WEATHER PREDICTION USING MACHINE LEARNING TECHNIQUES

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Sahil Ratra (151226)

Under the supervision of

Dr. Amit Kumar

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CERTIFICATE

Candidate's Declaration

I hereby declare that the work presented in this report titled **"Weather Prediction using Machine Learning Techniques"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of **Dr. Amit Kumar** (Assistant Professor, Department of Computer Science and Engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Sahil Ratra , 151226

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Amit Kumar

Assistant Professor

Department of Computer Science and Engineering and Information Technology

Dated:

ACKNOWLEDGEMENT

I wish to express our significant and true appreciation to Dr. Amit Kumar, Assistant Professor, Department of Computer Science and Information Technology, Jaypee University of Information Technology, who guided me into the complexities of this project non-chalantly with incomparable generosity. He always co-worked and assisted with the examination work. He likewise revealed unmistakable fascination and significant help in the field of Machine Learning for advancement of our task work.

Table of Contents

Abstract

Traditionally, climate estimation has dependably been performed by considering the environment as a liquid. The current condition of the air is inspected. The future condition of the environment is registered by comprehending numerical conditions of thermodynamics and liquid elements. Yet, this conventional arrangement of differential conditions that oversee the physical model is some of the time shaky under unsettling influences and uncertainties while estimating the underlying states of the air. This prompts an inadequate comprehension of the environmental forms, so it limits climate forecast up to 10 day period, on the grounds that past that climate estimates are essentially unreliable.But Machine learning is moderately hearty to most barometric unsettling influences when contrasted with customary techniques. Another favorable position of machine learning is that it isn't reliant on the physical laws of environmental procedures.

In this report, a reenacted framework is created to foresee different climate conditions utilizing Data Analysis and Machine learning procedures, for example, straight relapse and strategic relapse. The primary wellspring of information to be utilized for directed taking in is to be gathered. The current climate condition parameters ex. temperature and so on are utilized to fit a model and further utilizing machine learning methods and extrapolating the data, the future varieties in the parameters are broke down.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Weather prediction is the task of prediction of the atmosphere at a future time and a given area. In early days, this has been done through physical equations in which the atmosphere is consider as fluid. The current state of the environment is inspected, and the future state is predicted by solving those equations numerically, but we can not determine a very accurate weather for more than 10 days and this can be improved with the help pf scienceand technology

There are numerous kinds of machine learning calculations, which are Linear Regression, Polynomial Regression, Random Forest Regression, Artificial Neural Network and Recurrent Neural Network. These models are prepared dependent on the authentic information gave of any area. Contribution to these models are given, for example, in the event that anticipating temperature, least temperature, mean air weight, greatest temperature, mean dampness, and order for 2 days. In light of this Minimum Temperature and Maximum Temperature of 7 days will be accomplished

Machine Learning

Machine learning, is relatively robust to perturbations and does'nt need any other physical variables for prediction. Therefore, machine learning is much better opportunity in evolution of weather forecasting. Before the advancement of Technology, weather forecasting was a hard nut to crack. Weather forecasters relied upon satellites, data model's atmospheric conditions with less accuracy. Weather prediction and analysis has vastly increased in terms of accuracy and predictability with the use of Internet of Things, since last 40 years. With the advancement of Data Science, Artificial Intelligence, Scientists now do weather forecasting with high accuracy and predictability.

USE OF ALGORITHMS:

There are different methods of foreseeing climate utilizing Regression and variety of Functional Regression, in which datasets are utilized to play out the counts and investigation. To Train the calculations ¾ size of information is utilized and ¼ size of information is named as Test set. For Example, in the event that we need to anticipate climate of Austin Texas utilizing these Machine Learning calculations, we will utilize 6 Years of information to prepare the calculations and 2 years of information as a Test dataset.

On the as opposed to Weather Forecasting utilizing Machine Learning Algorithms which depends essentially on reenactment dependent on Physics and Differential Equations, Artificial Intelligence is additionally utilized for foreseeing climate: which incorporates models, for example, Neural Networks and Probabilistic model Bayesian Network, Vector Machines. Among these models Neural Network is widely utilized as it is efficient to catch more conditions of past weather report and future weather conditions.

In any case, certain machine learning calculations and Artificial Intelligence Models are computationally costly, for example, utilizing Bayesian Network and machine learning calculation in parallel.

To finish up, Machine Learning and Artificial Intelligence has enormously change the worldview of Weather estimating with high precision and predictivity. What's more, inside the following couple of years greater progression will be made utilizing these advances to precisely foresee the climate to avoid catastrophes like typhoon, Tornados, and Thunderstorms.

Machine learning has the following main algorithms:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised learning: This is a set of predictors. These predictors are independent variables. The objective of this learning algorithm is to predict from this set of independent variables. The prediction is for the outcome variable. This is a dependent variable. With the set of independent variables, a function is generated that facilitates the allocation of our inputs to the desired outputs. To achieve a certain precision in our training data, the machine is continuously trained. Examples of supervised learning are linear regression, logistic regression, KNN decision tree, random forest, etc.

Uncontrolled learning: in this algorithm, there is no particular goal or result that can be estimated or predicted. It is used to group into different groups, which is used for segmentation into different groups for specific interventions. Some examples of unsupervised learning are K-Means, Apriori's algorithm.

Reinforcement learning: certain decisions have been made when training the machine with this algorithm. It works so that the machine is exposed to conditions in any environment. The machine is continuously trained with the trial and error method. To make accurate business decisions, the machine learns from past experience by capturing the best possible knowledge. Some examples of learning by reinforcement are the Markov decision process.

1.2 Problem Statement

Heavy rainfall can lead to numerous hazards, for instance:

flooding, including danger to human life, harm to structures and framework, and loss of products and domesticated animals. avalanches, which can compromise human life, upset transport and interchanges, and cause harm to structures and foundation. Where overwhelming precipitation happens with high breezes, hazard to ranger service crops is high..

In the case of initial treatment of patients, the probability of survival has increased significantly with early diagnosis of breast cancer. With proper tumor classification, unnecessary treatment can be avoided. Each volume should be treated differently. Therefore, if there is no proper diagnosis then there is a high risk of death for the patient. Correct diagnosis of breast cancer and classification of tumors in benign and malignant tumors is an area of investigation.

For example if we consider an area affected by tropical cyclone the fundamental impacts of tropical cyclone incorporate heavy rain, strong wind, huge tempest floods close landfall, and tornadoes. The devastation from a tropical cyclone, for example, a sea tempest or hurricane, depends for the most part on its power, its size, and its area. Tropical tornados act to evacuate woods shade and additionally change the scene close beach front zones, by moving and reshaping sand ridges and causing broad disintegration along the drift. Indeed, even well inland, overwhelming precipitation can prompt mudslides and avalanches in rugged regions. Their belongings can be detected after some time by concentrate the convergence of the Oxygen-18 isotope inside caverns inside the region of typhoons' ways. So we are providing a better way to get accurate predictions.

As mentioned above, the benefits of identifying important features of mechanical learning, complex data sets, play an important role in forecasting of weather. Since the best results can be achieved with engineering learning algorithms,we should use these techniques to aware people from natural disasters. This is because learning engineering algorithms can provide more accurate results. Apart from this, the results are achieved at a short time and people get enough time to do preparations or to escape from that place .

1.3 Objectives

Our project aims to predict the Weather and Atmosphere conditions using the previous dataset of the weather forecasting with a focus on improving the accuracy of prediction. This will increase the accuracy of the weather prediction and we will get accurate results than the traditional methods. Our dataset consists of max and min. temperature of everyday from the specific location.

Classifications:

When gathering datasets to give to the models there are sure parameters which are called as ordered information which incorporates: snow, rainstorm, rain, mist, cloudy, for the most part overcast, halfway shady, scattered mists, and clear. These can be additionally ordered into four classes.

1. Rain, tempest, and snow into precipitation
2. For the most part shady, foggy, and cloudy into exceptionally shady
3. Scattered mists and somewhat shady into modestly shady
4. Clear as clear

Thus our aim is to provide accurate result in order to provide correct prediction of weather for future so in critical conditions people can be aware of upcoming natural calamities.

1.4 Methodology

The dataset utilized in this arrangement will be gathered from Weather Underground's complementary plan API web benefit. I will utilize the solicitations library to collaborate with the API to pull in climate information since 2015 for the city of Lincoln, Nebraska. When gathered, the information should be process and collected into an organization that is appropriate for information examination, and afterward cleaned.

Then we will concentrate on examining the patterns in the information with the objective of choosing fitting highlights for building a Linear Regression, Polynomial Regression. We will examine the significance of understanding the suppositions vital for utilizing a Linear and Polynomial Regression show and exhibit how to assess the highlights to fabricate a hearty model. This will finish up with a discourse of Linear and Polynomial Regression show testing and approval.

Atlast we will concentrate on utilizing Neural Networks. I will look at the way toward building a Neural Network show, deciphering the outcomes and, by and large precision between the Linear and Polynomial Regression demonstrate worked earlier and the Neural Network display.

We have a problem statement which comes under the category of Classification. It is a multiclass classification in which the classes given to us are

1. Rain, tempest, and snow into precipitation
2. For the most part shady, foggy, and cloudy into exceptionally shady
3. Scattered mists and somewhat shady into modestly shady
4. Clear as clear

Our aim is to classify the given data into the above given classes. In order to do so, we have to first analyze the data given to us. For analyzing the features, we are using different techniques.

