Symbiosis Skills and Professional University
Kiwale, Pune

PROJECT REPORT
On

SkyInsight Machine Learning Weather Forecasting System.

SUBMITTED BY

Rushikesh Pandhare.
Abhijeet Dhotre.
Rohit Gend.

REGISTERED BATCH: ML_08

UNDER THE GUIDANCE OF

TRAINER: AMRITA AJOTIKAR MA'AM.

STUDENT DECLARATION AND ATTESTATION BY TRAINER

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Name Of Student **Signature**

Rushikesh Pandhre

Abhijeet Dhotre

Rohit Gend

Signature of trainer

Amrita ma'am

CERTIFICATE

This is to certify that,

Rushikesh Pandhare,

Abhijeet Dhotre,

Rohir Gend,

Has completed and submitted the project entitled, "SkyInsight Machine Learning Weather Forecasting System", under the guidance of Amrita ma'am, to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a recordof Bonafede project work carried out by them and is worthy of consideration for the completion of certificate course in "Machine Learning".

Date:

Signature **Supervisor**
Miss Amrita ma'am

ACKNOWLEDGEMENT

We are profoundly grateful to trainer Amrita ma'am for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to express our deepest appreciation towards SYMBIOSIS SKILLS & PROFESSIONAL UNIVERSITY, Kiwale. Prof. Sumeet Sir and Prof. Baliram Sir whose invaluable guidance supported us in completing this project and also JP Morgan for funding and giving us this opportunity.

At last, we must express our sincere heartfelt gratitude to all staff members of Computer Science & Engineering Department who helped us directly or indirectly during this course of work.

Thanking you all!!!

INDEX

ABSTRACT

Weather, a dynamic and ever-changing facet of our natural world, has captivated human curiosity for centuries. This project delves into the rich history of weather patterns, offering a comprehensive examination of how our understanding of the atmosphere has evolved over time. Spanning epochs and cultures, this endeavor explores the development of meteorological knowledge, the impact of weather on societies, and the ways in which modern technology has transformed our ability to predict and respond to weather phenomena.

Beginning with ancient civilizations, we trace the roots of weather observation and interpretation. From the Babylonians' celestial observations to the Chinese's meticulous recording of rainfall, we uncover early attempts to understand atmospheric phenomena. We then follow the thread of weather science through the Renaissance, where figures like Galileo and Leonardo da Vinci made pioneering contributions to our understanding of meteorology.

The project highlights the pivotal role of the 19th century in the history of meteorology. This era witnessed the establishment of meteorological networks, the development of instruments like the barometer and anemometer, and the emergence of weather forecasting as a discipline. The advent of telegraphy and global data sharing during this time marked a significant turning point, enabling meteorologists to piece together a more complete picture of global weather systems.

The 20th century brought remarkable advancements, from the discovery of El Niño to the development of radar and satellite technology. These innovations revolutionized our ability to monitor and predict weather events, saving countless lives and mitigating damage from natural disasters. The project also explores the emergence of climate science and the recognition of anthropogenic climate change, underscoring the importance of weather history in addressing contemporary environmental challenges

KEYWORDS:Weather History, Climate Data, Historical Weather Records, Meteorological Archives, Weather Patterns, Temperature Trends, Precipitation History, Wind Patterns, Extreme Weather Events, Climate Change Analysis, Weather Data Sources, Weather Data Analysis.

1. Introduction:

1.1 Overview:

The "Weather History Project" is a comprehensive endeavor aimed at delving into the past to unravel the intricate tapestry of Earth's weather patterns. Weather is an integral aspect of our daily lives, influencing everything from agriculture and transportation to disaster preparedness and urban planning. Understanding historical weather data is crucial for gaining insights into long-term climate trends, assessing the impact of climate change, and making informed decisions about our environment.

1.2 Motivation:

Motivation for this project also springs from the recognition of climate change as one of the defining challenges of our era. As we confront the consequences of anthropogenic climate change, a deeper understanding of historical weather patterns becomes increasingly vital. By examining the past, we gain insight into the natural variability of weather, enabling us to distinguish between natural fluctuations and the human-induced shifts we now witness. In turn, this knowledge empowers us to make informed decisions for the future.

