International Journal of Advanced Engineering and Technology
www.allengineeringjournal.com
ISSN: 2456-7655
Received: 07-12-2022, Accepted: 24-12-2022, Published: 10-01-2023
Volume 7, Issue 1, 2023, Page No. 1-4

Efficient comprehension of weather pattern from the meteorological dataset acquired from airport sectors of Fujairah, UAE, using machine learning and data mining approaches

Amnah Saeed Sulaiman Aldhanhani, Muhammed Sirajul Huda Kalathingal, Shaher Bano Mirza*, Fouad Lamghari Ridouane
Fujairah Research Centre, Fujairah, United Arab Emirates

Abstract
Among the major challenge that climatic department encounters are to predict the weather properly. These forecasts are significant because they impact daily living as well as the economics of a county or even a region. Weather forecasting is especially vital since it is the first line of defense against natural catastrophes. They also aid in reducing deprivation and limiting the mitigation procedures that must be implemented following a natural disaster. Many academics recently suggested that machine learning algorithms may make reasonable weather forecasts despite having no deep understanding of climate science. Relatively high scientific methods and practices, such as machine learning algorithm implementations, are required for an efficient comprehension of weather patterns. In this work, we employed the random forest machine learning classifier to characterize the meteorological sets of data with approx. 80% accuracy.

Keywords: weather, temperatures, wind, humidity, classification, random forest, Fujairah, UAE

Introduction

Data analysis is essential for locating important information and making decisions. In many rapidly increasing industries, such as medicine, meteorology, entertainment, farming, and educational, data analysis is used for business development and to promote consumer delight. So, Weather and climate change have several consequences for human society. Rainfall is an indication of a weather component, and changes in climate affects the amount of these components. Weather conditions have a significant influence on energy supplies such as natural gas and electrical power. Climate changes across the years for example; monsoon or clean; hot or cold weather, have a significant influence on civilization in every way imaginable.

In most instances, traditional numerical depictions are challenging to comprehend and understand, and it is necessary to improve the interpretation of meteorological information. The present study's major goal is to create a weather categorization system. To categorize weather patterns, machine learning approaches such as random forest classifications and data analytics are mainly used. The machine learning algorithm random forest classification employs an ensemble learning approach in which 2 or even more machine-learning models are blended to create a single output. While the information is being prepared, many decision trees are built, and then the classifying strategies of each tree are decided.

Data mining is the process of finding useful data by looking through a significant amount of information using a logical technique. The goal of this method is to discover previously unknown patterns. Once recognized, these patterns may be utilized to assist entrepreneurs in making decisions [5]. As a consequence of information technology innovation, massive databases and records have been created in a variety of disciplines. Due to databases and information technology development, a mechanism for preserving and manipulating this essential data for future policy has been devised. Data mining is the technique of collecting useful information and patterns from massive volumes of data. It can also be known as extracting knowledge, information retrieval for understanding, statistical modeling, or information assessment [5].

There is a massive quantity of data accessible in the information sector. These data is meaningless until it is turned into useful information. These massive amount of data should be evaluated in order to extract useful information. In addition to the data extraction procedure, data mining necessitates other operations such as data preprocessing, integration, data processing, data analysis, pattern evaluation, and presentation of data. When these operations are finished, we will be able to use the obtained data for a number of objectives, such as fraudulent activities, market research, production management, scientific inquiry, and so on [4].

Analyzing data to forecast future values humidity and temperature data is one of the critical elements that might benefit the economy and community. Work has been conducted in this domain for many years. Temperatures, moisture content, and other meteorological parameters have been predicted using a number of approaches [3].

Proposed Approach

The approach employed in this study consists of data mining processes and the random forest machine learning classifier to evaluate meteorological sets of data. It is made up of numerous separate decision trees, as the name indicates. Each tree in the random forest produces a class prediction. The leaves in such tree structures indicate class labels, and the branches represent feature conjunctions that relate to those class labels. The model predicts the class that receives the most scores. It is one of the best ensemble methods for multidimensional data. Each tree is constructed using randomly selected vectors of equitable distribution.

