International Journal of Smart System

**INTERNATIONAL JOURNAL OF SMART SYSTEMS**
Volume 1 Issue 1 Year 2023 Pages 19 -27
e–ISSN XXXX–XXXX
Url: https://ijss.etunas.com/index.php/ijss/index

# Algorithm Implementation With Ensemble Learning On Weather Forecast

**Ira Ramadhaniyati[1]*, Marlianawati Khodijah [1], Moh. Ilham Sahrulkhan[1]**
1 Tadris Matematika, IAIN Syekh Nurjati Cirebon, Indonesia

*Correspondence to: *iraramadhaniyati11@gmail.com*

**Abstract:** Weather conditions are vital in the continuation of human life. Knowing weather conditions became crucial because almost all human life is connected to it. From agriculture, plantations, even to human activities. Therefore, the method of a weather forecast is required for information on today's weather conditions as well as in the future. The purpose of these weather forecasts is simply so that people can use this for survival. Normally we can tell weather conditions from rainfall, temperature and wind speed. The issue, however, is how to determine accurate weather predictions and can be easily used by the general public. In the study, selecting learning ensemble to calculate existing data groups by involving multiple algorithms to find average accuracy and determine which methods work most optimally. As for research results it is expected to be the basis for building weather forecast applications. Accuracy results at 81.21% and mse 18.79%.

**Keywords:** Weather, Ensemble Learning, Algorithm, Method, Prediction, Classification

**Article info:** Date Submitted: 12-Feb-2023 | Date Revised: 13-Feb-2023 | Date Accepted: 14-Feb-2023

## INTRODUCTION

Erratic weather is a major factor in human life. After all, many activities in daily life are influenced by weather factors. Weather forecasts are helpful in all areas. Like, in agriculture, to determine harvest seasons, to ensure flight schedules, to prevent flood disasters and other things. Hence, weather forecasts are of great importance. Predicting the weather is not easy, many factors should be noted in the predicting process. Start with temperature, air pressure and wind speed. So, in the process the prediction is aided by technology[1].

The forecast itself is an attempt made to determine the temperature, the air in the future. Weather forecasts are expected to minimize the impact of natural disasters, the saturation of agricultural activity and the security of flight schedules. Given the importance of weather forecasts for human activities, it is necessary to measure high accuracy in weather forecasts. Because many weather forecasts miss what is predicted. Some approaches to statistic air temperature forecasting like regressive (ar), ar-integrated moving average (arima) both methods have weaknesses, such as adverse weather predictions.

These weather identification is necessary and effective in various areas of society. This triggers researchers to find and explore more precise methods for determining weather forecasts. In research carried out by emery at (2020), ensemble learning is used, using multiple algorithms to achieve better predictions. Learning ensemble can improve accuracy, as for its application with three bases classifiers. The best model with highest accuracy is bagging with a classifier decision tree algorithm

(95.312%). The experiment showed that a classifier ensemble model was done better than a base that was only a classification.[2]

## RELATED WORK

After researchers have done a study of some studies, some are related to the study. Such studies include:

Research conducted by Ike Fitriyaning. "With the predictive events flooding" with ensemble machine learning using BP-NN and SVM in 2019. His research USES historical data methods, with his research predicting precipitation and water discharge best using bp-miss is a six-day prediction with training combinations testing.[3]

Research conducted by Eko Supriyadi. Called predicting weather parameters using deep learning long short term memory (LSTM) in 2019. His research used the method of deep learning LSTM, with his results only temperature and air humidity forecasting increase over time. The parameters of wind velocity and air pressure are decreasing as the research is conducted on the third day and increasing continuous next month.[4]

Research done by Diky Djafar Sidik. Called the stacking classic for predictive precipitation in 2019. The study used an ensemble stacking/inequality generalization method, with his research that the highest accuracy value lies in the Majalengka's datassets and based on the testing, stacking method succeeded in improving the classifier base production.[5]

