Air Temperature Prediction Using Machine Learning

Made By:

NIKHIL KUMAR : 1804002

SUMIT ASHISH : 1804008

SAUMYA SINGH : 1804049

Branch : ECE -A

7th Semester

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Under the Supervision of : Dr. Bikash Chandra Sahana

Introduction

- Mitigating climate change is one of the biggest challenges of humankind.
- Efforts to understand the influence of historical climate change, at global and regional levels, have been increasing over the past decade.
- In particular, the estimates of air temperatures have been considered as a key factor in climate impact studies on agricultural, ecological, environmental, and industrial sectors.
- Accurate temperature prediction helps to safeguard life and property, playing an important role in planning activities for the government, industry, and the public.

Our Aim

- The primary aim of our group is to review the different machine learning strategies for temperature forecasting, available in the literature, presenting their advantages and disadvantages and utilizing these to predict the Air temperature with best accuracy possible.

- Machine Learning techniques helps us to accurately predict temperatures based on a set of input features, which can include the previous values of temperature, relative humidity, solar radiation, rain and wind speed measurements, among others

Applications of air temperature prediction

- Air temperature forecasting has been a crucial climatic factor required for many different applications in areas such as agriculture, industry, energy, environment, tourism, etc.
  Some of these applications include:
  - short-term load forecasting for power utilities
  - air conditioning and solar energy systems development
  - adaptive temperature control in greenhouses
  - prediction and assessment of natural hazards
  - prediction of cooling and energy consumption inresidential buildings

Why are we using machine Learning

- The main objective of the algorithms developed using Machine Learning is to obtain a mathematical model that fits the data.
- Once this model represents accurately the known data, it is used to perform the prediction using new data.
- In this way, the learning process involves two steps:

1. *The estimation of unknown parameters in the model based on a given dataset*
2. *The output prediction based on new data and parameters obtained previously*.

- In this way, ML strategies find models between inputs and outputs, even if the system dynamics and its relations are difficult to represent.
- For this reason, this approach has been widely implemented in a great variety of domains, such as pattern recognition, classification, and forecasting problems.

ABOUT THE DATASET

- The Dataset was obtained from the Meteorological Centre Patna , India Meteorological Department.
- It contained 9 attributes which included attributes like wind direction , wind speed, visibility etc.
- This data was collected from January 2014 till January 2021 i.e. for a span of 7 years

ABOUT THE ATTRIBUTES

- **Wind Direction** : Wind direction is reported by the direction from which it originates. For example, a north or northerly wind blows from the north to the south. Wind direction is usually reported in cardinal direction, or in degrees.
- **Wind Speed** : In meteorology, wind speed is a fundamental atmospheric quantity caused by air moving from high to low pressure, usually due to changes in temperature. Wind speed is now commonly measured with an anemometer.

ABOUT THE ATTRIBUTES

- **Visibility** : In meteorology, visibility is a measure of the distance at which an object or light can be clearly discerned.
- **Wet bulb Temperature**: It is the lowest temperature that can be reached under current ambient conditions by the evaporation of water only.
- **Dew point Temperature**: The dew point is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form.

ABOUT THE ATTRIBUTES

- **Total cloud cover**: Total cloud cover is the fraction of the sky covered by all the visible clouds.
- **Wet bulb Temperature**: It is the lowest temperature that can be reached under current ambient conditions by the evaporation of water only.
- **Dew point Temperature**: The dew point is the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%. At this point the air cannot hold more water in the gas form.

