Mini Project Presentation

CLIMATE CHANGE ANALYSIS USING MACHINE LEARNING MODELS

SUBMITTED BY:-

ALLEN VARGHESE PAUL (201CV103)
RAKSHITH SAJJAN (201CV142)
SHANNON BRITNEY CARLO (201CV249)

PROJECT GUIDE:-

DR. SUBRAHMANYA KUNDAPURA
FACULTY OF ENGINEERING
NITK SURATHKAL

DEPARTMENT OF WATER RESOURCES AND OCEAN ENGINEERING

NITK SURATHKAL- 575 025

TABLE OF CONTENTS

Introduction

- **Climate change** refers to long-term shifts in temperatures and weather patterns.
- Such shifts can be **natural**, due to changes in the **sun's activity** or **large volcanic eruptions**. But in recent days human activities have been the main driver of climate change, primarily due to the burning of fossil fuels like coal, oil and gas.
- The main greenhouse gases that are causing climate change include **carbon dioxide** and **methane.**
- Energy, industry, transport, buildings, agriculture and land use are among the **main sectors** causing greenhouse gases.

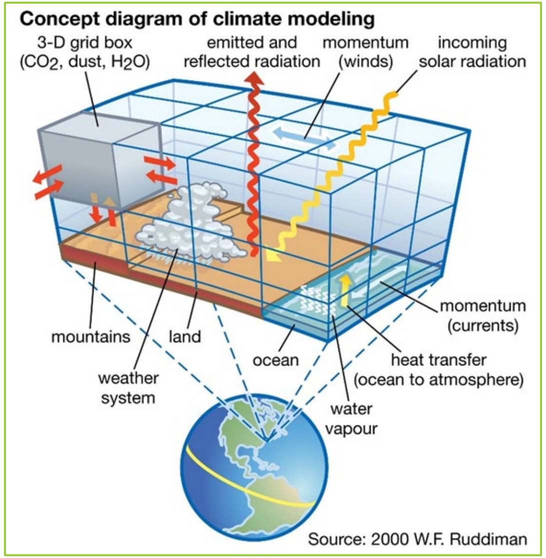
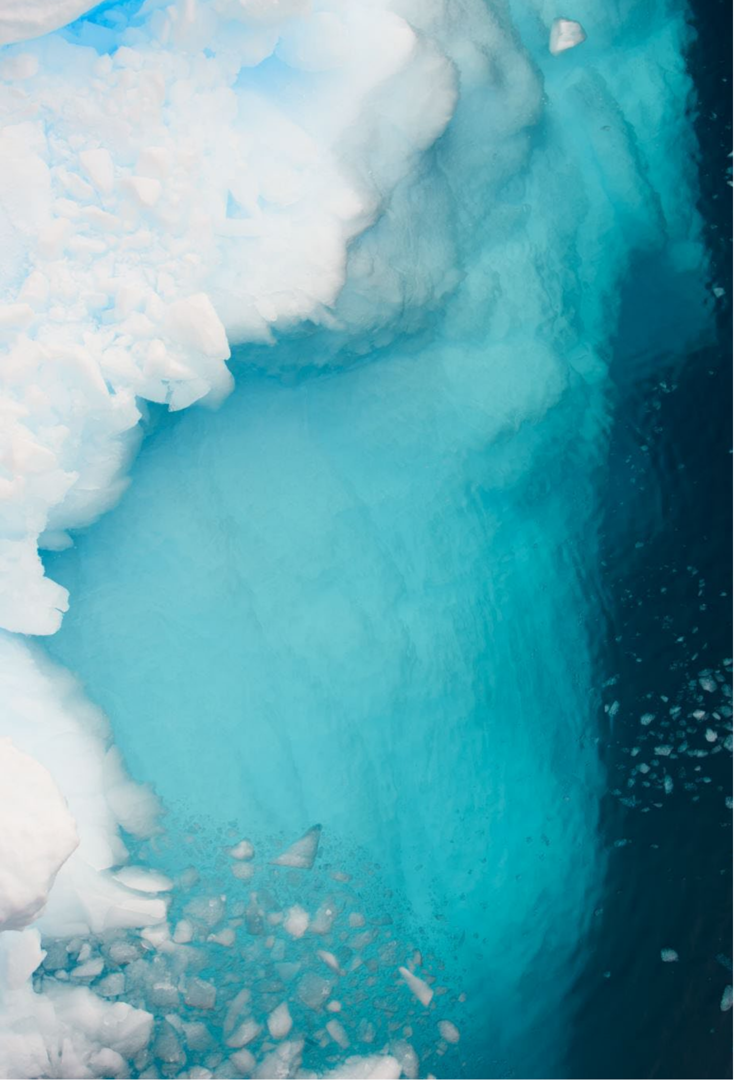


