Tribhuvan University

Institute of Science and Technology

A Project Report

On

"Weather Forecasting System Using Naïve Bayes, Decision Tree And Regression Algorithm"

Submitted To:

Office of the Dean

Institute of Science and Technology

Tribhuvan University

Kirtipur, Nepal

Under the supervision of

Mr. Bikash Balami

A project report submitted for the Partial Fulfillment of the Requirement of Bachelor of Science in Computer Science and Information Technology (BSc.CSIT) of 7th Semester of Tribhuvan University, Nepal

Submitted By:

Kushal Karki (TU Exam Roll No. 20991/075)

Pranita Dangol (TU Exam Roll No. 21003/075)

Prerana Upadhyay(TU Exam Roll No. 21005/075)

May, 2023

Date: …….

SUPERVISOR'S RECOMMENDATION

I hereby recommend that this project prepared under my supervision by **Kushal Karki, Pranita Dangol** and **Prerana Upadhyay** entitled "**Weather Forecasting System Using Naïve Bayes, Decision Tree And Regression Algorithm**" in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Information Technology is satisfactory to process for the future evaluation.

……..

Mr. Bikash Balami
Supervisor
Asian School of Management and Technology

Date: …….

CERTIFICATE OF APPROVAL

The undersigned certify that he has read and recommended to the Department of Computer Science and Information Technology for acceptance of report entitled "**Weather Forecasting System Using Naïve Bayes, Decision Tree And Regression Algorithm**" submitted by **Kushal Karki, Pranita Dangol** and **Prerana Upadhyay** in partial fulfilment for the degree of Bachelor of Science in Computer Science and Information Technology (BSc.CSIT), Institute of Science and Technology, Tribhuvan University.

……

ER. Anil Lal Amatya

Principal

……

Program Co-Ordinator

……

Mr. Bikash Balami

Supervisor

……

External Examiner

ACKNOWLEDGEMENT

We owe our gratitude to our respected supervisor, **Mr. Bikash Balami** for his enthusiasm, patience, insightful comments, practical advice and unceasing ideas that have guided and helped us tremendously at all times in our research and writing of this report. His immense knowledge, profound experience and professional expertise have enabled us to complete this project successfully.

Further, we would like to express our special thanks to the Department of Computer Science and Technology of our college for providing all the necessary resources and facilities for the project development activities. We also acknowledge and owe to the assistance provided by other supervisors and panels, particularly in our project presentation, which has improved our presentation skills as a result of their comments and advice. We are thankful and fortunate enough to get constant support from our seniors and every teaching staff, without their support existence of this project would not have been possible.

Special thanks to all the staff members of B.Sc. CSIT department and colleagues who directly or indirectly helped and encouraged us.

Thanking you,

Kushal Karki (TU Exam Roll No. 20991/075)

Pranita Dangol (TU Exam Roll No. 21003/075)

Prerana Upadhyay(TU Exam Roll No. 21005/075)

ABSTRACT

Weather forecasting is the application of science and technology to predict the state of the atmosphere. The topic of the project is “Weather Forecasting System using Naïve Bayes (Gaussian) and Regression algorithm”. The main object of the project is to predict the current weather condition on the basis of given user input and to compare the accuracy of different algorithms. Here we are using machine learning where we have trained the model first and then test.

Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. For example, it might be observed that if the sunset was particularly red, the following day often brought fair weather. However, not all of these predictions prove reliable. Here this system will predict weather based on parameters such as temperature and wind. User will enter current temperature and wind, System will take this parameter and will predict weather from previous data in database (dataset). The role of the admin is to add previous weather data in database, so that system will calculate weather based on these data. Weather forecasting system takes parameters such as temperature and wind and will forecast weather based on previous record therefore this prediction will prove reliable. This system can be used in Air Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

Contents

LIST OF ABBREVIATION

CSS	Cascading Style Sheet
DFD	Data-flow Diagram

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction

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

Machine learning can be used to process immediate comparisons between historical weather forecasts and observations. With the use of machine learning, weather models can better account for prediction inaccuracies, such as overestimated rainfall, and produce more accurate predictions. 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.

Weather forecasting is used to predict the state of the atmosphere for a given location. Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. For example, it might be observed that if the sunset was particularly red, the following day often brought fair weather. However, not all of these predictions prove reliable. Here this system will predict the weather based on parameters such as temperature and wind. The user will enter the current temperature and wind, System will take this parameter and will predict weather from previous data in a database (dataset). The role of the admin is to add previous weather data in the database temperature and wind so that the system will calculate weather based on these data. The weather forecasting system takes parameters such as temperature and wind and will forecast weather based on previous records therefore this prediction will prove reliable. This system can be used in Air Traffic, Marine, Agriculture, Forestry, Military, and Navy, etc.

1.2 Problem Statement

Weather forecasting is a prediction of what the weather will be like in the future. The purpose of this project is to extract the patterns for day-to-day weather prediction from historical weather data using data mining. In this project, a prototype of the system will be developed which includes the main components of the system such as training, analysis, and prediction. The traditional forecast process employed by most National Meteorological and Hydrological Services (NMHSs) involves forecasters producing text-based, sensible, weather-element forecast products (e.g. maximum/minimum temperature, cloud cover) using numerical weather prediction (NWP) output as guidance. The process is typically schedule-driven, product-oriented, and labor-intensive. Over the last hydro-meteorological forecasts and warnings to become much more specific and accurate.

1.3 Objectives

The major objective of our project is:

- To predict the weather using the Naïve Bayes (Gaussian) and Regression algorithm and to compare accuracy between them.

1.4 Scopes and Limitations

The application doesn't require any registration, and any resident is able to use the application. The application can be used by the user to predict future weather conditions.

The limitation of this application are:

- The prediction may not be 100% accurate.
- This system is based on assumptions, approximations, and natural conditions which render those predictions inaccurate.

1.5 Development Methodology

For developing our application we followed the Scrum development process, which is an agile development methodology based on iterative and incremental processes. To follow this method we divided our project workflow into a different sprints with each sprint to be completed within a week or depending on the task to be done. Thus by breaking our project plan to sprint, each of the team members could focus on individual tasks and complete on time.

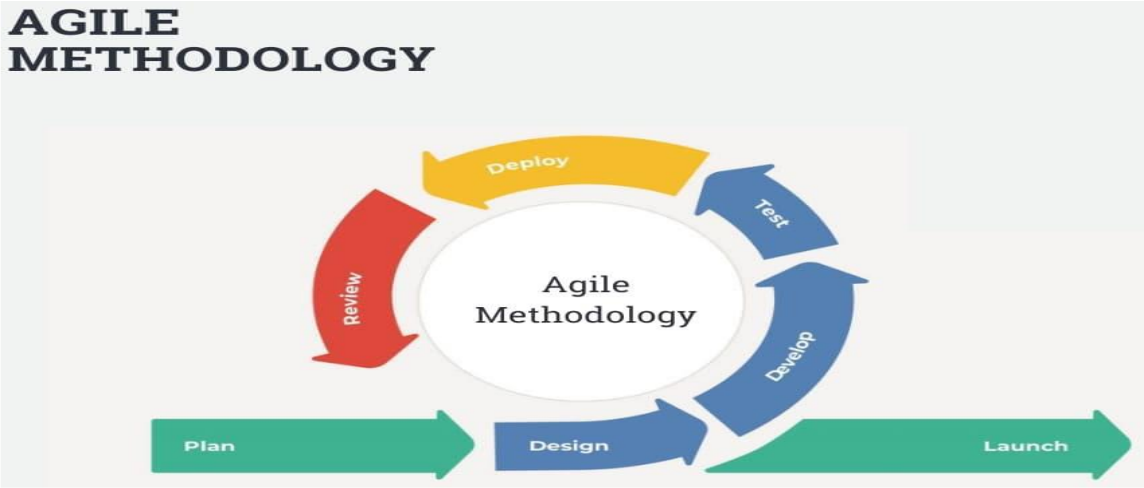

Figure 1.1 Agile Methodology

1.6 Report Organization

The Report Organization includes the contents about how the study is being organized and carried out. The report is divided into six chapters. **Chapter one** contains the background of the study, statement of the problem, objectives of the study, scope, and limitations of the study, development methodology, and organization of the study. **Chapter two** contains a background study and a review of related literature. **Chapter three** contains system analysis including requirement analysis and feasibility study, the way of data modeling, and process modeling of the system. **Chapter four** includes the design of the system including system architecture, database design, interface design, and the algorithm used for designing the system. **Chapter five** contains an implementation of the system using different tools and finally testing the system as a unit and as a whole. **Chapter six** contains conclusions and future recommendations from the study.

