International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Sep-2023)

http://dx.doi.org/10.12785/ijcds/140167

A Machine Learning-Based Approach for Meteorological Big Data Analysis to Improve Weather Forecast

Abderrahim El Mhouti[1], Mohamed Fahim[1], Asmae Bahbah[2], Yassine El Borji[3], Adil Soufi[4], Ayoub Aoulalay[1] and Chaimae Ouazri[1]

[1]*ISISA,FS, Abdelmalek Essaadi University, Tetouan, Morocco*
[2]*S2IPU,ENS, Abdelmalek Essaadi University, Tetouan, Morocco*
[3]*SOVIA,ENSAH, Abdelmalek Essaadi University, Tetouan, Morocco*
[4]*FSTH, Abdelmalek Essaadi University, Tetouan, Morocco*

Received 15 Dec. 2022, Revised 14 Jul. 2023, Accepted 10 Aug. 2023, Published 01 Sep. 2023

Abstract: Today, the web is a vast and valuable source of weather data. Every day, several petabytes of meteorological information are generated, leading to the weather big data. By using various Machine Learning (ML) techniques, weather big data is used for forecasting and decision making. However, processing such a large weather data is a challenge for ML algorithms and computing resources. Weather big data often includes a very large number of variables, which requires huge resources for the analysis and processing. As a result, ML techniques used produce forecasts that are not always efficient and take longer to forecast.

To improve the prediction precision in a minimum of time, this paper aims to investigate the influence of data sampling techniques on the accuracy of ML models used in weather data analysis. To this end, we used the dimensionality reduction technique ”Random Projection” (RP) combined with two ML classifiers (Decision Tree, Naïve Bayes) and applied it on weather big data collected from web sources using web scraping technique. The results of the conducted experimentation show that reducing the dimensionality of weather data considerably maximizes the performance of ML models and thus improves the accuracy of weather forecasts while reducing the processing resources.

Keywords: Machine Learning, Dimensionality reduction, Weather Big Data, Weather forecast

1. INTRODUCTION

Every day, several petabytes of weather information are generated and managed via the web in a variety of formats. This enormous amount of weather information, known as “weather Big Data”, is becoming increasingly important. These weather big data are used for various purposes such as weather prediction or weather forecasting [1]. These weather forecasts have a very important impact on the lives of people and organizations as they play a key role to make decision for weather management and for important sectors such as agriculture, transport, industry and tourism [2] [3] [4].

To predict weather conditions, several contributions of research woks have been conducted while relying on data mining methods and machine learning approach of environmental data sets in different regions and countries [5][6][7][8][9]. Nevertheless, the complex erratic and uncertain nature of weather leads to weather big data containing huge and gigantic information in an organized, semi-organized and unstructured way, which can make traditional weather forecasting tedious and difficult. These traditional prediction processes use expensive and complex computer resources to generate forecasts. These forecasts can sometimes be inaccurate and have various negative impacts on society.

Indeed, weather prediction requires the collection of a great deal of volume of data which is usually disorganized data containing modifiable parameters that vary according to the rapidly changing weather conditions [9]. This makes the analysis of this weather data for weather prediction a complex task [7]. In this sense, several big data solutions such as machine learning have been implemented for processing and analyzing this big data, but processing such a large amount and type of data (unstructured, unclassified, rapidly changing) still remains a challenge for machine learning algorithms and computing resources to deploy [10]. To this end, several techniques can be employed for big data analysis with machine learning algorithms, including data

E-mail address: a.elmhouti@uae.ac.ma

sampling and dimensionality reduction [11] [12] [13]. Thus, data sampling and dimensionality reduction, which focus on building features and representations of data using raw data [14], represent an important component of machine learning that can lead to improved performance for machine learning prediction models while reducing resources [15].

In this paper, the data sampling technique combined with machine learning algorithms is used to implement a new weather forecasting model founded on the analysis of weather big data collected from web pages sources. The essential aim of the present research work is to evaluate and understand the effect of data sampling and dimensionality reduction on improving weather forecasting and reducing processing and computational resources.

For this, the proposed approach includes a series of steps. At first, raw weather big data and features are collected from weather web sources employing web scraping method. After that, the data is pre-processed and the relevant features are selected using the PCA (Principal Component Analysis) technique. After selecting the relevant features with the PCA method, the data set obtained is employed as an input models created from machine learning principles. Then, the adopted experimentation used ”Random Projection” as a feature selection technique to create reduced data sets at well-defined reduction rates (75%, 50% and 25%). In addition, two machine learning classifiers were also used (firstly, Decision Tree and secondly, Naïve Bayes) to make weather forecasts using the full data set on the one hand (first experiment) and the reduced data sets created with ”Random Projection” on the other hand (second experiment).

The overview of the results obtained highlights the advantage of the approach designed in improving weather forecasting and reducing processing and computational resources. With data reduced to only 50% of the original data set, the Naïve Bayes algorithm achieved a better performance (83.9%) than all other cases, including the case where the whole data set is used (81.6%). With the Decision Trees classifier, the best performance (78.4%) was recorded when the full data set was used. This performance decreased slightly (to 78%) when only 50% of the original data set was used. However, by comparing this slight decrease in performance (0.4% decrease) with the considerable reduction in resources (50% of the original data set), it can be concluded that the sampling technique provided the advantage of reducing the resources used without affecting the weather forecast accuracy too much.

Compared to other related leading contributions, the importance of the proposed approach is that it allows machine learning algorithms to be a promising research avenue in the extraction of unstructured data representations at high abstraction levels and in the improvement of prediction in general.

The continuation of the present paper is coordinated according to this structure: section 2 gives an outline of weather forecasting works using various machine learning models and techniques. Sections 3 describes the designed approach and basic concepts adopted to ensure meteorological big data analysis and improve weather forecast. Section 4 contains how the proposed approach was evaluated and presents and discusses the evaluation results. And finally, section 5 outlines the present research conclusions and future work.