Fig. Concept diagram of climate modelling



Need to learn about climate change

- Climate change endangers animals. Due to Rising global temperatures, changing weather patterns, and human development can all have an immensely destructive impact on wildlife.

- Climate change threatens important ecosystems. Climate change also has an increasingly ruinous impact on important ecosystems and biomes, including rainforests, coral reefs, and polar ice caps.

- Climate change damages human health and infrastructure. Climate change has the potential to increase the frequency of natural disasters, and damage human health through food shortages, increased temperatures and pollution.

- Climate change will affect future generations. Global warming, air pollution, acidification of the ocean, destructive weather, and human activities will also have an increasingly harmful impact on future generations.

Factors Affecting Climate Change Analysis

- Pressure, wind speed, and humidity are the three main factors considered for climate change analysis.

- Studies reveal that specific humidity has also risen over the oceans, which is to be expected given that both the oceans and the air above them have warmed, allowing for greater water evaporation and gaseous retention in the atmosphere. The frequency and severity of heat waves are increasing along with the temperature and humidity levels.

- Excessive heating in tropical latitudes results in increased pressure at upper levels in the tropics as thunderstorms transport air to higher levels. Additionally, the increased heating/cooling contrast during the winter results in stronger pressure changes.

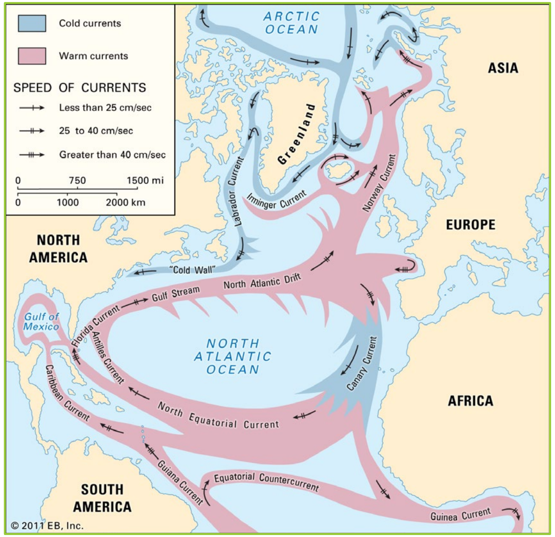


Fig. Global wind currents

Greenhouse effect

Heat from the host star passes through the atmosphere and warms the planet's surface, known as the greenhouse effect. Greenhouse gases function by absorbing wavelengths of radiation emitted by planets like the Earth while being transparent to wavelengths emitted by stars like the sun. Because matter radiates energy at a wavelength correlated with its temperature, the wavelengths are different.

The greenhouse effect is named after a greenhouse analogy. While the heat from sunlight is retained in both greenhouses and the greenhouse effect, the mechanics are different. Although their panels also restrict heat radiation and conduction, greenhouses typically retain heat by impeding airThe four main greenhouse gases are, in order of their percentage contribution to the total greenhouse effect on Earth: Water vapor (H2O), Carbon dioxide (CO2), Methane (CH4) & Tropospheric ozone (O3).

Impact of Climate Change on the Environment

- Climate change has a huge impact on the environment and plays a major role in maintaining balance and order. Oceans, ice, and weather are only a few of the environmental repercussions of climate change that are widespread and far-reaching.

- Droughts and heat waves have increasingly occurred at the same time since the 1950s. In India and East Asia, the frequency of extremely wet or dry episodes has increased throughout the monsoon season.

- Extreme weather causes harm and fatalities, and crop failures result in undernourishment. Warmer climates make it simpler to spread infectious diseases like dengue fever and malaria.

