EAI Endorsed Transactions
on Cloud Systems

Research Article EAI.EU

Predictive Analytics In Weather Forecasting Using Machine Learning Algorithms

Aastha Sharma[1,*], Vijayakumar V[1]

[1]SCSE, Vellore Institute of Technology, Chennai, India

Abstract

Agriculture is the backbone of every economy. In a country like India, which has ever increasing demand of food due to rising population, advances in agriculture sector are required to meet the needs. To add to it, the present economic conditions and government policies of India are such that it necessitates the adoption of Precision farming or smart farming. It will enable the farmers to maximize their crop yields and minimize the input costs as well as the losses due to reasons like uncertain rainfall, droughts etc. from this model. For Predicting weather forecasting we will use machine learning Algorithms like Linear Regression, Decision tree.

Keywords: Predictive Analytics, Weather Forecasting, Machine Learning Algorithms.

Received on 10 January 2019, accepted on 24 February 2019, published on 15 March 2019

doi: 10.4108/eai.7-12-2018.159405

*Corresponding author. Email:aastha.sharma2018@vitstudent.ac.in

1. Introduction

Machine Learning Technique is most robust technique for predicting weather forecasting. In past days we had to give instructions to System and then it gave result. but now we have machine learning algorithm so we can directly give inputs and feature and it generates result automatically. Just we need train the data then it generates model and features. [1]Most of the work related to machine learning for agriculture either solves the purpose of cultivating a crop and suggest weather data based on the statistical information.[2]Most of the work does not handle the planting of crops based on the climate.[3] plant diseases and insect pests causes significant reduction in quality as well as quantity of agricultural product so plant disease and insects pests forecasting is of great significance and quite necessary.

2. Machine Learning Algorithms

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y): **Y = f(X).**

2.1. Linear Regression

Linear regression is the most basic and frequently used predictive model for analysis. Regression estimates are generally used to describe the data and the elucidate relationship between one or more independent variables and dependent variables. Linear regression finds the best-fit through the points, graphically. The best-fit line through the points is known as the regression line.

OLS Model

Ordinary least square model is the most common estimate method that is used in linear model. It is used for getting best estimates. It minimizes the sum of square in the dependent variable. This helps us to find the relationship between dependent variable and independent variable. As it

calculated the distance between predicted value and actual value.

Advantage

- Simple mathematical representation.
- It doesn't take extra-large memory.
- It is very easy to clarify. Because it has numerical results.

Disadvantage

- It requires linearly spread data. If we have more features it doesn't provide accurate result.
- The linear regression model fails when we have non-linear data.

Algorithm Steps

- Import all libraries and read weather data.
- Define independent variable.
- Define dependent variable.
- Split and train data then test the data.
- Create linear regression model.
- Predict weather for future.

2.2. Decision Tree

It is a type of supervised learning algorithm that we mostly use for classification problem. it works for two dependent variable categorical and continuous dependent variables. In this type of algorithm, we split the population into two or more homogeneous sets. This is done because of most significant attributes/ independent variables to make as distinct groups as possible.

Data Shaping

It is a type of supervised learning algorithm that we mostly use for classification problem. it works for two dependent variable categorical and continuous dependent variables. In this type of algorithm, we split the population into two or more homogeneous sets. This is done because of most significant attributes/ independent variables to make as distinct groups as possible.

Label Selection

After shaping the data we select labels as features. We create labels for classify data. After labelling we move to splitting. We select first two columns from data for labelling.

Splitting

After labeling we splits the labels for finding best feature. Based on the best feature result we only get the accurate predicted result.

Advantage

- This algorithm handles both the continuous and categorical data.
- When we have non linear data decision tree will be useful. Because it splits the data set for creating more features.

Disadvantage

- If you have more features, your decision tree is probably going to be the deeper and bigger.
- It normally over fits a lot as it creates high-variance models.

Algorithm Steps

- Import all libraries and read weather data.
- Shape all data.
- Remove noisy data.
- Select labels.
- Classify and train the data.
- Predict the result.

.

2.3. Used Python Library

SKlearn

It is very useful library for machine learning modeling. It initially released on 2007. It includes lot of machine learning algorithms. In this library we use modules like DecisionTreeClassifier, train_test_split, accuracy_score.

Numpy

It is basically used in mathematical operations. It reads the data as numpy array for the manipulation purpose. It provides fast mathematical functions for calculation. For machine learning it is very common library.

Panda

This is the library make data analysis easier in python. This library also used to read and write the files. With data frame data manipulation can be easily done.

3. Proposed System Architecture

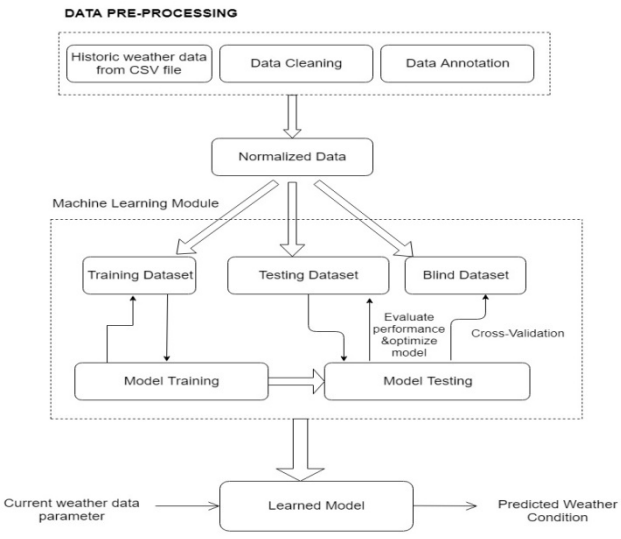

Figure 1. Architecture of Proposed Model

3.1. Methodology

Data Preprocessing

The more you preprocess the data set, the more accurate result you will get. basically, it is the process where we remove some unwanted or not useful, noisy data from the collected data. Also, if we don't remove any null value or empty field then we cannot get the proper results.