2. RELATED WORKS

There are several research works dealing with the meteorological data collection and analysis for forecasting and decision support purposes. However, in terms of the techniques adopted, the approaches proposed differ from one work to another. Some works focus on the collection of meteorological big data, others adopt artificial intelligence paradigms (Machine Learning models and also Deep Learning models) to analyze these data and the production of forecasts, while few of these works use dimensionality reduction techniques to improve forecasts. Finally, some of these works combine some or all of these techniques.

In the present section, a review of the literature is presented including research that focuses on weather forecasting based on weather big data and employing artificial intelligence models (in particular machine learning models), while focusing on contributions that adopt dimensionality reduction techniques to improve predictions.

In their study presented in [16], the authors are interested in the evaluation of the reduction dimensionality methods in the forecasting of solar radiation. Actual data issued from Weather Research and Forecasting model was used by the authors and experimented with three multivariate feature selection approaches using machine learning algorithms in order to obtain robust feature sets that increase prediction accuracy and deliver reliable outcomes. The study showed that the reduction in dimensionality (which can be up to 10% of the original variables) offers significant improvements in accuracy and facilitates the interpretation of results.

In the study presented in [17], the authors combined the reduction of dimensionality techniques and machine learning methods to improve solar radiation forecasting. In order to overcome the limitations of existing dimensionality reduction techniques (in particular PCA, SNMF, LPP and LOL), the authors propose the SLMVP (Supervised Local Maximum Variance Preserving) technique. It is a supervised nonlinear technique for dimensionality reduction and feature extraction. The proposed feature extraction method is compared with PCA, SNMF, LPP and LOL methods. The observed results demonstrate that compared to the PCA, SNMF, LPP, LOL approaches, the suggested SLMVP method yields smaller errors.

In [18], to forecast the weather, the authors used a stacked sparse autoencoder that is based on deep learning.

To obtain relevant data from weather big data, and to enhance the speed of the prediction approach, the designed model uses the PCA technique (Principal Component Analysis) in order to reduce the dimensionality and extraction of relevant features. The proposed model also incorporates the method of feature selection which is based on the algorithm of Binary Butterfly Optimization and a deep stack autoencoder to enhance prediction precision. The suggested model has been tested, and the results demonstrate that it offers better than the current models in terms of error rate, precision and computation time.

In the study presented in [19], the authors describe machine learning models used in the prediction of frost events for agricultural applications. The machine learning algorithms studied include convolutional neural networks, deep neural networks and also random forest models with time delays of 6 to 48 h. In this study, the authors used the feature extraction technique to provide dimensionality reduction of data from a weather station. The results obtained show promising accuracy for use in minimum temperature and frost prediction applications. The feature extraction and dimensionality reduction allowed for optimization of computer resources, reduction of model overfitting and higher accuracy in temperature predictions.

In [5], the authors propose an approach to climate expectations that uses climate data to create machine learning models capable of anticipating certain climate conditions. The authors first used the PCA (Principal Component Analysis) feature extraction method to extract irrelevant attributes from the data sets. Secondly, the authors employed 5 algorithms of machine learning to predict the mid-term temperature conditions and finally four performance indicators as well as the training time employed to identify the best fitting model. The outcomes revealed indicate that the model that uses the PCA dimensionality reduction method is the best fitting model with the best results.

In addition to the field of weather forecasting, dimensionality reduction techniques are widely used and tested in other areas of big data analysis using artificial intelligence models. Among the works interested in big data analysis using dimensionality reduction, we find the work presented in [20] and where the authors demonstrated a brand-new Spark-based hybrid parallel framework for feature reduction. The objective was to make the reduction of feature on distributed/shared memory clusters easier. The evaluation of the proposed framework demonstrated how much quicker the model is than traditional feature reduction methods. The designed method required more than 1 minute in order to select 4 features from the dataset. On the other hand, the comparative algorithm, took more than 2 hours to perform the same task. The authors concluded that the suggested approach can accomplish improved precision with less time and memory when compared to conventional feature reduction methods.

The work presented in [21] presents an example of practical application of the random projection feature extraction technique in the R programming language. The proposed implementation is compared with similar implementations. It has been tested with various formats of data, including images, text and data received from sensor. The experiments conducted and the results found show that the random projection method is a good choice for dimension reduction in multivariate big data analysis.

Another work presented in [22] proposes to study the performance of feature selection technique, feature extraction technique and also a combination of both techniques. The study conducted developed 21 rice yield prediction models for 8 sub-regions of Vietnam based on machine learning methods. The results of the study reveal that the combination of feature selection and feature extraction takes full advantage of both techniques combined, allowing the models based on this combination to be the best in 18 of the 21 models developed. These models which are based on feature selection and feature extraction improve the models based on no feature reduction by an average of 21% and up to 60% in terms of the root mean square deviation.

In [23], the authors conducted an extensive empirical comparison between the random projection technique and the random selection technique in a variety of data sets with different features. The study conducted resulted in the finding that random projection performs better in general than random feature selection in terms of classification precision in small samples, although random feature selection is also unexpectedly effective in several instances. These methods have the potential to enable learning from large dimensional data sets and are effective overall.

Thus, the review of related work shows that there are several contributions in the area of collecting weather big data and analyzing it using machine learning models to generate weather forecasts. In this study, we also found that dimensionality reduction, as an increasingly used technique for extracting relevant features from big data, was discussed in several scientific research works in various fields, including weather data analysis.

In addition, through the literature review conducted, we found that several works have attempted to use machine learning models to analyze huge volumes of weather data for various purposes. However, processing such a large amount of data is a challenge for machine learning algorithms and computing resources. This is because the collected weather data contains a very large number of variables, which is time-consuming and requires huge resources to analyze and process the data. Extracting complex models from huge volumes of weather data can then lead to inaccurate weather forecasts that can have catastrophic impacts on society.