- Many terrestrial and freshwater species are moving poleward and upward because of recent warming. Global greening has been caused by increased atmospheric CO2 levels and a prolonged growing season.

Artificial Intelligence

- AI refers to the development of computer systems able to perform tasks that require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

- AI researchers have adapted and integrated a wide range of problem-solving techniques – including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, probability and economics.

- Some uses of AI include generative or creative tools (ChatGPT and AI art), automated decision-making, and competing at the highest level in strategic game systems.

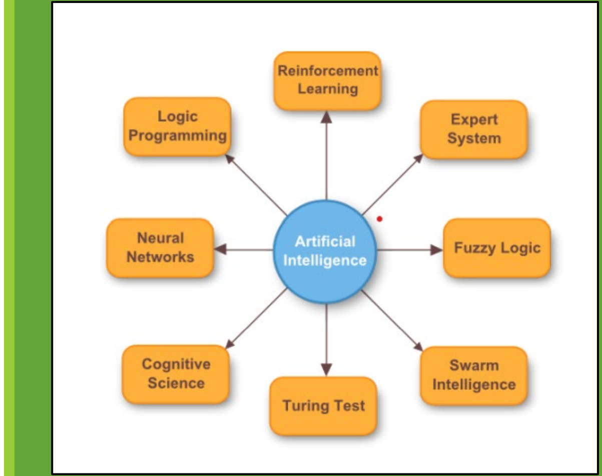


Fig. Artificial Intelligence

Machine Learning

- Machine learning is a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

- Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

- Machine learning (ML), reorganized as a separate field, started to flourish in the 1990s. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature.

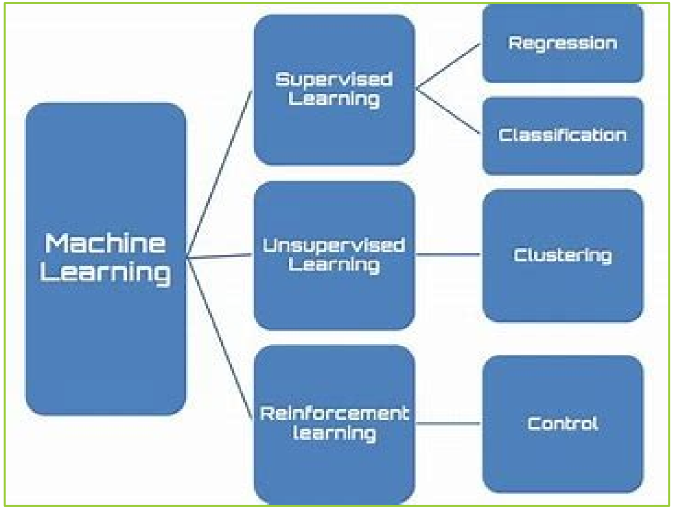
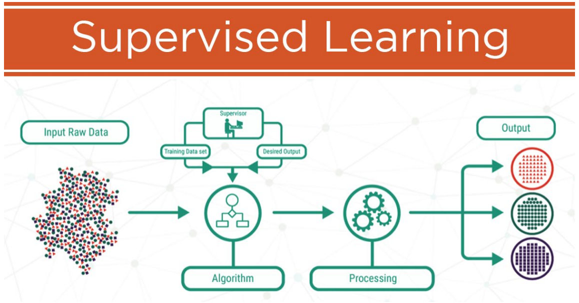


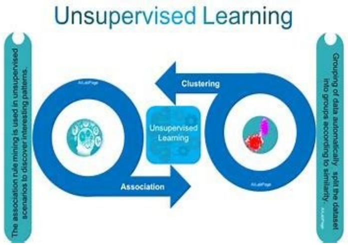
Fig. Machine Learning

Types of Machine learning

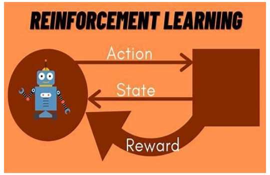
Supervised Learning - Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data and consists of a set of training examples. Types of supervised-learning algorithms include some major algorithms like active learning, classification and regression.



Unsupervised learning - Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labelled, classified, or categorized.