Ika Meila Pradipta, if they come. The genesis flooding with ensemble machine learning using BP-NN and SVM in 2020. With his research using the KNN subset of ensemble methods, with his research that the greater the accuracy of the classification, the greater the closeness between predictive value and actual value.[6]

Bian Suma. Application of machine learning in weather forecasts by 2020. Using the application of machine learning, his research suggests that defining weather data would be used in machine learning and the definitive model used by machine learning.[7]

## METHODS

The study USES learning ensemble methods that aim to minimize classification errors by using Kaggle's Datasset. The proposed method in research is naive bays, generalize linear model, deep learning, decision tree and random test.

Furthermore, it also applies a bagging method derived from boostrap aggregation that is another form of learning ensemble. Also use a few versions of learning methods based on applicable methods of interpretation based on the desired results derived from the average results of all later methods. By making a few copies of the training set that produced the decision tree of each series of ignition training that would later be obtained from the selection of child n pattern training where n shows an entire datasset size.[8]

A. Learning With Different Base Classifiers

This approach is very popular with the diversity of ensemble members coming from various kinds of scalding and not relying on different subsets from samples produced by sampling techniques. Classifier ensemble is a method that USES several classifications to enhance classification performance. This technique is more resistant to noise than a single classifier.

The approach to dealing with unbalanced datassets (qualities) is that of approach to data level, algorithm level and ensemble. Bagging is better than random sampling. At the deductions of the data

level are used to encompass multiple data resampling techniques and synthetics to improve the distribution of data class. At an algorithmic level with its main method of synchronizing existing algorithmic operations is used to classify classifier with the aim of being more conductive to the classification of minorities. Meanwhile, for ensemble methods, there are two algorithms used and popular. The first is the boosting, this alogarithm is used as a meta-technique for coping with the unbalance class. And bagging or boostrap aggregating, using sub-datassets (boostrap) to generate tear-training l (learning). Where I used to train the basis of unstable learning procedures during testing and retrieval used for classification and regression.[9]

The aim of learning ensemble with different base classifiers is to fix alogartimh classifying by without changing the data, which makes it only a two approaches, a data level approach and an allogartima level. Divide into two parts this approach makes it easier to focus on fixing objects, with a level approach focused on initial data and conscious processing, whereas an allogarithmic level is focused on repairs or increases allogarithms.

B. Training Data Collection And Test Data

This assessment has several phases used for data collection. Data and attributes are only partially used. Value attributes must go through a few steps in advance that is, initial data processing (preparation data) in order to get the quality data done by some of the techniques, as follows: date, observation, location, maxtamp, mindtamp, mindtamp, and so on.[10]

C. Data Cleaning

Cleaning data technique is a technique used to handle incomplete data. The data-cleaning process contains among other things: discard data duplicates, review inconsistencies and correct errors in the data. Cleaning data also does the encrichment process which is an existing "enriching" process with the relevant and necessary data for KKD (knowledge discovery database) for example, such as external information.[11]

Good data quality comes from a basic key by producing good quality, outlier data noise or noise, incomplete data of its attributes and inconsistent data in its application. The stage is used to eliminate naive bayes algorithms in ways that can handle incomplete or missing data and inconsistent data, while also to divide redundancy caused by data interrogation.

The cleaning process occurs when there is a double, incosystem or value and outlier missing data. This is done to prevent further performance of the classification process. In addition to that cleaning data is also used to remove unnecessary features when the classification process will be done.

D. Transformation data

The transformation of the data is changing the original data to become another form. With intent to normalize data, matching data with the assumption underlying analysis, double data analysis. As well as increasing algorithm accuracy. The transformation of the algorithm occurs when the original data is worth less than 10 and closer to zero. Naive bayes's algorithm had the ability to process positive, nominal, and ordinal data. Therefore the value of every attribute contained in the datasset must not be changed.

