Big Data Analytics For Weather Forecasting

University Name

Course Name

Student Name

Contents

[INTRODUCTION 3](#_Toc120579553)

[DATA COLLECTION 3](#_Toc120579554)

[Research Question 4](#_Toc120579555)

[Research Hypothesis 5](#_Toc120579556)

[DATA DESCRIPTION 5](#_Toc120579557)

[DATA PREPROCESSING 5](#_Toc120579558)

[Acquiring the dataset 5](#_Toc120579559)

[Import Packages Required 5](#_Toc120579560)

[Import the Datasets 6](#_Toc120579561)

[Datasets Merging 6](#_Toc120579562)

[Identifying and Handling Missing Values 6](#_Toc120579563)

[Encoding the Categorical data 6](#_Toc120579564)

[Splitting the Big Data Weather Forecasting Dataset 6](#_Toc120579565)

[Data visualization 7](#_Toc120579566)

[Results and Conclusion 9](#_Toc120579567)

[References 11](#_Toc120579568)

# INTRODUCTION

Weather forecasting is a mechanism that involves the application of technology and science to predict the future atmospheric conditions that a specific location is likely to experience in a given time (Neal, Fereday, Crocker, & Comer, 2016). This is most significant in many ways as farmers get to know and prepare for their planting and harvesting times, event planners get to know the best season to arrange for a specific event to be successful, and many more (Fathi, Haghi Kashani, Jameii, & Mahdipour, 2021). In this project, we are addressing weather forecasting using various classification techniques including, Random forest, Support Vector Machines and Logistic Regression and the Big Data collected. This is done in order to find out whether is possible to make valuable predictions of meteorological conditions which is only achieved based on the previously seen meteorological data. The goal of classification is thus, provided with a set of weather measurements, the classification technique should be able to predict which meteorological condition should occur.

# DATA COLLECTION

Data collection is the second step after topic selection. After we selected and agreed on our topic, “Big Data Analytics for weather forecasting,” the next step was to find the dataset related to the study topic. Data collection took place in one way in our case. We performed a google search over the internet to see historical data related to our study topic. Since this paper project aims at implementing a Big Data project, the data collected had to meet the 5V model requirements of Big Data.

Data collected comes from the Kaggle website, a worldwide data science platform where almost all datasets for practice and challenges events on data science and Big Data take place. We found an existing repository on Kaggle with weather forecasting data already collected and which was to be used for the challenge. We downloaded the dataset into our pc to work on the pc as the working directory. The link to the dataset is provided at <https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data>

.

**Why is Our Big Data in-terms of 5V model?**

Big Data is an enormous quantity of information that cannot be effectively managed using traditional methods. Since then, the idea of "Big Data" has emerged to describe the immediate availability of vast data stores. It has three primary characteristics:

* How quickly information is processed.
* How much information is listed?
* What kind of information is saved (including processed and unprocessed data from many sources)?

The primary goals of Big Data are to enhance a company's or system's responsiveness to a vast amount of collected data, boost productivity, and refine knowledge of customer behavior to deliver targeted marketing and develop novel markets.

The storage, analysis, and visualization of big data can be challenging, which might slow down or even halt subsequent procedures or findings. Big data analytics studies large datasets to discover previously unknown patterns and associations. Companies and organizations can benefit from this data by acquiring a competitive edge through more profound insights. Big data implementation needs to be carefully planned and carried out for this reason. This paper will first examine the concept of "big data" and the various ways it has been characterized. Second, how can it best be utilized, and what are the hallmarks of big data? In the third section, we cover Big Data models and architecture. As a fourth question, what are some ways in which big data could be used to benefit both the machines and people who choose to keep working with it? We'll talk about what the future holds for Big data.

Our data collected is considered to be a Big Data in terms of 5V model because of the following characteristics;

1. Its Volume - the scale and quantity of our big data that we will process and evaluate
2. Its Value - Our Big data's value, the most crucial "V" from a business's perspective, stems from the insights and patterns it reveals, which in turn improve efficiency and strengthen ties with customers/those depending on it.
3. Its Variety – represents Data types, including raw data, semi-structured data, and unstructured data, can be found in a wide variety. This is what is entailed in our data collected.
4. Its Velocity - Its all about data acquisition, storage, and management rates often measured in terms of "velocity," which is defined as "the rate at which information is acquired, stored, and managed by an organization." From data acquisition up to management, it resembles Big Data.
5. Veracity - When it comes to making decisions, top-level executives place a premium on veracity, which is defined as the "truth" or accuracy of data and information assets.