In this sense, we have concluded that by combining dimensionality reduction techniques with machine learning

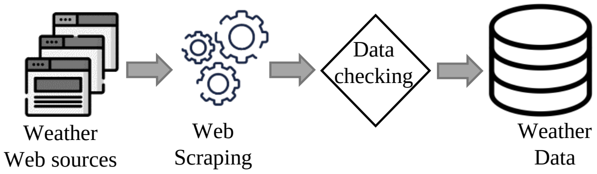


Figure 1. Weather Big Data collection steps

models, we can improve the predictive performance of these models compared to the use of standard classifiers. Furthermore, in large weather datasets, it is quite possible that reducing the amount of data can produce predictions as good as the original data while reducing processing time and resources.

3. PROPOSED APPROACH AND BASIC CONCEPTS

As mentioned earlier, the collection of weather data from web sources leads to a vast amount of data called weather big data. This data is typically used as input for machine learning models to produce weather forecasts. However, this data, which typically includes a large number of variables, is becoming increasingly complex to use in Machine Learning techniques, which can lead to a degradation in forecast accuracy and increasing forecast time.

In this context, this work aims at analyzing the effect of data sampling on the improvement of the performance of the Machine Learning approach in the production of weather forecasts. To this end, this work proposes a prediction approach based on the reduction of dimensionality of big weather data collected from web sources via the web scraping method.

The data used to implement and test the proposed approach is collected from web sources using the web scraping technique. Web scraping technique is one of the main sources for extracting unstructured big data from the web. Web scraping represents a high potential big data approach [24] allowing the conversion of unstructured data sourced from the web pages into structured data for storage and analysis [25]. Fig. 1 illustrates the steps involved in collecting weather big data.

Once the weather data set is created from data scraped from web sources, and before it is used for prediction employing machine learning techniques, this data set will undergo a dimensionality reduction. The used dimensionality reduction procedure is Random Projection.

Random Projection is a decomposition method employed to bring down the dimensionality of highly dimensional data. This method is characterized by its power and efficiency in producing very low error rates, especially when we reduce a large-dimensional data space to a medium dimensional space of data. Fig. 2 illustrates the steps of the proposed meteorological prediction approach.

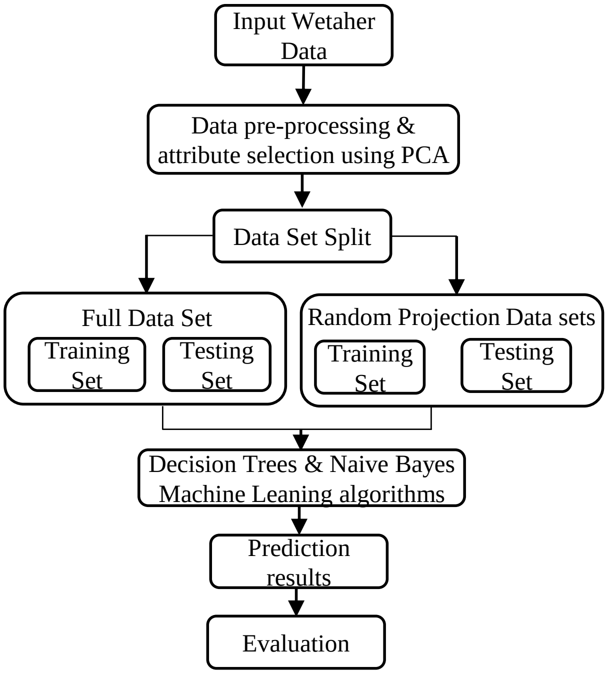

Input Wetaher Data
Data pre-processing & attribute selection using PCA
Data Set Split
Full Data Set
Training Set
Testing Set
Random Projection Data sets
Training Set
Testing Set
Decision Trees & Naive Bayes Machine Leaning algorithms
Prediction results
Evaluation

Figure 2. Implementation process of the suggested prediction approach

To make weather forecasts, the input weather big data has a large number of characteristics such as:

- date;
- time of day;
- air temperature;
- precipitation;
- air pressure;
- wind;
- air density;
- cloud cover;
- radiation surface;
- humidity;
- insolation;
- snowfall and snow depth;
- frost;
- etc.

Some of these parameters are even aggregated and can be broken down. For example :

- air temperature is broken down into maximum, minimum and average temperature, ...
- the wind is broken down into surface wind, wind at a given height, ... and for each type of wind there is the speed, the direction, ...
- etc.

The first step of the proposed approach is then preprocessing data and the extraction of features using the PCA method. The aim is to remove irrelevant information from the data set and to convert the unstructured data set into a structured data set.

PCA is an unsupervised feature extraction technique for converting irrelevant data sets into low-dimensional data with the least amount of error [5]. We chose the PCA method because it is a statistically rigorous approach to simplify the data and generate a new collection of variables called the principal component. Using this method, each generated component is proportionally related to the original variables. Another advantage of PCA is that the principal components of each data set are orthogonal to each other, so that there is no redundant data after preprocessing. The output of the data pre-processing step with PCA is a complete data set, but also reduced data sets obtained using the Random Projection dimensionality reduction method with well-defined reduction rates.

Once the data set is obtained, the next step is to use it by classifiers for evaluation. For this, two types of evaluation are proposed. The first evaluation is conducted by creating machine learning models using the complete data set. The second evaluation is conducted by creating machine learning models using reduced data sets obtained by the Random Projection dimensionality reduction method.

The selected classifiers are used to learn a model from the data sets obtained in the previous steps. Thus, the Decision Tree and Naïve Bayesian classifiers are used for the experimental evaluation. We chose the Decision Tree as the first algorithm because it is a more flexible algorithm that requires little data preparation and can be employed in tasks requiring both classification and/or regression operations. In addition, we chose the Naïve Bayesian algorithm because it is a probabilistic algorithm that has the advantage of using relatively little training data to estimate the parameters needed for classification.

Finally, the evaluation of the prediction results obtained following the application of the Decision Tree and the Naive Bayes algorithms on the different data sets (full data set and reduced data sets) will allow us to analyze the effect of data sampling on the improvement of the performance related to machine learning algorithms in the production of weather forecasts.