The training of model can be done in many ways. It depends on how the data is prepared for further processing. The data can be used directly depending on the situation or the data can be used to form a histogram. After these modifications, we choose a particular model on which we will train our data. This model can be: Linear regression, Logistic Regression, SVM, Neural Networks, Decision Tress, K-Nearest Neighbors etc. Parameter tuning can also be done in order to increase our accuracy.

Once the model is trained, we can test our data by applying our algorithms on the Test Data. With the help of this we can find the learning ability of our algorithm

CHAPTER 2

LITERATURE SURVEY

2.1 Collection of Dataset FromWeatherUnderground:

The information utilized in this arrangement will be gathered from Weather Underground's complementary plan API web benefit. I will utilize the solicitations library to connect with the API to pull in climate information since 2015 for the city of New Delhi, India. When gathered, the information should be process and totaled into a configuration that is appropriate for information investigation, and afterward cleaned.

Weather Underground gives various web benefit API's to get to information from however, the one we will be worried about is their history API. The history API gives an outline of different climate estimations for a city and state on an explicit day.

Making Requests to the API :

To make solicitations to the Weather Underground history API and process the returned information we will make utilization of a couple of standard libraries and some well known outsider libraries.

After installing libraries, we will define data into data variables and rearrange them from the first day of the year to the targeted date.

Now we are heading toward data retrieval and setting up our pandas DataFrame.Since we have a sizable records rundown of DailySummarywe will utilize it to work out a Pandas DataFrame. The Pandas DataFrame is an extremely helpful information structure for some, programming errands which are most prominently known for cleaning and handling information to be utilized in machine learning undertakings .

After that We will derive the features and classify the data into classes as we have stated above.

The most important thing is Data cleaning. The notorious saying, "refuse in, waste out", is as fitting as ever with regards to machine learning. Be that as it may, the

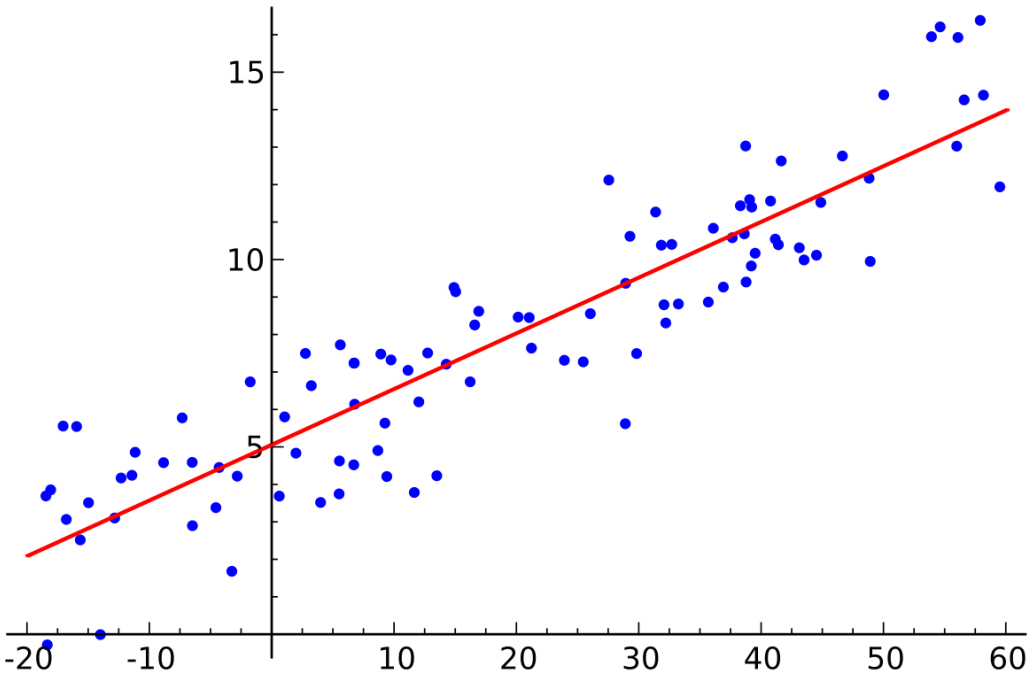
information cleaning some portion of an examination venture isn't only a standout amongst the most vital parts it is likewise the most tedious and relentless.

2.1.1 Linear Regression Algorithm:

Before recognizing what is Linear Regression, let us get ourselves acclimated with regression. Regression is a technique for demonstrating an objective esteem dependent on free indicators. This strategy is for the most part utilized for estimating and discovering circumstances and end results connection between factors. Regression methods for the most part vary dependent on the quantity of autonomous factors and the kind of connection between the free and ward factors.

Linear Regression

Basic linear regression is a kind of regression examination where the quantity of autonomous factors is one and there is a straight connection between the independent(x) and dependent(y) variable. The red line in the above diagram is alluded to as the best fit straight line. In view of the given information focuses, we attempt to plot a line that models the focuses the best. The line can be displayed dependent on the straight condition demonstrated as follows.



y = a_0 + a_1 * x ## Linear Equation

The intention of the linear regression calculation is to locate the best qualities for a_0 and a_1. Before proceeding onward to the calculation, we should view two critical ideas you should know to more readily comprehend linear regression.

2.1.2 Polynomial Regression Algorithm:

Polynomial regression models are typically fit utilizing the strategy for least squares. The least squares technique limits the difference of the fair estimators of the coefficients, under the states of the Gauss– Markov hypothesis. The least squares technique was distributed in 1805 by Legendre and in 1809 by Gauss. The primary plan of an examination for polynomial regressiom showed up in a 1815 paper of Gergonne. In the twentieth century, polynomial regression assumed a vital job in the improvement of regression investigation, with a more noteworthy accentuation on issues of structure and induction. All the more as of late, the utilization of polynomial models has been supplemented by different techniques, with non-polynomial models having preferences for a few classes of issues

A polynomial term is the equation that can convert a straight linear model into curve.

This makes it a pleasant and direct approach to display curves without demonstrating confused nonlinear models.

One normal example inside machine learning is to utilize linear models prepared on nonlinear elements of the information. This methodology keeps up the by and large quick execution of direct techniques while enabling them to fit an a lot more extensive scope of information.

For instance, a straightforward direct relapse can be reached out by developing polynomial highlights from the coefficients. In the standard straight relapse case, you may have a model that resembles this for two-dimensional information:

On the off chance that we need to fit a paraboloid to the information rather than a plane, we can join the highlights in second-arrange polynomials, with the goal that the model resembles this:

The perception is this is as yet a linear model: to see this, envision making another variable

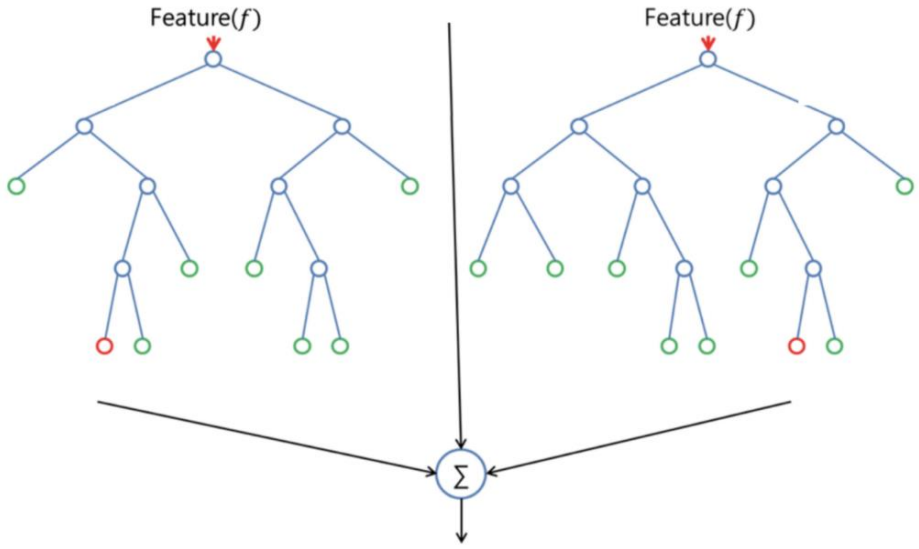
With this re-marking of the information, our concern can be composed. We see that the subsequent polynomial regression is in a similar class of linear models we'd considered above and can be unraveled by similar methods.

By considering, direct fits inside a higher-dimensional space worked with these premise capacities, the model has the adaptability to fit an a lot more extensive scope of information.

2.1.3 Random Forest Regression:

Random Forest is user friendly machine learning algorithm that produces, an extraordinary outcome not even without hyper-parameter tuning. It is much better and helpful from other algorithm, since it's simplemethod andit can be use for both classification and regression tasks.

One uniqueness of random forest is, that we can utilize it for both issues: characterization and regression. Underneath you can perceive how arandom forest would look like with two trees:



2.1.4 Artificial Neural Network:

An artificial neural network (ANN) is a computerized framework which depends on the structure and elements of natural neural systems. The data that we get from the system shows an effect on neural network with the different data we enter neural system changes or we can say it starts learning from the previous values and we get accurate predictions after a clean data processing.

ANNs are the nonlinear model which process all previous data and explore a structure or design layout to give proper and accurate results.

ANN is also known as a neural network.

An ANN has very few bugs but it is a better alternative to approach for more accurate predictions which can predict just by watching the information indexes. Thusly, ANN is usedforirregular predictions. These kinds of apparatuses help gauge the most practical and perfect strategies for touching base at arrangements while characterizing processing capacities or disseminations. ANN takes information tests instead ofinformationindexes, which saves both time and cash. ANNs are viewed as

genuinely basic scientific models to upgrade existing information investigation advances.

