Analyze Weather Condition using Machine Learning Algorithms

Kavya.A[1], Divya.L[2], Kavya.N[3], Rajendra.M[4]

[1]Student, Dept. of computer science, Atria Institute of Technology, Karnataka, India
[2] Student, Dept. of computer science, Atria Institute of Technology, Karnataka, India
[3] Student, Dept. of computer science, Atria Institute of Technology, Karnataka, India
[4] Assistant Professor, Dept. of computer science, Atria Institute of Technology, Karnataka, India

---

Abstract – *Weather prediction is a very important in meteorology and is one of the most challengeable problems both scientifically and technologically all over the world from the last century. Historical weather data is used to train the model. The dataset consists of every feature over time, in a day. The dataset used in this project is part of a database contains following features: Temperature, Humidity, Pressure, Wind speed and so on. These are some of the dataset that is used to build our project. Weather prediction is done considering every feature over time, in a day. Classification is done by using empirical statistical technique by using Simple linear Regression, gradient booster and Random forest tree classifier .This method trains the data as per the data collected and predicts accurately and the various models are compared. The challenge here is the complexity and accuracy in the prediction.*

Key Words: Simple linear regression, Gradient booster, Random forest, Weather data analysis

1. INTRODUCTION

In present era weather forecasting and analysis has become a challenging problem around the world from the last century. The reason behind are the two main factors: Firstly, it is useful for many human activities like agriculture sector, tourism and natural disaster prevention. Secondly, due to various technological advances like the growth of computational power and ongoing improvements in measuring systems. All over the world, major challenges faced by meteorologist are the accuracy and prediction of weather. On the other hand, researchers had tried to predict different meteorological parameters by utilizing different data mining techniques[1]. While some of these techniques are more precise than others. Over the past few decades the availability of climate data has been increased. Such sources of climate data like observational records, understudy data, etc. makes it more important to find tools with higher accuracy rate to analyze different patterns from massive data. Therefore, meteorological data mining is a form of mining which is concerned with finding hidden patterns inside massive data available. So, the information extracted can be transformed into practical knowledge. This knowledge plays a vital role to understand the climate change and prediction. Having Knowledge of meteorological data is the key for variety of application to perform analysis and prediction of rainfall and it also does good job for prediction of temperature, humidity and irrigation system. In this research, we have used machine learning algorithms to predict the weather and gathered useful knowledge on historical weather data. 2016 weather data is used to train the model. There are 12 attributes in weather data class namely, summary, precipitation type, temperature, Apparent Temperature, Humidity, Wind Speed, Wind Bearing, Visibility, Pressure, Data type. We start with Data preprocessing where we handle the null values in the data and handle the outliers (we need to manage the data which are not within the range). The next step is Explanatory data analysis (Cleaning the data) where we perform visualization step and correlation step between each attribute and output (always varies between +1 and -1) and we plot the graphs for all the attributes in order to visualize then we get the important features. The last and final step is prediction and this prediction is done using the machine learning algorithm. For this climate variation prediction we use python as a programming language. The goals for data analysis are those which involve weather variations that affect our daily runtime changes in min and max temperature, humidity level, rainfall chances and speed of wind. This knowledge can be utilized to support many important areas which are affected by climate change includes Agriculture, Water Resources, Vegetation and Tourism. Studies shows that human society is affected in different ways by weather affects. For example, water resources are the main sources of irrigation in production of agriculture crops and the amount of rain is one of them that affect the crops abruptly due to climate change. It is also directly related to the different human activities. Moreover, poor growth and low quality is due to negative effects of weather resulting in failure of high production. Hence, changes in weather conditions are risky.

2. PROPOSED MODEL

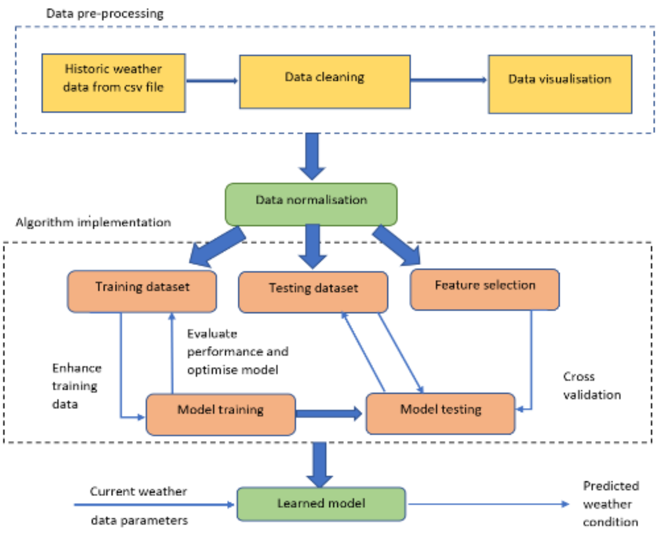

Fig 1: Block diagram

The proposed approach consists of two parts: data pre-processing and algorithm implementation. Figure 1 presents the block diagram of the proposed approach.

2.1 DATA PREPROCESSING

Raw data may contain noisy, incomplete and inconsistent values which may lead to error while implementing. In order to avoid these errors we first pre-process the data. Data pre-processing is an important step to transform the raw data into understandable format and is divided into 4 main tasks: historic weather data from CSV file, data cleaning, data visualization and data normalization.

