IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Weather Prediction using Scikit-Learn

Sudhnya Kashikar[1*], Sumedha Patil[2], Ameya Vedantwar[3], Shivani Katpatal[4], Sofia Pillai[5]

[1,2,3,4,5]Dept. of Computer Science & Engineering G.H. Raisoni College of Engineering, Nagpur, Maharastra, India

Corresponding Author: kashikar_sudhnya.ghrcecs@raisoni.net

DOI: https://doi.org/10.26438/ijcse/v7i4.3640 | Available online at: www.ijcseonline.org

Accepted: 12/Apr/2019, Published: 30/Apr/2019

***Abstract*—** Weather is the most important factor in terms of farming and agriculture. It continuous, data-intensive, multidimensional, and chaotic process. These properties of weather make its forecasting a formidable challenge. The most technologically challenged problems of the last century are weather forecasting. The harvest of crops is dependent on this factor. To make an accurate weather prediction is one of the major challenges that is being faced all over the world. Scientists have tried their best to forecast environmental characteristics using a number of methods and some of these methods are more accurate than others. Weather forecasts provide critical information about future weather. Every year notorious weather harms the life, property and many government activities which is usually heavily funded is destroyed, as a result weather forecasting would help government to plan out things in advance to prepare its citizens for the worst of the weather. There are many different methodologies that have come into observation regarding weather prediction. This paper describes one of the many techniques used for prediction of weather which will be beneficial for the farmers, agricultural and scientists. It will help them to better understand the weather for yielding crops and for studying environment too.

Keywords- Regression, Pandas, Scikit Learn, Numpy**.**

I. INTRODUCTION

Agriculture is one of the major revenues producing sectors of India and a source of survival. Numerous seasonal, economic and biological patterns influence the crop production but unpredictable changes in these patterns lead to a great loss to farmers. These risks can be reduced when suitable approaches are employed on data related to soil type, temperature, atmospheric pressure, humidity and crop type. Whereas, crop and weather forecasting can be predicted by deriving useful insights from these agricultural data that aids farmers to decide on the crop they would like to plant for the forthcoming year leading to maximum profit.

Climate prediction is considered as the most important issue both theoretically and scientifically by the world in the most recent decade. This eventually resulted into a great demand for developing models which assist towards effective prediction of the weather data. A decent number of the meteorologists have made estimations for the climate prediction using various model's dependent on time arrangement. In most of the models, the analysis of weather data is carried out by considering a few variables for the assessment of the data. However, the attributes of the weather play a considerable role in weather forecasting. The advancements in the last decade in science and technology helped in proposing dynamic approaches for the prediction of weather by using empirical approaches.

With the increase in the number of weather stations, huge data is available on daily basis, weekly, monthly and yearly basis and the data is stored exponentially. This data is stored and is made available for effective analysis of weather prediction, catastrophe forecasting and for the usage of other departments. The analysis of the related data from this massive data is of crucial importance, and needs mining techniques.

In most of the cases, the prediction is carried out indirectly using statistical methods, which take the advantage of associations between local-scale precipitation and those large-scale atmospheric variables. The approach is considered as statistical downscaling.

Two machine learning algorithms were implemented: linear regression and a variation of functional regression. A corpus of historical weather data for Stanford, CA was obtained and used to train these algorithms. The input to these algorithms was the weather data of the past two days, which include the maximum temperature, minimum temperature, mean humidity, mean atmospheric pressure, and weather classification for each day. The output was then the maximum and minimum temperatures for each of the next seven days.

II. LITERATURE REVIEW

A. Survey on Crop and Weather Forecasting based on Agriculture related Statistical Data:
By Pushpa Mohan, Dr. Kiran Kumari Patil

Clustering based techniques and supervised algorithms are utilized for managing the collected statistical data. Support Vector Machine (SVM), neural networks, Linear Regression are some of the technologies used.

There are various frameworks that use different systems to control information, to determine bits of knowledge and help decision making for agriculturists. Be that as it may, the significant concern is that they either concentrate on one product expectation or estimate any parameter like either yield or cost. This plan is utilized to estimate the climate, yield and cost of real products of Karnataka dependent on verifiable information. Particularly, for Mysore locale, since they are the biggest maker of espresso, ragi, and coarse oats and furthermore the biggest rice delivering area in Karnataka. The measurable information and anticipated yield are open for the agriculturists through an independent easy to use application. This guides agriculturist to settle on the harvest they might want to plant for the coming year, which causes them to acquire most extreme cost for their items. The factual information and anticipated yield are open for the ranchers through an independent easy to understand application. This guides rancher to settle on the yield they might want to plant for the prospective year and encourages them to get most extreme cost for their items.