E. Naive Bayes Algorithm

An algorithm briefly involves squinting measures that address a problem. The naive bayes algorithm is a classification method that has a higher accuracy than any other classification method[12]. Furthermore, the algorithm is very easy to use in managing a large enough data set. Naive bayes was a statistical and probability method of classification. This design, could predict the probability of membership in data that would be in a particular class. This was based on the bayes

theorem presented by Thomas bayes (a statistician). Combined with naive, assuming that the existence of each variable is free from the other. Naive bayes's algorithm predicts events of a future future, based on events such as had occurred in the past.

Here is bayes's equation to calculate the probability of an event[13]:

$$\mathrm{P}(\mathrm{H}|x) = \frac{\mathrm{P}(x|\mathrm{H})P(\mathrm{H})}{\mathrm{P}(x)}$$

F. Algorithm Decision Tree

Classification is the process of data analysis that produces models to describe the classes contained by the data (classifier) [14]. They would be represented in the shape of the decision tree (resembling a flowchart). This is helpful in the process of making decisions with various considerations and risks. With this, we can make the right decision.

Decision trees are used to classify and predict patterns from data (target attribute and variable) that are summarized from previous data summaries. This decision tree is quite popular and easy to understand and analyze. Decision tree consists of three structures. First, the root (top knot) that is the starting point of a decision tree. Second, the intermediate knot (a twig) associated with a test or various action options (made of branches). Third, a leaf knot that contains the possible result of a particular action to draw a final conclusion or target from the decision tree. To come to the end conclusion, to read from the root to the conclusion leaf.

G. Random forest Algorithm

In accordance with his name random forest as his name is, there are a large number of individual decision trees that operate as ansambel. Each random forest tree embossed class and classroom predictions with the most votes to model predictions. "The visualization of the random forest model made predictions. The fundamental concept behind random forest is a simple but powerful one folk wisdom [13]. In the science of data speaking, the reason that random forest models work properly is: a large number of relatively non-correlates models operating as forest will surpass one of the individual constituent models. The low correlation between the model is the key.

The Random Forest Algorithm, as the name implies, consists of a number of individual decision trees that act as an ensemble. All random forest trees make class predictions, and the class with the most votes becomes the model prediction. Predictions are made by visualizing the Random Forest model. The basic concept behind Random Forest is the wisdom of the masses, which is simple yet powerful. From a data science perspective, the reasons why the Random Forest model works so well are: A large number of relatively uncorrelated (trees) models where acting as forest is superior to each individual configuration model. The low correlation between the models is important.

H. Algorithm Deep learning

Deep learning, also known as deep structured learning, is part of a large family of machine learning techniques based on artificial neural networks with typical learning. Learning can be monitored, partially monitored, or unsupervised In areas such as computer vision, machine vision, speech recognition, and natural language processing, deep learning architectures such as highly confident neural networks, iterative neural networks, and convolutional neural networks. used. Speech recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, materials inspection, and board game programming have produced results that match or even exceed human professional performance. Artificial Neural Network (ANN) inspired by information processing and communication nodes distributed in biological systems. The ANN is different from the biological brain. Neural networks in particular tend to be static and symbolic, but the biological brains of most living organisms are dynamic and analogous.

I. Generalized linear model algorithm

In statistics, the generalized linear model (GLM) is a flexible generalization of generalized linear regression, which allows response variables with errors in the model distribution other than the normal distribution. GLM generalizes linear regression by allowing a linear model is associated with the response variable via a link function and takes the magnitude of the variance of each measurement as a function of the predicted value. General linear model formulated by John Nelder and Robert Wedderburn to integrate various other statistical models, including linear regression, logistic regression, and Poisson regression. We propose a least weighted quadratic method for maximum likelihood estimation of the model parameters. Maximum likelihood estimation is still common and is the standard method for many statistical calculation packages. Other approaches have been developed, including the Bayesian and quadratic are optimal for a stable response distribution.