ANNs is a three layer structure which is interconnected. The main layer consists info neurons. These neurons transfer data to the second layer, which transer yield neurons to other layer .

Preparing a fake neural system includes looking over permitted models for which there are a few related calculations.

CHAPTER 3

SYSTEM DEVELOPMENT

3.1 Analysis

The learning procedure starts with the perception of information, so examples can be discovered in information and prevalent choices can be taken later on which depend on the precedents gave. The principle point is to enable PCs to learn without human help or collaboration and modify their activities as needs be.

The amount and size of malignant growth databases are expanding quickly, yet most are not dissected to discover covered up and profitable learning. Machine learning procedures can be utilized to find shrouded connections and examples. Models created utilizing machine learning systems enable specialists to settle on exact choices.

Accordingly, we utilize programmed learning strategies, for example, random forests, linear regression, polynomial regresstion, and so on to prepare our machine. The gadget adjusts to the predefined information record and gains from the predetermined parameters. From that point forward, machine learning strategies have turned out to be precise in a few fields before. Along these lines, the utilization of machine learning is helpful for the conclusion of malignant growth. Collapsing neural systems work surprisingly better than linear regression, polynomial regression and Random Forests. This is on the grounds that, at every one of the dimensions, the weights proceed to return and attempt to diminish the mistake.

The most critical piece of our task is the examination of data when programmed learning methods are utilized. To dissect the pictures, we utilize a few descriptors, for example, nearby double examples, ORB, edge nearness measurements without parameters (PFTAS), GLCM. These element extractors help remove the usefulness of each picture. Subsequent to seeing these element vectors, we can at long last train our machine in like manner. At last, this will enable us to get an exact determination of the forecast.

The most extreme exactness was accomplished when Parameter Free Threshold Adjacency Statistics was utilized as the component extractor and SVM was utilized as the Machine Learning Algorithm. The best outcome was accomplished when parameters were tuned in like manner. With the end goal to tune parameters and gets quick outcomes, Grid Search technique was utilized. In network seek strategy, a scope of parameters is given to the classifier and the calculation at long last takes up the best blend of all the given parameters. The best arrangement of parameters is taken with the end goal that it gives the most extreme exactness.

3.2 System Design

The record has just been separated into train set and test set. Each information has just been labeled. First we take the trainset organizer.

We will train our model with the help of histograms. The feature so extracted is stored in a histogram. This process is done for every data in the train set. Now we will build the model of our classifiers. The classifiers which we will take into account are Linear Regression, Polynomial Regression, Random Forest and Neural Networks. With the help of our histogram, we will train our model. The most important thing to in this process is to tune thee parameters the accordingly, such that we get the most accurate results.

Once the training is complete, we will take the test set. Now for each data variable of test set, we will extract the features using feature extraction techniques and then compare its values with the values present in the histogram formed by train set. The output is then predicted for each test day. Now in order to calculate accuracy, we will compare the predicted value with the labeled value. The different metrics that we will use are confusion matrix, accuracy score, f1 score etc.

3.3 Model Development

Our strategy for model improvement is exploratory. The objective of our undertaking is to ensure the conclusion of malignancy with greatest accuracy. This must be accomplished by exploring different avenues regarding distinctive systems from a specific field. We have considered the programmed learning descriptors and algorithms.

The Machine Learning Algorithms that we are using are:

- Linear Regression
- Polynomial Regression
- Random Forest
- Neural Networks

Subsequently our point is to locate the best mix which will furnish us with greatest precision. Along these lines this task is absolutely test based. In addition parameter tuning is a noteworthy piece of any Machine Learning Algorithm. Regardless of whether the calculation works exceptionally solid in specific conditions, at that point too because of terrible determination of parameters, the precision could be low. In this manner we likewise need to center around the right arrangement of parameters. Hence parameter tuning must be done in whichever show we pick.

Parameter tuning should either be possible physically or by utilizing the lattice seek technique. Network looking is the procedure in which information is checked with the end goal to discover ideal parameters for some random model. Contingent upon the kind of model that we are utilizing, tuning of specific parameters is vital. Framework seeking applies to a solitary model sort as well as number of models. Network looking can be connected in machine learning with the end goal to ascertain the best parameters for its utilization in some random model. It very well may be computationally greatly costly and may set aside a long opportunity to keep running on the machine. Matrix Search constructs a model on every conceivable parameter mix. At that point it repeats through every parameter blend lastly stores a model for each mix.

CHAPTER 4

ALGORITHMS

4.1 Simple Linear Regression

Simple linear regression is a factual strategy that enables us to abridge and consider connections between two ceaseless (quantitative) factors: One variable, signified x, is viewed as the indicator, logical, or free factor. The other variable, indicated y, is viewed as the reaction, result, or ward variable.

Since alternate terms are utilized less much of the time today, we'll use the "predicator" and "reaction" terms to direct to the factors experienced in this course. Alternate terms are referenced just to give you knowledge of them that you should experience them in different fields. Straightforward simple linear regression gets its descriptor "basic," since it concerns the investigation of just a single indicator variable. Conversely, different straight relapse, which we examine later in this course, gets its descriptor "numerous," on the grounds that it concerns the investigation of at least two predicator factors.

The previous data is given in Table 1 .We can observe a positive relationship between X and Y. There can be a relation observed between X and Y, Higher the estimation of X more accurate will be the prediction of Y.

Table 1. Example data.

X	**Y**
5.00	2.25
4.00	3.75
3.00	1.30
2.00	2.00
1.00	1.00

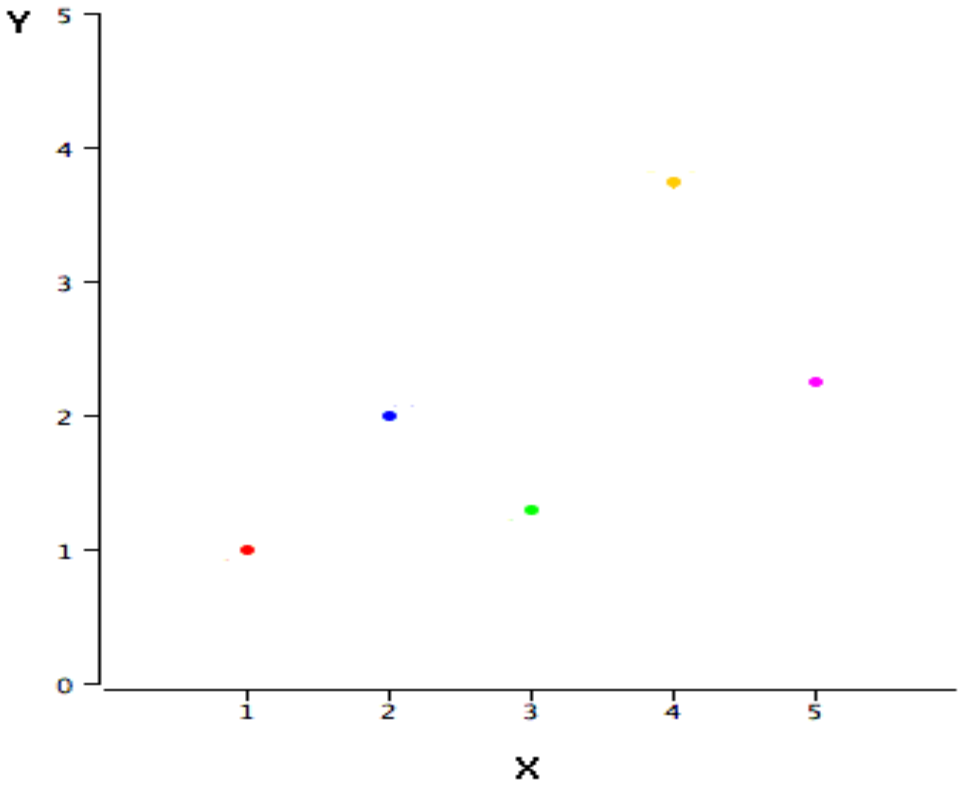


Figure 1.Scatter Graph of information.

Linear regression consist of exploring the best straight line through the points also known as regression line . The dark line in Figure 2 is the regression line and consists of the expected score on Y for every estimation of X. The vertical lines from the focuses to the regression line speak to the blunders of expectation. As we can see in the graph, the red point is very very close to the regression line, its acuuracy is better. Conversely, the yellow point is very far from the line.than the regression line and subsequently its mistake of forecast is vast.

.

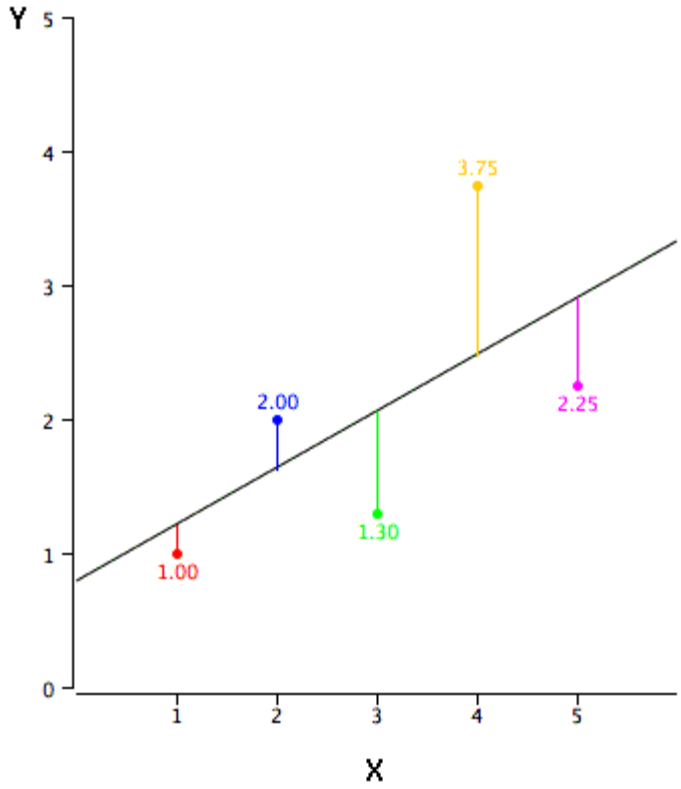


Figure 2. A scatter graph of the data with regression line.