# Research Question

Is it possible for us to make valuable predictions of meteorological conditions only based on previously seen/collected meteorological data?

# Research Hypothesis

1. The meteorological condition for pressure in the country Canada increases from the year 2012 to 2017.
2. The maximum humidity value attained between the year 2012 and 2017 is 70.0 for Israel.
3. The “sky is clear” weather condition was the most dominant condition that we could describe.
4. The year 2016 was the year with highest temperatures as one of the meteorological conditions.

# DATA DESCRIPTION

This dataset contains historically hourly weather measurement data of 36 cities collected from 2012 to 2017. These five years of data results in approximately 45,000 measurements of each town and for every air pressure, humidity, wind speed, temperature, wind direction, and the like. In our dataset folder, there are seven files of CSV. Apart from the city attribute file, all other files contain more than 45,000 measurements of 36 cities. This dataset needs to be merged to work on one running set of data later in the project implementation.

# DATA PREPROCESSING

This is the first step that marks the initiation of classification-building techniques. In most cases, the real-world dataset we collect is incomplete, inconsistent, and inaccurate. This means that this dataset, in any case, must contain errors/ outliers as a result which were caused during data entry processes. Or, we may find that the dataset needs specific attribute values/trends. To work on all this, data preprocessing is a step that helps the analysts bring the dataset back to completeness, consistency, and accuracy. This is achieved by performing data cleaning, formatting, organizing, and structuring the raw data collected. This is the only preparation for implementing the classification techniques. The following are the various steps implemented during data preprocessing to achieve a prepared dataset ready for classification;

## Acquiring the dataset

Having collected data, loading the dataset into the current working environment is the first step of data preprocessing. We must acquire relevant data sets to build appropriate models using classification techniques. This data is composed of data gathered from multiple sources. These are combined in one proper format to form a dataset. In our case, data collected will only be relevant if it contains information such as humidity, rainfall, temperature, and so on for different latitudes and longitude. In simple terms, a weather forecasting dataset must contain weather-related data. Several online sources have been provided where one can download historical data to use. Apart from downloading data, one can also use Python APIs and many ways to scrape data from online platforms. However, in our case, we downloaded data from Kaggle, our data source for this project. The link to our dataset has been provided in the introduction section of this report paper. Our dataset was stored in comma-separated values (CSV) format.

## Import Packages Required

Before working on the downloaded data set, various libraries have to be imported to make it possible to load the data and start working over it. The following are the core python data preprocessing libraries that will be used in our project implementation;

Pandas – loading different file formats, data manipulation, and analysis.

Matplotlib – for 2D plotting of figures

Seaborn - alternative 2D plotting library to matplotlib.

## Import the Datasets

In this step, the downloaded dataset must be imported into the working environment. Therefore, we first set the current directory as the working directory, then using panda’s library “the read\_csv()” function, import the CSV dataset files.

## Datasets Merging

Since we had downloaded large sets of data that came in more than five files, we had to merge all these files to start working on one primary dataset. In our case, we combined the datasets files into one using the reduce() function from the functools class. After joining, the dataset was too huge to work with since our machine disk space could not tolerate the memory required; we decided to take a sample from each dataset before we merged to work with it.

## Identifying and Handling Missing Values

It is pivotal to identify and handle the missing values correctly. Failure to work on these results in inaccurate and faulty conclusions and inferences from the data used. Therefore, to get accuracy in our classification techniques employed, we had to find out the missing values and work on them. (Jia & Zhang, 2021). In this project, we handled the missing values by calculating the mean for the numeric data variables like latitudes, years, and temperature and replacing the missing values with the mean calculated. On the other hand, for categorical variables like weather conditions, the missing values were replaced by finding the mode in the whole column and using the mode to replace the missing values. This practice added variance to our dataset; hence it yields better results as compared to dropping the missing values rows, which would have led to data loss (Chu, Ilyas, Krishnan, & Wang, 2016).