Furthermore, the Weather History Project is driven by the desire to bridge the gap between science and the general public. Weather is a topic that touches everyone, yet it often remains shrouded in complexity. Through this project, we aim to demystify meteorology, making it accessible to all. By presenting historical weather data in an engaging and understandable manner, we hope to foster a greater appreciation for the beauty and significance of the atmospheric phenomena that surround us daily.

In conclusion, the motivation behind the Weather History Project is a testament to the enduring human quest for knowledge, a commitment to addressing pressing global challenges, and a passion for sharing the wonders of the natural world. As we embark on this journey through the annals of weather history, we invite you to join us in uncovering the remarkable stories hidden within the ever-changing skies.

1.3 PROBLEM STATEMENT:

The "Weather History Project" aims to collect, analyze, and present historical weather data for a specific geographical location, with the goal of providing valuable insights into past weather patterns and trends

1.4 OBJECTIVE:

i. To improve long-term weather forecasting by analyzing historical data for patterns and trends that can inform predictive models.

ii. To create interactive visualizations or maps that allow users to explore historical weather data in a user-friendly way.

iii. To provide data and tools for climate scientists and researchers to conduct such analyses.

iv. This data can be valuable for climate research, understanding long-term weather patterns, and studying the impact of climate change.

METHODOLOGIES OF PROBLEM SOLVING

2.1 Algorithm:

(a) Linear Regression:

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. The range of the coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data between two variables.

(b) Decision Tree:

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision trees are composed of three main parts—decision nodes (denoting choice), chance nodes (denoting probability), and end nodes (denoting outcomes). Decision trees can be used to deal with complex datasets, and can be pruned if necessary to avoid overfitting.

(c) Random Forest;

Random forest is solid choice for nearly any prediction problem (even non-linear ones). It's a relatively new machine learning strategy (it came out of Bell Labs in the 90s) and it can be used for just about anything. It belongs to a larger class of machine learning algorithms called ensemble methods.

(d) KNN:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well –

(i) **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

(ii) **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data

(e) Naive Bayes:

Naïve Bayes algorithms is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other. We have the following three types of Naïve Bayes model under Scikit learn Python library −

(i) Gaussian Naïve Bayes: It is the simplest Naïve Bayes classifier having the assumption that the data from each label is drawn from a simple Gaussian distribution.

(ii) Multinomial Naïve Bayes: Another useful Naïve Bayes classifier is Multinomial Naïve Bayes in which the features are assumed to be drawn from a simple Multinomial distribution. Such kind of Naïve Bayes are most appropriate for the features that represents discrete counts.

(iii) Bernoulli Naïve Bayes: Another important model is Bernoulli Naïve Bayes in which features are assumed to be binary (0s and 1s). Text classification with 'bag of words' model can be an application of Bernoulli Naïve Bayes

In this project we had implemented Gaussian Naïve Bayes.

(f) Support Vector Machines:

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following step − First, SVM will generate hyperplanes iteratively that segregates the classes in best way.Then, it will choose the hyperplane that separates the classes correctly.

2.2 HYPOTHESIS TESTING:

The Pearson's Chi-Square statistical hypothesis is a test for independence between categorical variables. In this article, we will perform the test using a mathematical approach and then using Python's SciPy module. We start by defining the null hypothesis (H0) which states that there is no relation between the variables. An alternate hypothesis would state that there is a significant relation between the two.

If our calculated value of chi-square is less or equal to the tabular(also called critical) value of chi-square, then H0 holds true.

2.3 ROC CURVE:

An ROC curve shows the relationship between sensitivity and specificity for every possible cut-off. The ROC curve is a graph with:

The x-axis showing 1 – specificity (= false positive fraction = FP/(FP+TN))

The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN))

Thus every point on the ROC curve represents a chosen cut-off even though you cannot see this cut-off. What you can see is the true positive fraction and the false positive fraction that you will get when you choose this cut-off.

2.4 LIBRARIES USED:

- **Pandas:** It provides fast, expressive, and flexible data structures to easily (and intuitively) work with structured (tabular, multidimensional, potentially heterogeneous)

- **Numpy:** It has advanced math functions and a rudimentary scientific computing package. Numpy is a popular array – processing package of Python. It provides good support for different dimensional array objects as well as for matrices.