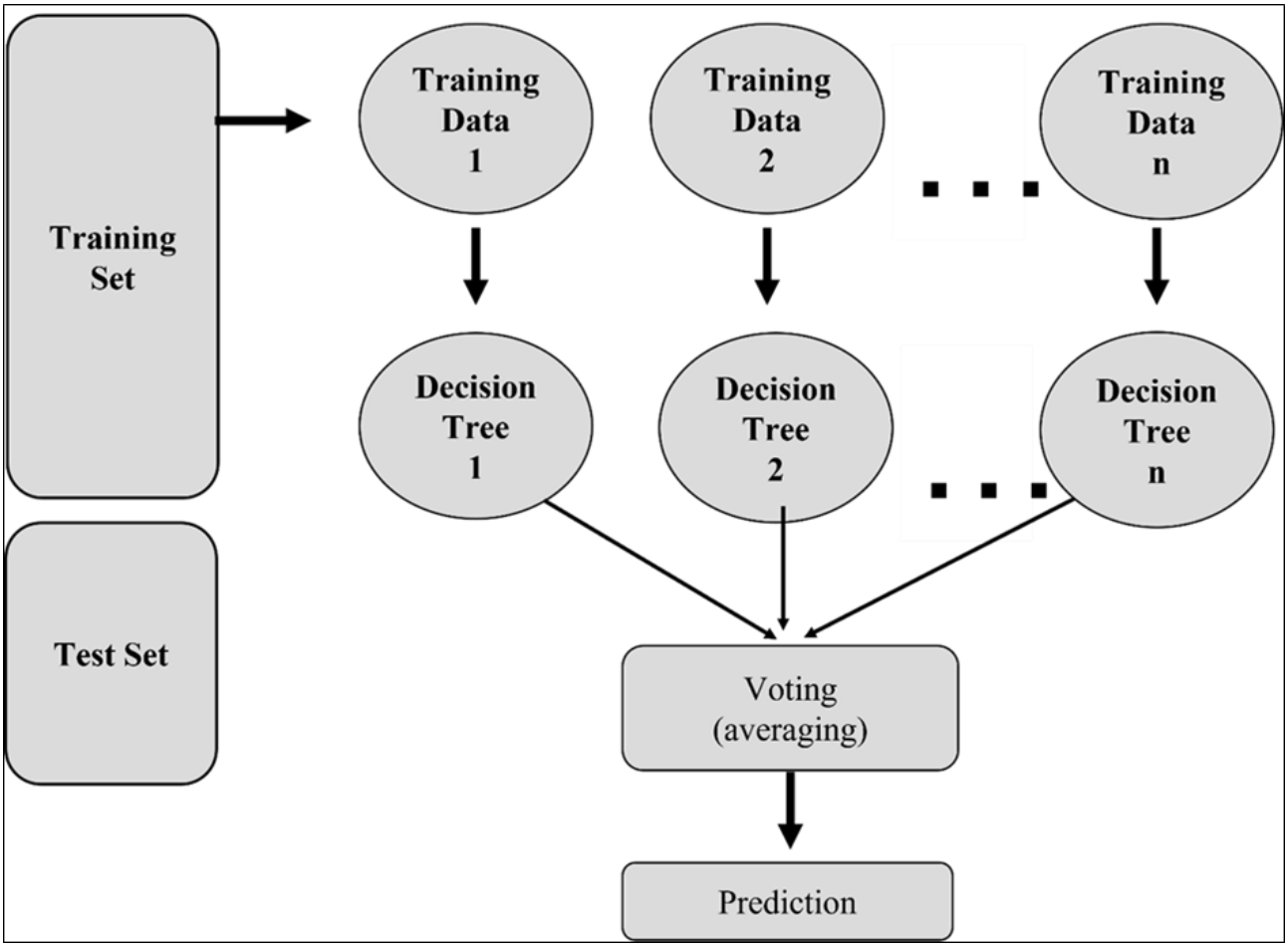

Fig 1: Random Forest architecture describing the algorithm work in which random samples were selected from a given data set to establish decision trees for each sample and getting the prediction results from each decision tree were obtained. After that voting for each prediction result was carried out and then the most predicted result as the last prediction was chosen**.**

Methodology

Data Collection

Data for the proposed study related to Fujairah airport, UAE was collected from the Windfinder web portal (https://www.windfinder.com/forecast/fujairah) and retrieved after every 6 hours since 2010 to 2022. The information was in CSV format and included information such as temperature (degree C), time (am/pm), weather description (classes), lowest temperature (tempL), pressure in Hg (Baro), wind (km/h), Wdirect (angle), and humidity. This information was employed as input for the random forest classifier.