CHAPTER 2: BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Background study

Weather forecasting is the science of predicting future weather conditions by analyzing past and current weather patterns. The use of weather forecasting systems has become increasingly important in recent years due to changing climate, with the potential for extreme weather events such as hurricanes, floods, and wildfires.

Other popular mobile-based weather forecasting apps include Accuweather, Appy Weather, Google Feed, Over drop, and Today Weather. Although the mentioned app is doing a respectable job when it comes to quick prediction, some do need account for registration. This gives rise to our project Weather Forecasting System.

The main task of a Weather Forecasting System is to provide the user the information about future weather conditions and to find out which model is more accurate. Weather Forecasting System is mainly done by implementing Naïve Bayes and Regression Algorithm.

Naïve Bayes Algorithm and Regression Model is used to solve many different problem statements, and it is quite fast in training a model since they completely works on probability, so the conversion happens quickly. The implementation of this algorithm was a big challenge because its time complexity, and space complexity all have been considered. This project helps to implement this algorithm properly and gives the best result to the user who wants to get information about future weather conditions.

2.2 Literature Review

Weather forecasting has been one of the most challenging difficulties around the world because of both its practical value in popular scope for scientific study and meteorology. Weather is a continuous, dynamic, multidimensional chaotic process, and data-intensive and these properties make weather forecasting a stimulating challenge. On a worldwide scale, large numbers of attempts have been made by different researchers to forecast Weather accurately using various techniques. But due to the nonlinear nature of Weather, prediction accuracy obtained by these techniques is still below the satisfactory level.

Physicist Piero Paialunga in 2021 (Weather forecasting with Machine Learning, using Python*)* discuss based on forecasting the average temperature using traditional machine learning algorithms: Auto Regressive Integrated Moving Average models (ARIMA) and he found that these methods are extremely easy to adopt as they don't require any specific computational power like Deep Learning methods (RNN, CNN..) In this model climate has been deefined as "complex system". That is unsolvable in analytical ways to solve this climate challenge machine learning algorithms has been used like Auto Regressive Integrated Moving Average models (ARIMA) with python framework [1]. Weather forecast using the Naïve Bayes classifier algorithm has been implemented by Dheemant Bhat using python where he trained and tested the model on a sample weather forecast dataset [2].

A weather forecasting application using python has been implemented by C K Gomathy in 2022 where it successfully predicted the rainfall using linear regression but here this is not very accurate only sometimes any way it depends upon the climate changes from season to season. Prediction of weather in weather forecasting using python has been implemented. Traditional forecast process employed by most NMHSs involves forecasters producing text-based, sensible, weather-element forecast products (e.g. maximum/minimum temperature, cloud cover) using numerical weather prediction (NWP) output as guidance [3]. In 1950 metrologist Jule Charney of MIT, Agnar Fjortoff, and mathematician John Neumann published "Numerical Integration of the Barotropic Vorticity Equation,". The paper reported the first weather forecast by an electronic computer. It took twenty-four hours of processing time to calculate a twenty-four-hour forecast. Weather forecasting in Python Django has been implemented by (Yugesh Verma, 2021) which involves predicting things like cloud cover, rain or snow, wind speed, and temperature, where forecasting the weather by looking at current conditions, the motion of air and clouds, historical patterns,

pressure changes, and computer models [4]. Weather forecasting using time series data and algorithm done by Aman Kharwal on 2022 using python [5]. Innovations and New Technology for Improved Weather Services by John L. Guiney where Different innovations and technologies have imerged like Internet, wireless communication, digital database forecasting, next-generation workstations, nowcasting systems. This article provides overview of different key innovations technical advancements and IT systems which can help improvising weather forecasting [6]. Weather Forecasting Model using Artificial Neural Network, Author links open overlay panel Kumar Abhishek , M.P. Singh , Saswata Ghosh , Abhishek Anand [7]. Imran Maqsood, Muhammad Riaz Khan, and Ajith Abraham. An ensemble of neural networks for weather forecasting, Neural Comput & Applic (2004) [8]. Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring Weather Data December 2018International Journal of Computer Applications 182(34):20-24DOI:10.5120/ijca2018918265 Authors: Munmun Biswas BGC Trust University Bangladesh Tanni Dhoom Premier University Sayantanu Barua BGC Trust UniversityBangladesh [9].

CHAPTER 3: SYSTEM ANALYSIS

3.1 System Analysis

System analysis includes evaluation of a system to achieve certain objectives during the development phase. Including analyzing its functional requirement which is an overview of how a system operates or should behave in any particular situation along with non-functional requirements regarding its usability, availability as well reliability. System analysis also emphasizes feasibility analysis to determine if a system is feasible or not. It basically includes looking at the wider system and figuring out how the system works in order to achieve a specific goal.

3.1.1 Requirement Analysis

A requirement analysis is a description of a prediction system to be developed, laying out functional and non-functional requirements.

i. Functional Requirement

Functional requirements are the requirements that describe the functionalities of the system elements. The system should be able to produce the minimum, maximum and average data of a particular weather parameter when it is requested by the user. The system should be able to provide the information about the temperature, pressure, rainfall etc. To show the functional requirement of the system, we have built an use case diagram. In the below use case diagram we have two actors user and admin. In this diagram it shows how we can get the weather updates.

The data from the user are taken as inputs. Then the data is processed. Data is implemented in the trained module and are tested. After testing the data, the weather is predicted and executes the output.

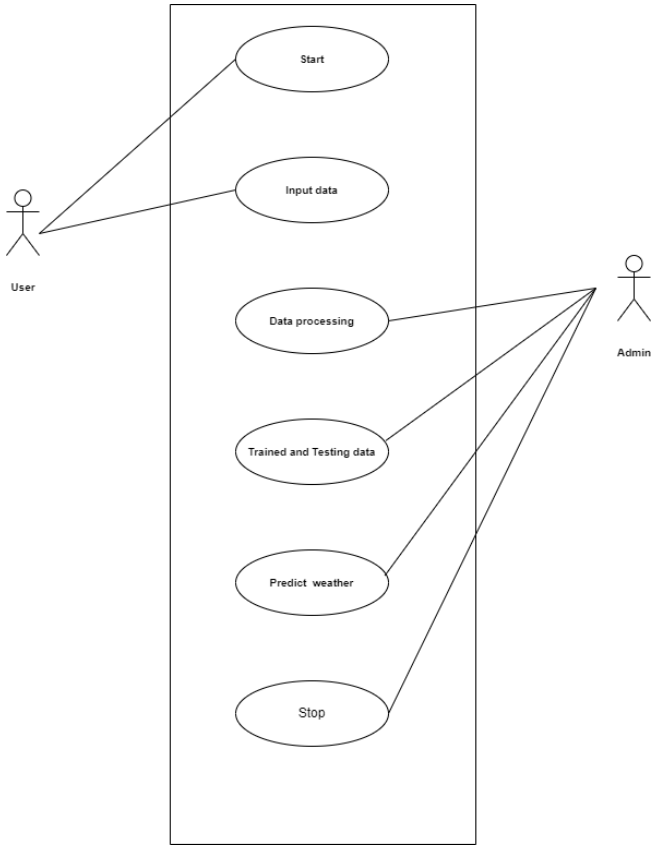

Figure 3.1 Use Case Diagram for Weather Forecasting System

ii. Non-functional Requirement

Non-functional requirements describe the system properties and constraints whereas these may not describe directly what the system should do or perform.

