***A PROJECT ON***

# “PUNE TEMPERATURE PREDICTION”

SUBMITTED IN

PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF

DIPLOMA IN BIG DATA ANALYSIS



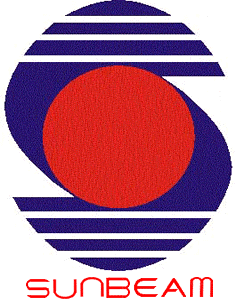
**SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE**

Submitted By:

Aditya Dwivedi (80695)

Pranjali Vanjari (80267)

**Mr.Nitin Kudale Mrs.Manisha Hingne** Centre Coordinator Course Coordinator



**CERTIFICATE**

This is to certify that the project work under the title ‘Pune Temperature Prediction’ is done by Aditya Dwivedi & Pranjali Vanjari in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P Mrs. Manisha Hingne

**Project Guide** **Course Coordinator**

Date:

# ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket P.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Aditya Dwivedi

DBDA September 2023 Batch,

SIIT Pune

Pranjali Vanjari

DBDA September 2023 Batch,

SIIT Pune

**TABLE OF CONTENTS**

1. **Introduction**
   1. Introduction And Objectives
   2. Why this problem needs To be Solved?
   3. Dataset Information

## Problem Definition and Algorithm

* 1. Problem Definition
  2. Algorithm Definition

## Experimental Evaluation

* 1. Methodology/Model
  2. Exploratory Data Analysis

## Results And Discussion

1. **GUI**
2. **GitHub link**

## 7.Future Work And Conclusion

* 1. Future Work
  2. Conclusion
     1. **Introduction**
        1. **Introduction And Objectives:**

Weather prediction is a critical task that influences various sectors, including agriculture, transportation, tourism, and disaster management. Accurate weather forecasting can help individuals and organizations make informed decisions, plan activities, and mitigate risks associated with adverse weather conditions.

In this project, we aim to develop a machine learning model to predict weather patterns in Pune, a major city in India. Pune experiences a diverse range of weather phenomena throughout the year, including monsoon rains, hot summers, and cool winters. By leveraging historical weather data and advanced machine learning techniques, we seek to build a model that can forecast various weather parameters such as temperature, humidity, precipitation, and wind speed.

## Why this problem needs To be Solved?

The problem of weather prediction in Pune, and in general, is significant for several reasons:

1. Risk Management: Accurate weather forecasting helps individuals, businesses, and governments mitigate risks associated with adverse weather conditions. It allows farmers to plan crop cultivation and irrigation, helps transportation companies optimize routes and schedules, and enables disaster management agencies to prepare for extreme weather events such as floods, storms, and heatwaves.
2. Resource Allocation: Effective weather prediction aids in the efficient allocation of resources. For example, knowing when and where precipitation is likely to occur can help water resource managers manage reservoirs and water distribution systems more effectively. Similarly, energy companies can optimize electricity generation from renewable sources like solar and wind by forecasting weather patterns accurately.
3. Public Safety: Timely and accurate weather forecasts are crucial for public safety. They enable authorities to issue warnings and advisories to the public, reducing the risk of weather-related accidents and injuries. For instance, advance notice of severe weather events such as hurricanes, tornadoes, or heavy rainfall allows people to take necessary precautions and evacuate vulnerable areas if needed.
4. Economic Impact: Weather affects various sectors of the economy, including agriculture, tourism, construction, and retail. Accurate weather forecasting helps businesses make informed decisions, reduce operational disruptions, and minimize economic losses caused by weather-related events. For instance, retailers can adjust inventory levels based on expected weather conditions to meet consumer demand more effectively.
5. Environmental Sustainability: Weather prediction is essential for environmental monitoring and conservation efforts. It enables scientists to study climate change trends, track the spread of pollutants and contaminants, and assess the impact of human activities on ecosystems. By accurately predicting weather patterns, we can better understand and address environmental challenges such as air and water pollution, habitat destruction, and biodiversity loss.

## Dataset Information.

## Pune.csv

It has 23 columns.