## RESULT AND DISCUSSION

Well-prepared tables and or figures must be of significant feature of this section, because they convey the major observations to readers. Any information provided in tables and figures should no longer be repeated in the text, but the text should focus on the importance of the principal findings of the study. In general, journal papers will contain three-seven figures and tables. Same data can not be presented in the form of tables and figures. The results of the study are discussed to address the problem formulated, objectives and research hypotheses. It is higly suggested that discussion be focused on the why and how of the research findings can happen and to extend to which the research findins can be applied to other relevant problems.

In statistics, the generalized linear model (GLM) is a flexible generalization of linear regression ordinary, which allows a response with the model distribution error other than the normal distribution. GLM generalizes linear regression by allowing a linear model is associated with the response variable via a link function and takes the magnitude of the variance of each measurement as a function of the predicted value. General linear model created by John Nelder and Robert Wedderburn to integrate various other statistical models such as linear regression, logistic regression, and Poisson regression. We propose a minimum re-weighting method for the maximum likelihood estimation value of the model parameters. Maximum likelihood estimation is still common and is the standard method for many statistical calculation packages. Other approaches have been developed, including the Bayesian and quadratic are optimal for a stable response distribution.

### 1. Model Evaluation

A. Accuracy

Binary classification evaluations compare two binary attribute assignments methods, one of which is usually the standard method and the other is studied. There are many metrics that can be used to quantify classification or predictor performance. Different fields have different preferences for certain metrics because of different goals.

Polymatrix is the predictive result on the issue of classification. The number of correct and incorrect predictions is then summarized by calculating the value and reaching each class. It gives insight not only into mistakes made by the classifier but a more important kind of mistakes made.

Table 1. Confussion Matrix

| | Class 1: Positif | Class 2: Negatif |
|---|---|---|

| | | |
|---|---|---|
| Class 1: Positif | TP | FN |
| Class 2: Negative | FP | TN |

Explanation
Class 1 = Positif
Class 2 = Negative
True positive (TP), False negative (FN), True negative (TN), False positive (FP).
Formula for calculating accuracy as follows:

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

B. Means Square Erorr (MSE)

In statistics, either Mean Square Of Error (MSE) or Mean Squared Of Deviation (MSD) breaks down equal to the balance of the policy to take into account total unobserved balance. Measuring quadratic rifts fragmented miss is-that is, flat squares of opposition distances of value guesswork and the actual ideal. MSE is a risk benefit, coupled with values expected to break the square mistakes of loss. The fact that the MSE approached was always very positive (and not zero) was because of randomness or because the scales had not predicted the facts that would make for a more accurate guess.

$$MSE = \frac{1}{2}\sum_{i=1}^{N}(fi - yi)^2$$

**Explanation :**

Where n is the data amount

$fi$: Returnd value by model

$yi$: Actual score on data I

## 2. Results and discussions

A. Research results

Climate change and weather make the problem encountered approach the entire environment that categorizes and predicts. The excessive variable varrable makes it quite difficult and unfathomable. Climate change and weather is global warming made by individuals who shape it increasingly difficult to open up weather problems. The results of this research could be into measuring the precision phase and MSE towards the weather. After this could be produced a forecast application or weather group, by monitoring the precision of this most part of the analysis. Here are the chart of the correct results and the MSE from various algorithms.