- **Usability:** Users can easily and effectively learn and use a system.
- **Security**: Assures all data inside the system or its part will be protected against malware attacks or unauthorized access.
- **Portability**: The system can be easily launched within one environment or another.
- **Scalability**: Under high workloads the system will still meet the performance requirements.

3.1.2 Feasibility Study

While working on this project we came across different phases and were able to analyze the feasibility accordingly. The feasibility study of this project is divided into four major categories depending upon the distribution of time and availability of resources. They are given as follows:

i. Technical Feasibility

This is a web-based system that uses the Naïve Bayes algorithm and Regression model as its core algorithm. We will be able to construct this system using our existing knowledge of technologies. This is not a large system that might bring complexities in the near future. It is a simple system that can be accessed by the internet and can be used on a regular basis.

ii. Operational Feasibility

This system aims at detecting almost accurate changes in the weather along with accuracy of model. It tries its best to update the temperature on an hourly basis. It's a simple system with simple features. A person with basic knowledge of the internet can easily use this system.

iii. Economic Feasibility

This system will be constructed at a minimum price. Almost every resource and data set required for the system will be acquired from the internet. We will be able to do this project using our understanding of available languages and technologies. The dataset required for this project will be gathered from an online source. Once the datasets are collected, we will be able to train and test the model using a decent available machine.

iv. Schedule Feasibility

The time given for the completion of this project was a whole semester. So, we had enough time for completion for this project.

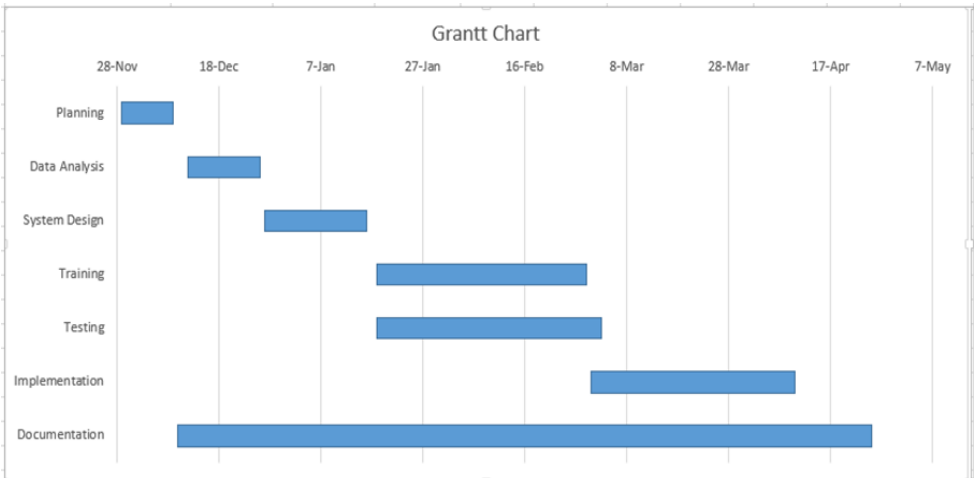

Figure 3.2 Gantt Chart the working Schedule of the System

The above Gantt chart displays the overall timeline of the project. It presents the sequential breakdown of the task involved in the project with the time taken for each task. The Planning and Analysis Phase was carried out in parallel with the Data Collection. After that

coding was done followed by testing and deployment in their mentioned time. And finally, documentation was done and the final report was prepared.

3.1.3 Analysis

The Weather forecasting system is a system that aims to detect the weather condition and update the present condition of the weather and compare the accuracy between the algorithms. The system will use a naïve algorithm and regression to model the datasets and predict the current temperature condition.