B. Future Weather Prediction Using Genetic Algorithm and FFT for Smart Farming:

In this paper, advanced farming approach utilizing key advancements like genetic algorithm and FFT. Farmers should be enlisted through android versatile application or from the work area, having Internet connection. Cloud storage is utilized to store farmer details and climate information. Present climate conditions are acquired from the farmer's area utilizing Internet and GPS facilitates for future weather forecast. The model proposed using Data Acquisition Network, Low-Power Wide Area Networks (LPWAN), Weather Forecast Service, Open Weather offers web services, Data Processing System, etc. would help farmers in arranging the pre-post agricultural exercises. A message will be sent about weather and product harm alarms through sms and email to the farmers.

Following is the proposed system algorithm-:
Step 1. Client Register and then login to the system
Step 2. Continuously update weather condition using android
Step 3. Load and Initialize dataset
Step 4. Train the dataset
Step 5. Cloud storage will receive user request
Step 6. Apply genetic algorithm
Step 7. Apply Fast Fourier Transform for better result
Step 8. Predict forecast using FFT

C. IoT-based System to Forecast Crop Frost
BY M.´ANGELGUILL´EN-NAVARRO,FERNANDO PERE˜ N´IGUEZ-GARC´IADEPRAQUEL MART´INEZ-ESPA˜NA

In the paper, we propose a framework dependent on IoT advancements to approach the yield ice issue. This issue is very stressing among the ranchers of the Murcia locale (Spain) as agrarian protection is progressively confined and the monetary misfortune is ending up progressively detectable. There are several techniques / systems to combat frost, however, there is no economic and reliable alert system that warns or helps to forecast when there is going to be a frost so that farmers can activate the anti-frost systems. The proposed system is composed by three components, the data acquisition network the weather forecast service which acts as a complementary source of information and the data processing system. The results demonstrate the viability of our system to accurately forecast the occurrence of frosts in a certain area and opens new research challenges that need to be addressed before obtaining a fully operational forecasting system.

III. PROPOSED METHODOLOGY

A. Machine learning:

1. Being humans we evolved by observing and understanding the patterns & trends in our surrounding which helped us to respond the given situation in a right way.
2. But humans are bound by the limits of finite data computation and understanding where computers outperforms humans largely this is where the concept of machine learning came into existence.
3. It is a branch of Artificial Intelligence which helps the user to harness the computational skills of a machine to analyse and understand the patterns in any given dataset.

B. Supervised Machine Learning:

1. Supervised machine learning includes labelled datasets which gives computer a exact idea whether the given output is correct or not.
2. Every dataset must be divided into two major parts one of which is, training dataset it helps the computer to analyse the trends in any given data and the understand the given outcome through it.
3. Once the computational training is over it needs to be tested against custom values from the testing dataset every dataset must be divided into the training and testing zones which would help us to understand the accuracy levels of our model.

C. Linear Regression:

1. It is a statistical model which helps to understand the given set of inputs and binds a relationship between them which leads to a predictive outcome.

2. It runs on a basic linear equation consisting of y = a + bx where x is the explanatory variable and y is the dependant variable, b is the slope variable and a is the value of y when x is 0.
3. It consist of two major sides i.e. predictive variable and explanatory variable.
4. Explanatory variable or independent variable these are the fixed values which does not rely on any other entity in our dataset to change their state. Every independent variable helps the predictive variable to understand the next outcome, like the wind speed on a given day is independent to the precipitation occurrence on that day.
5. Predictive variable or dependent variable as the name suggests it is the variable or condition we are trying to predict. According to our understanding about independent variable, the predictive variable's state is dependent on the changes occurred in the explanatory variables, like the precipitation occurrence for a given day is dependent on the wind speeds.