The regression line is used for preditions and the values that has been given in graph are the actual data set given to the model.We use colored vertical lines and regression line for thecomparisionoferror

The prediction error for a point is the subtraction of the value of the point from the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the prediction errors(Y-Y'). For example, the second point has a Y of 2.00 and a predicted Y (called Y') of 1.635. Therefore, its prediction error is 0.36.

Table 2. Example data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
5.00	2.25	2.910	-0.660	0.436
4.00	3.75	2.485	1.265	1.600
3.00	1.30	2.060	-0.760	0.578
2.00	2.00	1.635	0.365	0.133
1.00	1.00	1.210	-0.210	0.044

You may have seen that we didn't indicate what is implied by "best-fitting line." By far, the most regularly utilized criterion for the best-fitting line is the line that limits the whole of the squared mistakes of expectation. That is the criterion that was utilized to discover the line in Figure 2. The last column in Table 2 depicts the squared prediction errors. The sum of the squared prediction errors shown in Table 2 is lesser than it would be for any other regression line.

The formula for a regression line is

$$Y'=bX+A$$

where Y' is the score prediction, the slope of line is b while A is the Y intercept. The equation is

$$Y' = 0.425X + 0.785$$

If X = 1,

Y' =1.21.

If X = 2,

Y' = 1.64.

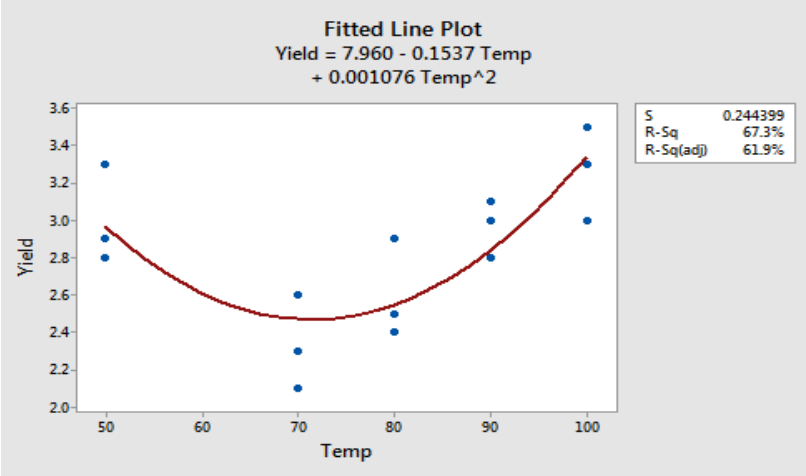
4.2 Polynomial Linear Regression

It is a type of linear regression in which the relationship between the input factors x and the yield variable y is displayed as a polynomial. Albeit polynomial regression fits a nonlinear model to the information, as a measurable estimation issue it is linear, in the feeling that the regression work is linear in the obscure parameters that are evaluated from the information. Thus, polynomial regression is viewed as a unique instance of linear regression.Once in a while, a plot of the residuals versus a predictor may recommend there is a nonlinear relationship. One approach to attempt to represent such a relationship is through a polynomial regression demonstrate. A model for such a single predictor, *X*, is:

$$Y=\beta 0+\beta 1X+\beta 2X2+\ldots+\beta hXh+\epsilon,$$

h is **degree** of the polynomial. For lower degrees, the relationship has an explicit name (i.e., h = 2 is called quadratic, h = 3 is called cubic, h = 4 is called quartic, etc). In spite of the fact that this model takes into consideration a nonlinear relationship among Y and X, polynomial regression is as yet viewed as linear regression since it is linear in the regression coefficients, β1,β2,...,βh.

Figure 3. A Sample Polynomial regression plot with best fitted line



With the end goal to evaluate the condition above, we would just need the reaction variable (Y) and the predictor variable (X). In any case, polynomial regression models may have other predictor factors in them too, which could prompt connection terms. So as should be obvious, the essential condition for a polynomial regression display above is a generally simple model, however you can envision how the model can develop contingent upon your circumstance!

Generally, we actualize indistinguishable investigation techniques from done in numerous linear regressions.

4.3 Random Forest Regression

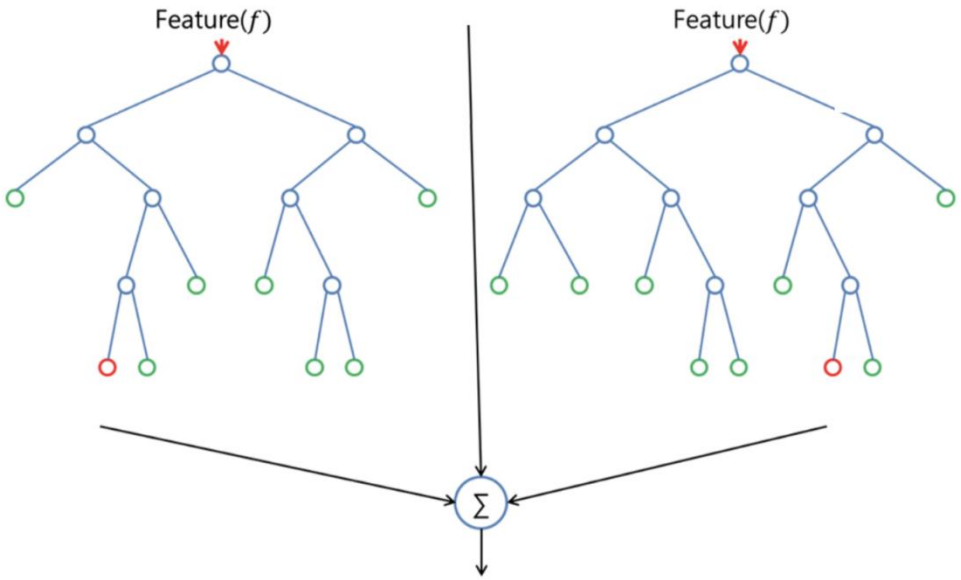
Random Forest is an adaptable, simple to utilize machine learning algorithm which produces, even without hyper-parameter tuning, a consistent outcome more often than not. It is likewise a standout amongst the most utilized algorithms, as it is straightforward and can very well be utilized for both classification and regression tasks.

Random Forest is a type of supervised learning algorithm. As is evident from its name, it makes a forest and makes it by a random way of selection. The “forest" it assembles, is an outfit of Decision Trees, prepared with the "bagging" strategy more often than not. The general thought of the bagging technique is that a mix of learning models expands the general outcome.

To state it in simple words: It forms various decision trees and clubs them all together to get a more accurate prediction.

Random Forest (RF) is an extremely adaptable and simple to utilize Machine Learning (ML) calculation. It creates exceptionally precise outcomes even without the high degree of hyper-parameter tuning. Irregular Forest (RF) is additionally a standout amongst the most utilized Machine Learning (ML) calculations. This is on the grounds that it is extremely basic and can likewise be utilized for both characterization and relapse tests.

.



Random Forest as two tree .

Fundamentally there are two phases in Random Forest (RF) calculation. First is irregular timberland creation. Second is to play out an expectation from the recently made irregular woods classifier. The entire procedure can be given as:

:

i) Select randomly “K” features from the total “m” features. Here k << m.

ii) Now Among these “K” features, using the best split point calculate the node “d” .

iii) Then split the node into its daughter nodes by using the best split.

iv) Repeat all the steps from a to c until “l” number of nodes has been finally reached.

Now build the forest by repeating steps a to d for “n” number times in order to create “n” number of trees.