Unsupervised Learning
The association rule mining is used in unsupervised scenarios to discover interesting patterns.
Clustering
Unsupervised Learning
Association
Grouping of data automatically split the dataset into groups according to similarity.

Reinforcement learning - Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment to maximise some notion of cumulative reward.



Literature review

Using data from 1901 to 2015 across India at the meteorological divisional level, Praveen B. and Talukdar S.'s work evaluates and predicts long-term spatiotemporal changes in rainfall. The Mann-Kendall (MK) test and Sen's Innovative trend analysis were used to assess the rainfall trend, while the Pettitt test was used to identify the abrupt change point in time. The ANN-MLP (Artificial Neural Network-Multilayer Perceptron) was used to predict India's rainfall for the next 15 years.

“Monthly prediction of air temperature in Australia and New Zealand with machine learning models” by S. Salcedo-Sanz, R. C. Deo, L. Carro-Calvo, and B. Saavedra Moreno is also a prediction-based article. In this idea, the temperature is the only focus. The physical factors are not focused on which are responsible for Global Warming. In this project, SVR and multilayer-perceptron methods are used.

“Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes” by F. Mekanik, M.A. Imteaz, S. Gato-Trinidad, A. Elmahdi is also another prediction-based idea. In this project regression and artificial neural networks are used to predict the rainfall. This idea is focussing on only the rainfall. It is not focussing on temperature or greenhouse gases.

“Development and Analysis of ANN Models for Rainfall Prediction by Using Time-Series Data” by Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, and AK Upadhyay is also used as a reference. In this project regression, mean square error and MRE are used. This idea also focused only on rainfall but not on temperature or greenhouse gases.

“Application of Artificial Neural Networks to Rainfall Forecasting in Queensland, Australia” by John Abbot and Jennifer Marohasy has also been taken for reference. In this project, Artificial Neural Networks are used to observe and forecast rainfall. This idea also doesn’t give any explanation for global warming.

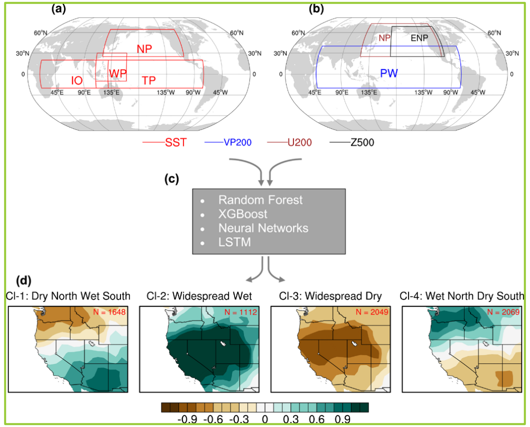


Fig. Analysis of climate change

Data and Study Area

The data used consists of the various parameters which affect the temperature of a specific region. The data is taken from the years 1982 – 2022 daily. The various climate changes in Surathkal are noted and taken into consideration while recording the data. The table below shows the data as well as the source of the data collected.

Surathkal is one of the major localities in the northern part of Mangalore city located on NH-66 in the Dakshina Kannada district, Karnataka. Surathkal is located at 12°58'60 N 74° 46' 60E. The maximum and minimum temperature in a year varies between 37 °C and 25 °C. But ambient temperature occasionally touches 40 °C during the summer season (usually March, April, and May) recorded in the 21st century. It has an average elevation of 22m above mean sea level. It receives about 95% of its total annual rainfall between May to September but remains extremely dry from December to March. Humidity is approximately 75% on average and peaks during June, July and August.

