International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 28-31
Published Online April 2018 in IJEAST (http://www.ijeast.com)

IMPLEMENTATION OF METEOROLOGICAL DATA ANALYSIS USING TECHNIQUES FOR WEATHER PREDICTION

Ms. Nikita Gupta
Department of IT
PCCOE, Pune, Maharashtra, India

Ms. Rashmi Narayanan
Department of IT
PCCOE, Pune, Maharashtra, India

Ms.Anagha Chaudhari
Department of IT
PCCOE, Pune, Maharashtra, India

***Abstract*— Weather data extraction is a form of data mining concerned to find hidden patterns within the weather data available to a large extent, so that feedback can be converted into knowledge in our daily lives in order to use it in another time. Work has been done in this constrain since years. Meteorological data mining uses finding hidden patterns in available meteorological data and transforming it into usable knowledge. In this study, we have three main algorithms Time-series, K-means and Naïve forecast which are used for weather prediction. Meteorological data analysis considers real time data while making predictions and predicts weather forecasts. The system is be scalable, portable and should work on variety of data. We propose a system by using R programming language for analysis of weather data using Microsoft Azure HDInsight for good long term predictions. We are discussing the application of different data to predict or to associate or to classify or to cluster the pattern of meteorological data.**

Keywords— **Meteorological data, Time-Series, Naïve Forecast, K-means clustering, Azure, HDInsight, Hive, RTVS, R, Shiny**

I. INTRODUCTION

Data analytics (DA) also known as analysis of data, is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Real-time analytics involves dynamic analysis and reporting based on data entered into a system within a time unit before actual time of using the data. Weather data is an important source for testing the results of our analysis in an accurate way. Microsoft Azure is used as a cloud platform, which provides services like HDInsight for all Hadoop related technologies.

This system can use real time weather data stored on a Hadoop cluster and this available data is cleaned, summarized and queried using hive on azure.

This data is then analyzed based on the trend and other parameters using R Programming language. It provides a good long term predictions. By using clustering, Naïve forecast and K-means technique we can acquire weather data and can find hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction of climate condition.

MICROSOFT AZURE HDINSIGHT

- Microsoft Azure provides services like HDInsight for all Hadoop related technologies.
- We plan to build a system that uses real time weather data stored on a Hadoop cluster and this available data is cleaned, summarized and queried using hive on azure.
- This data is then analyzed based on the trend and other parameters using R Programming language.
- Microsoft Azure HDInsight is an Apache Hadoop distribution powered by the cloud[3]. This means that it handles any amount of data, scaling from terabytes to petabytes on demand, with any number of nodes at any time.
- Scale to petabytes on demand[3].
- Crunch all data-structured, semi-structured, unstructured.

Hive

Apache Hive is a data warehouse system for Hadoop, which enables data summarization, querying, and analysis of data by using HiveQL[3]. Hive can be used to interactively explore your data or to create reusable batch processing jobs. Microsoft Azure provides Hive editor through cluster dashboard by providing cluster credentials. All Hive queries can be performed and submitted and results can be seen through job history[3]. Table location and directories can be viewed though file browser. By downloading Microsoft Azure SDK you can create Hive applications which allows you to query on the dataset loaded on HDInsight cluster. This is the done by configuring the name of the HDInsight cluster in the .hql file[3].

II. COMPARISON BETWEEN FORECAST AND PRECISION

PREDICTION: An estimation of any event happening (in past, present or future). Predicted values are calculated for observations in the sample[1].

FORECASTING: Forecasting pertains to out of sample observations, whereas prediction pertains to in sample observations. It is always associated with a time dimension in the future i.e estimation for some specific future duration or over a period of time [1]

III. WHY R PROGRAMMING?

Analysis of data using R programming improve time complexity through the algorithms being implemented.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 28-31
Published Online April 2018 in IJEAST (http://www.ijeast.com)

IV. WHY MICROSOFT AZURE?

By using this cloud customers are given new services in a user friendly way and on a software like microsoft.