Date	Temperature	Time	classes	TempL	Baro(inHg)	Wind	Wdirect	Hum
2010-12-08 00:00:00	19	"12 am"	"Clear."	19	30	8.078	310	71
2010-12-08 06:00:00	18	"6 am"	"Sunny."	18	30.03	8.078	300	54
2010-12-08 12:00:00	26	"12 pm"	"Sunny."	26	30	9.321	110	52
2010-12-09 00:00:00	22	"12 am"	"Clear."	22	29.95	11.807	260	45
2010-12-10 18:00:00	24	"6 pm"	"Clear."	24	29.92	0	0	44
2010-12-11 06:00:00	18	"6 am"	"Sunny."	18	29.90	8.078	300	53
2010-12-14 06:00:00	24	"6 am"	"Partly sunny."	24	29.96	17.4	280	68
2010-12-14 18:00:00	24	"6 pm"	"Duststorm."	24	29.97	14.914	300	43
2010-12-15 00:00:00	22	"12 am"	"Duststorm."	22	29.96	19.264	310	52
2010-12-15 06:00:00	21	"6 am"	"Duststorm."	21	30.02	12.428	310	56
2010-12-15 12:00:00	26	"12 pm"	"Passing clouds."	26	30.00	9.321	80	46
2010-12-15 18:00:00	24	"6 pm"	"Clear."	24	30.02	2.486	10	60
2010-12-17 06:00:00	20	"6 am"	"Haze."	20	30.03	10.564	310	48
2010-12-18 00:00:00	19	"12 am"	"Clear."	19	30.01	5.593	310	68
2010-12-18 18:00:00	24	"6 pm"	"Clear."	24	30.03	6.836	60	55
2010-12-19 06:00:00	19	"6 am"	"Scattered clouds."	19	30.04	4.35	300	65
2010-12-19 18:00:00	23	"6 pm"	"Passing clouds."	23	30.05	1.24	110	61
2010-12-20 06:00:00	19	"6 am"	"Sunny."	19	30.12	6.836	310	71
2010-12-21 12:00:00	25	"12 pm"	"Sunny."	25	30.06	9.321	50	46

Fig 2: Airport weather dataset collected after every 6 hours since 2010-2022 including temperature range, time, classes, lowest temperature (TempL), Baro (in Hg), wind, Wdirect and humidity.

Data Processing

Applying Python programming, the information collected file was processed and translated into the appropriate format. The data was then divided into training and testing, with 80% for training and 20% for testing.

Table 1: Attribute selected for the classification depicting that Time, Date and classes were under variable type "String" While remaining were in "Numeric" form.

Time (am/pm)	String
Humidity (%)	Numeric
Barometric Pressure (inHg)	Numeric
Wind Speed	Numeric
Wind direction	Numeric
Temperature	Numeric
Minimum Temp	Numeric
Date	String
Classes	String

Table 2: Different output classes

1.	Clear
2.	Sunny
3.	Partly Sunny
4.	Dust storm
5.	Passing cloud
6.	Haze
7.	Scattered Clouds
8.	Overcast
9.	Light rain. Overcast
10.	Light rain passing clouds
11.	Drizzle overcast
12.	lots of rain passing clouds
13.	Fog
14.	thunderstorms passing clouds
15.	thunderstorms passing partly sunny
16.	Thunderstorm scattered cloud
17.	Light rain partly sunny
18.	Thundershowers overcast
19.	Thundershowers partly sunny
20.	thundershowers passing clouds
21.	Hail overcast
22.	rain overcast
23.	drizzle low clouds
24.	light rain low clouds
25.	thunderstorms overcast
26.	sprinkles overcast
27.	drizzle partly sunny
28.	lots of rain passing clouds
29.	thundershowers low clouds
30.	sprinkles passing clouds
31.	thundershowers low clouds
32.	sprinkles passing clouds
33.	rain passing clouds
34.	rain partly sunny
35.	low clouds
36.	hail passing clouds

Different output classes defining that whether a specific attribute is retained or not at a specific target scale. The decision for each attribute may have multiple possibilities. This experiment involves 36 categories of output classes according to the different study areas and different target scales (Table 2).

Model Training

For measuring selected features, a Gaussian Classifier random forest model featuring 100 trees and Gini impurity was chosen. Following training, test dataset was used to determine accuracy. To comprehend the parameters of high relevance in developing the model, feature important scores were also produced.

The random forest classifier is one of the most used algorithms for determining the relevance of features. The feature significance may be calculated using the Random forest approach as the average impurity reductions obtained from all selection trees in the forest. This is true regardless of whether the dataset is linear or non-linear.