4.4 Logistic Regression

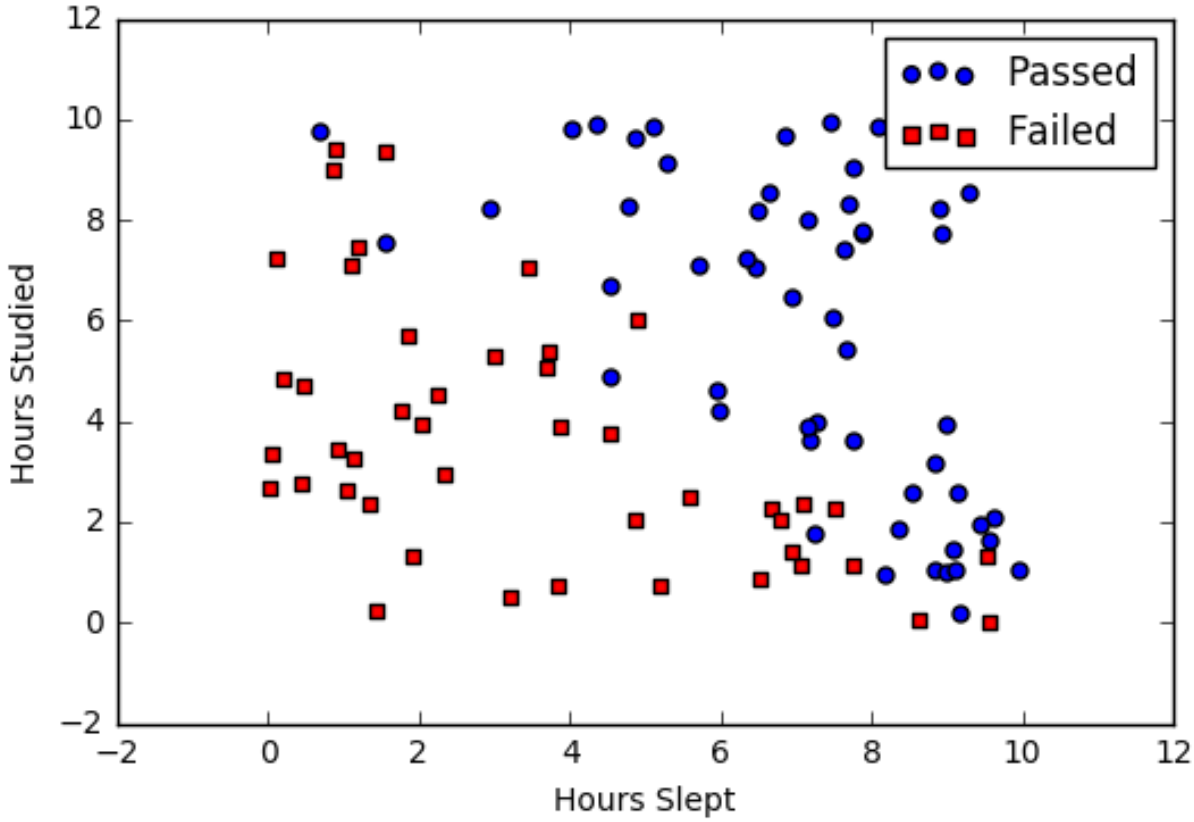
Logistic regression is a classification algorithm used to allocate perceptions to a discrete arrangement of classes. Dissimilar to linear regression which yields persistent number qualities, logistic regression changes its yield utilizing the logistic sigmoid capacity to restore a likelihood esteem which would then be able to be mapped to at least two discrete classes.

Let's assume we're given information on understudy test results and our objective is to foresee whether an understudy will pass or fall flat dependent on number of hours dozed and hours spent considering. We have two highlights (hours rested, hours contemplated) and two classes: passed (1) and fizzled (0).

Studied	Slept	Passed
4.84	9.62	1
8.61	3.22	0
5.42	8.22	1
9.21	6.33	0

Table : Number of ours slept versus number of hours studied.

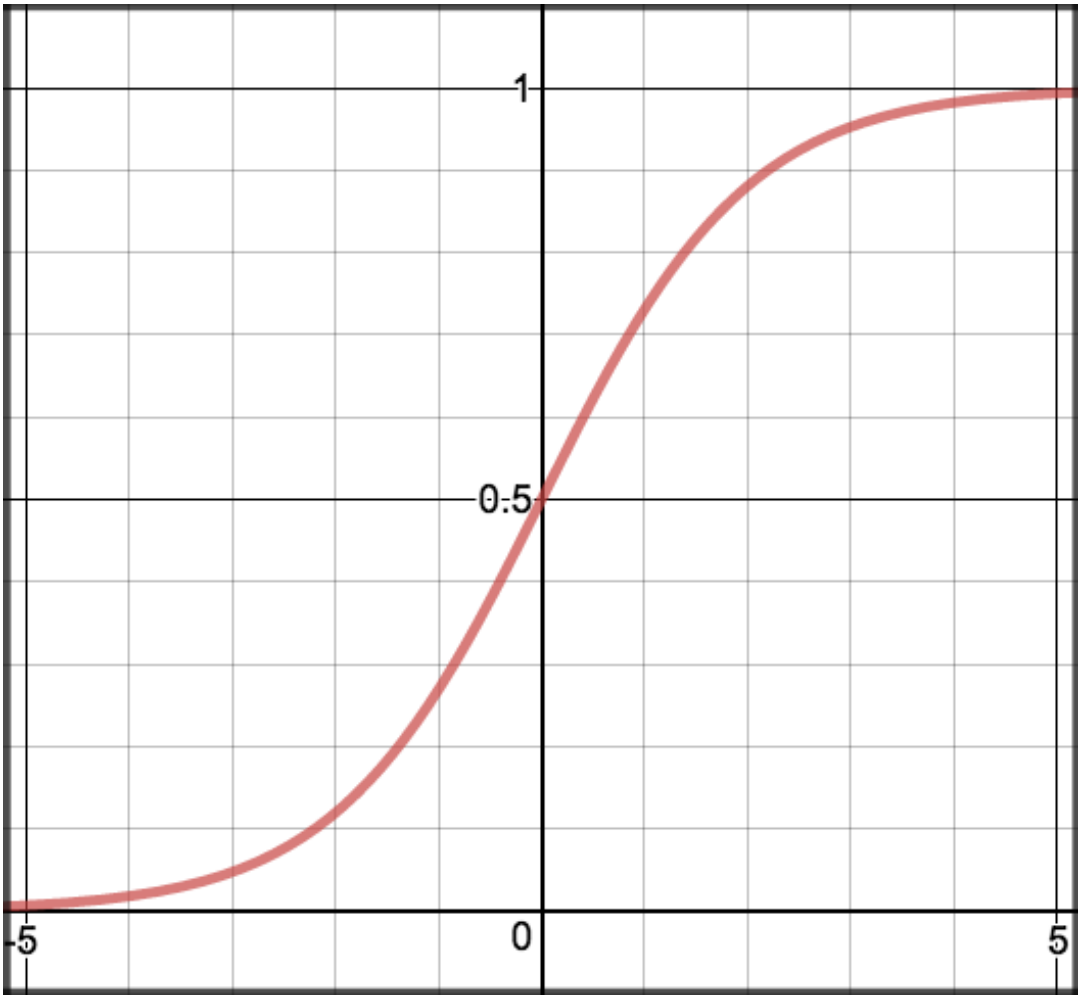
Graphically we could represent our data with a scatter plot.



Sigmoid Function

So as to delineate qualities to probabilities, we utilize the sigmoid function. The capacity maps any genuine incentive into another incentive somewhere in the range of 0 and 1. In machine learning, we utilize sigmoid to outline the probabilities.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \ldots(1)$$



Graph: Sigmoid Function

4.5 Neural Networks

A neural system (NN) is a worldview of data handling that is roused by the working of organic sensory systems, for example, our cerebrum, which forms data. The primary component of this worldview is the extraordinary novel structure of our data preparing framework. This framework comprises of countless interconnected preparing components (neurons) cooperating. Their fundamental objective is to take care of particular issues. Neural systems (NN) more often than not learn by precedent.

A Neural Networks (NN) is arranged for a specific application. This incorporates information characterization or example acknowledgment through a precise learning process. Learning in the organic frameworks by and large includes acclimations to the principle synaptic associations that typically exist between the neurons. Same is the situation with Neural Networks (NN).

Initiation Functions: Neuron can't learn with just a straight capacity that is appended to it. Any non-straight enactment capacity will dependably give it a chance to pick up as per the distinction as for mistake. Consequently initiation capacities are required.

Different types of activation functions that we will use in this project are:

4.4.1 Linear: This function is a line or can also be called linear. Therefore, the output of these functions will not be confined to any range.

Equation can be given as: f(x) = x

Range can be given as: (-infinity to infinity)

It never helps with the complexity or various different parameters of the usual data that is generally fed to the neural networks.

4.4.2 Logistic: The Sigmoid Function or the Logistic Function curve looks like a solid S-shape.

The reason why we mainly use logistic function is because of its it existence between (0 to1). Hence, it is particularly used for models where the output to be predicted is probability. Since the probability exists only in the range of 0 and 1, logistic is the right choice. The function is also differentiable. Thus we can find

the slope of the logistic function curve at any two given points. The logistic function is monotonic but its derivative is not. The softmax function can be said as a more generalized logistic activation function as it is used for multiclass classification.

4.4.3 tanh: tanh is also similar to logistic sigmoid but better. The range of this function is from (-1 to 1) and it is also sigmoidal (s - shaped). The advantage in tanh is that the negative inputs will be strongly mapped negative and zero inputs will be mapped close to zero in the tanh graph. This function is also differentiable. The function is aslo monotonic whereas its derivative is not. tanh is generally used in classification between two classes.

4.4.4 Rectified Linear Unit (ReLU): The Regulated Linear Unit (ReLU) is currently the most used activation function in the world. Since then, it has been used in almost all convolutional neuronal networks or deep learning. Its range can be specified as follows: [0 to infinity] The function is monotonic and also its derivative. However, the problem is that all negative values become zero immediately, which quickly reduces our model's ability to properly adjust or train the given data. This means that as soon as the negative entries that are passed to the ReLU activation function, the value in the graph immediately to zero, the resulting graph is affected because the negative values are not assigned accordingly

CHAPTER 5

PERFORMANCE ANALYSIS

5.1 Dataset

The dataset being used for our prediction models comprises of weather records of the city in focus collected over a period of time using various different parameters like temperature, humidity, atmospheric pressure, and so on. Till date it consists of a record of weather over a period of 20 years (1997-2016).

Characteristics

The data enclosed in our dataset is classified into the following categories:-

i) Temperature
ii) Atmospheric pressure
iii) Humidity
iv) Fog
v) Dew point

Temperature is a measure of the degree of hotness or coldness of the surroundings. It, like all weather conditions, varies from instance to instance. Similarly, atmospheric pressure and humidity, that plays a vital role in predicting whether an area will receive precipitation or not, is also included in the dataset. Details about fog and dew point are included in the dataset as well, as they only contribute to improving the accuracy of the predictions made by the prediction models.