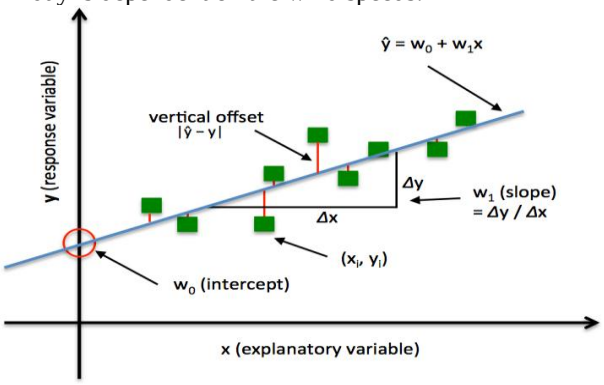

Fig.1. Linear Regression

D. Line of best fit:

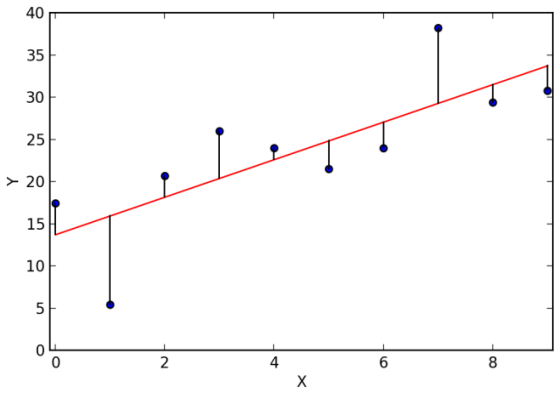

Fig.2. Line of best fit

1. The line of best fit helps the programmer to improvise the accuracy results of a given model.
2. It helps the user to find how close the observed value is to the actual values, the difference between the actual & observed value is the error, we are trying to minimize the error.
3. The line where all the points are equidistant and closest would be the ideal predictive point.

IV. DATA COLLECTION/ TOOLS/ PLATFORMS

A. Data Collection:

1. For data collection we used Kaggle for clean pre-processed data, this website conducts various workshops on machine learning and other topics.
2. Our dataset consist of various atmospheric parameters such as temperature, humidity, dew point, pressure etc.

B. Implementation tools

1. Python
2. Pandas
3. Numpy
4. Scikit Learn

C. Implementation tools

1. Python 3.5 is an open sourced integrated development and learning environment for the python interpreter.
2. It is a multiparadigm programming language which supports object-oriented, imperative, functional and procedural with features like self-memory management.
3. The major advantages of Python is its extensibility of application:
   For an instance a Java language forces the user to download entire set of libraries at the time of installation but on the contrary Python goes with the philosophy to only install that libraries that user actually wants to implement in his/her program.
4. Python's utility ranges from Web & Internet Development, Desktop GUIs, Software developments to Actual Scientific uses.
5. Scientific uses are the domain we are targeting for our project machine learning itself is a mathematical puzzle so we needed a language which is simpler & explicit enough to harness the power of Machine learning in the most efficient way possible. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

D. Pandas:

1. It is a library designed and developed for R-programming language which in background makes use of NumPy array for the implementation of pandas data objects.

2. The main features of pandas are:
   a. Data Frame object for data manipulation with integrated indexing.
   b. Data alignment and integrated handling of missing data.
   c. Reshaping and pivoting of data sets.
   d. Data structure column insertion and deletion.
   e. Data set merging and joining.
   f. Provides data filtration.

E. *Numpy:*
1. NumPy (Numerical Python) is a linear algebra library in Python. It is a very important library on which almost every data science or machine learning Python packages such as Mat−plotlib (plotting library), Scikit-learn, etc depends on to a reasonable extent.

F. *Scikit-Learn:*
1. It is a open-sourced machine learning library consisting of various algorithms consisting of Regression, Classification and Clustering supports.
2. It provides cross-validation functionality to check the accuracy of the unknown datasets.

V. DESIGN AND IMPLEMENTATION

A. Project development:

The project consist of various parts consisting of:
1. Data Collection
2. Data Pre-processing
3. Dividing the dataset into training and testing parts.
4. Assigning a pre-built model by Scikit Learn
5. Optimizing the output graphs for the best fit of co-ordinates.

B. Data Collection:
1. We collected the dataset from Kaggle which is a competition and workshop holding site it helped us by providing metrological data of Jaipur city which was the basis of our search.
2. It consisted of various data points consisting of date of reading, precipitation, dew point, and many more.
3. Every entity is divided into its min mean & max version so that the algorithm gets the comprehensive understanding of the weather data of that day.