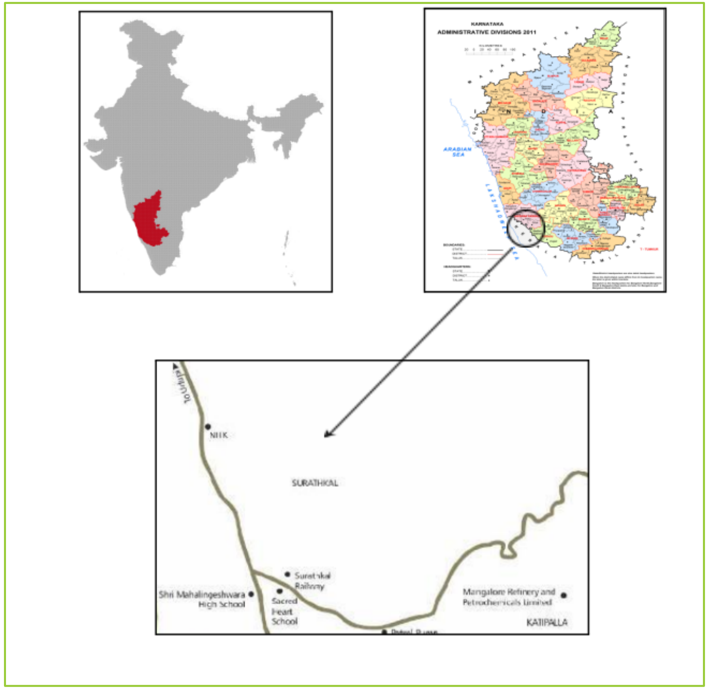

KARNATAKA
ADMINISTRATIVE DIVISIONS 2011
ARABIAN SEA
To Udupi
NITK
SURATHKAL
Surathkal Railway
Shri Mahalingeshwara High School
Sacred Heart School
Petrol Bunk
Mangalore Refinery and Petrochemicals Limited
KATIPALLA

Fig. Study Area

Methodology

Data cleansing, model construction, and model testing comprise the traditional methods for generating the model. By doing this, the process efficiency will be improved, and less time and other resources will be used in the process.

The suggested model, which uses machine learning to forecast rainfall using the parameters for climate change acquired in the dataset

1. Data pre-processing and Exploratory Data Analysis

2. Model training and validation

3. Model performance evaluation

4. Selection of best model and result analysis

Data Pre-Processing

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn.

The steps in data pre-processing are:

1. Getting the dataset
2. Importing libraries
3. Importing datasets
4. Finding Missing Data
5. Encoding Categorical Data
6. Feature scaling

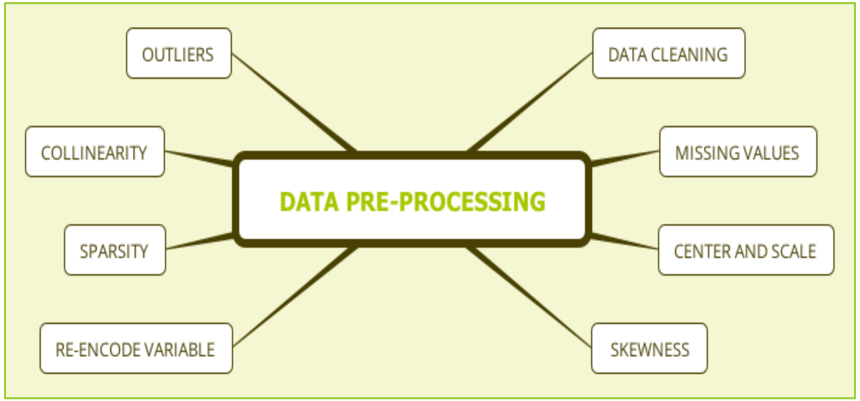


Fig. Data pre-processing

Model Training & Validation

Model training is at the heart of the data science development lifecycle where the data science team works to fit the best weights and biases to an algorithm to minimise the loss function over the prediction range.

The steps in model training are:

1. Splitting the dataset

2. Selecting Algorithms to test

3. Hyperparameter tuning

4. Fit and tune models

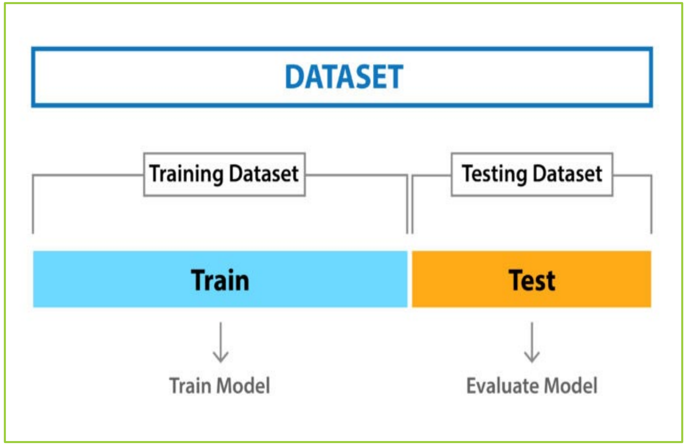


Fig. Model training

Model training produces a mathematical representation of the relationship between the data features and a target label when supervised learning is utilised. It develops a mathematical representation among the data features itself in unsupervised learning. The first phase in machine learning is model training, which produces a functional model that can subsequently be tested, validated, and deployed.

The different models used

- Linear regression - Here try to predict one output variable using one or more input variables. The representation of linear regression is a linear equation, which combines a set of input values (x) and predicted output(y) for the set of those input values. It is represented in the form of a line.

- Decision Tree - employs a tree-like structure to organise decisions and the potential outcomes and repercussions of those actions. Each internal node in this diagram represents a test on an attribute, and each branch is the result of the test. A decision tree's outcome will be more accurate the more nodes it contains.

- Random Forest - Random Forest is the ensemble learning method, which consists of many decision trees. Each decision tree in a random forest predicts an outcome, and the prediction with the most votes is considered as the outcome.

- SVM - An SVM outputs a map of the sorted data with the margins between the two as far apart as possible. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences.

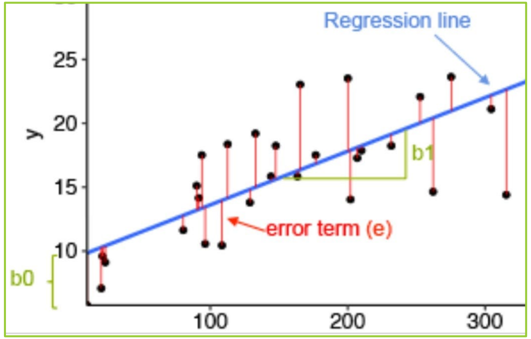


Fig. Linear regression

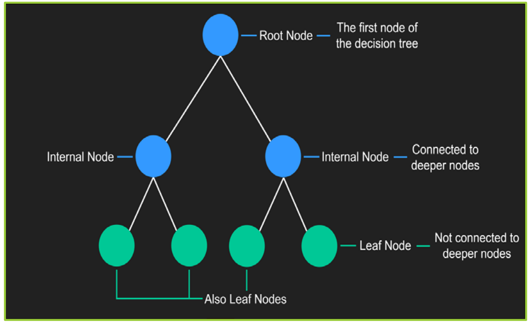

Root Node
The first node of the decision tree
Internal Node
Internal Node
Connected to deeper nodes
Leaf Node
Not connected to deeper nodes
Also Leaf Nodes

Fig. Decision Tree

- KNN – The k-nearest neighbours' algorithm, also known as KNN is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. The average of the k nearest neighbours is taken to make a prediction with continuous values.

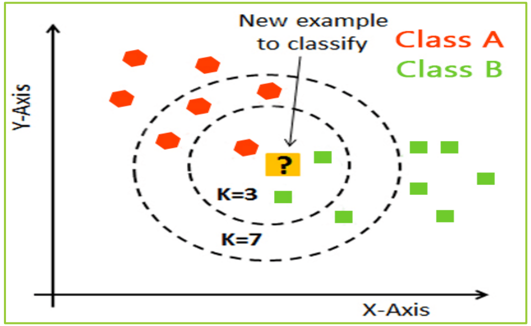


Fig. KNN

- Gradient Boosting Regressor – It is a technique used in creating models for prediction. Gradient boosting presents model building in stages, just like other boosting methods, while allowing the generalisation and optimization of differentiable loss functions.

- ADA Boosting Regressor – It makes 'n' a number of decision trees during the data training period. As the first decision tree/model is made, the incorrectly classified record in the first model is given priority. Only these records are sent as input for the second model. The process goes on until we specify several base learners we want to create.

