathematically it is defined as:

$$\frac{TP}{TP + FN}$$

Where TP= True positive, FN= False Negative

F1-score: This is the harmonic mean of precision and recall metrics, it is the overall correctness the model has achieved, mathematically it is defined as

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

1. Random Forest Classifier

For the random forest classifier accuracy, precision, recall, and f1-score were generated, as it is shown in figure 4.1 below, and the random forest classifier achieved a total accuracy of 79.50%.

```
[[  0   0   0   0  11]
 [  1   3   1   0  26]
 [  0   0 144   0  11]
 [  0   0   5   1   0]
 [  6  12   2   0 143]]
              precision    recall  f1-score   support

     drizzle       0.00      0.00      0.00        11
         fog       0.20      0.10      0.13        31
        rain       0.95      0.93      0.94       155
        snow       1.00      0.17      0.29         6
         sun       0.75      0.88      0.81       163

    accuracy                           0.80       366
   macro avg       0.58      0.41      0.43       366
weighted avg       0.77      0.80      0.77       366

Accuracy score of Random Forest Classifier: 79.50819672131148
```

2. Gradient Boosting Classifier

The Gradient Boosting Classifier achieved an accuracy of 80.87%, Recall, precision and f1-score were generated. As shown in figure 4.2 below

```
[[  0   1   0   0  10]
 [  1   1   1   0  28]
 [  0   1 143   0  11]
 [  0   0   4   2   0]
 [  1  11   1   0 150]]
              precision    recall  f1-score   support

     drizzle       0.00      0.00      0.00        11
         fog       0.07      0.03      0.04        31
        rain       0.96      0.92      0.94       155
        snow       1.00      0.33      0.50         6
         sun       0.75      0.92      0.83       163

    accuracy                           0.81       366
   macro avg       0.56      0.44      0.46       366
weighted avg       0.76      0.81      0.78       366

Accuracy of GradientBoostingClassifier model is: 80.8743
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 June 2022 doi:10.20944/preprints202206.0428.v1

3. Gaussian Naive Bayes model
The Gaussian Naive Bayes model achieved an 84.15% accuracy, recall, f1-score, and precision were also calculated as shown in figure 4.3 below

```
[[  0   0   0   0  11]
 [  0   0   0   0  31]
 [  0   0 141   2  12]
 [  0   0   2   4   0]
 [  0   0   0   0 163]]
              precision    recall  f1-score   support

     drizzle       0.00      0.00      0.00        11
         fog       0.00      0.00      0.00        31
        rain       0.99      0.91      0.95       155
        snow       0.67      0.67      0.67         6
         sun       0.75      1.00      0.86       163

    accuracy                           0.84       366
   macro avg       0.48      0.52      0.49       366
weighted avg       0.76      0.84      0.79       366

Accuracy score of GaussianNaive Bayes model: 84.15300546448088
```

4. Training Decision Tree
The Training Decision Tree achieved an accuracy of 80.87%, Recall, precision and f1-score were generated. As shown in figure 4.4 below

```
[[ 1  2  1  0  7]
 [ 6  8  1  0 16]
 [ 0  1 140  4 10]
 [ 0  0  3  3  0]
 [ 18 23  9  0 113]]
              precision    recall  f1-score   support

     drizzle       0.04      0.09      0.06        11
         fog       0.24      0.26      0.25        31
        rain       0.91      0.90      0.91       155
        snow       0.43      0.50      0.46         6
         sun       0.77      0.69      0.73       163

    accuracy                           0.72       366
   macro avg       0.48      0.49      0.48       366
weighted avg       0.76      0.72      0.74       366

Accuracy score of Training Decision Tree : 72.40437158469946
```

Performance comparison of the models

Performance comparison of the models was done with the accuracy score, Gaussian Naive Bayes model is the best performing model with an accuracy score of 84.15%.

Classifier	Accuracy (%)
Random Forest	79.50%
Gradient Boosting Classifier	80.87%
Gaussian Naive Bayes model	84.15%
Training Decision Tree	72.40%

conclusions and future work
This paper proposed a model for forecasting weather conditions, the weather condition forecasted are drizzle, fog, rain, snow, and sun. The objective of this research was to develop a weather forecasting model, evaluate the performance of the models developed and review related work on weather forecasting.
Related work on forecasting was reviewed in chapter 2, A weather forecasting model was proposed and developed in chapter 3, and the performance was evaluated in chapter 4, the model result shows significant success at predicting various weather conditions from my analysis it was deducted that the Gaussian Naive Bayes model is the most accurate of the implemented algorithms, the result from the model provided accurate prediction and useful guidance for meteorologist in their operational forecasting duties.

Despite the good performance of the model demonstrated, it is scalable, and it could be improved upon, only one dataset from settle was used for training and testing in this study, however, further studies can improve the performance by using datasets from multiple cities, the dataset can contain more attributes, and a more advanced technique could also be applied to the data pre-processing in the attempt to use a larger dataset with more attributes.
The model can also be implemented on a neural network, Adaboost or other algorithms which have shown to have high predictive accuracy with the goal of improving accuracy and efficiency. Further development can also be done to the model using various approaches as it is deemed necessary, the model can be implemented on a system to forecast real-time weather conditions and provide useful guidance for meteorologists in their operational weather forecasting, though a lot of work will have to be done.

References

A H M, J., Hossain, M. and Mohammad, A., 2020. Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee. arXiv preprint arXiv, 2008 (10789).

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Das, M. and Ghosh, S.K., 2014, December. A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables. In *2014 9th International Conference on Industrial and Information Systems (ICIIS)* (pp. 1-6). IEEE.

Fang, W., Xue, Q., Shen, L. and Sheng, V.S., 2021. Survey on the application of deep learning in extreme weather prediction. *Atmosphere*, *12*(6), p.661.

Gagne, D., McGovern, A., Haupt, S., Sobash, R., Williams, J. and Xue, M., 2017. Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. Weather and Forecasting, 32 (5), 1819-1840.

Haupt, S.E., Cowie, J., Linden, S., McCandless, T., Kosovic, B. and Alessandrini, S., 2018, October. Machine learning for applied weather prediction. In *2018 IEEE 14th international conference on e-science (e-Science)* (pp. 276-277). IEEE.

Hewage, P., Trovati, M., Pereira, E. and Behera, A., 2021. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, *24*(1), pp.343-366.

Ho, T.K., 1995, August. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

kaggle, 2022. WEATHER PREDICTION [online]. Kaggle.com. Available from: https://www.kaggle.com/datasets/ananthr1/weather-prediction/code [Accessed 10 May 2022].

McCarthy, P., 2007, June. Defining the impact of weather. In *Proc. 22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction, American Meteorological Society, Utah*.

Murugan Bhagavathi, S., Thavasimuthu, A., Murugesan, A., George Rajendran, C., A, V., Raja, L. and Thavasimuthu, R., 2021. Weather forecasting and prediction using hybrid C5.0 machine learning algorithm. International Journal of Communication Systems, 34 (10).

Petre, E.G., 2009. A decision tree for weather prediction. Bul. Univ. Pet.–Gaze din Ploieşti, 61(1), pp.77-82.

Singh, N., Chaturvedi, S. and Akhter, S., 2019, March. Weather forecasting using machine learning algorithm. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 171-174). IEEE.

Wang, B., Lu, J., Yan, Z., Luo, H., Li, T., Zheng, Y. and Zhang, G., 2019, July. Deep uncertainty quantification: A machine learning approach for weather forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2087-2095).

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S. and Zhou, Z.H., 2008. Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), pp.1-37.

Zhang, Y. and Haghani, A., 2015. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, pp.308-324.