V. WHY SHINY?

Shiny makes it easy to build interactive web apps straight from R. Build useful web applications with only a few lines of code—no JavaScript required. Shiny applications are automatically "live" in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser. Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility. Works in any R environment (Console R, Rgui for Windows or Mac, ESS, StatET, RStudio, etc.) Pre-built output widgets for displaying plots, tables, and printed output of R objects[11].

VI. DATA SET DESCRIPTION

- The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event.
- The data set consists of past 60 years data, which has rainfall as an attribute.
- Granularity: Annual as well as monthly rainfall for different regions of India.
- Rainfall unit : mm
- This data set contains monthly rainfall detail of 36 meterological sub-divisions of India[10].

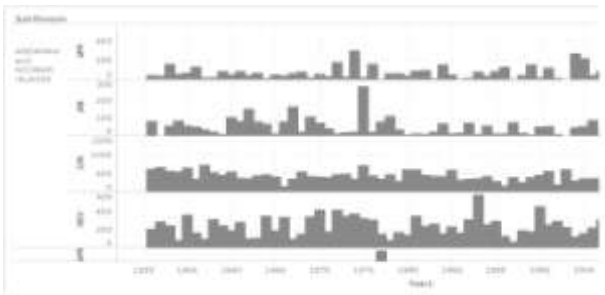

Fig 1 : Graphical representation of dataset for Andaman and Nicobar Islands

In Fig 1, the graphical representation of dataset for Andaman and Nicobar Islands is shown using tableau.

VII. PREDICTION APPROACHES

1. TIME SERIES

Time series is used for understanding the trends in the historical dataset which is available with us. Time series is a series of data points in which each data point is associated with a timestamp[1]. A simple example is the price of a stock in the stock market at different points of time on a given day. Another example is the amount of rainfall in a region at different months of the year which will be the output of this study [3].

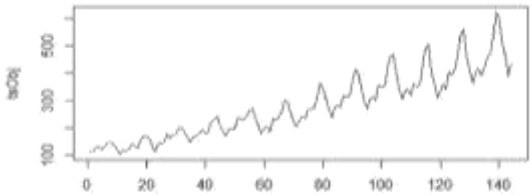

Fig. 2: An example of Time-series plot

R language uses many functions to create, manipulate and plot the time series data. The data for the time series is stored in an R object called **time-series object**. It is also R data object like a vector or data frame. Using the **ts ()** function the time series object is created.

2. K-MEANS

The data given by x are clustered, which aims to partition the points into *k* groups such that the sum of squares from points to the assigned cluster centers is minimized. At the minimum, all cluster centres are at the mean of their sets of data points which are nearest to the cluster centre.

K-Means allows to cluster all items with similar properties which helps to study different behavior or patterns in data. In this case it helps us to understand the rainfall trend in data. K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

Input:

1. K: the number of clusters,
2. D: a data set containing n objects.

Output: a set of k clusters.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers;
2. Repeat
3. (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. Update the cluster mean;
5. Until no change [1];

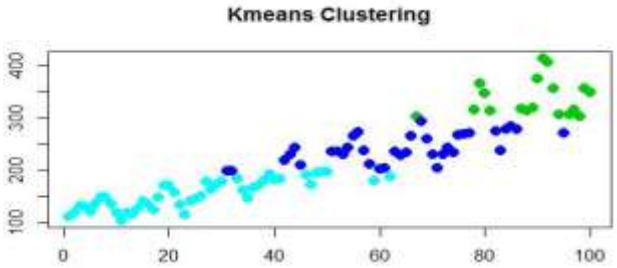

Fig 3 : An example of K-means Plot

K-Means Clustering for this dataset can be done as follows.

- kmeasOut<-kmeans(tsObj,4) where 4 is number of clusters.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 28-31
Published Online April 2018 in IJEAST (http://www.ijeast.com)

3. NAÏVE FORECAST

A **forecast** is the mean or median of simulated futures of a time series. The very simplest forecasting method is called a **naive** forecast[6]. This can be done for many time series including most stock price data, and even if it is not a good forecasting method. For example, if you want to forecast the sales volume for next March, you would use the sales volume from the previous March. This is implemented in the snaive() function, meaning, **seasonal naive**.

naive() returns forecasts and prediction intervals for an ARIMA(0,1,0) random walk model applied to x. snaive() returns forecasts and prediction intervals from an ARIMA(0,0,0)(0,1,0)m model where m is the seasonal period[3].