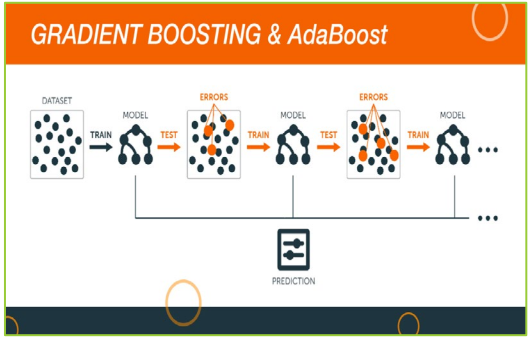


Fig. Gradient Boosting Regressor

Model Performance & Evaluation

Accuracy - Accuracy is, simply put, the total proportion of observations that have been correctly predicted. TP represents the number of True Positives. This refers to the total number of observations that belong to the positive class and have been predicted correctly. TN represents the number of True Negatives. This is the total number of observations that belong to the negative class and have been predicted correctly. FP is the number of False Positives. FN is the number of False Negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Root Mean Squared Error (RMSE) - The Root-Mean-Square Error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model and the values observed.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

Selection of Best Model and Result Analysis

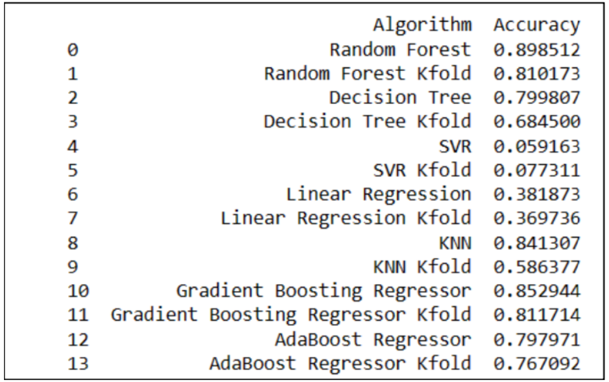
- Model selection is the process of choosing one among many candidate models for a predictive modelling problem.

- There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources.

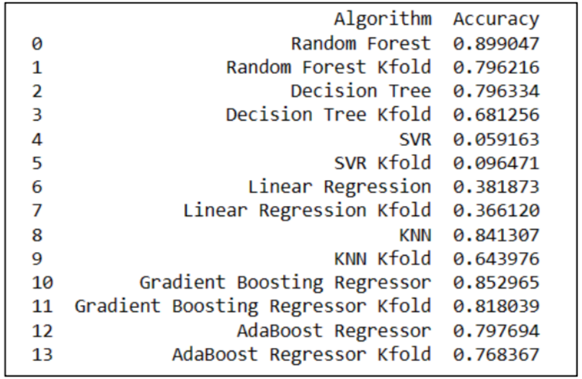
- Model selection is a process that can be applied both across different types of models (e.g., logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters.

- Given the statistical noise in the data, the incompleteness of the data sample, and the restrictions of each unique model type, all models have some predictive inaccuracy. With the parameters provided in the dataset, we may predict climate change using the best-trained model with the chosen hyperparameters and significant variables.