C. Data Preprocessing:
1. Pre-processing is the most important factor of machine learning work cycle, as data is the key element for a robust model and perfect prediction in the course of time.
2. All the white spaces, irrational entries and false data was removed by variable python functions.
3. In the below graph we are trying to understand the distribution of pressure across the range of values we have in our data set.

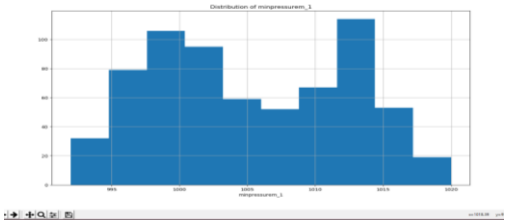

Fig.3. Distribution of minimum pressure

4. In the below graph we are trying to understand the distribution of pressure across the range of values we have in our data set.

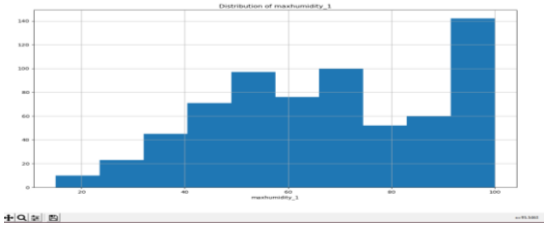

Fig.4. Distribution of maximum humidity

5. We plotted all the graphs using Matplotlib functionality provided by Python built in libraries.

D. Dividing the dataset into training and testing parts:
1. We used Scikit learn's function named as train_test_split for this purpose.
2. It divides the dependent variables and the independent variables into 2 parts consisting of 80% training data and 20% testing data out of the total dataset.

E. Assigning a pre-built model by Scikit Learn:
1. Scikit learn has a plethora of built-in models for machine learning purpose where every model serves a very specific purpose.
2. Since we intend to predict the future cases we are making use of regression algorithm named as Linear Regression.
3. The Scikit learn's library must be imported before initiating the application process. So Scikit learn defines Linear Regression in the form of Linear Model.
4. The Fit function will fit the given csv file's entities on a linear repressor which would lead to a generation of model.
5. Once the model is generated we can predict the next upcoming temperatures with the help of predict function.

F. Optimizing the output graphs for the best fit of co-ordinates:

1. The errors in our model can be understood by the scikit-learn's built-in function. The error helps us to optimize the model further for better accuracy by minimizing the error.

VI. CONCLUSION

The code was analyzed and tested. Based on the test results, the whole system performed according to the designed aim and objectives of the project. The weather prediction model system was able to forecast weather for next 7 days to 1 month with accuracy and efficiency.

Output:
The Explained Variance: 0.94
The Mean Absolute Error: 1.07 degrees Celsius
The Median Absolute Error: 0.81 degrees Celsius

REFERENCES

[1]. Pushpa Mohan, Dr. Kiran Kumari Patil," Survey on Crop and Weather Forecasting based on Agriculture related Statistical Data", International Journal of Innovative Research in Computer and Communication Engineering, Bangalore, India, Vol. 5, Issue 2, February 2017, pp no. 2320-9801. [1]
[2]. Sneha S. Gumaste, Anilkumar J. Kadam, "Future weather prediction using genetic algorithm and FFT for smart farming", India.[2]
[3]. M. Manikandan , R. Mala," Optimal Prediction of Weather Condition Based on C4.5 Classification Technique", International Journal of Computer Sciences and Engineering, Vol.-6, Issue-10, Oct 2018, pp no. E-ISSN: 2347-2693 [3]
[4]. K.P. Mangani, R. Kousalya, "Big Data Approach for Weather Based Crop Insurance", IJSRNSC Volume-5, Issue-3, June 2017, pp no. E-ISSN: 2347-2693 [4]
[5]. Amit Palve, Ajit Patil, Amol Potgantwar, "Big Data Analysis Using Distributed Approach on Weather Forecasting Data", Volume-5, Issue-3, June 2017, India, pp no. ISSN: 2321-3256[5]
[6]. L. Shaikh, K. Sawlani,"A Rainfall Prediction Model Using
[7]. Articial Neural Network", IJSRNSC Volume-5, Issue-1,
[8]. April 2017, pp no. ISSN: 2321 3256.[5]