RangeIndex: 45503 entries, 0 to 45502

dtypes: float64(10), int64(8), object(5)

## Problem Definition and Algorithm:

* + - 1. **Problem Definition**

The problem is to predict the temperature in Pune, India, based on historical weather data. It's a regression task, aiming to forecast a continuous value. Data includes historical temperature readings along with other relevant features like humidity, wind speed, and weather conditions. Common regression algorithms like Linear Regression, Random Forest, or Gradient Boosting can be used.. Evaluation is typically done using metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE). Preprocessing involves cleaning data, feature engineering, and scaling. After training the model, it's evaluated on a test set, and if satisfactory, it's deployed for real-time predictions, with ongoing monitoring and maintenance to ensure performance.

## Algorithm Definition:

**Linear regression:** is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

**Ridge Regression:** Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

**Lasso Regression:** Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.Lasso Regression uses L1 regularization technique,It is used when we have more number of features because it automatically performs feature selection.

**Random forest:** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**Decision Tree:** algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**XGBoost:** or extreme gradient boosting is one of the well-known [gradient](https://analyticsindiamag.com/gradient-descent-everything-you-need-to-know-with-implementation-in-python/) [boosting](https://analyticsindiamag.com/gradient-descent-everything-you-need-to-know-with-implementation-in-python/) techniques (ensemble) having enhanced performance and speed in tree- based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. It is open- source software. Earlier only [python and R packages](https://analyticsindiamag.com/python-vs-scala-for-apache-spark/) were built for XGBoost but now it has extended to Java, Scala, Julia and other languages as well.

## Experimental Evaluation:

* + - 1. **Methodology:**

The objective of this project is to predict the temperature of Pune city on the basis of historic data. The data set is contained from Kaggle and has 1 csv file having previous years data.

**Loading in raw data:**

df = pd.read\_csv("pune.csv")

df.head()

**Preprocessing:**

Deleting unnecessary columns:

df.drop(['date\_time', 'moonrise', 'moonset', 'sunrise', 'sunset'], axis = 1, inplace = True)

Finding The Missing Data (NaN Values):

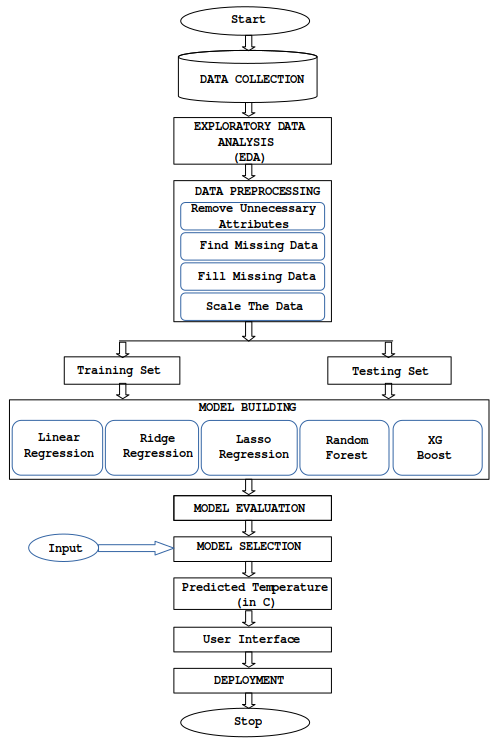
df.isna().sum()

### Filling The Missing Data

Filling the missing data:

filled\_df = df.fillna(df.mean())

## Flow Diagram :

****

**3.2 Exploratory Data Analysis**

The comparison of tempC dependent variable with other independent variables such as maxtempC, mintempC, pressure, humidity, HeatIndexC, uvIndex from the exploration point of view:

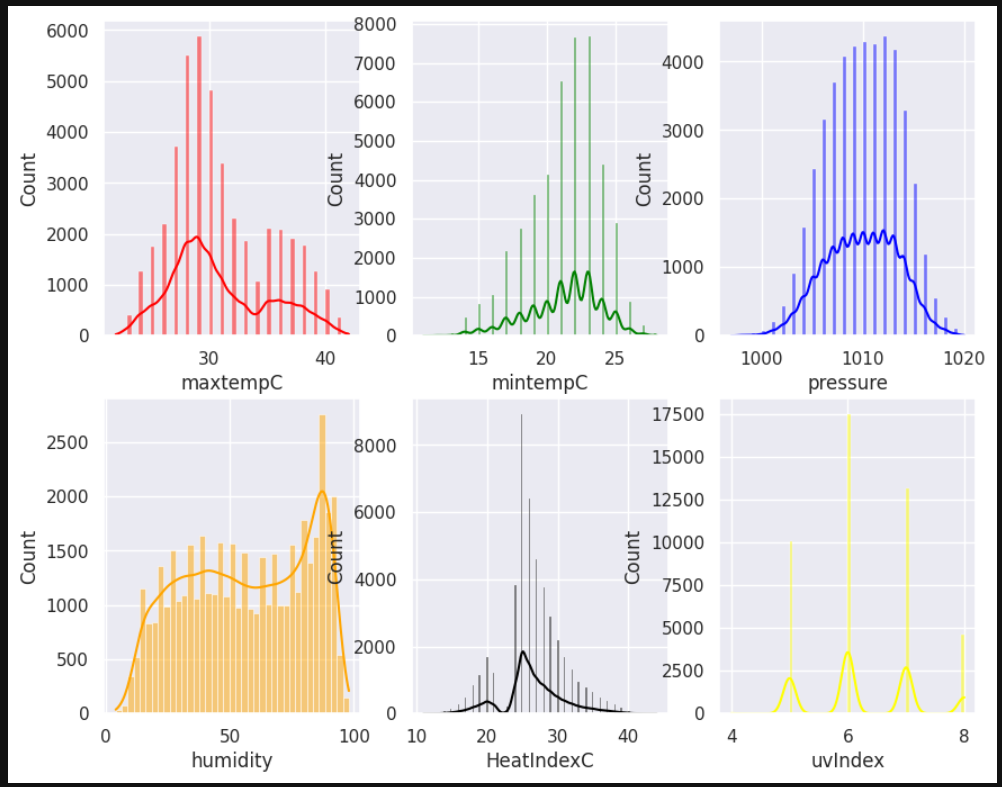


Fig 2: other independent vaiables vs tempC

Analysis shows the variation of different independent variables across different years with the tempC column.

## Results and discussion:

Linear regression, Lasso regression, ridge regression, random forest, decision tree and gradient boosting machine algorithm were used to predict the

Pune temperature. Among the given algorithms, Linear Regression algorithm was the best performing one as it provided the highest R2 score of 0.95.

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(x\_train, y\_train)

y\_pred = model.predict(x\_test)

from sklearn.metrics import mean\_absolute\_error

mae=mean\_absolute\_error(y\_test,y\_pred)

from sklearn.metrics import r2\_score

R2=r2\_score(y\_test,y\_pred)

MAE: 0.73

R2 Score: 0.95

## GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

**6.GitHubLink:** https://github.com/B2-80267/DBDA\_Project

## 7.Future work And Conclusion:

## 7.1 Future Work:

1. Integrate high-resolution data from various sources.
2. Experiment with ensemble learning and hyperparameter tuning.
3. Further explore feature engineering and deep learning architectures.
4. Develop methods for uncertainty estimation and localized predictions.
5. Integrate with decision support systems for urban planning and disaster management.

## 7.2 Conclusion:

In conclusion, the Pune temperature prediction project has successfully developed a machine learning model capable of forecasting temperatures with reasonable accuracy based on historical weather data. Through data preprocessing, feature selection, and model training, the project has demonstrated the potential to provide valuable insights into future temperature trends in Pune.

While the current model shows promising results, there are opportunities for further enhancement and refinement. Future work may involve integrating additional data sources, exploring advanced modeling techniques, and tailoring predictions to specific localities within Pune. By continuing to improve prediction accuracy and reliability, the project aims to support various stakeholders in making informed decisions related to climate-sensitive sectors such as agriculture, tourism, and infrastructure development.

Overall, the Pune temperature prediction project lays the groundwork for ongoing research and collaboration in the field of climate science and machine learning, contributing to efforts aimed at addressing the challenges posed by climate variability and change in the region.