Section 4 below details the experimental setup adopted and Section 5 below presents and discusses the prediction results obtained by the algorithms used on the basis of the two types of data sets (full original data set and after reduced data sets).

A. Random Projection Technique

For high-dimensional data analysis, the Random Projection technique is thought to be the most effective, underutilized feature extraction method. This technique is characterized by data independent projection aspect, simpler computation specification and distance preserving property property [21]. The idea behind this technique is to project the data onto a seemingly random subspace, that preserves the Euclidean distances that separates pairs of points further to a projection [23] [26].

The random projection method is based on the Johnson-Lindenstrauss lemma. This theory states that the tiny set of high-dimensional points can be merged into a more manageable subspace, and it also roughly preserves the distance with a higher probability. According to the the said lemma, for a set of N points in p dimensions, the Euclidean distances involving two data points can be preserved by a linear translation into a q-dimensional random subspace up to a factor 1 ± ϵ if the estimated number of dimensions is as follows:

$$q \geq \frac{\log n}{\epsilon^2} \tag{1}$$

where ϵ is a small constant such that $0 < \epsilon < 1$ [27]. This means that the original dimensionality is not relevant for distance preservation. What is important is the total of points that are projected, but also the precision with which the distances will be preserved.

Consider the data set : $X = \{x_1, x_2, ..., x_n\}$. Each point is a vector of dimension p such that $x_i \in R^p$. We want to reduce the data to a space of dimension q such that $1 \leq q < p$. Therefore, dimensionality reduction by random projections consists of the following steps [23]:

- Step 1: place the data in a matrix p×n where n is the number of data points and p is the dimensionality of the data;
- Step 2: create a q×p random projection matrix R^* by means of the MATLAB function randn (q, p);
- Step 3: to project the data downward into a random projection space, multiply the original data by the random projection matrix:

$$X^*_{q\times n} = R^*_{q\times p} * X_{p\times n} \tag{2}$$

These actions therefore amply demonstrate that the data translation into a random projection area is a straightforward matrix multiplication while maintaining the distance.

B. Decision Tree Algorithm

Considered as a supervised machine learning algorithm, Decision Tree adopts a tree structure to continuously seg-

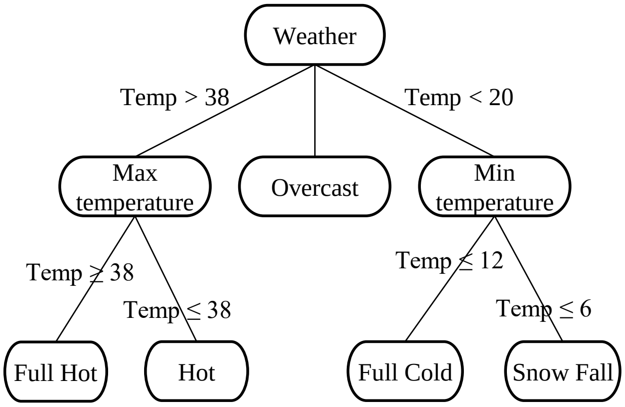


Figure 3. Decision tree for weather forecasting [32]

ment data based on a certain parameter. This algorithm is one of the most powerful solution and famous machine learning algorithms. It is employed in numerous real-world applications and is utilized to tackle regression and classification difficulties [2].

To implement the decision tree model, the following steps should be followed [5] [28]:

- Step 1: Sort out the best component in the given data set that can be designated as the root hub;
- Step 2: Subsets are created by separating the accessible data set for preparation;
- Step 3: A singular subset must have comparative qualities for a component;
- Step 4: Apply the above steps several times until you find the terminal hub for each part of the tree. The terminal hubs will contain the expected qualities.

This algorithm was used in this study as it has been used in several weather forecasting models and has been shown to have high predictive performance capabilities [2] [29] [30] [31]. The weather data is simply visualized as a tree. The tree consists of several branches except for one branch with certain conditions (as if, otherwise). Fig. 3 below shows an example of a decision tree for weather forecasting (predicting weather based on temperature). Four types of weather are represented in the decision tree: Hot, Full Hot, Full Cold and Snow Fall. The top node is considered the root node.

C. Naive Bayes Algorithm

Considered as one of the most popular and efficient algorithms of classification tasks, the principle of Naïve Bayes algorithm is based on the Bayesian theorem related on probability theory established by the Reverend Thomas Bayesian. For this theorem, hypothesis' likelihood depends on both newly available information and existing knowledge. This may serve as a comprehension tool of how the probability of a theory being true is impacted by new evidence. Naïve Bayesian algorithms have been used in many fields, namely biology, transportation, medicine, agriculture, meteorology, etc. In contrast to general Bayesian networks algorithm, Naïve Bayesian classifiers are characterized by their simplicity of construction, thus requiring little background knowledge in the domain. Furthermore, naive networks classifiers have a very constrained spatial and temporal complexity [33].

Naïve Bayesian handles large data sets well and can deal with unrelated data. The Bayes' theorem equation is formulated as follows (where Y is the class label and X represents the input vector including the data) [34]:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{3}$$

where :

- P(Y|X): it stands for the prior probability of the hypothesis Y given that the X is true;
- P(X|Y): it stands for the likelihood of the X given that the hypothesis Y is true;
- P(Y): it stands for the prior probability of the hypothesis Y;
- P(X): it represents the prior probability that the X is true.

This formula appears simple but has enormous practical significance in multiple applications. It is at the heart of Bayesian statistics, and is used to calculate the probability of a given event based on earlier empirical data-based probabilistic estimates. Indeed, the probabilities P(X|Y), P(Y) and P(X) are often easier to calculate when the requirement is the probability P(Y|X) [33]. However, with Naïve Bayesian classifier, the classification task requires several indices to determine which class is suitable for the sample under analysis [34]. Therefore, the previous formula must be adjusted in the following equation:

$$P\left(Y_j \mid X_1, \ldots, X_n\right) = \frac{P(Y)P\left(X_1, \ldots, X_n \mid Y_j\right)}{P\left(X_1, \ldots, X_n\right)} \tag{4}$$

The Y_j variable constitutes the class, whereas the $X_1, ..., X_n$ variables constitute the data to classify characteristics.