Results and Discussion

Feature importance was employed to evaluate features for modeling, troubleshooting, and analyzing data. The feature importance stage provided a set of features as well as an assessment of their significance. Once the relevance of characteristics was determined, they were chosen accordingly. To choose the most important features, feature selection and feature importance strategies were used. It is worth noting that choosing critical features leads in models with optimal computational load while assuring lower classification error due to the variability supplied by less characteristic features. The parameter, feature importance specifies the relevance of each feature in the sequence in which the features were ordered in training dataset.

The relevance of a feature may be quantified on a scale of 0 to 1, with 0 signifying zero importance and 1 suggesting that the characteristic is extremely necessary. Feature importance levels can also be negative, indicating that the feature is unfavorable to model performance. The model developed in this study was to predict the weather classes with an efficiency of 0.863 in this research, and the most essential feature for forecast was time, with most of the features we mentioned except date having excellent relevance in the final forecast (Table 3). The model's comparative accuracy was may be due to the model's various classes and the availability of diverse datasets. The following was printed representing the feature importance.

Table 3: Tabular representation of selected features along with feature importance score

Features	Feature Importance Score
Time	0.187606
Humidity	0.149934
Pressure	0.143388
Wind Speed	0.132637
Wind direct	0.126943
Temperature	0.095887
Minimum Temp	0.095107
Date	0.068498

To get a better idea about the significant features, visualization plot for feature importance using Random Forest Classifier was developed (Figure 3). Feature important score and feature are depicted on X and Y axis respectively, portraying the feature "Time" having significant feature important score while feature "date" with the lowest score.

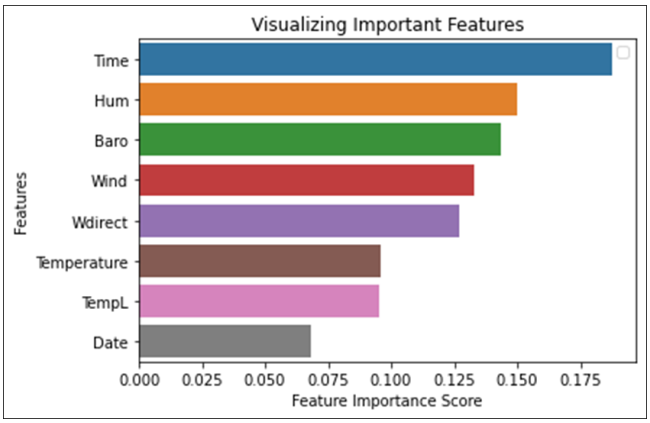

Fig 3: Graphical representations of the top 8 features after feature ranking on in machine learning. The x-axis represents the feature importance score after ranking and y-axis represents top 08 features.

The higher the value of the coefficient (either positive or negative), the greater the impact of the related attribute on the result. We can see from the graph above that feature like Time, Humidity Barometric pressure, Wind speed and Temperature are important in decision making since they have a high significance score.

Conclusion

In this work, the Random Forest machine learning technique was used with an airport dataset of nine meteorological parameters gathered over a 12-year period. The algorithm's output was considered to be rather acceptable, falling into the class of recommended methods for categorization and weather forecasting tasks. So when size of the dataset or the quantity of parameters in the dataset is large, Random Forest is strongly preferred above other decision trees. Further modifications to the algorithm's output can be achieved by applying suitable filters to the dataset during the pre-processing stage.

Conflicts of interest

The authors declare that they have no conflict of interests.

Author Contribution

ASSA involved in the overall conceptualization, experimental work, and preliminary drafting the manuscript. MSHK, technical support in experimental work; SBM, verification and final drafting; FLR, overall support.

Reference

1. TV, kanth R, VV SSS, B, NR. Analysis of Indian Weather Data Sets Using Data Mining Techniques, 2014, 89-94. https://doi.org/10.5121/csit.2014.4510
2. Kothapalli S, Totad SG. A real-time weather forecasting and analysis. *IEEE International Conference on Power, Control,* Signals and Instrumentation Engineering, 2018, 1567–1570. ICPCSI 2017. https://doi.org/10.1109/ICPCSI.2017.8391974
3. Badhiye SSBS, NCP, VWB. Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach International Journal of Emerging Technology and Advanced Engineering Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach, 2012:2(1). www.ijetae.com
4. TutorialsPoint. (2014). About the Tutorial.https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf
5. Ramageri BM. DATA MINING TECHNIQUES AND APPLICATIONS. In Indian Journal of Computer Science and Engineering, 2010, 1. https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications
6. Mathew A, Mathew J. Weather Forecasting Using the Random Forest Algorithm Analysis,2022:4(1). https://doi.org/10.5281/zenodo.6361990