2.1.1 DATASET

The first step of data pre-processing is to collect the raw data, that is, historic weather data in a csv file format. This data is loaded into the system for pre-processing.
We have used the daily historical weather data of 1 year (2016) for analysis. Following procedure includes data cleaning, data visualization, data normalization, feature selection, model training, model testing.

2.1.2 DATA CLEANING

This is the second step of the implementation where we clean the data by removing the outliers and the null values. Outliers are the extreme values which fall outside the observation. If we don't perform this step, it may lead to the inaccurate training and it may also take a longer training durations.

2.1.3 DATA VISUALIZATION

In this step the data will be visualized through graphs such as histograms, boxplot etc. A **histogram** is a graph that represents the probability distribution of a dataset. A **histogram** can be seen as a bar which are vertical and every vertical bar represents a value. The heights of the bars indicate the frequencies or probabilities for the different values or range of values. A **boxplot** is a way for visualizing the data distribution based on a five number summary ("maximum", first quartile (Q1), median, third quartile (Q3), and "minimum").It will give the information about the values of the outliers.

2.1.4 DATA NORMALIZATION

It is transformed to the required format. The proposed approach divides the dataset into 2 sets namely categorical data and numerical data. The numerical data is one which has numeric data type and then the data is cleaned. The categorical data is one which has string data type it is converted into numerical data type using mapping.

2.2 ALGORITHM IMPLEMENTATION

Once the data is pre-processed it is used for implementation. Following are the steps that are carried to implement an algorithm:

2.2.1 TRAINING AND TESTING DATASET

In order to train the model we must divide the dataset into testing and training dataset. This division may be in the proportion 70:30 or 80:20, where the proportion completely depends on the accuracy we need. The normalized data is used for dividing the dataset. The testing data is not used for training the model.

2.5 FEATURE SELECTION

In this step the necessary features are selected for the implementation. The features that help in predicting the correct values are selected by applying the variance threshold.

2.6 MODEL TRAINING

In this step the model is trained by the simple linear regression algorithm[2], gradient booster[3] and random forest classifier[4] algorithm. The model is trained only using the training data.

2.7 MODEL TESTING

In this step the model is tested for the accuracy. The accuracy of the model is tested using the error rate while testing. Low error rate indicates high accuracy.

3. CLASS DIAGRAM

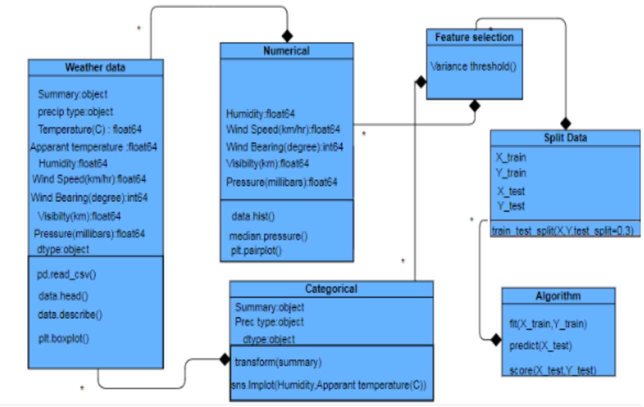

Fig 2. Class diagram

There are 12 attributes in weather data class namely, summary, precipitation type, temperature, Apparent Temperature, Humidity, Wind Speed, Wind Bearing, Visibility ,Pressure, Data type .The method which is used to read the weather data is read_csv(). The method head()which is mentioned in the Weather data class is used to check whether the data loaded successfully or not. And the method describe () is used to view some basic statistical details like percentile, mean, standard deviation etc. of a series of numeric values .When this method is applied to a series of string, it returns a different output such as percentile, include, exclude, return type. The method boxplot() is used to know the outliers. We divide the Weather data class into two categories namely numerical and categorical data. Numerical class contains the numerical data and categorical class contains the categorical data. The Numerical data features are Temperature, Apparent Temperature, Humidity, Wind speed, wind bearing, visibility and pressure. Categorical class contains the categorical features summary, precipitation type, data type. This class features are mapped to numerical features. We perform the feature selection by implementing the method called variance_threshold( ) and then we split the dataset as X_train, Y_train , X_test ,Y_test in the proportion 70:30 finally we train the model by applying the algorithm and we predict the score of each algorithms.

4. RESULTS AND CONCLUSION

After executing the proposed approach to the historic weather data of 2016, the score and accuracy is as shown in the table 1. From the table we can conclude that gradient booster algorithm has a better accuracy and low error rate, next is the random forest followed by the linear regression.

Algorithm	**Score**	**Error**
Linear regression	99.106595	0.0000370
Gradient booster	**99.956499**	**0.000018**
Random forest	99.818849	0.000075

Table 1: comparison of accuracy and error

REFERENCES

[1] Aishwarya Dhore, Anagha Byakude, Bhagyashri Sonar, Mansi Waste," Weather prediction using the data mining Techniques" IRJET Volume: 04 Issue: 05 | May -2017

[2] Saishruthi Swaminathan-" Linear Regression-Detailed View", February 26, 2018 [Online]: https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

[3] Richard S. Zemel Toniann Pitassi- "A Gradient-Based Boosting Algorithm for Regression Problems" [Online]: http://papers.nips.cc/paper/1797-a-gradient-based-boosting-algorithm-for-regression-problems.pdf

[4] Niklas Donges, "The Random Forest Algorithm" [Online]: http://papers.nips.cc/paper/1797-a-gradient-based-boosting-algorithm-for-regression-problems.pdf