4. EXPERIMENTAL SETUP

In this present section, we study the experimentation of the suggested approach. To do so, we first describe the data set used in this experimentation. Then, we describe and analyze the results of the experimentation conducted.

TABLE I. WEATHER DATA SET OVERVIEW

Year	Mont	Day	Hour	Temp	Preci	Wind	Hum	Press	Visib	...
…	…	…	…	…	…	…	…	…	…	
2022	1	1	0	11	0	11		1030	10	
2022	1	1	1	13	0	24	74	1029,4	10	
2022	1	1	2	13	0	22		1029	10	
2022	1	1	3	13	0	26		1028	10	
2022	1	1	4	12,2	0	19	77	1028,9	10	
2022	1	1	5	12	0	19		1028	10	
2022	1	1	6	12	0	17		1028	10	
2022	1	1	7	11,2	0	15	81	1029,3	10	
2022	1	1	8	11	0	13		1029	10	
2022	1	1	9	11	0	17		1030	10	
2022	1	1	10	12,8	0	17	77	1031,1	28	
2022	1	1	11	15	0	17		1031	10	
2022	1	1	12	17	0	17		1031	10	
2022	1	1	13	18,3	0	17	56	1031	40	
2022	1	1	14	19	0	7		1030	10	
2022	1	1	15	19	0	6		1029	10	
2022	1	1	16	18,1	0	11	62	1029,6	30	
2022	1	1	17	17	0	11		1029	10	
2022	1	1	18	15	0	9		1030	10	
2022	1	1	19	13,8	0	6	78	1030,4	28	
2022	1	1	20	13	0	6		1030	10	
2022	1	1	21	12	0	9		1030	10	
2022	1	1	22	11,2	0	9	85	1030,9	17	
…	…	…	…	…	…	…	…	…	…	

TABLE II. CONDUCTED EXPERIMENTATIONS

	Decision Tree classifier	**Naïve Bayes classifier**
Experiment -1- (Full Data Set)	100% of the Data set	100% of the Data set
Experiment -2- (Reduced Data Set)	75% of the Data set	75% of the Data set
	50% of the Data set	50% of the Data set
	25% of the Data set	25% of the Data set

A. *Data Set Description*

To test the proposed approach, the data set used consists of meteorological data extracted from a number of websites presenting meteorological information on cities in Morocco. It is a weather directory that contains the weather data for more than 260 cities and villages in Morocco.

This data set is obtained using the web scraping technique. To achieve the scraping of the information from the web source, the source code of the HTML pages is extracted to correctly analyze where the tags and data are located.

The file format used is CSV (Comma Separated Values). The file contains 2500 rows and 35 columns. An extract of the data set used is shown in Table 1.

The meteorological dataset used in this study is for the city of Tangier (in the north of Morocco) and was collected from meteorological web sources on an hourly scale and covers the period from January 2018 to January 2022.

The last column of the dataset contains the meteorological classification corresponding to the data in each row. Thus, the dataset contains 7 weather classifications, namely: sun, partly cloudy, cloudy, fog, drizzle, rain and snow.

For the analysis of these data, the data from the two data sets (the full data set and after the reduced data sets) were divided into test data and training data. For each data set, we used 20% of the data for testing, and 80% for training the models for the two machine learning classification algorithms (Decision Tree and Naïve Bayesian).

B. *Experimentation Methodology*

The objective of the conducted experiment is to study and analyze the influence of dimensionality reduction on the performance of machine learning classifiers. After preprocessing the weather data with the PCA method by removing attributes deemed irrelevant and undesirable, two kinds of data sets were created: the full data set and after the reduced data sets obtained using the ”Random Projection” method with reduction rates. The data sets are then used by two machine learning classifiers: Decision Tree and also Naïve Bayes to evaluate whether the dimensionality reduction of the of the weather data maximizes the performance of the models created (and thus improves the weather forecast). This led us to conduct two experiments:

- In the first experiment, the full data set (100% of the original data set) is used to develop machine learning models for weather prediction using the two classifiers: Decision Tree and Naive Bayes;
- In the second experiment, reduced data sets (with reduction rates of 25%, 50% and 75%) obtained with the Random Projection method are used to develop machine learning models for weather prediction using the same classifiers: Decision Tree and Naive Bayes.

The experiments of the two machine learning classifiers are presented in Table 2 according to the different types of data set used.

In the first experiment (Table 2), the machine learning models were run 20 times (for each classifier) and the average accuracy was retained.

Similarly, in the second experiment (Table 2), for each reduced data set, the machine learning models were run 20 times (for each classifier) and the average accuracy was retained.

Thus, the two experiments resulted in a total of 160 runs (40 runs in the experiment n°1 and 120 runs in the experiment n°2). We multiplied the number of runs and took the average accuracy because we believe that repeating the runs will give more confidence in assessing the accuracies of the models created and avoid any biased or random results.

In the following section, the average training accuracy values of the two classifiers in the two experiments are

TABLE III. ACCURACY OF THE DECISION TREE CLASSIFIER WITH THE FULL DATA SET AND THE REDUCED DATA SETS

Data Set percentage	Decision Tree classifier accuracy
25% of the original Data set	72.5 %
50% of the original Data set	78%
75% of the original Data set	77.9%
100% of the original Data set	78.4%

presented and discussed.

5. RESULTS AND DISCUSSION

Following the experimentation conducted, a variety of results were obtained by training the models with 80% of the data and testing them with 20% of the data in the different cases. This section presents and evaluates the developed models.

The first results concern the evaluation of the Decision Tree classifier applied on the full data set and on reduced data sets obtained by reducing the dimensionality with the Random Projection method with the following reduction rates: 25%, 50% and 75%. For each data set, the execution of the machine learning model was repeated 20 times and the average accuracy was used.