i. Workflow Diagram

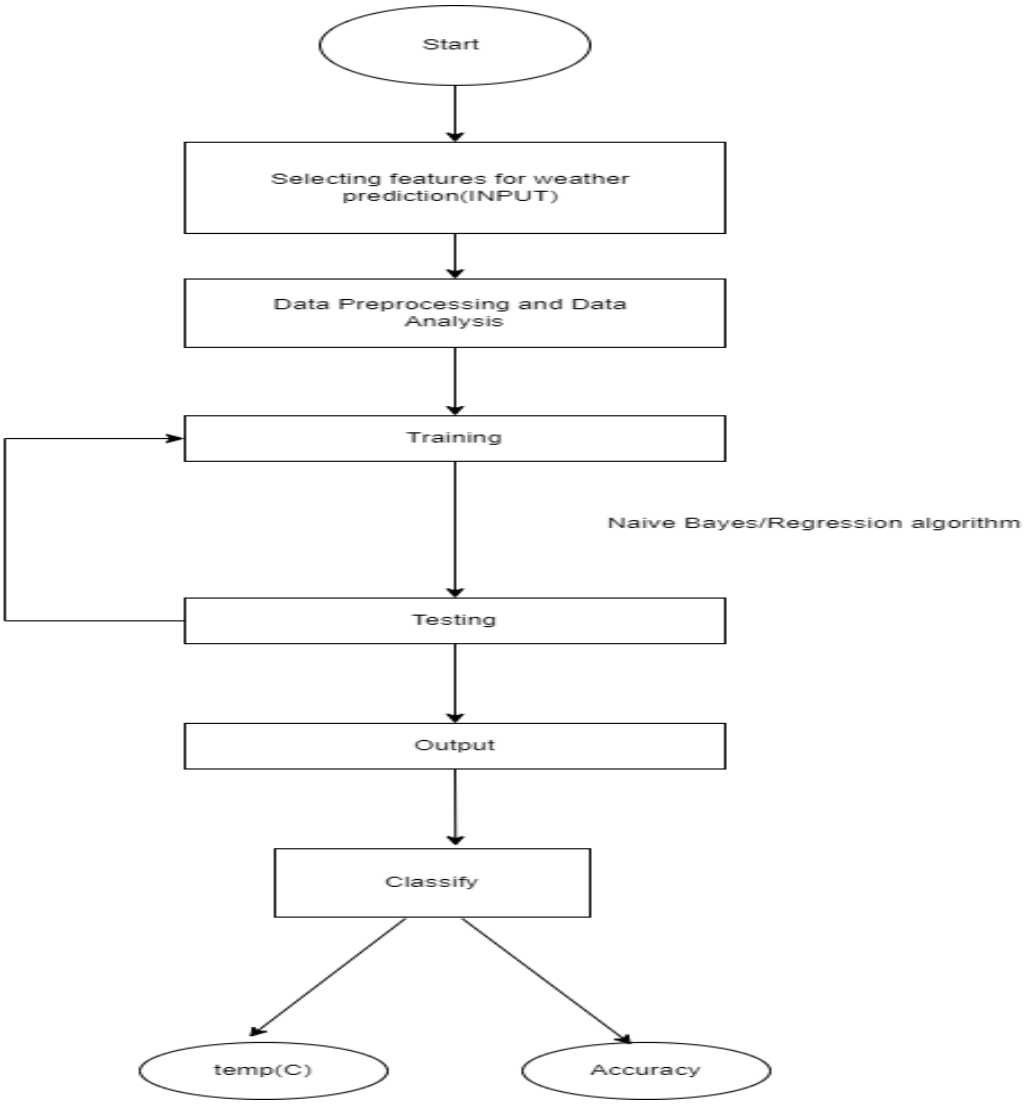

Figure 3.3 Workflow Diagram of Weather Forecasting System

ii. Database Design

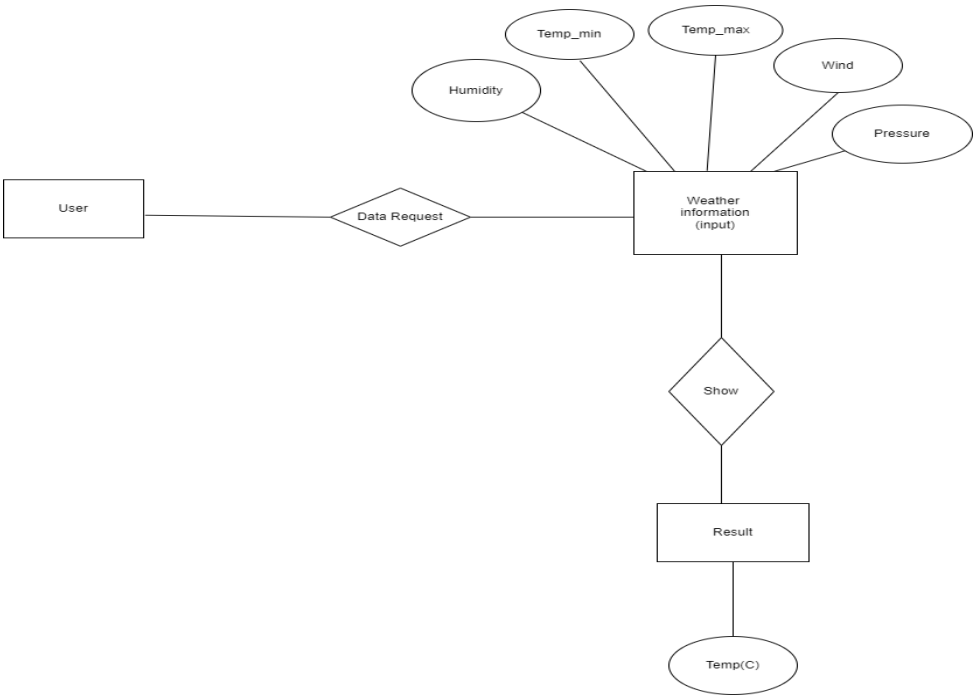

Figure 3.4 Database Design of Weather Forecasting System

An database design describes interrelated things of interest in a specific domain of knowledge. A basic database design is composed of entity types (which classify the things of interest) and specifies relationships that can exist between entities (instances of those entity types). Here user first request data by giving the inputs such as Max Temp(C), Min Temp(C), Cloud Cover, Humidity etc then the result will be shown as temp(celcius).

iii. Process Modeling using DFD

The process model is a core diagram in structured analysis and design. A data-flow diagram is a way of representing a flow of data through a process or a system (usually an information system). The DFD also provides information about the outputs and inputs of each entity and the process itself. A data-flow diagram has no control flow there are no decision rules and no loops.

In data flow diagram level 0 the whole system is represented as a single process. A level 1 DFD notates each of the main sub-processes that together form the complete system. We can think of a level 1 DFD as an "exploded view" of the context diagram.

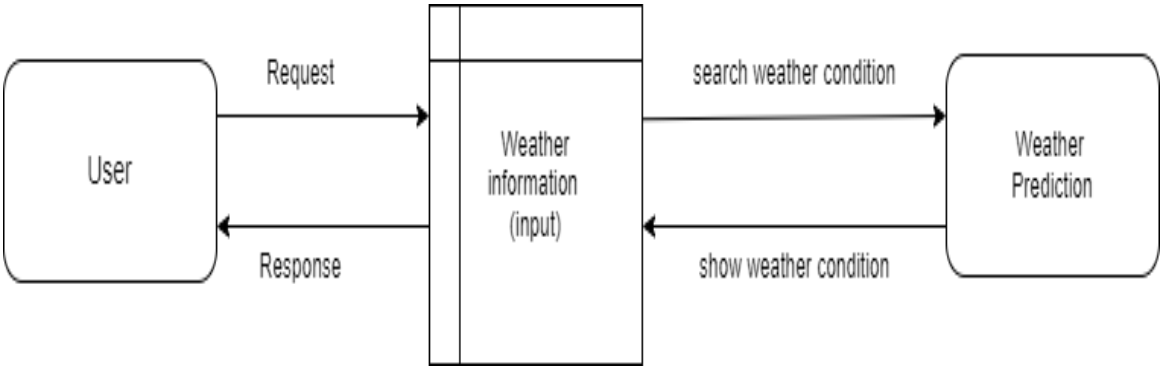

Figure 3.5 Data Flow Diagram Level 0

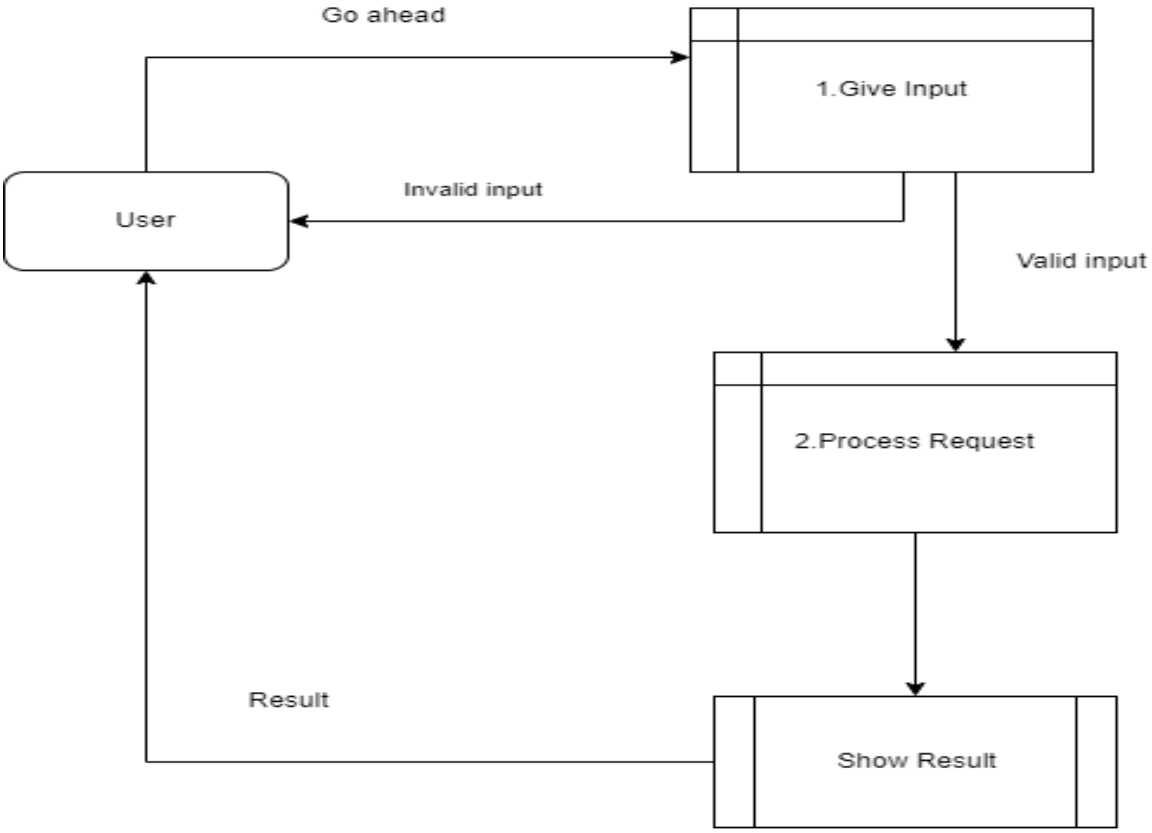

Figure 3.6 Data Flow Diagram Level 1

CHAPTER 4: SYSTEM DESIGN

4.1 System Design

System Design includes the means and methodologies to improve the management and control of the software development process. It includes structuring and simplifying the process using several diagrams and standardizing the development process by specifying the required activities and techniques to be implemented.

The process of weather prediction involves various steps. First of all the weather datasets is obtained from online site. These dataset are then preprocessed to remove noise. After preprocessing, the dataset are used to train a model. Then we test the datasets and find out the accuracy of the model. And finally we compare the accuracy of different algorithms.