- **Matplotlib:** Matplotlib helps with data analyzing, and is a numerical plotting library. Matplotlib can create such quality figures that are really good for publication. Figures you create with Matplotlib are available in hardcopy formats across different interactive platforms.

- **Seaborn:** It provides a high-level interface for drawing attractive and informative statistical graphics.

- **Sk-learn:** It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

- **Scipy:** The SciPy library, a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics, and much more. Matplotlib, a mature and popular plotting package that provides publication-quality 2-D plotting, as well as rudimentary 3-D plotting.

SOFTWARE REQUIREMENT SPECIFICATIONS

3.1 Assumptions And Dependences:

3.1.1 Assumptions:

i. The end user device should be a laptop.

ii. Additionally, the end user has an active internet connection in his/her laptop.

3.1.2 Dependencies:

i. The system browser is dependent on the end user device.

ii. The prediction and analysis purpose are dependent on the types of algorithms used.

3.2 Performance Requirements:

i. Accuracy: The system can predict with varying accuracy between 50 to 60% using one of the Algorithms c which gives maximum accuracy right now, but later on, as the number of responses will increase the accurarcy will also increase.

ii. Privacy: Data will be totally secured and will not be leak as no personal details are asked.

3.3 System Requirements:

3.3.1 Database Requirement:

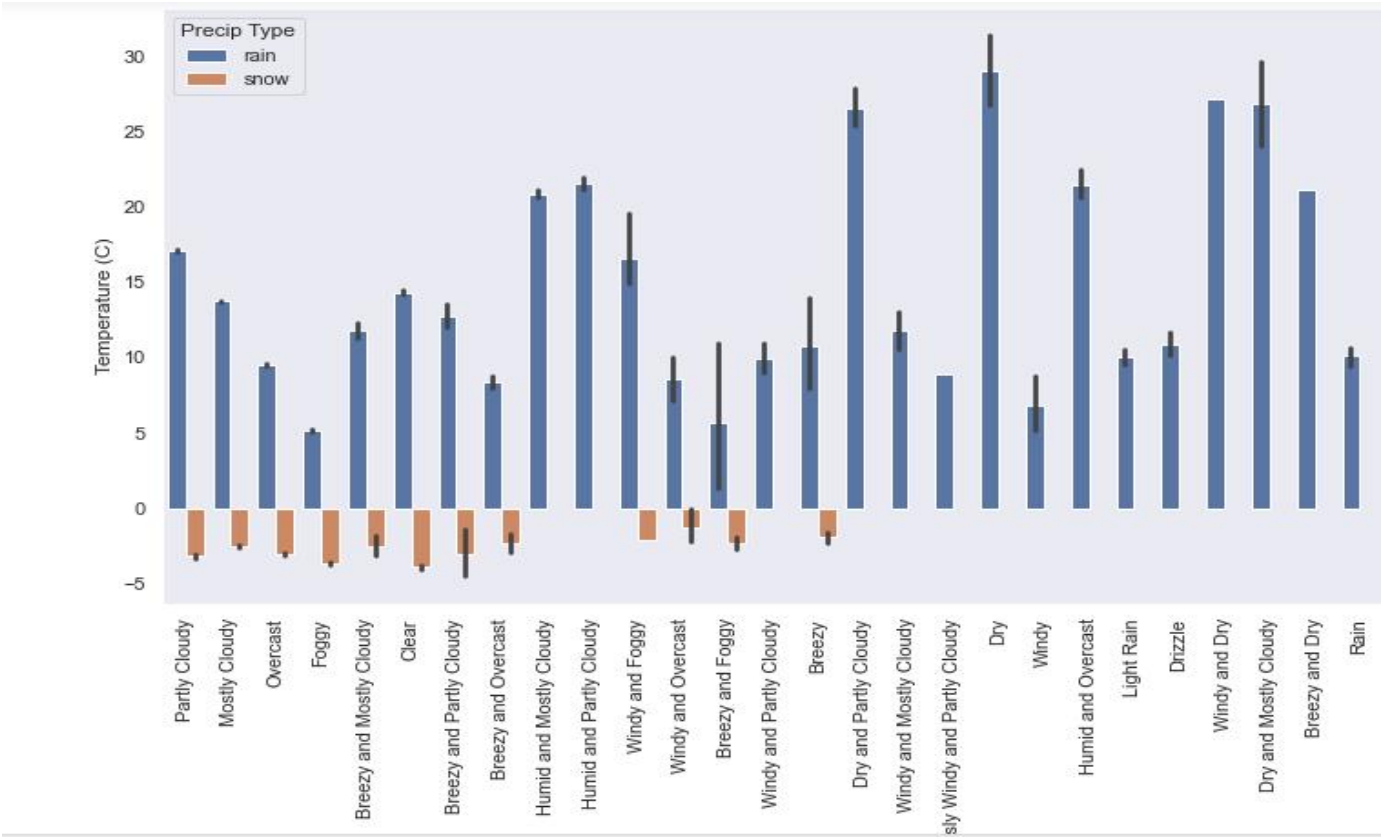
- Ms-Excel

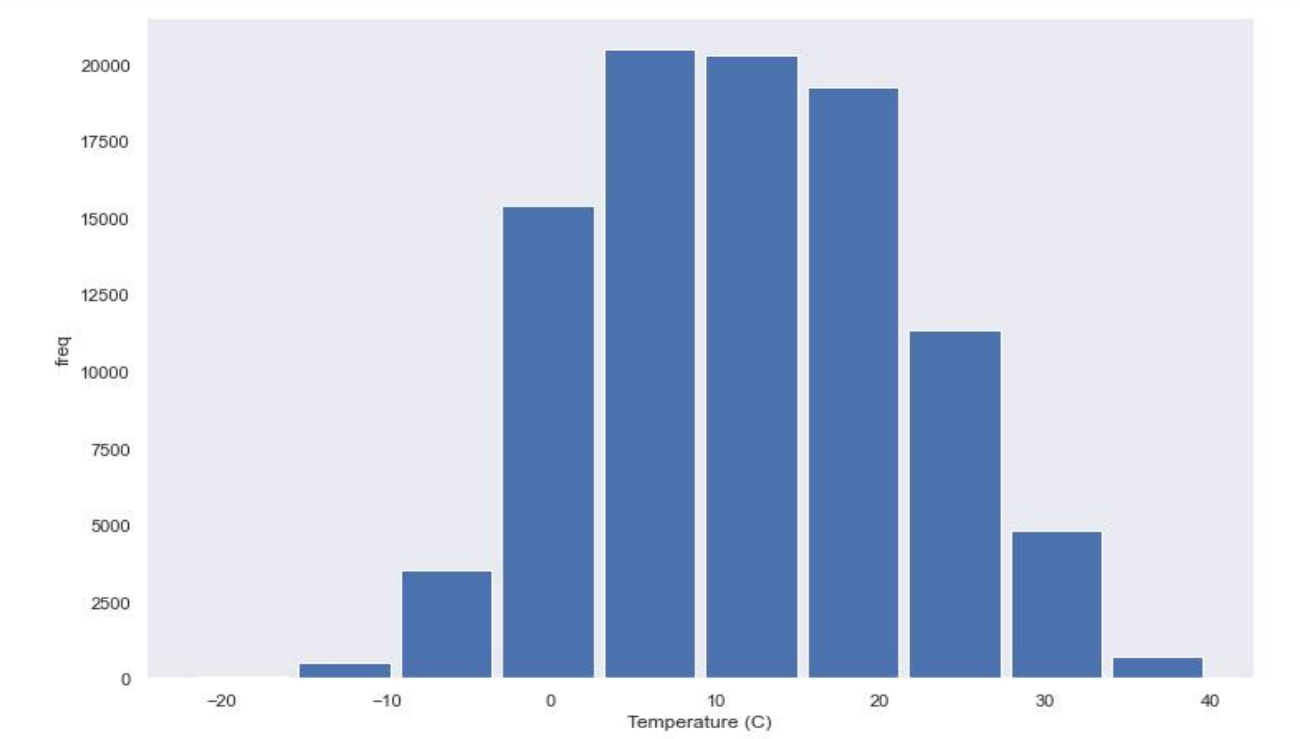
3.3.2 Software Requirement:

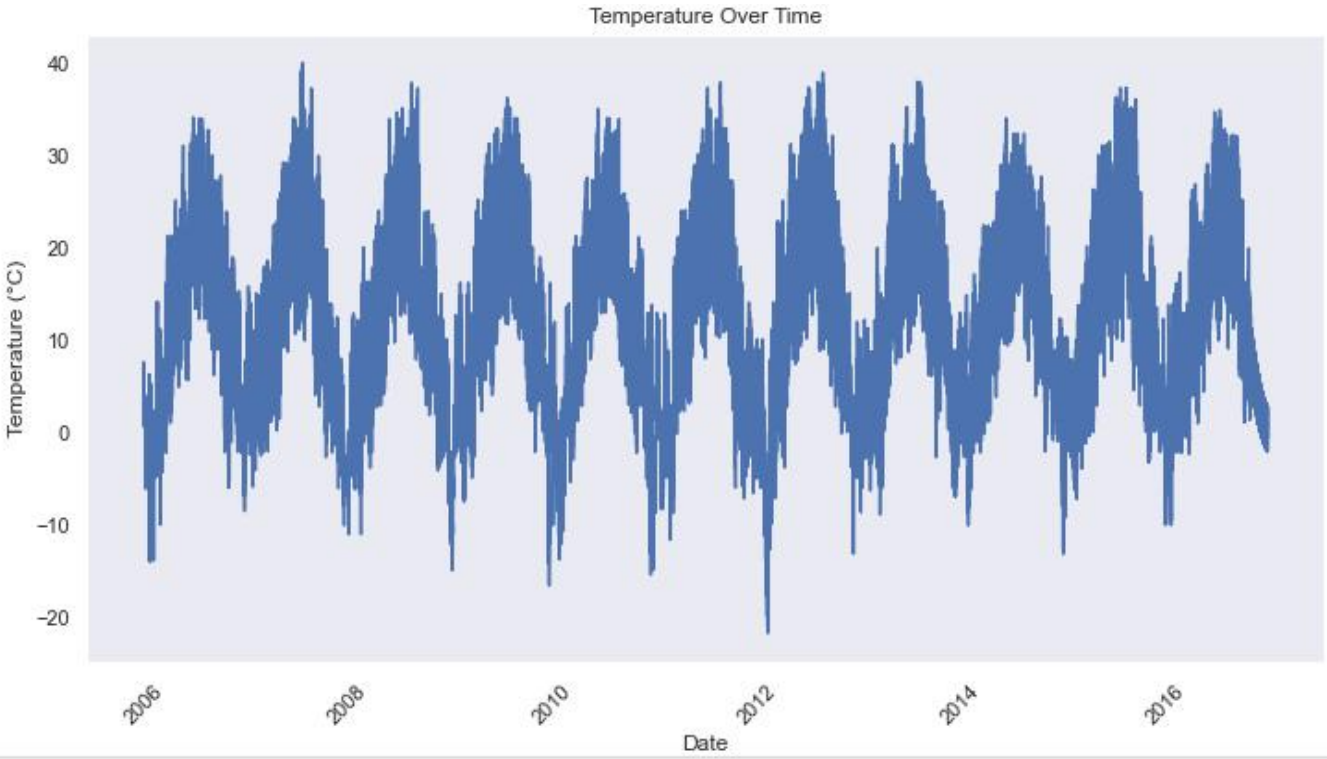
- Jupyter Notebook
- Os Windows11
- Programming Language- Python.