The measure of performance of the generated models is obtained by using the accuracy metric. The accuracy represents the total proportion of observations that were correctly predicted mathematically and is defined as follows:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

where the TP denotes “True Positive”, the FP stands for “False Positive”, the TN signifies “True Negative” and the FN refers to “False Negative”.

The Table 3 presents the accuracy of the Decision Tree classifier used with the different data sets (the full data set and the reduced data sets employing Random Projection technique).

The Fig. 4 illustrates the accuracy of the Decision Tree classifier used with the different data sets (full data set and reduced data sets with the Random Projection technique).

The average accuracy of the Decision Tree classifier when using the full data set is around 78.4%. This accuracy decreased slightly when the classifier used 75% of the original weather data set. However, with only 50% of the original weather data set, the accuracy returns to increase again to 78% with a negligible difference (0.4%) to the accuracy obtained with the full data set. What is also noticeable is that the accuracy dropped sharply when only 25% of the original data set is used.

Superficially, it appears that the Random Projection sampling technique did not contribute significantly to enhance the performance of the used Decision Tree classifier. However, the in-depth analysis of these results allows us to conclude that, even with the slight decrease in performance of the model generated with 50% of the data compared to the one generated with the full dataset, the Random Projection sampling technique has allowed a considerable reduction in terms of processing resources used.

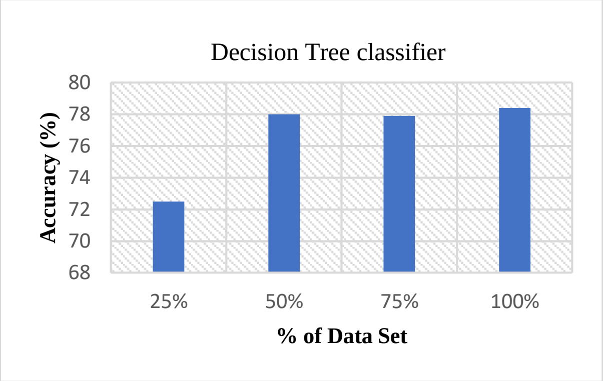


Figure 4. Measuring the accuracy of the Decision Tree classifier with the full data set and reduced data sets

TABLE IV. ACCURACY OF THE NAÏVE BAYES CLASSIFIER WITH THE FULL DATA SET AND THE REDUCED DATA SETS

Data Set percentage	Naive Bayes classifier accuracy
25% of the original Data set	80.9%
50% of the original Data set	83.9%
75% of the original Data set	83.5%
100% of the original Data set	81.6%

Specifically, this implies that with large amounts of data, the data sampling technique provides the advantage of saving processing resources without overly influencing the performance of the generated prediction models. This proves the potential of using sampling approaches for processing meteorological big data by reducing the processing resources without much affecting the weather forecast.

As for the evaluation of the Naïve Bayes classifier, it is also applied to the full data set and to the reduced data sets obtained by sampling the data with the Random Projection method with reduction rates of 25%, 50% and 75%.

As with the Decision Tree classifier, the execution of the models generated with Naïve Bayes was repeated 20 times for each data set (complete or reduced) and the average accuracy was used. The performance measure of the generated models is also calculated using the accuracy metric. Table 4 presents the accuracy of the used Naïve Bayes classifier used following the different data sets (full data set on the one hand and reduced data sets with the Random Projection technique on the other hand).

The Fig. 5 illustrates the accuracy of the used Naïve

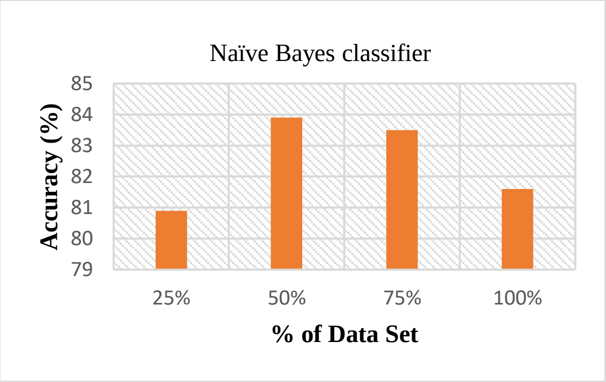


Figure 5. Measuring the accuracy of the used Naïve Bayes classifier with the full data set and the reduced data sets

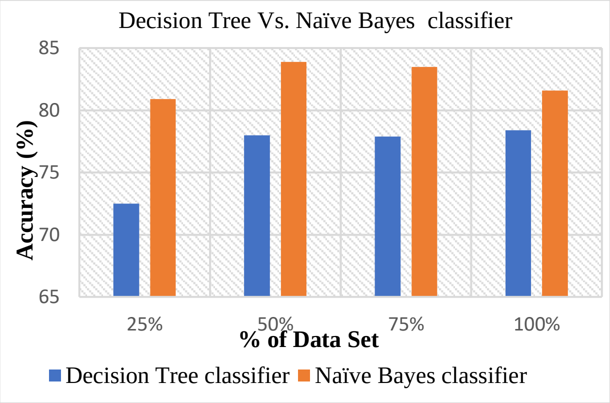


Figure 6. Comparison of the accuracy of Decision Tree and Naïve Bayes classifiers with and without Random Projection

Bayes classifier and this according to the different types of data sets (full data set on the one hand and reduced data sets with the Random Projection technique on the other hand).

When the full data set is employed, the accuracy of the Naïve Bayes classifier is around 81.6%. However, the scenario of this experiment has a different behavior than the one found with the Decision Tree classifier, and the Random Projection sampling technique brought advantages to the performance of the Naïve Bayes classifier. Indeed, the accuracy increased to 83.5% when the data set is reduced to 75% of the original data. With the data set further reduced to only 50%, the performance becomes even the best among all cases and the accuracy of the generated model reaches 83.9%.