All the data gathered in the dataset was collected from Wunderground that has an easy to use API, which makes data collection all the more simpler.

Given below is a tabular representation of the data collected in the dataset:

Date and time	Precipitation	Atmospheric pressure	Humidity	Fog	Temperature
18-11-1996 11:00	0	934	2	0	18
18-11-1996	0	936	3	0	19

12:00					
18-11-1996 1:00	0	932	4	0	20
18-11-1996 2:00	0	934	4	0	19
18-11-1996 3:00	0	934	3	0	17
18-11-1996 4:00	0	936	2	0	16

5.2 Test Metrics

5.2.1 Scikit-Learn library in Python

Scikit-learn is a free machine learning library for Python. It highlights different algorithms like support vector machine, random forests, and k-neighbors, and it likewise underpins Python numerical and scientific libraries like NumPy and SciPy.

The library has functions like accuracy_score(), RandomForestRegressor() and many other very useful regression functions that enable us to make accurate predictions.

Given below is a snapshot of the use of the library in the code.

```
26
27 #fitting logistic regression to the training set
28 from sklearn.linear_model import LogisticRegression
29 classifier = LogisticRegression(random_state=0)
30 classifier.fit(X_train, y_train)
31
```

Figure: Implementation of Scikit-Learn Library for Logistic Regression

```
16
17 #Dataset and test set split
18 from sklearn.cross_validation import train_test_split
19 X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.25, random_state = 0)
20
```

Figure: Implementation of Scikit-learn to split dataset into test set and train set

5.2.2 Pandas

Pandas is an open source, BSD-authorized library giving superior, simple to-utilize information structures and information investigation apparatuses for the Python programming language.

Pandas has been heavily utilized in the development of this project. Given below are a few snapshots from the code.

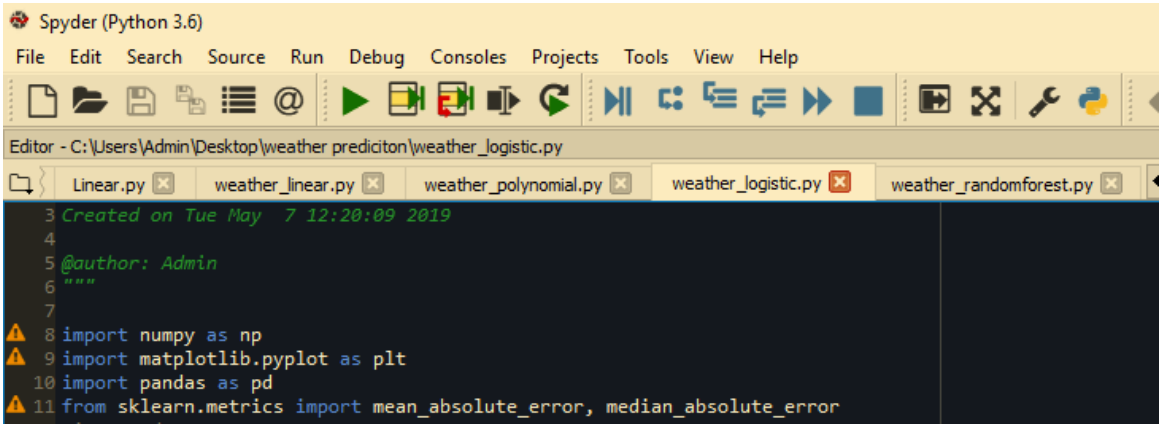


Figure: Pandas being imported to be used in the code

```
#import dataset
dataset = pd.read_csv('dataset.csv')
X = dataset.iloc[:,5:7].values
y = dataset.iloc[:,8].values
```

Figure: Using the Pandas object to import the dataset

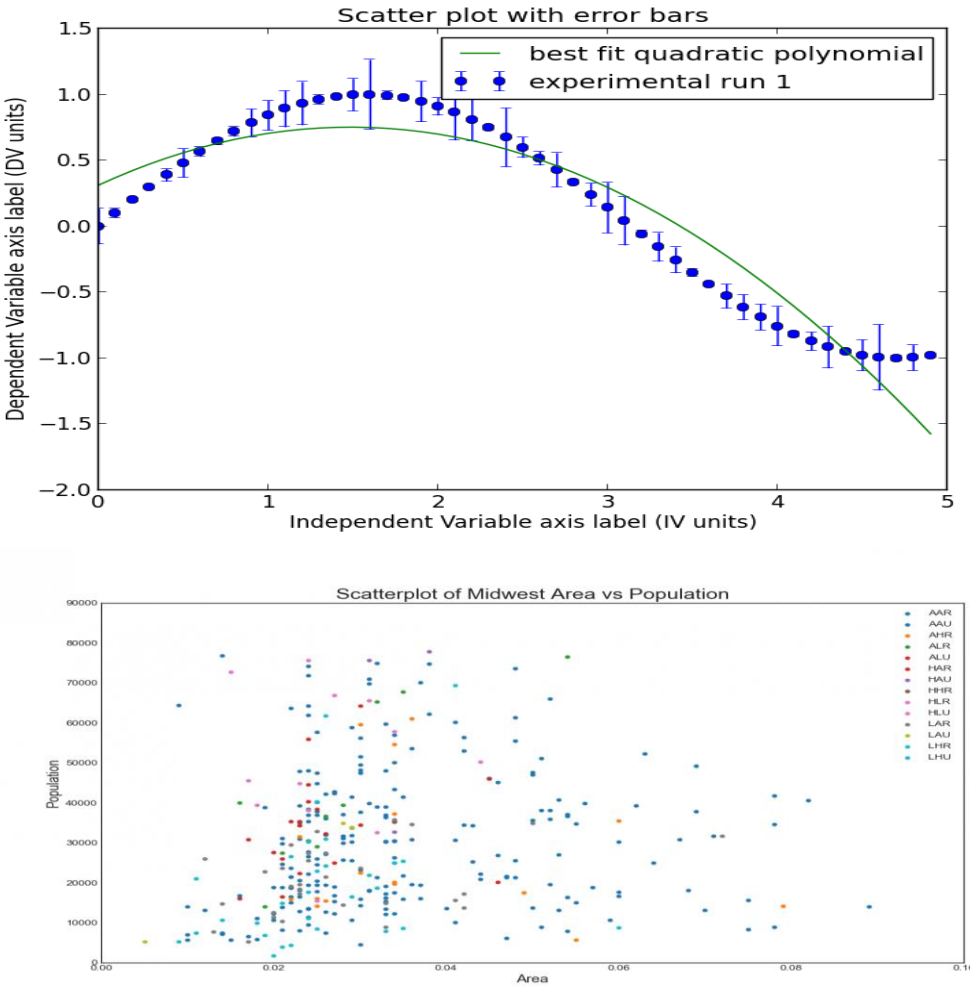
5.2.3 Matplot Library

Matplotlib is a plotting library for the Python programming language and its numerical science expansion NumPy. It gives an item situated API to implanting plots into applications utilizing broadly useful GUI toolboxs like Tkinter, wxPython, Qt, or GTK+. There is additionally a procedural "pylab" interface dependent on a state machine (like OpenGL), intended to intently look like that of MATLAB, however its utilization is discouraged. SciPy utilizes Matplotlib.

Matplotlib was initially composed by John D. Seeker, has a functioning improvement community, and is circulated under a BSD-style permit. Michael Droettboom was selected as matplotlib's lead designer presently before John Hunter's demise in August 2012, and further joined by Thomas Caswell.

Following are the examples of graphs that can be plotted using the matplot library:

Figure: Types of graphs plotted using matplot library



5.2.1 Confusion Matrix

A confusion matrix is a procedure which abridges the execution of an order calculation. It is a synopsis of forecast results on a characterization issue. Characterization precision can be deceiving there are an unequal number of perceptions in each class or if there are in excess of two classes in the dataset. Ascertaining a disarray lattice gives a superior thought of what the characterization demonstrate is getting right and the sorts of blunders it is making.

The quantity of off base forecasts and right expectations are outlined with check esteems and are separated by each class. This goes about as the way to the disarray grid. The manners by which the order display is befuddled when it makes forecasts is appeared by the disarray lattice. It gives an understanding into the mistakes being made by your classifier. It is this breakdown that conquers the restriction of utilizing characterization exactness alone.

Calculation of Confusion Matrix

The method for computing a confusion matrix is demonstrated as follows.

An arrangement of test information or an approval informational index is required with the normal outcome esteems. The forecast is made for each line in the test informational collection. From the normal outcomes and conjectures, coming up next are considered: The quantity of erroneous estimates for every classification. The quantity of right expectations for each class sorted out by the class gave. These numbers are sorted out into a table as pursues:

It is normal from the side: each line of the framework compares to an anticipated class. Forecast at the best: Each section of the table relates to a genuine class. Right and off base grouping numbers are finished in the table. The line esteem is normal for this class and the anticipated section an incentive for this class is loaded up with the aggregate number of right forecasts for a class.

Additionally, the request expected for that class esteem and the anticipated an incentive for this segment class is loaded up with the aggregate number of erroneous expectations for a class. Practically speaking, a parallel classifier like this can complete two kinds of mistakes: it is erroneously credited to a man who has not showed up in the predefined classification or is wrongly ascribed to a man who has not showed up in the predefined class. Deciding these two sorts of blunders is frequently a region of intrigue. A disarray framework is an advantageous method to show this kind of data.