Naïve method takes two inputs (tsobject of dataset, h)
Where tsobject is the time series object of the datset and h is the prediction period depending upon the value given by the user.

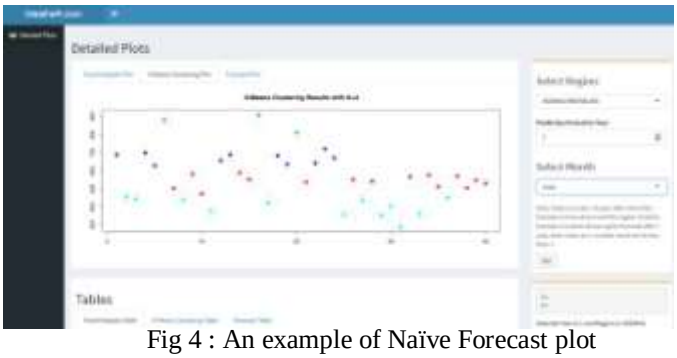

Fig 4 : An example of Naïve Forecast plot

The above Fig 4, is an example of naïve forecast plot. For forecasting methods, you can set the second argument h, which specifies the number of values you want to forecast; as shown in the code below, they have different default values. The resulting output is an object of class forecast [8].

```
naive(y, h = 10)
snaive(y, h = 2 * frequency(x))
```

VIII. PROPOSED ARCHITECTURE

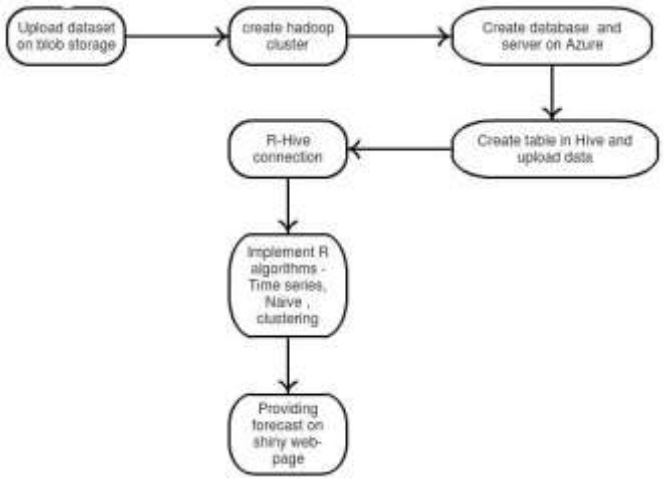

Fig 5: Proposed System Architecture

- In Fig 5, as first step we will be using three algorithms in R for forecasting purpose.
- The very first and important algorithm is Time series that is used for analyzing the trend in the given data its object known as time series object is used an input for other algorithms[7].
- Next, is naïve method used for forecasting of weather based on the input of years you want the forecasting of! Then, after naïve we will be doing clustering and classification so the final output is understandable by the user.
- We will upload dataset on Azure through Azure blob storage from local file System.
- We will connect HDInsight cluster with R code for accessing hive tables and operations. Next we will make Hive- R connection through RODBC(Open database connectivity) connections[7].
- Once connection has been established, we can read dataset from hive into R and then perform further operations and analysis on R.

IX. RESULTS AND ANALYSIS

1. TIME SERIES

Fig 6 : Output of Time series

In Fig 6, the expected output of time series is shown. The time series will show the trends in the dataset proposed in fig 1 above. The output is for 60 years. On the x- axis years are there and on y-axis there is rainfall in mm.