This discrepancy in accuracy raises a major question about the importance of using the full data set for the analysis of meteorological big data for forecasting purposes, since sampling these big data can provide even better model performance and consume fewer processing resources.

Indeed, the achievement of better weather forecasts with at least two reduced data sets (75% and 50%) is a good indication that data sampling techniques can contribute to the improvement of weather forecasts while saving processing resources. Furthermore, the existence of low performance with too small data sets (case where the data set is reduced to only 25%) can be justified by the fact that with these small data sets some valuable data is lost.

We therefore conclude that, when used in the processing of meteorological big data, sampling techniques could provide valuable improvements in weather forecasting. However, the tuning of the classifier for the reduction of the original data set (i.e. choice of the reduction threshold) remains another aspect to be studied.

Fig. 6 shows a comparison of the accuracy of both Decision Tree classifier and Naïve Bayes classifier with different data sets (the full data set and the reduced data sets with Random Projection).

In Fig. 6, we can easily see that, with different data sets, the overall accuracy of the Naïve Bayes model is more important than that of the Decision Tree model. With a data set reduced to only 25%, the gap between the performance of the Bayes model and that of the Decision Tree model is wider (about 8.5%). But as the data increases, this gap becomes tiny (about 3% with the full data set, which is negligible). This comparison highlights the contribution of the dimensionality reduction approach (in particular Random Projection) in improving the performance of Machine Learning models in the analysis of big weather data and thus improving weather forecasting results.

6. CONCLUSION AND FUTUR WORK

Weather forecasting plays a decisive role in the lives of individuals and organizations. Today, these weather forecasts are in many cases made using machine learning models that use big weather data from a variety of circumstances related to the environment over a long period of time and that include weather parameters that vary over time. Applying traditional machine learning models to this complicated and poorly structured data can lead to inaccurate weather forecasts.

This study analyzed the contribution of large data sampling techniques, in particular the Random Projection method, on improving the performance of Machine Learning models used in weather forecasting. Thus, two algorithms of machine learning (Decision Trees and also Naive Bayes) are used with a full data set and with reduced data sets (at 25%, 50% and 75%) with the Random Projection method to predict the weather.

With reduced data sets (75% and 50%), the Naive Bayes classifier performed better (over 83%) compared to the case where the full data set is used (81%). This proves that sampling weather data can provide even better model performance with less processing resources, which raises the question of whether it is worth using the full data set for weather big data analysis.

Furthermore, when using the Decision Tree classifier with reduced data sets (75% and 50%), the difference in performance compared to the case where the full data set is used is only about 3%, which is negligible. Therefore, the Decision Tree based model can also be considered as an efficient weather forecasting model since it minimizes the processing resources without affecting the weather forecasting performance too much.

As can be concluded, dimensionality reduction in general is a useful tool to enhance the performance of machine learning models used in weather forecasting while reducing processing resources.

However, weather forecasting remains an open challenge for future research on several levels. Weather forecasting by taking other attributes and using other dimensionality reduction methods are the research work we are considering in the future.

References

[1] M. Fahim, A. E. Mhouti, T. Boudaa, and A. Jakimi, "Modeling and implementation of a low-cost iot-smart weather monitoring station and air quality assessment based on fuzzy inference model and mqtt protocol," *Modeling Earth Systems and Environment*, pp. 1 – 18, 2023.

[2] I. Oshodi, "Machine learning-based algorithms for weather forecasting," *International Journal of Artificial Intelligence and Machine Learning*, 2022.

[3] C.-H. Lee, S. Lin, C.-L. Kao, M.-Y. Hong, P. Huang, C.-L. Shih, and C.-C. Chuang, "Impact of climate change on disaster events in metropolitan cities -trend of disasters reported by taiwan national medical response and preparedness system." *Environmental research*, vol. 183, p. 109186, 2020.

[4] V. Ramachandran, R. Ramalakshmi, B. P. Kavin, I. Hussain, A. H. Almaliki, A. A. Almaliki, A. Y. Elnaggar, and E. E. Hussein, "Exploiting iot and its enabled technologies for irrigation needs in agriculture," *Water*, 2022.

[5] S. Sen, S. K. Saha, S. Chaki, P. Saha, and P. Dutta, "Analysis of pca based adaboost machine learning model for predict mid-term weather forecasting," *Computational Intelligence and Machine Learning*, 2021.

[6] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *Journal of Big Data*, vol. 8, pp. 1–11, 2021.

[7] S. Madan, P. Kumar, S. Rawat, and T. Choudhury, "Analysis of weather prediction using machine learning & big data," *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 259–264, 2018.

[8] S. Vannitsem, J. B. Bremnes, J. Demaeyer, G. R. Evans, J. R. Flowerdew, S. Hemri, S. Lerch, N. M. Roberts, S. E. Theis, A. Atencia, Z. B. Bouallègue, J. Bhend, M. Dabernig, L. D. Cruz, L. Hieta, O. Mestre, L. Moret, I. O. Plenkovi'c, M. J. Schmeits, M. Taillardat, J. V. den Bergh, B. V. Schaeybroeck, K. Whan, and J. Ylhaisi, "Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world," *arXiv: Atmospheric and Oceanic Physics*, 2020.

[9] M. Fathi, M. H. Kashani, S. M. Jameii, and E. Mahdipour, "Big data analytics in weather forecasting: A systematic review," *Archives of Computational Methods in Engineering*, vol. 29, pp. 1247–1275, 2021.

[10] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[11] C.-W. Tsai, C.-F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data*, vol. 2, pp. 1–32, 2015.

[12] G. B. Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *An International Journal on Information Fusion*, vol. 28, pp. 45 – 59, 2015.

[13] J. Zakir, T. Seymour, and K. Berg, "Big data analytics," *Issues in Information Systems*, vol. 16, pp. 81 – 90, 2015.

[14] P. M. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, pp. 78 – 87, 2012.

[15] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. A. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, pp. 1–21, 2015.