This grid can without much of a stretch be utilized for issues in two classes where it is straightforward, however it can likewise be connected to issues with at least 3 class esteems, adding more lines and segments to network perplexity.

Accuracy

Accuracy is one of the measures to evaluate classification models. The precision is the fraction of the predictions given by the classification model. The precision has the following definition: Accuracy = Total no. of the correct forecasts / From predictions

For the binary classification, the accuracy can be calculated as negative and positive in the following way:

Accuracy=((TP+TN)/(TP+TN+FP+FN))

TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Table 5.2 Representation of Confusion Matrix

	Predicted Value	**Predicted Value**
Real Value	True Positive(TP) Reality: Rain ML model predicted: Rain	False Positive(FP) Reality: No Rain ML model predicted: Rain
Real Value	False Negative(FN) Reality: Rain ML model predicted: No Rain	True Negative(TN) Reality: Benign ML model predicted: No Rain

Precision

The precision determines how often it is correct when the model predicts positive. Accuracy helps determine when the cost of false positives is high.

Precision = TP / (TP + FP)

where TP is the number of real positives and FP the number of false positives. Precision refers to the ability of the classifier not to designate a positive sample as negative.

Recall

Recall it helps to determine how much the false negative cost is.

Recall = TP / (TP + FN)

Where TP is true positive and the number of FN is false negative number. Recall refers to the classification capability to find all the classified samples.

F1 Score

F1 is a measure of purity of the test. It checks both accuracy and memory. This is considered right when F1 score is 1 and there is a total failure of 0.

F1 = 2 * (Precision * Recall) / (Precision + Recall)

5.3 Test Setup

The test process is already in-built in our system. The testing process taking place just after the model is trained. After the completion of the training process, we analyze each data entry in the test set. In order to analyze each entry, we use descriptors to extract features. Now we compare these feature values with the feature values which were initially retained using the train set. The comparison is done according to the Machine Learning model used and finally the output for each entry is received. Since each data entry is already labeled, we can compute accuracy by comparing the predicted value with the received values.

5.4 Result

The results of the implementation of the project are demonstrated below.

Multiple Linear Regression:

This regression model has high variance, hence turned out to be the least accurate model. Given below is a snapshot of the actual result from the project implementation of multiple linear regression**.**

S.No	Actual Value	Predicted Value
1.	0	0.0459157
2.	0	0.0423579
3.	0	0.0474239
4.	1	0.8654278
5.	0	0.0325468
6.	0	0.0023542
7.	0	0.1236582

Table 5.4.1: Actual vs Predicted Values

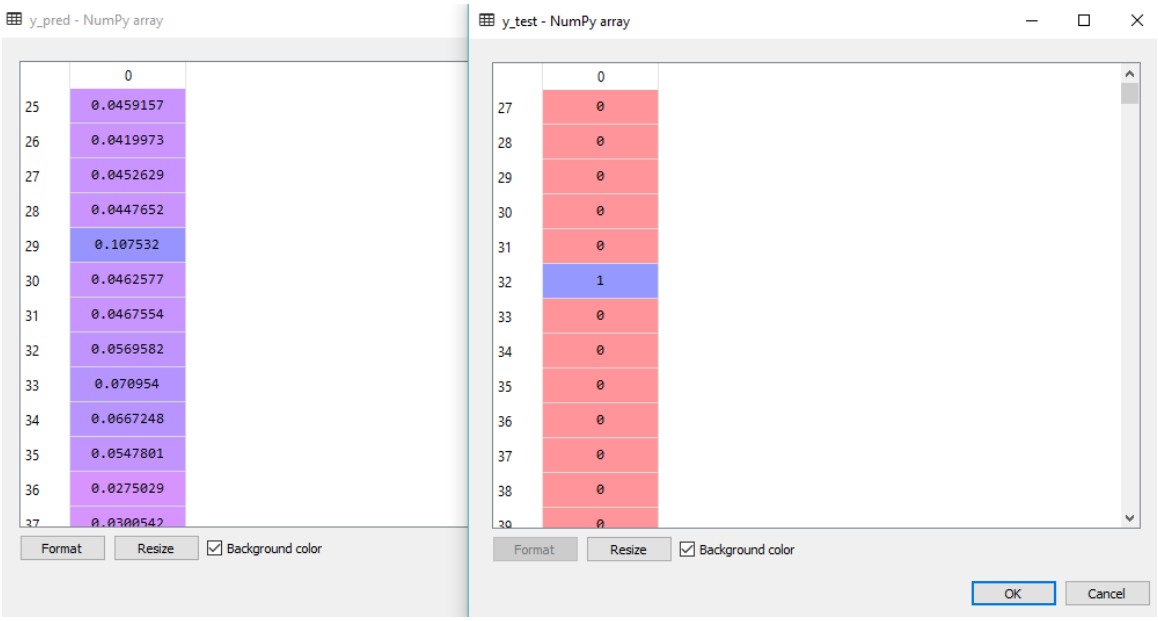


Figure 5.4.1: Predicted and actual values using Multiple linear regression.

Polynomial Linear Regression:

This regression model is much more accurate than the multiple linear regression model, hence it made predictions that were more closer to the actual result that linear regression. Below is a snapshot of its implementation in the code, and the result it displayed.

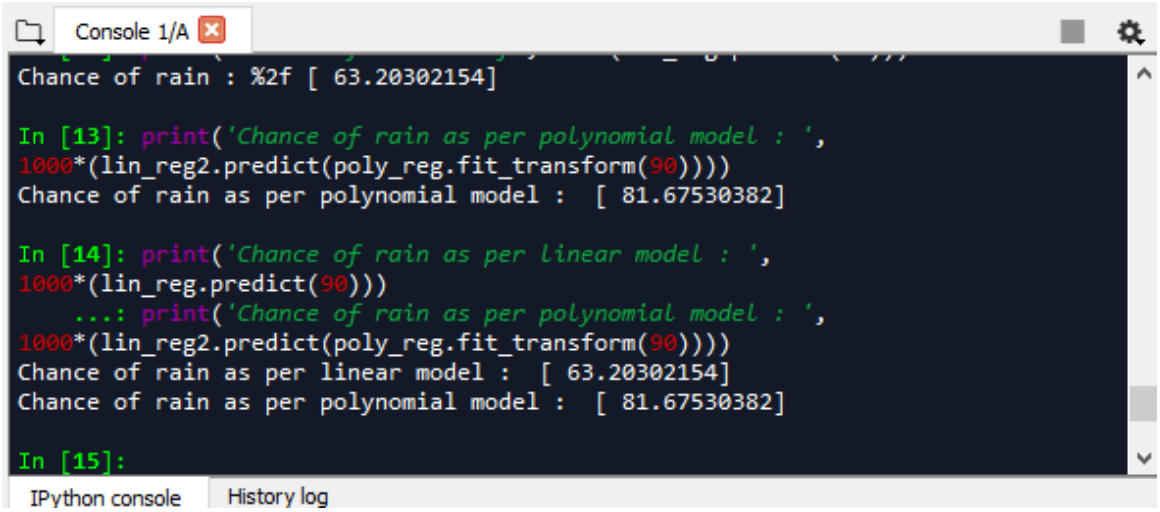


Figure 5.4.2: Comparison of results from linear regression and Polynomial regression

S.no	Actual Value	Predicted Value
1	0	0.0214568
2	0	0.2669756
3	1	0.8165476
4	0	0.0165959
5	1	0.6326548
6	1	0.7656548
7	0	0.0436597

Table 5.4.2 : Actual vs Predicted values from Polynomial regression.

Logistic regression:

This regression technique is used to classify the predictions. Here, I used binary logistic regression. The result of this regression technique was justified using the confusion matrix. The accuracy was 97%, as per the confusion matrix. Below is a snapshot of the same.

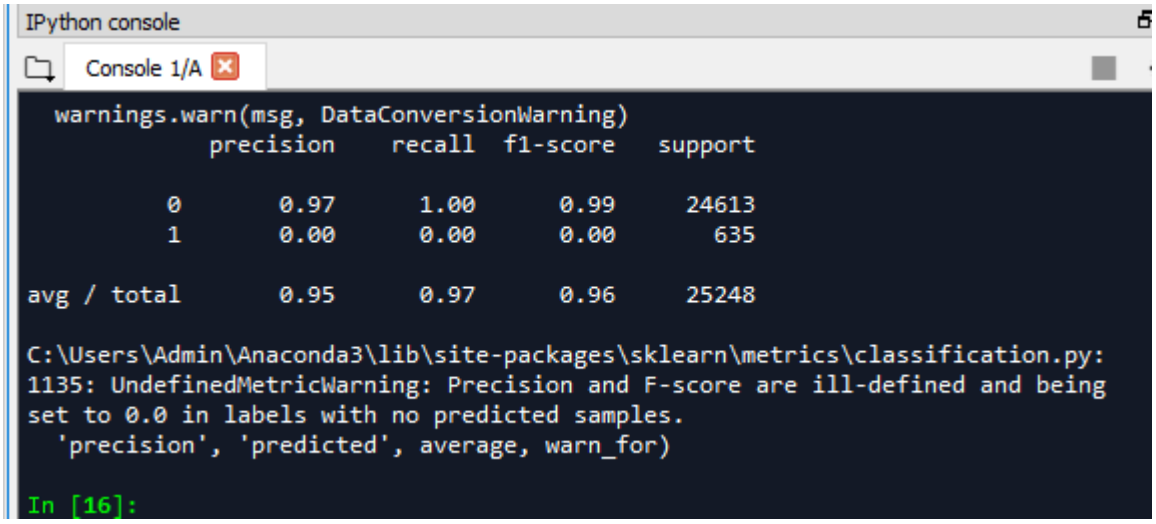


Figure 5.4.3: Confusion matrix for Logistic Regression.