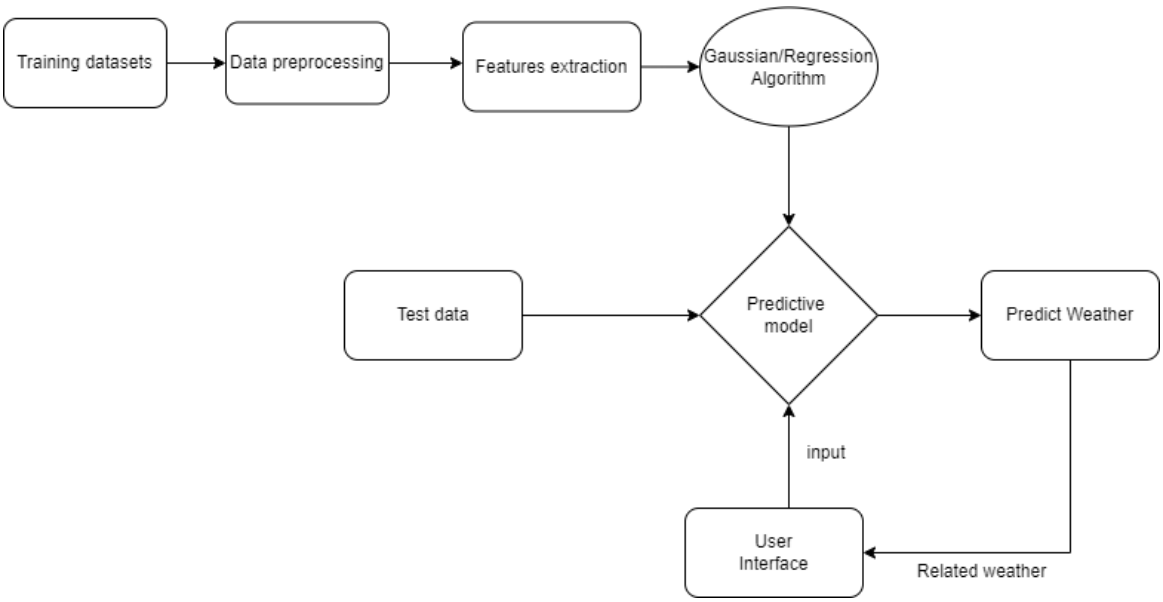

Figure 4.1 System Design of Weather Forecasting System

4.2 Algorithm Used

4.2.1 Naïve Bayes Classifier

Naive Bayes Classifier is a very popular supervised machine learning algorithm based on Bayes' theorem. It is simple but very powerful algorithm which works well with large datasets and sparse matrices, like pre-processed text data which creates thousands of vectors depending on the number of words in a dictionary. It works really well with text data projects like sentiment data analysis, performs good with document categorization projects, and also

it is great in predicting categorical data in projects such as email spam classification. It is used to solve many different problem statements, and it is quite fast in training a model since Naive Bayes classifier completely works on probability, so the conversion happens quickly. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Here in our model we are using gaussian algorithm.

Figure 4.2 Types of Naïve Bayes Classifier

Gaussian Naive Bayes (GNB) is a classification technique used in Machine Learning (ML) based on the probabilistic approach and Gaussian distribution. Gaussian Naive Bayes assumes that each parameter (also called features or predictors) has an independent capacity of predicting the output variable. If we assume that X's follow a Gaussian or normal distribution, we must substitute the probability density of the normal distribution and name it Gaussian Naïve Bayes. To compute this formula, you need the mean and variance of X.

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

In the above formulae, sigma and mu is the variance and mean of the continuous variable X computed for a given class c of Y.

4.2.2 Multiple Linear Regression

Linear Regression is a method that describes the relationship between a dependent variable and a set of independent variables. The equation of the line is given as Y=wx+b. It provides an estimate of weather using various atmospheric variables like cloud cover, humidity, wind, and average temperature to predict weather. An estimate of weather is easy to

determine at any given point since the regression method uses the previous correlation between the various atmospheric variables. Therefore, our equation will look like:

$$y_i = \beta_0 + \beta_1 x_{i1} \dots \dots + \beta_p x_{ip} + \varepsilon_i$$

where,

y_i = The predicted variable

β_0 = The intercept

β_1 = Measures the change in y_i with respect to x_{i1}

β_p = Measures the change in y_i with respect to x_{ip}

$x_{i1} \dots \dots x_{ip}$ = Predictor variable and ε_i the error

4.2.3 Decision Tree Regression

Decision tree is a widely used classification model. It can extract a tree-type classification model from the given training samples. The decision tree is composed of nodes and directed edges. There are two types of nodes, internal nodes and leaf nodes. Each internal node in the tree records which attribute is used for classification, each branch represents the output of a judgment result, and each leaf node represents the result after the final classification. One internal node represents a feature or an attribute; one leaf node represents a classification. The decision tree learning algorithm is usually a recursive selection of the optimal feature, which is based to segment the training data. This is the best classification process for each sub-data set, which corresponds to the division of the feature space and the construction of the decision tree. Entropy of the decision tree: Entropy is a concept borrowed from information theory quantifying randomness or disorder. The set with high entropy value is quite diversified, and the decision tree expects to find a segmentation that can reduce the entropy value, so as to eventually increase the isotropy in the group. The decision tree is a complex tree generated by taking full account of all data points, so there may be overfitting. Therefore, the decision tree should be pruned to reduce the complexity of the tree and the probability of overfitting. Decision tree pruning refers to deleting all the child nodes of a subtree and using the root node as a new leaf node.

4.2.4 Random Forest Regression

In this study, we use a random forest machine learning algorithm to predict temperature. Random forest algorithm is a supervised learning algorithm in machine learning that uses ensemble learning methods for regression. Ensemble learning is a technique that combines predictions from multiple machine learning algorithms to provide more accurate predictions than a single model. The random forest algorithm is the best algorithm for analyzing large amounts of data. Due to its high predictive accuracy, this algorithm is the most accessible and provides details about the importance of variables for classification and regression. In contrast to ANNs and SVMs, the training method for random forest algorithms is simple. The main parameter that needs to be adjusted is the number of trees. Artificial Intelligence and Support Vectors Performs a faster training process compared to machine based models. This algorithm includes a rule based approach and does not require data normalization.

The random-forest classification algorithm the machine learning technique that employs ensembling learning to combine two or more machine learning models to produce a new single model. It works by training the dataset with many decision trees and then presenting the categorization modes of the various trees. Random-Forest believed as the one and only finest ensemble-classifiers in the high-dimensional data. Random-forests are a set of trees predictors in which the values of the randomly sampled vectors with the equal distribution across all trees in the forest is used to build each tree in it. Each trees are trained with replacement on a variety subset of the training datas. To estimate error and variable relevance, the remaining of training data are employed. The number of votes from all of the trees is used to assign classes, and the average of the results is utilized for regression.

CHAPTER 5: IMPLEMENTATION AND TESTING

5.1 Implementation

The system is implemented using different tools and technologies like frontend tools for user interface and backend tools for establishing a connection with the database. It includes updated and recent technology to make it compatible with different browsers and for smooth functionality. The code base for system implementation is well-structured and clean code.

5.1.1 Tools Used

i. Frontend tools

- **HTML**

  HTML is the most basic building block of the web. It defines the meaning and structure of web content. It was used to develop the web page.

- **CSS**

  CSS is the style sheet language for describing the presentation and design of web pages including colors, fonts and layouts. It was used to enable the distinction between presentation and content, including colors, layout and fonts.

  .

ii. Backend tools

- **Python**

  Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation via the off-side rule. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object oriented and functional programming.

- **Django**

  Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel.