[16] Ó. G. Hinde, G. Terrén-Serrano, M. A. Hombrados-Herrera, V. Gómez-Verdejo, S. Jiménez-Fernández, C. Casanova-Mateo, J. Sanz-Justo, M. Martínez-Ramón, and S. Salcedo-Sanz, "Evaluation of dimensionality reduction methods applied to numerical weather models for solar radiation forecasting," *Eng. Appl. Artif. Intell.*, vol. 69, pp. 157–167, 2018.

[17] E. García-Cuesta, R. Aler, D. Pozo-Vázquez, and I. M. Galván, "A combination of supervised dimensionality reduction and learning methods to forecast solar radiation," *Applied Intelligence*, 2022.

[18] M. R. G and R. Dharavath, "Dssae-bboa: deep learning-based weather big data analysis and visualization," *Multimedia Tools and Applications*, vol. 80, pp. 27 471 – 27 493, 2021.

[19] C. J. Talsma, K. C. Solander, M. K. Mudunuru, B. Crawford, and M. Powell, "Frost prediction using machine learning and deep neural network models," *Frontiers in Artificial Intelligence*, vol. 5, 2023.

[20] A. H. Ali and M. Z. Abdullah, "A novel approach for big data classification based on hybrid parallel dimensionality reduction using spark cluster," *Comput. Sci.*, vol. 20, 2019.

[21] R. Siddharth and G. Aghila, "Randpro- a practical implementation of random projection-based feature extraction for high dimensional multivariate data analysis in r," *SoftwareX*, vol. 12, p. 100629, 2020.

[22] H. T. Pham, J. L. Awange, and M. Kuhn, "Evaluation of three feature dimension reduction techniques for machine learning-based crop yield prediction models," *Sensors (Basel, Switzerland)*, vol. 22, 2022.

[23] S. Mylavarapu and A. Kabán, "Random projections versus random selection of features for classification of high dimensional data," *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pp. 305–312, 2013.

[24] R. S. Chaulagain, S. Pandey, S. R. Basnet, and S. Shakya, “Cloud based web scraping for big data applications,” *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 138–143, 2017.

[25] E. Vargiu and M. Urru, “Exploiting web scraping in a collaborative filtering- based approach to web advertising,” *Artif. Intell. Res.*, vol. 2, pp. 44–54, 2012.

[26] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” in *Knowledge Discovery and Data Mining*, 2001.

[27] A. K. Menon, “Random projections and applications to dimensionality reduction,” 2007.

[28] Y. Song and Y. Lu, “Decision tree methods: applications for classification and prediction,” *Shanghai Archives of Psychiatry*, vol. 27, pp. 130 – 135, 2015.

[29] P. Benáek, A. Farda, and P. těpánek, “Postprocessing of ensemble weather forecast using decision tree–based probabilistic forecasting methods,” *Weather and Forecasting*, 2023.

[30] D. Chauhan and J. Thakur, “Boosting decision tree algorithm for weather prediction,” 2014.

[31] R. Kumar, “Decision tree for the weather forecasting,” *International Journal of Computer Applications*, vol. 76, pp. 31–34, 2013.

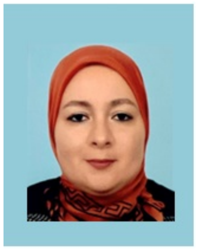
[32] Bhatkande and R. G. . Hubballi, “Weather prediction based on decision tree algorithm using data mining techniques,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, pp. 483–487, 2016.

[33] S. Wu, J. Zhu, and Y. Wang, “Weather forecasting using naïve bayesian,” 2012.

[34] A. U. Azmi, A. F. Hadi, Y. S. Dewi, I. M. Tirta, F. Ubaidillah, and D. Anggraeni, “Naive bayes classifier (nbc) for forecasting rainfall in banyuwangi district using projection pursuit regression (ppr) method,” *Advances in Computer Science Research*, 2022.

Abderrahim El Mhouti is a computer science professor at the Faculty of Science at the Abdelmalek Essaadi University in Tetouan, Morocco. He obtained a PhD in computer science in 2015 from the same university. Its research interests include: Big Data analytics, machine learning, cloud computing and educational technologies. Mr. El Mhouti has a number of articles in his field of research.

Mohamed Fahim is a computer science professor at the Faculty of Science and Technologies at the Abdelmalek Essaadi University in Tetouan, Morocco. He obtained a PhD in computer science in 2019 from the Moulay Ismail University, Morocco. Its research interests include: Big Data analytics, NLP, machine learning and educational technologies. Mr. Fahim has a number of articles in his field of research.



Asmae Bahbah holds a Master degree in Applied Mathematics in 2015. She is PhD student at the laboratory of Computer Science and University Pedagogical Engineering at the Abdelmalek Essaadi University in Morocco. She works on the computerized modeling of mathematical problem solving processes in pupils and students. She is active in several scientific events and national and international conferences.

Yassine El Borji is a computer science professor at the National School of Applied Sciences at the Abdelmalek Essaadi University in Tetouan, Morocco. He obtained a PhD in computer science in 2016 from the same university. Its research interests include: gamification, serious games and mixed reality, simulations, machine learning, integrated data models and data exchange formats.

Adil Soufi is a computer science professor at the Faculty of Science and Technologies at the Abdelmalek Essaadi University in Tetouan, Morocco. He obtained a PhD in computer science from the same university. His research fields include machine learning, e-learning, modeling, and fitting for epidemic models. He has published several articles in his areas of research.

Ayoub Aoulalay is a computer science PhD student at Abdelmalek Essaadi University, Tetouan, Morocco. He obtained a Master degree in Embedded systems and robotics from the same university. His research fields include: Machine Learning, Deep Learning. He has published several articles in his areas of research.

Chaimae Ouazri is a computer science PhD student at Abdelmalek Essaadi University, Tetouan, Morocco. She obtained a Master degree in computer science from the same university. Its research interests include: Big data analytics, Internet of Thing, machine learning, deep learning.