So, it is very important process to develop the model.

Normalization

It is also known as machine learning module. Here we train the collected dataset, test the dataset and then generate the new model, again for cross validation we blind the dataset.

Learn Model

This is the last process, In this phase we learn from model and predict the result. Learning model is important we have evaluated proper result. Here we get the artefact model from the training process.

4. Results

4.1. Scatter plots

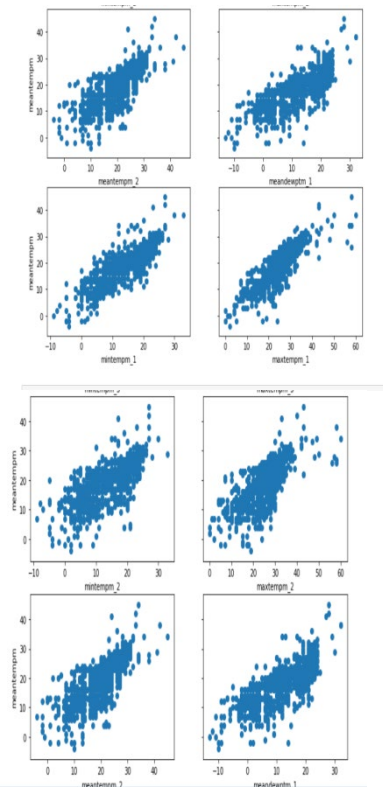

Figure 2. Scatter plots for linear regression model

4.2. OLS model Results

Out[12]:

OLS Regression Results

Dep. Variable:	meantempm	**R-squared:**	0.982
Model:	OLS	**Adj. R-squared:**	0.982
Method:	Least Squares	**F-statistic:**	5979.
Date:	Fri, 06 Jul 2018	**Prob (F-statistic):**	0.00
Time:	20:08:07	**Log-Likelihood:**	-2443.9
No. Observations:	987	**AIC:**	4906.
Df Residuals:	978	**BIC:**	4950.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
mindewptm_3	0.0821	0.022	3.665	0.000	0.038	0.126
maxdewptm_2	-0.1758	0.027	-6.533	0.000	-0.229	-0.123
mindewptm_2	-0.1459	0.029	-5.081	0.000	-0.202	-0.090
maxtempm_3	0.1630	0.021	7.908	0.000	0.123	0.203
meandewptm_1	-0.1103	0.052	-2.118	0.034	-0.212	-0.008
mindewptm_1	0.2859	0.044	6.537	0.000	0.200	0.372
mintempm_1	0.7105	0.135	5.260	0.000	0.445	0.976
maxtempm_1	0.8414	0.126	6.673	0.000	0.594	1.089
meantempm_1	-0.7120	0.254	-2.806	0.005	-1.210	-0.214

Figure 3. OLS model results

4.3. Calculated Errors

```
The Explained Variance: 0.85
The Mean Absolute Error: 2.10 degrees celcius
The Median Absolute Error: 1.30 degrees celcius
```

Figure 4. Errors for Estimates

4.1. Accuracy Measurement of Decision Tree

Out[38]:

	maxhumidity_3
0	84
1	92
2	92
3	80
4	80

Out[44]:

	maxhumidity_3
920	79
651	96
972	93
513	93
464	63

Figure 5. ***Select*** features **Figure 6.** Trained data

```
In [53]: accuracy_score(y_true = y_test, y_pred = predictions)
Out[53]: 0.9090909090909091
```

Figure 7. Accuracy measurement of decision tree

5. Future Scope

As for future scope we can't able to use linear regression when it comes to huge amount of data set and as its doesn't give accurate result. So, for predicting huge volume of dataset we can develop a neural network system for more better results and accurate prediction of the weather forecasting. Also we connect analysing process to IOT technology. Because without data we can not perform analysis and prediction because IOT is major source of data. So IOT will generate data from devices which helps to take initiative to improve decision making.

6. Conclusion

Machine learning algorithms plays a major role in predictive analytics, which uses the current and past historical data sets to discover knowledge from it and by using that data it the predict future occurrences. In this paper we have proposed two algorithm such as linear and decision tree for weather forecasting and prediction. we have concluded that linear regression is best when predicting weather forecast which have dependent dataset because already we have linear data for linear regression but for decision tree, we must give the label manually and the main Disadvantage of the decision tree is If you have more features, your decision tree is probably going to be the deeper and bigger and other one is that It normally over fits a lot as it creates high-variance models.

References

[1] Mark Holmstrom, Dylan Liu, Christopher Vo "Machine Learning Applied to Weather Forecasting" Stanford University(Dated: December 15, 2016).

[2] Priyanka P. Shinde, "Big Data Predictive Analysis:Using R AnalyticalTool" Assistant Professor, Department of MCA Government College of Engineering Karad, Karad, Maharashtra, India.

[3] *Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar,"*Predictive Analysis of Diabetic Patient DataUsing Machine Learning and Hadoop" Dept. of Computer Science and Engineering Rajarambapu Institute of Technology Sakhrale, Sangli Dist.

[4] Hina Gulati,"Predictive Analytics Using Data Mining Technique" Computer Science and Engineering Amity University, Noida, INDIA.

.