- **Jupyter Notebook**

  It provides an interactive computing environment that allows users to run code and see the results immediately. We used this platform to train and evaluate our model.

iii. Diagram Table

- **Draw.io**

  It is a popular online diagramming tool used to create various types of diagrams. We used this tool to create following diagram:

  - Use case Diagram
  - Class Diagram
  - Data flow Diagram
  - Database Design

- **Microsoft Project**

  It is a project management software application developed by Microsoft. The project schedules and progress are visually represented through Gantt chart provided by Microsoft project, which show tasks and their dependencies.

5.1.2 Data Collection

The data was collected through online sites like Kaggle. Where we can get different real time datasets for testing and training the model.

5.1.3 Methodology

The dataset utilized in this arrangement has been gathered from Kaggle. The dataset was created by keeping in mind the necessity of such historical weather data in the community. The datasets contain hourly weather data from 01-01-2009 to 01-01-2020. The data of each city is for more than 10 years. This data can be used to visualize the change in data due to global warming or can be used to predict the weather for upcoming days, weeks, months, seasons, etc. Furthermore, this data can also be used to make visualization which would help to understand the impact of global warming over the various aspects of the weather like precipitation, humidity, temperature, etc. In this

project, we are concentrating on the temperature prediction with the help of various machine learning algorithms and various regressions. By applying various regressions on the historical weather dataset we are predicting the temperature like first we are applying Gaussian Naïve Bayes, Multiple Linear regression, then Decision Tree regression, and after that, we are applying Random Forest Regression.

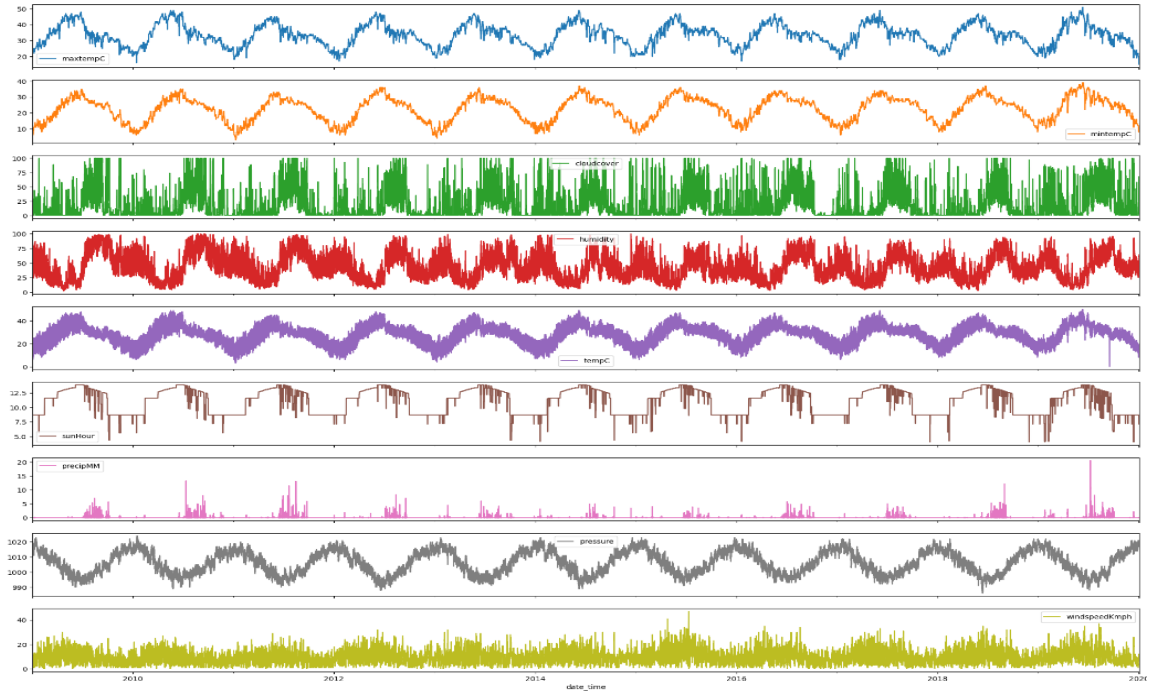

Figure 5.1 Plot for each factor for 10 years

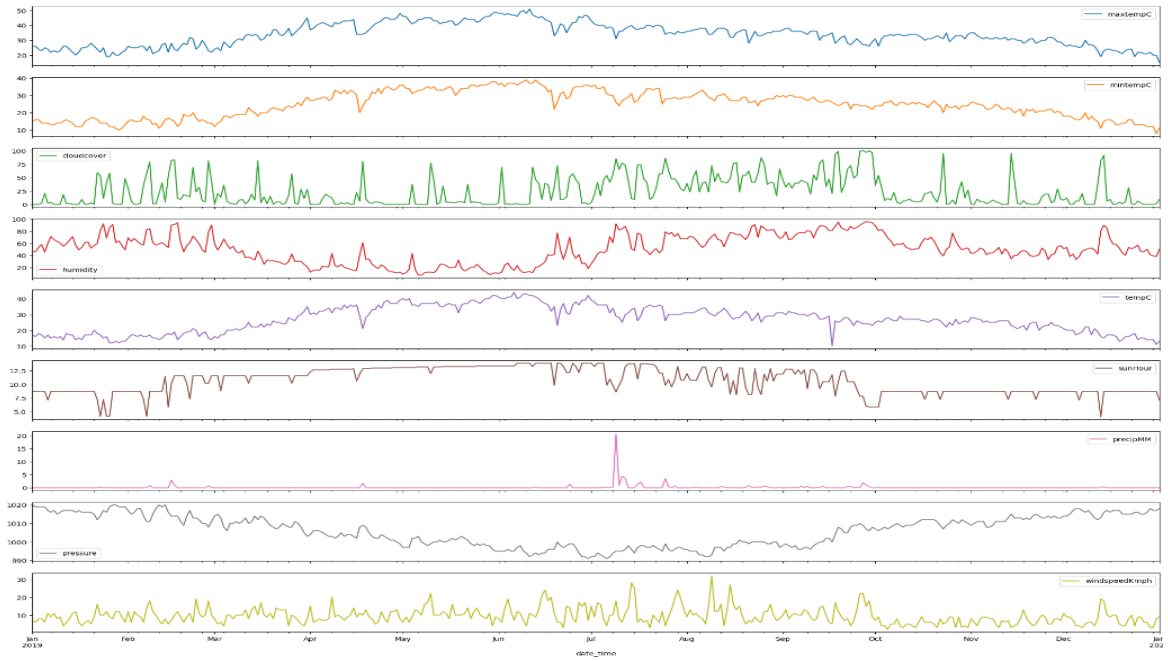

Figure 5.2 Plot for each factor for 1 years

5.1.4 Experimentation

The record has just been separated into a train set and a test set. Each information has just been labeled. First, we take the trainset organizer. We will train our model with the help of histograms and plots. 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 Gaussian Naïve Bayes, Linear Regression, Decision Tree Regression, and Random Forest Regression. With the help of our histogram, we will train our model. The most important thing in this process is to tune these parameters 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 the 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 the 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 confusion matrix, R2 score, etc.