## Encoding the Categorical data

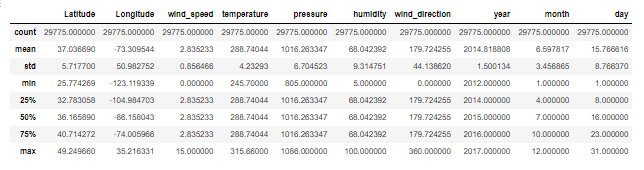
The categorical data variables refer to the information with specific categories. In our dataset, there is one categorical variable, the weather condition variable. Since the classification techniques we want to work with are machine learning models, these models are primarily based on mathematical equations. Therefore, they can’t work with data in such a format. Therefore, keeping categorical data in the equation without encoding will cause specific issues; hence only numbers are needed in the equations. Consequently, we encoded our weather condition variable into numbers for our models to understand the data and work with. This is achieved by using the label Encoder() function from the sci-kit learn library (Shah, Xue, & Aamodt, 2021).

## Splitting the Big Data Weather Forecasting Dataset

After implementing the steps mentioned above, data splitting in data preprocessing is the next move. To enforce any technique in machine learning, every dataset must be partitioned into two sets – the training set and the test set. The training set denotes the subset of data used for training the model. Here, we are already aware of the output. On the other hand, a test subset of data is used for testing the model trained. The model uses this test set to predict outcomes. Data is usually split into 70:30 or 80:20, indicating that 70% or 80% is for training while 30% or 20% is for the testing set. In our case, the splitting done was 70:30. However, it is essential to note that the splitting process continuously varies according to the shape and size of the dataset. This is achieved using the train\_test\_split() function in the sci-kit learn library.

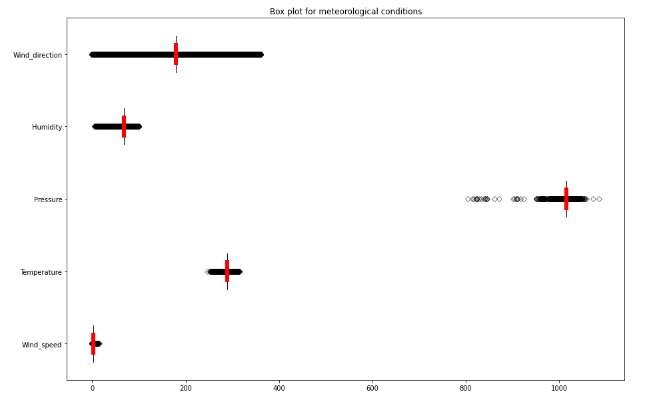
# Data visualization

Data description table.

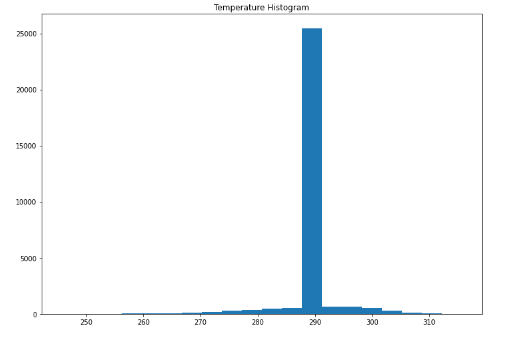


The table above shows the mean, maximum, minimum, quartiles, and standard deviation. Using the max, min, and mean values, we can find the outliers in the data if found in the dataset, then we work on removing them. By looking keenly, we find out that wind speed has outliers since the min value of 0.00, the max value of 15.00, and the mean value of 2.83 have a wide range.