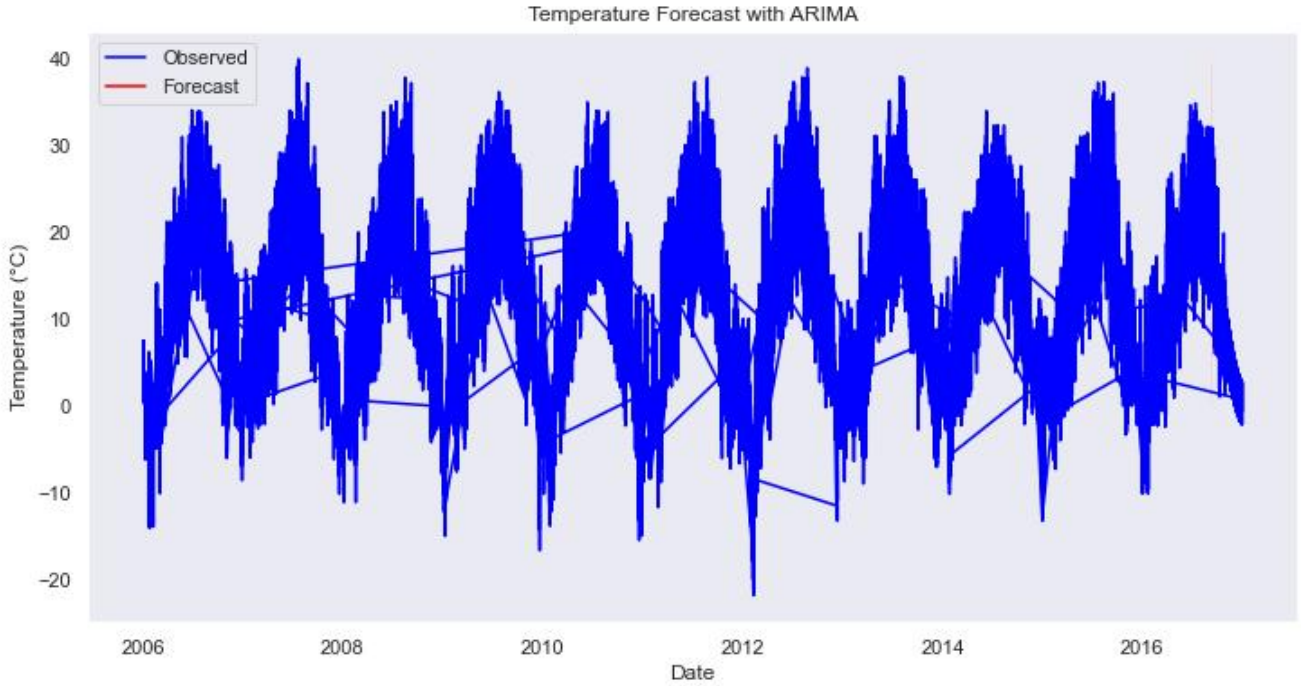
3.3.3 Hardware Requirement:

- Any Device With Brower Support.


Precip Type
rain
snow
Temperature (C)
30
25
20
15
10
5
0
−5
Partly Cloudy
Mostly Cloudy
Overcast
Foggy
Breezy and Mostly Cloudy
Clear
Breezy and Partly Cloudy
Breezy and Overcast
Humid and Mostly Cloudy
Humid and Partly Cloudy
Windy and Foggy
Windy and Overcast
Breezy and Foggy
Windy and Partly Cloudy
Breezy
Dry and Partly Cloudy
Windy and Mostly Cloudy
sty Windy and Partly Cloudy
Dry
Windy
Humid and Overcast
Light Rain
Drizzle
Windy and Dry
Dry and Mostly Cloudy
Breezy and Dry
Rain


freq
20000
17500
15000
12500
10000
7500
5000
2500
0
−20
−10
0
10
20
30
40
Temperature (C)


Temperature Over Time
Temperature (°C)
40
30
20
10
0
−10
−20
2006
2008
2010
2012
2014
2016
Date


Temperature Forecast with ARIMA
Observed
Forecast
Temperature (°C)
40
30
20
10
0
−10
−20
2006
2008
2010
2012
2014
2016
Date

Hypothesis Testing

We performed Chi Square hypothesis testing, Z testing, T testing and ANOVA on the dataset, to check relation between :

i. As Daily Summary Frequency has a negligible correlation with the other features and it is apparent that the final target is not going to be affected by this, removing it.

ii. Average tem. of sample is same as Average humuduty of population.

iii. Average Tem. of sample is not same as Average Tem. of population.

iv. Using Anova test if all summary having same wind speed.

v. Implement one tail Z test to check if Apparent temp of overcast summary is higher than Apparent temp.of foggy summary.

vi. The Temperature of Overcast Summary is less than or equal to the Temperature of Foggy Summary.

This will help us to understand dependant and independent variables.