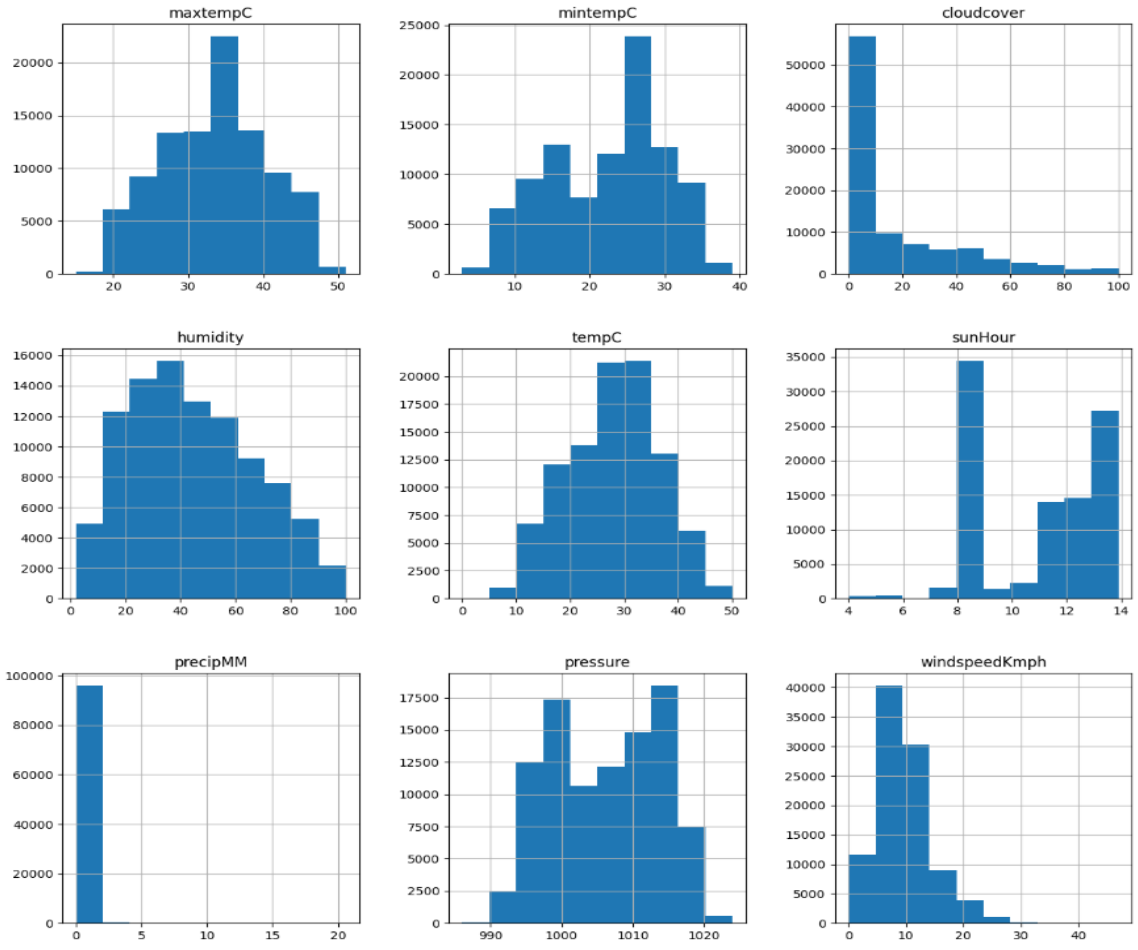

Figure 5.3 Experimentation Datasets

5.1.5 Algorithms for Gaussian classifier

Gaussian Distribution is also called Normal Distribution. The normal distribution is a statistical model that describes the statistical distributions of continuous random variables in nature. The normal distribution is defined by its bell-shaped curve. The two most important features of the normal distribution are a mean (**μ**) and a standard deviation (**σ**). The mean is the average value of a distribution, and the standard deviation is the "width" of the distribution around the mean.

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

5.1.6 Multiple Linear Regression

Linear Regression is a method that describes the relationship between a dependent variable and a set of independent variables. The equation of the line is given as Y=wx+b. It provides an estimate of weather using various atmospheric variables like cloud cover, humidity, wind, and average temperature to predict weather. An estimate of weather is easy to determine at any given point since the regression method uses the previous correlation between the various atmospheric variables. Therefore, our equation will look like:

$y_i = \beta_0 + \beta_1 x_{i1} \dots\dots + \beta_p x_{ip} + \varepsilon_i$

where,

y_i = The predicted variable

β_0 = The intercept

β_1 = Measures the change in y_i with respect to x_{i1}

β_p = Measures the change in y_i with respect to x_{ip}

$x_{i1} \dots\dots x_{ip}$ = Predictor variable and ε_i the error

5.1.7 Decision Tree Regression

Decision tree is a widely used classification model. It can extract a tree-type classification model from the given training samples. The decision tree is composed of nodes and directed edges. There are two types of nodes, internal nodes and leaf nodes. Each internal node in the tree records which attribute is used for classification, each branch represents the output of a judgment result, and each leaf node represents the result after the final classification. One internal node represents a feature or an attribute; one leaf node represents a classification. The decision tree learning algorithm is usually a recursive selection of the optimal feature, which is based to segment the training data. This is the best classification process for each sub-data set, which corresponds to the division of the feature space and the construction of the decision tree. Entropy of the decision tree: Entropy is a concept borrowed from information theory quantifying randomness or disorder. The set with high entropy value is quite diversified, and the decision tree expects to find a segmentation that can reduce the entropy value, so as to eventually increase the isotropy in the group. The decision tree is a complex tree generated by taking full account of all data points, so there may be overfitting. Therefore, the decision tree should be pruned to reduce the complexity of the tree and the probability of overfitting. Decision tree pruning refers to deleting all the child nodes of a subtree and using the root node as a new leaf node.

5.1.8 Random Forest Regression

In this study, we use a random forest machine learning algorithm to predict temperature. Random forest algorithm is a supervised learning algorithm in machine learning that uses ensemble learning methods for regression. Ensemble learning is a technique that combines predictions from multiple machine learning algorithms to provide more accurate predictions than a single model. The random forest algorithm is the best algorithm for analyzing large amounts of data. Due to its high predictive accuracy, this algorithm is the most accessible and provides details about the importance of variables for classification and regression. In contrast to ANNs and SVMs, the training method for random forest algorithms is simple. The main parameter that needs to be adjusted is the number of trees. Artificial Intelligence and Support Vectors Performs a faster training process compared to machine based models. This algorithm includes a rule based approach and does not require data normalization.

The random-forest classification algorithm the machine learning technique that employs ensembling learning to combine two or more machine learning models to produce a new

single model.It works by training the dataset with many decision trees and then presenting the categorization modes of the various trees. Random-Forest believed as the one and only finest ensemble-classifiers in the high-dimensional data. Random-forests are a set of trees predictors in which the values of the randomly sampled vectors with the equall distribution across all trees in the forest is used to build each tree in it. Each trees are trained with replacement on a variety subset of the training datas. To estimate error and variable relevances, the remaining of training data are employed. The number of votes from all of the trees is used to assign classes, and the average of the results is utilized for regression.

5.2 Testing

Testing is the process of evaluating the systems functionality, performance, and accuracy to ensure that it meets the requirements and objectives of the system.

5.2.1 Unit Testing

Unit testing was done by taking the small sample of the datasets. We performed unit testing on every activity from preprocessing to the final results. Finally, whole system worked fine.

5.2.2 System Testing

First of all, we tested sample datasets and classification tasks and then integrate them to form the whole system. Then, we tested the system which gave correct results.

5.3 Result Analysis

Our project main goal is to predict the correct weather in term of given correct inputs. We have separate the entire datasets into 2 parts. Training and testing datasets. To analyze the accuracy of our project, we split out datasets into 80-20 on testing and training.

5.3.1 Gaussian Naïve Bayes

This naïve bayes model has high mean absolute error, hence turned out to be the least accurate model. Given below is a actual result from the project implementation of gaussian naïve bayes.

Actual	Prediction	diff
34	34	0
25	23	2
34	32	2
28	20	8

28	30	-2
27	29	-2
18	12	6
22	23	-1
27	23	4
29	27	2

5.3.2 Random Forest Regression

This regression model has low mean absolute error, hence turned out to be the more accurate model among all the models. Given below is a actual result from the project implementation of random forest regression.

Actual	**Prediction**	**diff**
34	33.940000	0.060000
25	24.430000	0.570000
34	34.360000	-0.360000
28	26.350000	1.650000
28	28.170000	-0.170000
27	26.990000	0.010000
18	15.445000	2.555000
22	21.350000	0.650000
27	25.924333	1.075667
29	28.790000	0.210000

5.3.3 Multiple Linear Regression

This regression model has second high mean absolute error, hence turned out to be the least accurate model. Given below is a actual result from the project implementation of multiple linear regression.