Box plot to check outliers in the dataset

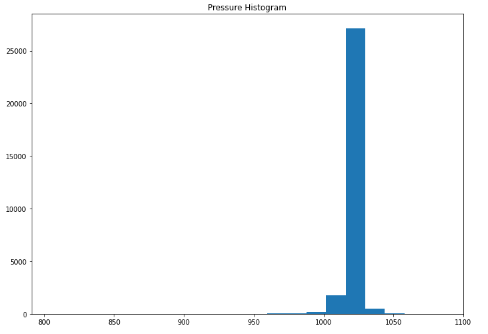


Temperature histogram to show data distribution



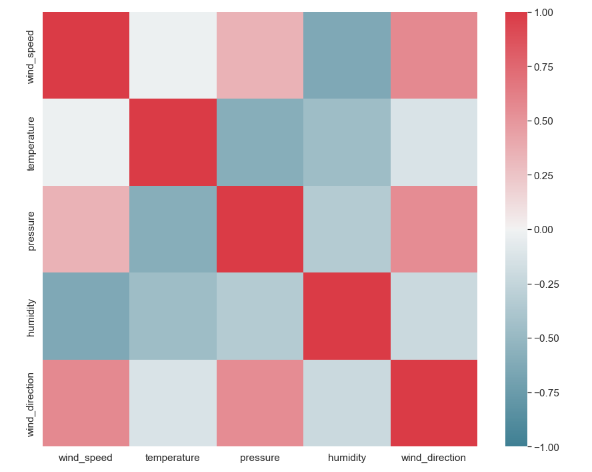
The data distribution is skewed to the right. This means that the distribution of data is not normal distribution.

Pressure Histogram to show data distribution



The data distribution is skewed to the right. This means that the distribution of data is not normal distribution.

Correlation matrix for the selected features of our dataset



There is a negative high correlation between wind speed and humidity. There is a positive medium correlation between wind speed and pressure features. Our dataset has a medium to a high positive correlation between wind speed and wind direction features. There is a positive medium correlation between wind direction and pressure features. The above diagram shows the correlation of the data features as explained above.

# Results and Conclusion

Null hypothesis 1 : The meteorological condition for pressure in the country Canada increases from the year 2012 to 2017.

Alternate hypothesis: The meteorological condition for pressure in the country Canada at first increases from year 2012 up to year 2013, then reduces from year 2014 to 2017 at a constant value.

Null hypothesis 2: The maximum humidity value attained in Israel between the year 2012 and 2017 is 70.0.

Alternate hypothesis: The maximum pressure attained in Israel between the range of years between 2012 and 2017 is at 75.0

Null hypothesis 3: The “sky is clear” weather condition was the most dominant condition that we could describe.

Alternate hypothesis: The “sky is clear” weather condition was the most dominant condition that we could describe.

Null hypothesis 4: The year 2016 was the year with highest temperatures as one of the meteorological conditions.

Alternate hypothesis: The year 2014 was the year with highest temperatures as one of the meteorological conditions.

We also implemented three classification techniques: Support Vector Machine (SVM), Logistic regression, and Random forest. The following interpretation was made from the results attained. The overall we can interpret is that 42% of the SVM and Logistic predictions are correct, and 43% of Random forest predictions are accurate. For an ideal model, RMSE/MAE=0 and R2 score = 1 (Chicco, Warrens, & Jurman, 2021). By checking the three techniques we worked on, it’s clear that none of the models worked better. To improve this, I propose working on other classification techniques like Naïve Bayes Decision trees and many more in future research. Model hyperparameter tuning is also an alternative to improve the model performance. However, I spent more time cleaning the data due to the large dataset. We have seen that both SVM and logistic regression techniques are the best. In future, I propose that we should consider working on a more balanced dataset to have better results. (Jain & Jain, 2017).

# References

Chicco, D., Warrens, M. J., & Jurman, G. (2021). *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. .* PeerJ Computer Science, 7, e623.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. . *In Proceedings of the 2016 international conference on management of data*, 2201-2206.

Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E. (2021). Big data analytics in weather forecasting: . *A systematic review. Archives of Computational Methods in Engineering*, 1-29.

Jain, H., & Jain, R. (2017). Big data in weather forecasting: Applications and challenges. *In 2017 International conference on big data analytics and computational intelligence*, 138-142.

Jia, B. B., & Zhang, M. L. (2021). *Multi-Dimensional Classification via Decomposed Label Encoding. .* IEEE Transactions on Knowledge and Data Engineering.

Neal, R., Fereday, D., Crocker, R., & Comer, R. E. (2016). A flexible approach to defining weather patterns and their application in weather forecasting over Europe. *Meteorological Applications*, 389-400.

Shah, D., Xue, Z. Y., & Aamodt, T. (2021). *Label Encoding for Regression Networks.* In International Conference on Learning Representations.

# Appendix

“Big Data Analytics For Weather Forecasting.ipynb” file contains project code

“Big Data Analytics For Weather Forecasting.pptx” file contains the project presentation.

“dataset” folder that contains the collected data.