Table 2. Performance Accuracy Classification Algorithm

| Metodhs | Accuracy | MSE |
|---|---|---|
| Naïve bayes | 77.22% | 22.78% |
| Decision tree | 79.46% | 20.54% |

| | | |
|---|---|---|
| Random forest | 82.38% | 17.62% |
| Deep learning | 82.92% | 17.08% |
| Generalized linier model | 84.06% | 15.94% |

Table 2 shows the spectacle all measured methods of accuracy and MSE. From the chart above the highest accuracy of the GLM algorithm (linear model) is 87.06% and MSE at 17.08%, followed by deep learning at 82.92% and MSE 17.62%. To simplify reading the spectacle, can be seen in the form of this graph below:

Table 3 is the highest among several methods of method. The results of this study can be superior to others, along with the highest accuracy and lowest mse like the table below:

Table 3. Learning Esemble Research

| | |
|---|---|
| Accuracy | 81.21% |
| MSE | 18.79% |

B. Study discussion

Analysis is done, quantifying a differentiated matrix model to rate inequality and mse to think it's classfication erorr. It is most popular to assess the success of the part algorithm on dealing with the problem, because before constructing the prediction app, it should measure the power of the algorithm to be used, as part of this analysis by analogizing five algorithms and adding all others called esemble under bagging, the result being a variety of thought is 81.21 to achieve MSE 18.79%. Meaning esambel's impact on this study is lower than the gated GLM (generalized linear model), which may be used for a dataster-based event, for algorithms used by various characteristics, such as the decision tree group, its impact on accuracy, and deep learning above the average esambel learning, naive bayes below the average esambel learning.

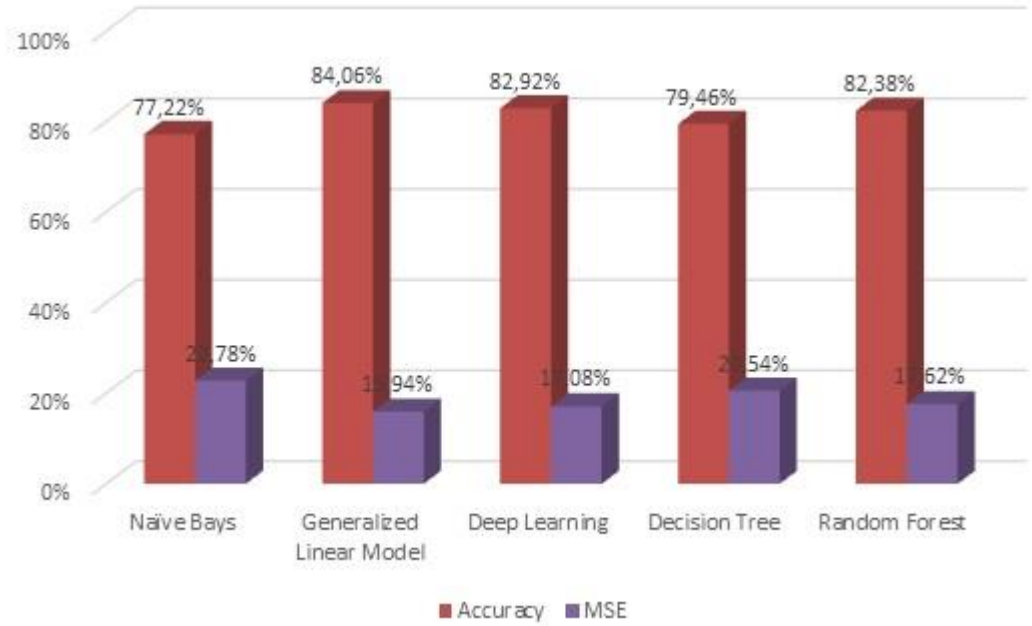


Figure 1: Comparative algorithm

## CONCLUSION

At the deductions of the data level are used to encompass multiple data resampling techniques and synthetics to improve the distribution of data class. Training Data Collection And Test Data This assessment has several phases used for data collection. Data Cleaning, cleaning data technique is a technique used to handle incomplete data. Good data quality comes from a basic key by producing good quality, outlier data noise or noise, incomplete data of its attributes and inconsistent data in its application. The stage is used to eliminate naive bayes algorithms in ways that can handle incomplete or missing data and inconsistent data, while also to divide redundancy caused by data interrogation. Transformation data, the transformation of the data is changing the original data to become another form. With intent to normalize data, matching data with the assumption underlying analysis, double data analysis. Based on the impact of this study with the esambel learning approach, it can be taken some formulas by the precision and MSE of learning 'esemble systems as 81.21% as accuracy and 18.79% on MSE. To the best deciding tree grass is random forest 82.38% for precision and 17.62 is MSE. The deep learning algorithm has its performance is 82.92% like precision and 17.08% for MSE. The largest spectator algorithm escorted by another 84.06% is an accuracy and 15.94% to MSE.