2. K-MEANS

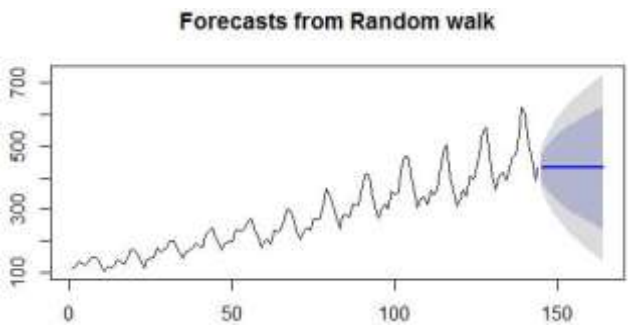

Fig 7: Output of Time series

In Fig 7, the expected output of k-means is shown. The k-means algorithm will cluster the complete data set for the selected region based on rainfall value. The k value is the

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 28-31
Published Online April 2018 in IJEAST (http://www.ijeast.com)

number of clusters. On the x- axis 5 years timespan is represented and on y-axis there is rainfall in millimeter.

3. NAÏVE FORECAST

Fig 8 : Output of Naïve Forecast

In Fig 8, the expected output of Naïve Forecast is shown. The Naïve forecast algorithm gives forecasts ranges depending on the value of the number of years for which you want to forecast. On the x- axis 20 years timespan is represented and on y-axis there is rainfall in millimeter.

4. EXPECTED OUTPUT

Fig 9 : Output available to the user

In Fig 9, the expected output is displayed to the end user. The predicted output is in user understandable format.

X. FUTURE WORK

One growing area of research in the area of forecast systems is API of dataset can be created, a new way to interact with custom geo data. Important industry applications like customer feedback analysis, sales prediction; academic applications like determining the trend in education system, etc can use this service engine. For future enhancement, the system can work on Linux OS providing Linux cluster is available. It can be used for various smart city applications by giving forecast by identifying the user's location.

XI. CONCLUSION

The above mentioned work proposes a new intelligent and effective forecast system approach to develop a Weather prediction system framework for processing large data sets. As MS Azure is a recent platform, any new R&D is a surge to giving customers new services in a user friendly way and on a software like Microsoft. Therefore, now this system would give predictions for datasets that show basic trend in its behavior. It helps to study the Microsoft Azure Platform which is diverse in technologies.

XII. REFERENCE

[1] Nikita Gupta, Rashmi Narayanan, Anagha Chaudhari "Implementation and analysis of data mining techniques for weather prediction" in the ICSTM Volume 6, Issue 11, November 2017.

[2] Meghali A. Kalyankar and Prof. S. J. Alaspurkar , "Data Mining Technique to analyse the Metrological Data", IJARCSSE Volume 3, Issue 2, February,**2013**

[3] Mugdha Kulkarni, Priyusha Nair, Shruti Kulkarni, Swati Shekapure "Meteorological Data Analysis Using HDInsight With RTVS", International Journal In Engineering And Technology,**2015.**

[4] Sarah N. Kohail, Alaa M. El-Hales, "Implementation of data mining techniques for meteriological data analysis", IJICT Volume 1, NO. 3,**2011**

[5] A. R. Chaudhari, D. P. Rana, R. G. Mehta, "Data Mining with Meteorological Data", International Journal Of Advanced Computer Research Volume-3 No-3 Issue-11, September, **2013**

[6] E.Manjula, S.Djodiltachoumy, "Analysis of Data Mining Techniques for Agriculture Data", International Journal Of Computer Science And Engineering Communications, Vol.4, Issue.2,**2016.**

[7] Ms.P.Shivaranjani, Dr.K.Karthikeyan, "A Review of Weather Forecasting Using Data Mining Techniques", International Journal Of Engineering And Computer Science,Volume 5 Issue 12,**2016**.

[8] Fahad Sheikh , S. Karthick, D. Malathi, J. S. Sudarsan and C. Arun, "Analysis of Data Mining Techniques for Weather Prediction", Indian Journal Of Science And Technology, Vol 9(38), October **2016.**

[9] M. Sri Saranya and S. Vigneshwari, "Analysis of Weather Datasets Using Data Mining Techniques", International Journal Of Controltheory And Applications,Volume 10, Issue 14, **2017**.

[10]http://www.tutorialspoint.com/data_mining/

[11] https://www.researchgate.net

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 2, Issue 12, ISSN No. 2455-2143, Pages 28-31
Published Online April 2018 in IJEAST (http://www.ijeast.com)