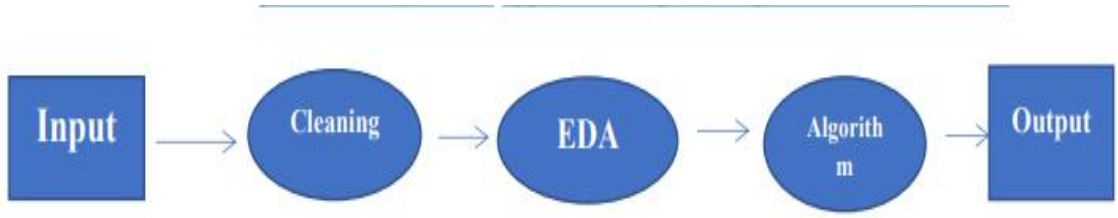
Feature Engineering

After knowing which were the independent and dependent variables, feature engineering techniques were used for X and y. it is important step as the accuruacy will be dependent on X and y variable. It was try and error method because to get the highest accuracy we needed to check the dependency.

Implementation of Algorithms

Before implementing the algorithms the data was splitted for train and test purpose. 80% of the data was for train and remaining 20% was for test. We also plotted heatmap and pairplot to get a better idea. Algorithms we used in this were Linear Regression, Naïve Bayes Algorithm, KNearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine. In all algorithms we plotted confusion matrix and overviewed classification report. On bases of y-test and y-pred we calculated accuracy.

In DFD level-2 show that Give dataset as Input, data cleaning and EDA perform, Algorithm apply and display output in accuracy result.



Comparision of Algorithms via Accuracy

After implementing all the algorithms we created a dataframe of all accuracy with respect to the algorithms. So that we can clearly overlook which has highest accuracy. And the highest accuracy is 94% and the algorithm is Random Forest.

Algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.805	0.802	0.805	0.802
Gaussian Naive Bayes	0.756	0.763	0.756	0.741
Decision Tree	0.894	0.894	0.894	0.894
KNN	0.898	0.898	0.898	0.898
Random Forest	0.939	0.939	0.939	0.938

After implementing all time series models we create a dataframe of all forecasted with StandardError and Confidance Interval value of the our data variable.

Model	Apparent Temperature (C)	Humidity	Pressure (millibars)	Visibility (km)
Forecast Value	20.00954018	0.620682098	1015.13117	15.49117468
stderr	19.71284809	0.627859618	1015.23439	15.46932565
conf_int	19.50782547	0.63268234	1015.224034	15.45299119

Advantages

i. Weather history data is essential for studying long-term climate patterns and trends. By analyzing historical weather data, scientists can better understand.
ii. Examining past weather records allows researchers to identify and analyze extreme weather events such as hurricanes, droughts, heatwaves, and floods. This information can be used to improve disaster preparedness and response strategies..
iii. Understanding past weather patterns helps optimize crop yields and reduce risks associated with adverse weather conditions.
iv. By understanding past weather-related losses, insurers can more accurately predict and manage their exposure to weather-related claims..

Limitations

v. Limited historical weather data: Historical weather data might not be available for all locations and time periods, especially in remote or less-monitored regions.
vi. Data accuracy: Historical weather data can contain errors or inaccuracies, which can affect the reliability of your analysis.
vii. Statistical limitations: Drawing meaningful conclusions from historical data may require advanced statistical methods, and results can be influenced by the chosen analysis techniques.
viii. Causation vs. correlation: Establishing causation between weather variables and other phenomena can be challenging and may require additional data and expertise.

Application

Climate Research and Analysis:
Studying historical weather data to analyze long-term climate trends.
Identifying patterns and changes in temperature, precipitation, and other weather parameters over time.
Weather Forecasting Improvement:
Using historical weather data to improve forecasting models and accuracy. Identifying past weather events that are similar to current conditions for more accurate short-term predictions.

Future scope

ix. Weather history data can be instrumental in studying long-term climate trends.
x. Weather history can aid in disaster preparedness and response. Governments and organizations can use historical weather data to assess the frequency and severity of natural disasters like hurricanes, floods, and wildfires, enabling them to better plan for and mitigate the impact of such events.
xi. Conservationists and environmental organizations can use weather history data to study the impact of climate change on ecosystems and wildlife. This information can guide conservation efforts and habitat restoration projects.

Conclusion

xii. Analyzing the historical weather data to identify trends, patterns, and anomalies.

xiii. Generating statistical summaries and visualizations to illustrate historical weather conditions.

xiv. We build all Supervised Machine learning algorithms such as KNN Classifier, Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, and Multinomial Logistic regression and checked their accuracy and we get the best-supervised algorithm model i.e. Random Forest Classifier and its accuracy is 93.9%.