## REFERENCES


[1] R. J. Yuniar, D. R. S, and O. Setyawati, "Perbaikan Metode Prakiraan Cuaca Bandara Abdulrahman Saleh Dengan Algoritma Neural Network Backpropagation," *J. EECCIS*, vol. 7, no. 1, p. pp.65-70, 2013.

[2] A. M. Siregar, Tukino, S. Faisal, A. Fauzi, and I. Kadori, "Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning," *J. Pengkaj. dan Penerapan Tek. Inform.*, vol. 13, no. 2, pp. 138–147, 2020, doi: 10.33322/petir.v13i2.998.

[3] I. Fitriyaningsih and Y. Basani, "Flood Prediction with Ensemble Machine Learning using BP-NN and SVM," *J. Teknol. dan Sist. Komput.*, vol. 7, no. 3, pp. 93–97, 2019, doi: 10.14710/jtsiskom.7.3.2019.93-97.

[4] E. Supriyadi, "Prediksi Parameter Cuaca Menggunakan Deep Learning Long-Short Term Memory (Lstm)," *J. Meteorol. dan Geofis.*, vol. 21, no. 2, p. 55, 2021, doi: 10.31172/jmg.v21i2.619.

[5] D. D. Sidik and T. W. Sen, "Penggunaan Stacking Classifier Untuk Prediksi Curah Hujan," *IT Soc.*, vol. 4, no. 1, pp. 21–27, 2019, doi: 10.33021/itfs.v4i1.1180.

[6] M. I. Pradipta, "Klasifikasi Curah Hujan Menggunakan Metode Ensemble Subset K-Nearest Neighbor," p. 67, 2020.

[7] B. Suma, "Penerapan Machine Learning Di Dalam Bandung September 2020," *Univ. Pas. Bandung*, no. September, 2020.

[8] A. Saifudin and R. S. Wahono, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, no. 1, pp. 28–37, 2015.

[9] R. T. Prasetio and Pratiwi, "Penerapan Teknik Bagging Pada Algoritma Klasifikasi Untuk Mengatasi Ketidakseimbangan Kelas Dataset Medis," *J. Inform.*, vol. II, no. 2, pp. 395–403, 2015, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/118.

[10] Y. Pristyanto, "Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset," *J. Teknoinfo*, vol. 13, no. 1, p. 6, 2019, doi: 10.33365/jti.v13i1.184.

[11] Jasmir, "Implementasi Teknik Data Cleaning dan Teknik Roughset pada Data Tidak Lengkap dalam Data Mining," *Semin. Nas. APTIKOM*, pp. 99–106, 2016.

[12] D. Xhemali, C. J. Hinde, and R. G. Stone, “Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages,” *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.

[13] D. Nofriansyah, K. Erwansyah, and M. Ramadhan, “Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL ( Studi Kasus di CV. Sumber Utama Telekomunikasi),” *J. Saintikom*, vol. 15, no. 2, pp. 81–92, 2016.

[14] Z. Azmi and M. Dahria, “Decision Tree Berbasis Algoritma Untuk Pengambilan Keputusan,” *Saintikom*, vol. 12, pp. 157–164, 2013, [Online]. Available: http://demo.pohonkeputusan.com/files/DECISION TREE BERBASIS ALGORITMA UNTUK PENGAMBILAN KEPUTUSAN.pdf?i=1.