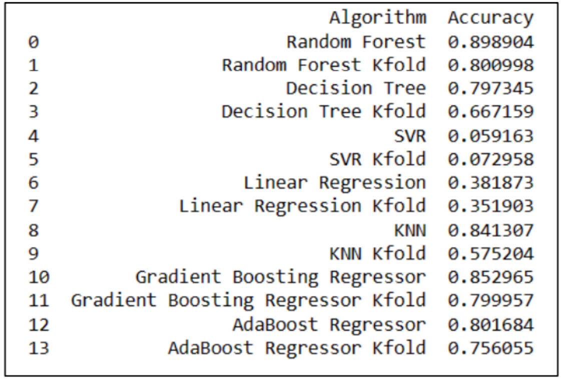
Result



```
                            Algorithm  Accuracy
0                       Random Forest  0.898512
1                 Random Forest Kfold  0.810173
2                       Decision Tree  0.799807
3                 Decision Tree Kfold  0.684500
4                                 SVR  0.059163
5                           SVR Kfold  0.077311
6                   Linear Regression  0.381873
7             Linear Regression Kfold  0.369736
8                                 KNN  0.841307
9                           KNN Kfold  0.586377
10        Gradient Boosting Regressor  0.852944
11  Gradient Boosting Regressor Kfold  0.811714
12                 AdaBoost Regressor  0.797971
13           AdaBoost Regressor Kfold  0.767092
```



```
                            Algorithm  Accuracy
0                       Random Forest  0.899047
1                 Random Forest Kfold  0.796216
2                       Decision Tree  0.796334
3                 Decision Tree Kfold  0.681256
4                                 SVR  0.059163
5                           SVR Kfold  0.096471
6                   Linear Regression  0.381873
7             Linear Regression Kfold  0.366120
8                                 KNN  0.841307
9                           KNN Kfold  0.643976
10        Gradient Boosting Regressor  0.852965
11  Gradient Boosting Regressor Kfold  0.818039
12                 AdaBoost Regressor  0.797694
13           AdaBoost Regressor Kfold  0.768367
```



```
                            Algorithm  Accuracy
0                       Random Forest  0.898904
1                 Random Forest Kfold  0.800998
2                       Decision Tree  0.797345
3                 Decision Tree Kfold  0.667159
4                                 SVR  0.059163
5                           SVR Kfold  0.072958
6                   Linear Regression  0.381873
7             Linear Regression Kfold  0.351903
8                                 KNN  0.841307
9                           KNN Kfold  0.575204
10        Gradient Boosting Regressor  0.852965
11  Gradient Boosting Regressor Kfold  0.799957
12                 AdaBoost Regressor  0.801684
13           AdaBoost Regressor Kfold  0.756055
```

Fig. Accuracy table of all models

It is observed that the Random Forest, Gradient Boosting Regressor as well as the KNN model have the highest accuracies. The Random Forest model has an accuracy of around 90% in every trial. The Gradient Boosting regressor has an accuracy of around 85%. The KNN model has an average accuracy of around 84%. These can be improved by hyperparameter tuning. The other models have low accuracy due to the problem of overfitting which resulted in few models having higher accuracies. If hyperparameter tuning is done, then the accuracies of all models can be brought above 80%.

Conclusion

The primary goal of this project is to assess various Machine Learning models to predict temperature changes in the Surathkal area of Mangalore, Karnataka. The parameters used for this analysis include wind speed, humidity, and atmospheric pressure. Seven Machine Learning algorithms were employed in the project, including linear models such as Linear Regression and Support Vector Machine (SVM), non-linear models like Decision Tree and K-Nearest Neighbors (KNN), and ensemble models such as Gradient Boosting Regressor, ADA Boosting Regressor, and Random Forest. Their K-fold variations have also been taken into consideration.

The results indicated that the Random Forest model had the highest accuracy, approximately 90%, making it suitable for climate change analysis with some hyperparameter tuning. As the parameters for climate change analysis are hydrological, it is crucial to account for overfitting. The success of the Random Forest model suggests that non-linear models should be preferred for this type of analysis.

Random Forest is a versatile model that can handle binary, categorical, and numerical features, making it suitable for both regression and classification tasks. Additionally, it requires minimal pre-processing, eliminating the need for rescaling or data transformation. Overall, the project demonstrates the potential of Machine Learning models, particularly the Random Forest model, in predicting temperature changes and contributing to climate change analysis.

References

Praveen, B., Talukdar, S., Shahfahad et al. Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Sci Rep 10, 10342 (2020)

Salcedo-Sanz, S., Deo, R.C., Carro-Calvo, L. et al. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. Theor Appl Climatol 125, 13– 25 (2016)

R. Adhikari, RK Agrawal. et al Forecasting strong seasonal time series with Artificial Neural Network. Journal of Scientific and Industrial Research, vol. 71, pp. 657-666 (2012)

S. Singh, J. Gill, et al. Temporal Weather Prediction using I.J. Intelligent Systems and Applications, vol.6 (12), pp. 55-61 (2014)

B.M. Al-Maqaleh, A.A. Al-Mansoub and F.N. Al-Badani, et al Forecasting using Artificial Neural Network and Statistics Models‖, I.J. Education and Management Engineering, vol.3, pp. 20-32 (2016)

Abbot, J., Marohasy, J. et al Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv. Atmos. Sci. 29, 717–730 (2012)