	Precision	Recall	F1-score	Support
0	0.97	1.00	0.99	24613
1	0.00	0.00	0.00	635
Avg/total	0.95	0.97	0.96	25248

Table 5.4.3: Confusion Matrix for Logistic regression.

Random Forest Regression:

Out of all the regression techniques, Random Forest was the one with the maximum accuracy. Random forest is extremely versatile and widely used because of this feature. Given below is a snapshot of the result generated, compared to the actual data. The random forest was populated with 300 decision trees.

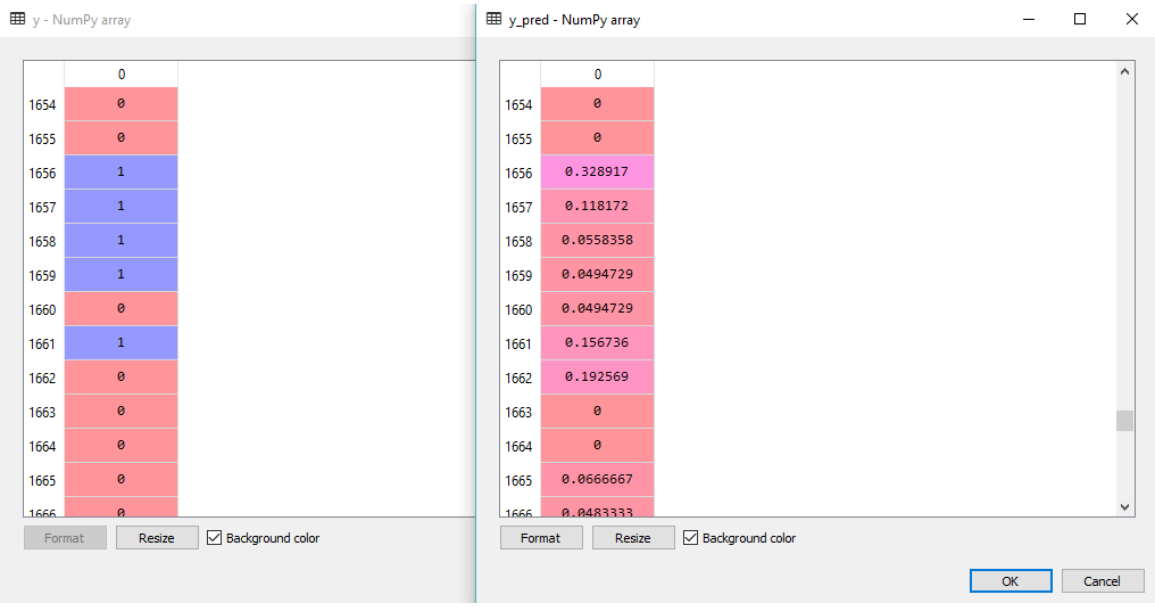


Figure 5.4.4: Actual and predicted values using Random Forest Regression.

S.No	Actual Values	Predicted Values
1	0	0
2	0	0
3	1	0.324546
4	1	0.121647
5	1	0.564642
6	1	0.195487
7	0	0

Table 5.4.4: Actual vs Predicted values from Random Forest Regression.

CHAPTER 6

CONCLUSION

All the machine learning models: linear regression, various linear regression, polynomial linear regression, logistic regression, random forest regression and Artificial neural systems were beaten by expert climate determining apparatuses, in spite of the fact that the error in their execution reduced significantly for later days, demonstrating that over longer timeframes, our models may beat genius professional ones.

Linear regression demonstrated to be a low predisposition, high fluctuation model though polynomial regression demonstrated to be a high predisposition, low difference model. Linear regression is naturally a high difference model as it is unsteady to outliers, so one approach to improve the linear regression model is by gathering of more information. Practical regression, however, was high predisposition, demonstrating that the decision of model was poor, and that its predictions can't be improved by further accumulation of information. This predisposition could be expected to the structure decision to estimate climate dependent on the climate of the previous two days, which might be too short to even think about capturing slants in climate that practical regression requires. On the off chance that the figure were rather founded on the climate of the past four or five days, the predisposition of the practical regression model could probably be decreased. In any case, this would require significantly more calculation time alongside retraining of the weight vector w, so this will be conceded to future work.

Coming to the Logistic Regression, it proved vital to classify whether a day would be rainy or not. Its significance was proven by the accuracy of the results, where it predicted the classification right, more often than not.

Figure 6.1: Logistic Regression Code

```
#fitting logistic regression to the training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state=0)
classifier.fit(X_train, y_train)

#predict
y_pred=classifier.predict(X_test)

#confusion matrix
from sklearn.metrics import confusion_matrix
cn=confusion_matrix(y_test, y_pred)

from sklearn.metrics import classification_report
print(classification_report(y_test,y_pred))
```

Talking about Random Forest Regression, it proves to be the most accurate regression model. Likely so, it is the most popular regression model used, since it is highly accurate and versatile. Below is a snapshot of the implementation of Random Forest in the project code:

```
#random Forest
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators=500,random_state=0)
regressor.fit(X,y)
y_pred=regressor.predict(X)
```

Figure 6.2: Random Forest Regression code.

ANN with backpropagation utilizes an iterative procedure of preparing where, it more than once contrasts the watched yield and focused on yield and computes the mistake. This blunder is utilized to rearrange the estimations of loads and predisposition to show signs of improvement yield. Subsequently this technique attempts to limit the blunder. In this manner, Artificial Neural system with Backpropagation algorithm is by all accounts most fitting strategy for estimating climate precisely.

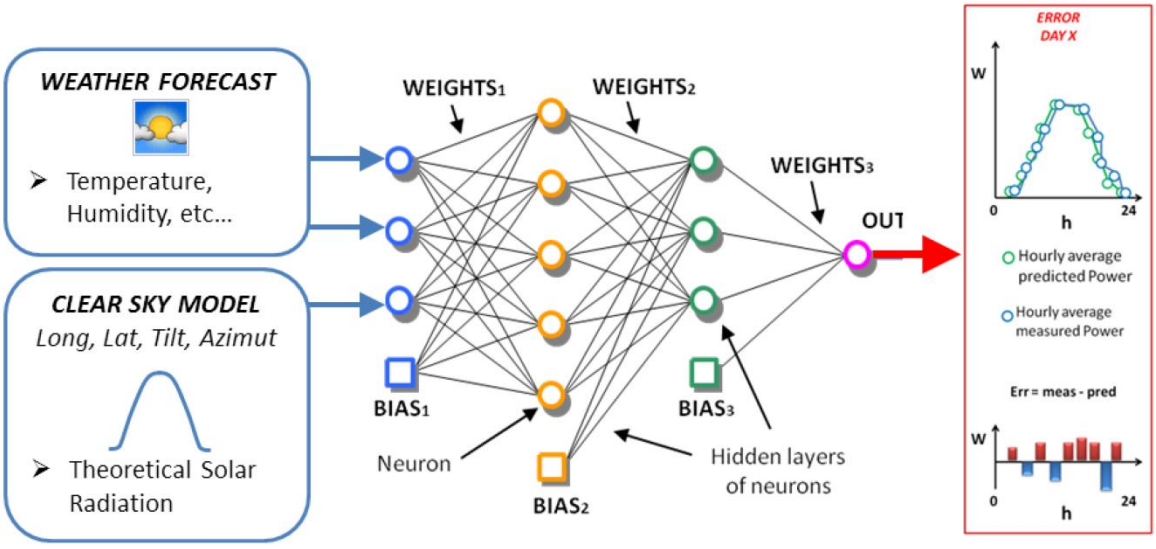


Figure 6.3: Diagrammatic representation of weather prediction using ANN.

The climate Forecasting has a major test of foreseeing the precise outcomes which are utilized in numerous ongoing frameworks like power offices, air terminals, the travel industry focuses, and so forth. The trouble of this determining is the mind boggling nature of parameters. Every parameter has an alternate arrangement of scopes of qualities. This issue is tended to by ANN. It acknowledges every single complex parameter as info and produces the clever examples while preparing and it utilizes similar examples to create the gauges.

REFERENCES

[1] Mohammad Wahiduzzaman, Eric C. J. Oliver, Simon J Wotherspoon, Neil J. Holbrook, “A climatological model of North Indian Ocean tropical cyclone genesis, tracks and landfall”.

[2] Jinglin Du, Yayun Liu , Yanan Yu and Weilan Yan, “A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms”

[3] Prashant Kumar, Atul K. Varma, “ Atmospheric and Oceanic Sciences Group, EPSA, Space Applications Centre (ISRO), Ahmedabad, IndiaAssimilation of INSAT-3D hydro-estimator method retrieved rainfall for short-range weather prediction”

[4] Prashant Kumar, C. M. Kishtawal, P. K. Pal, “Impact of ECMWF, NCEP, and NCMRWF global model analysis on the WRF model forecast over Indian Region”

[5] H. Vathsala, Shashidhar G. KoolagudiPrediction, “Model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches”

[6] Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning applied to weather forecasting”, Stanford University, 2016.

[7] Gyanesh Shrivastava, Sanjeev Karmakar, Manoj Kumar Kowar, “ Application of Artificial Neural Networks in Weather Forecasting: A Comprehensive Literature Review”, International Journal of Computer Application, 2012.

[8] Meera Narvekar, Priyanca Fargose, “Daily weather forcasting using Artificial Neural Network”, International Journal of Computer Application, 2015.