Actual	**Prediction**	**diff**
34	33.209030	0.790970
25	25.275755	-0.275755
34	31.975338	2.024662
28	20.496727	7.503273
28	28.401085	-0.401085
27	25.764618	1.235382
18	18.713759	-0.713759
22	24.203460	-2.203460

27	28.580807	-1.580807
29	24.922430	4.077570

5.3.4 Decision Tree Regression

This regression model has low mean absolute error, hence turned out to be the more accurate model. Given below is a actual result from the project implementation of decision tree regression.

Actual	Prediction	diff
34	34.0	0.0
25	25.0	0.0
34	34.0	0.0
28	28.0	0.0
28	28.0	0.0
27	27.0	0.0
18	14.0	4.0
22	18.0	4.0
27	26.0	1.0
29	29.0	0.0

5.3.5 Mean absolute error and R2-score

The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the datasets. It measures the average of the residuals in the dataset.

The R2 score is a very important metric that is used to evaluate the performance of a regression based machine learning model.

Algorithms	Mean absolute error	R2-score
Random Forest Regression	1.2972674009608605	0.9510226386177125
Decision Tree Regression	1.657514560757678	0.8992529086993096
Multiple Linear Regression	2.508334289021174	0.8603753521958108

Figure 5.4 Mean absolute error and r2 score

CHAPTER 6: CONCLUSION AND FUTURE RECOMMENDATIONS

6.1 Conclusion

We successfully predicted the weather condition using the gaussian naïve bayes and various regressions algorithm but here this is not very accurate. We have compare the accuracy of the four algorithm, among all the algorithm random forest regression is the most accurate and gaussian naïve bayes algorithm is the least accurate algorithm. Here we are taking the datasets of certain years. we define the system's requirement specification and the actions that can be taken on these objects. we comprehend the problem domain and created a system model that represents the operations that can be performed on the system. we created the user interface which is designed to be user-friendly and allow users to easily access and view the weather forecasts. Finally, the system is build and tested in accordance with the test cases.

6.2 Future Recommendations

Further improvements can yet be implemented. The app may use a vital addition in addition to UI and UX improvements. To make the software more accountable, other aspects like data collection methods might be improved.

Incorporating more AI and machine learning: Artificial Intelligence (AI) can analyze vast amounts of data and provide more accurate predictions.

Appendices

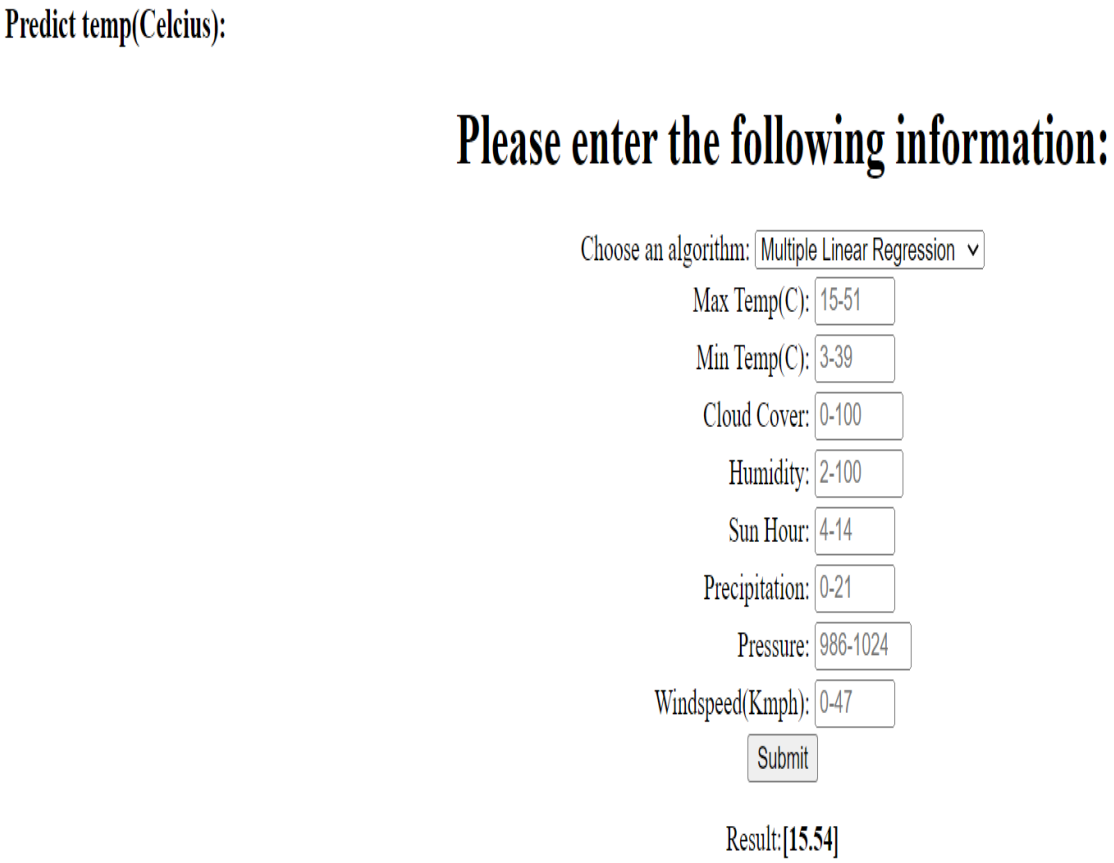

Figure 6.1 User Interface

References

[1] P. P. Paialunga, "Weather forecasting with Machine Learning using Python," 2021. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-from-scratch-with-python-942708211470.

[2] D. Bhat, "Weather forecasting using Naive Bayes classifier algorithm in python," [Online]. Available: https://www.kaggle.com/code/dheemanthbhat/naive-bayes-classifier/notebook.

[3] C. K. Gomathy, "Weather forecasting application using python (rain pridiction)," 2022. [Online]. Available: .https://www.researchgate.net/publication/360620450_WEATHER_FORECASTING_APPLICATION_USING_PYTHON .

[4] Y. Verma, "Weather forecasting in Python Django," 2021. [Online]. Available: https://projectworlds.in/weather-forecast-project-in-python-django-with-source-code/ .

[5] A. Kharwal, "Weather forecasting using time series data and algorithm," 2022. [Online]. Available: https://thecleverprogrammer.com/2022/10/17/weather-forecasting-using-python/ .

[6] J. L. Guiney, "Innovation and New Technology for Imporved weather services," [Online]. Available: https://public.wmo.int/en/bulletin/innovations-and-new-technology-improved-weather-services .

[7] M. S. S. G. A. A. Kumar Abhishek, "Weather Forecasting Model using Artificial Neural Network," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221201731200326X .

[8] M. R. K. a. A. A. Imran Maqsood, "An ensembel of neural networks for weather forecasting," 2004. [Online]. Available: https://link.springer.com/article/10.1007/s00521-004-0413-4.

[9] M. Biswas, "Weather Forecast Prediction: An Integrated Approach for Analyzing and Measuring weather data," 2018. [Online]. Available: https://www.ijcaonline.org/archives/volume182/number34/30251-2018918265.

Supervisor Log

S.N	Date	Activity	Signature
1.	22/12/2022	Discussed about the idea of Weather forecasting system	
2.	28/12/2022	Submitted Proposal for 1st time and got feedback	
3.	04/01/2023	Second Submission of Proposal	
4.	08/01/2023	Third Submission of Proposal	
5.	11/01/2023	Reviews regarding the Proposal	
6.	16/01/2023	Submitted Proposal	
7.	21/01/2023	Proposal Defense	
8.	28/02/2023	Reviews regarding project and documentation	
9.	05/03/2023	Small Demo	
10.	21/04/2023	Discussed About Diagrams	
11.	23/04/2023	Submitted Report 1st Time	
12.	01/05/2023	Submitted Report 2nd Time	
13.	06/05/2023	Mid Defense	
14.	26/05/2023	Final Defense	