MACHINE LEARNING ALGORITHMS USED FOR MAKING THE PREDICTIONS

1. Linear Regression
2. K-Nearest Neighbors Algorithm
3. Support Vector Machines (SVMs)
4. Decision Trees
5. Extreme Gradient Boosting (XGBoost)
6. Random Forest Regressor

LINEAR REGRESSION

- **Linear Regression** is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).
- So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

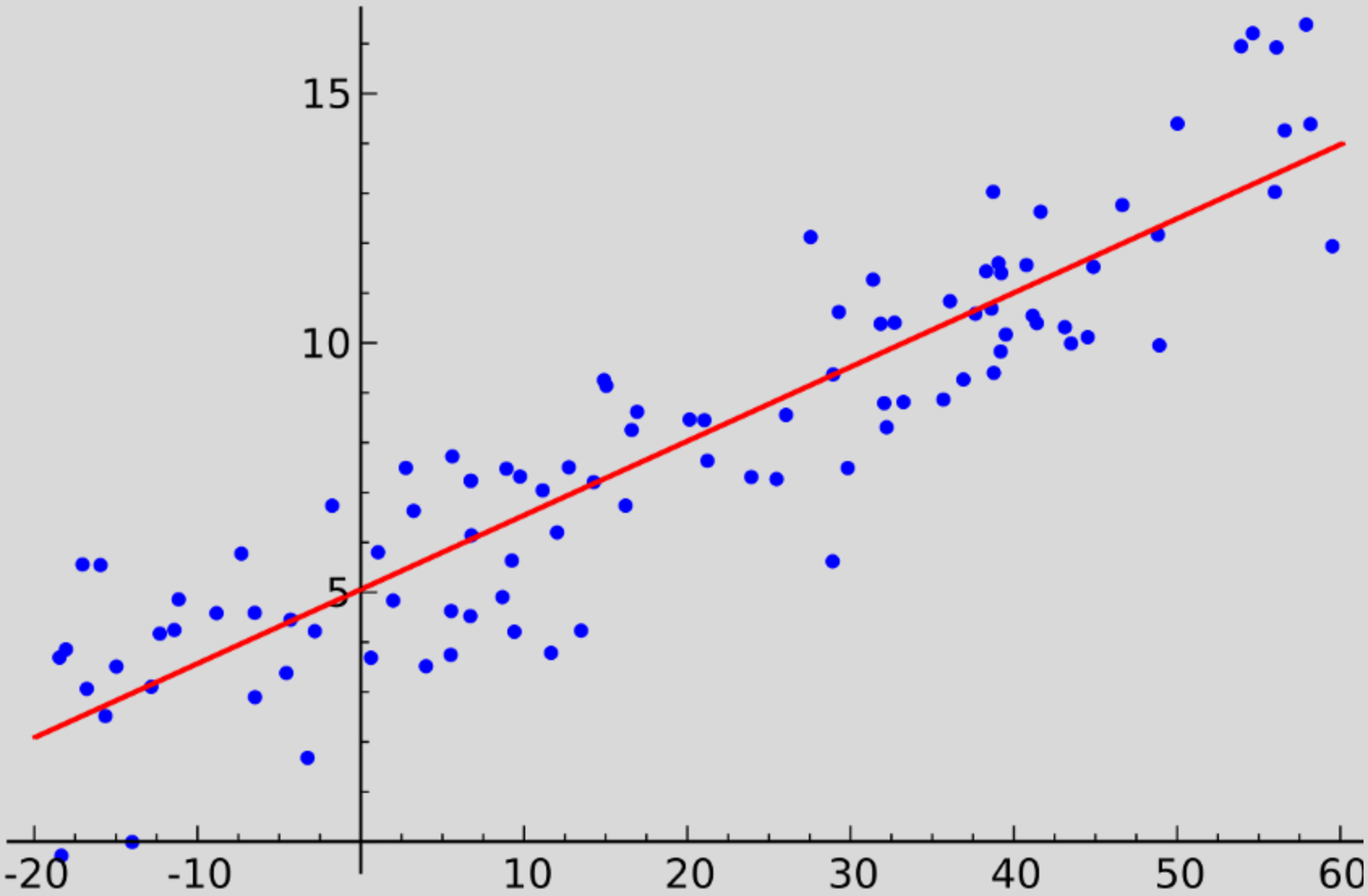


FIGURE REPRESTING LINEAR REGRESSION

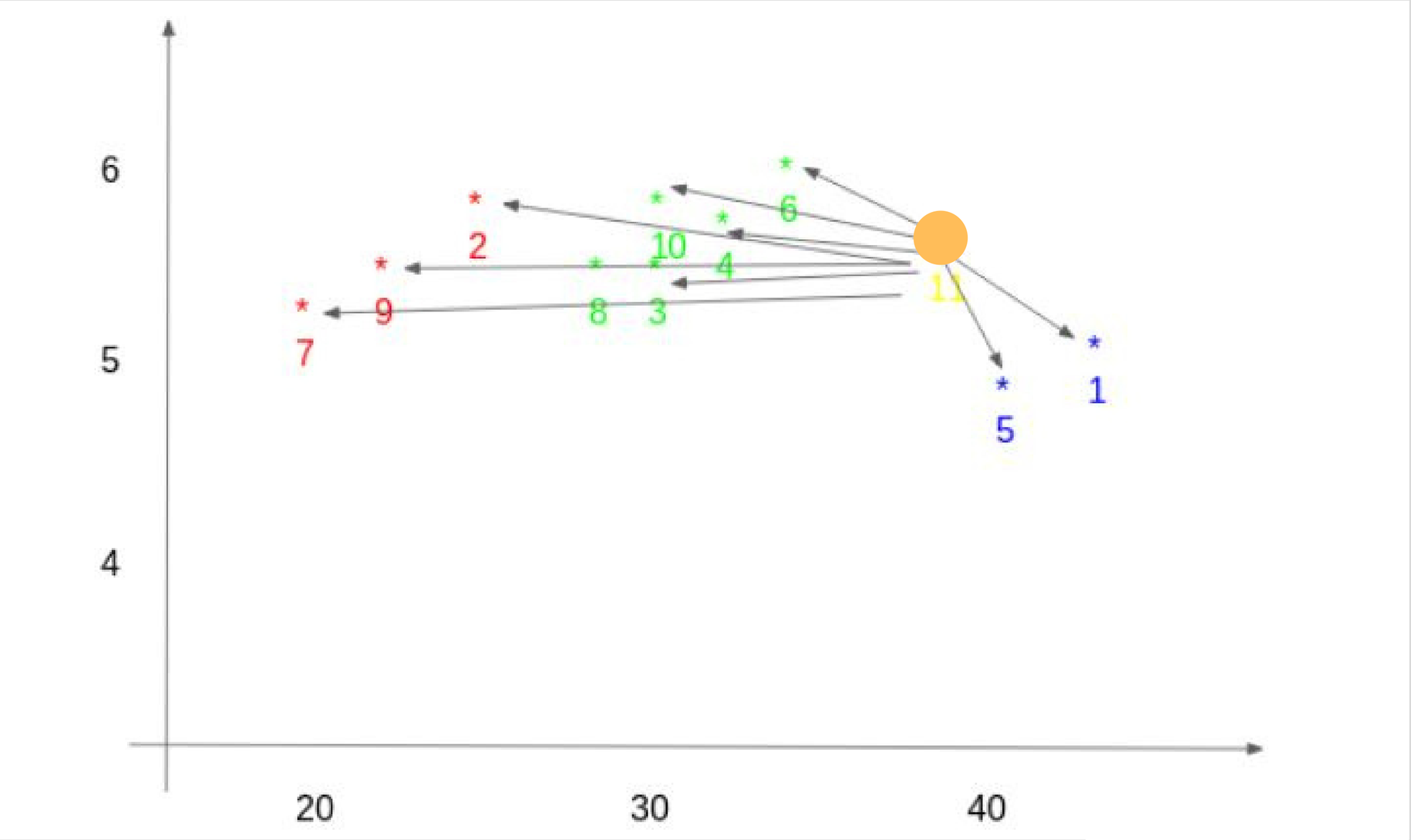
K-Nearest Neighbors Algorithm for Regression

- The K-Nearest Neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

- KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

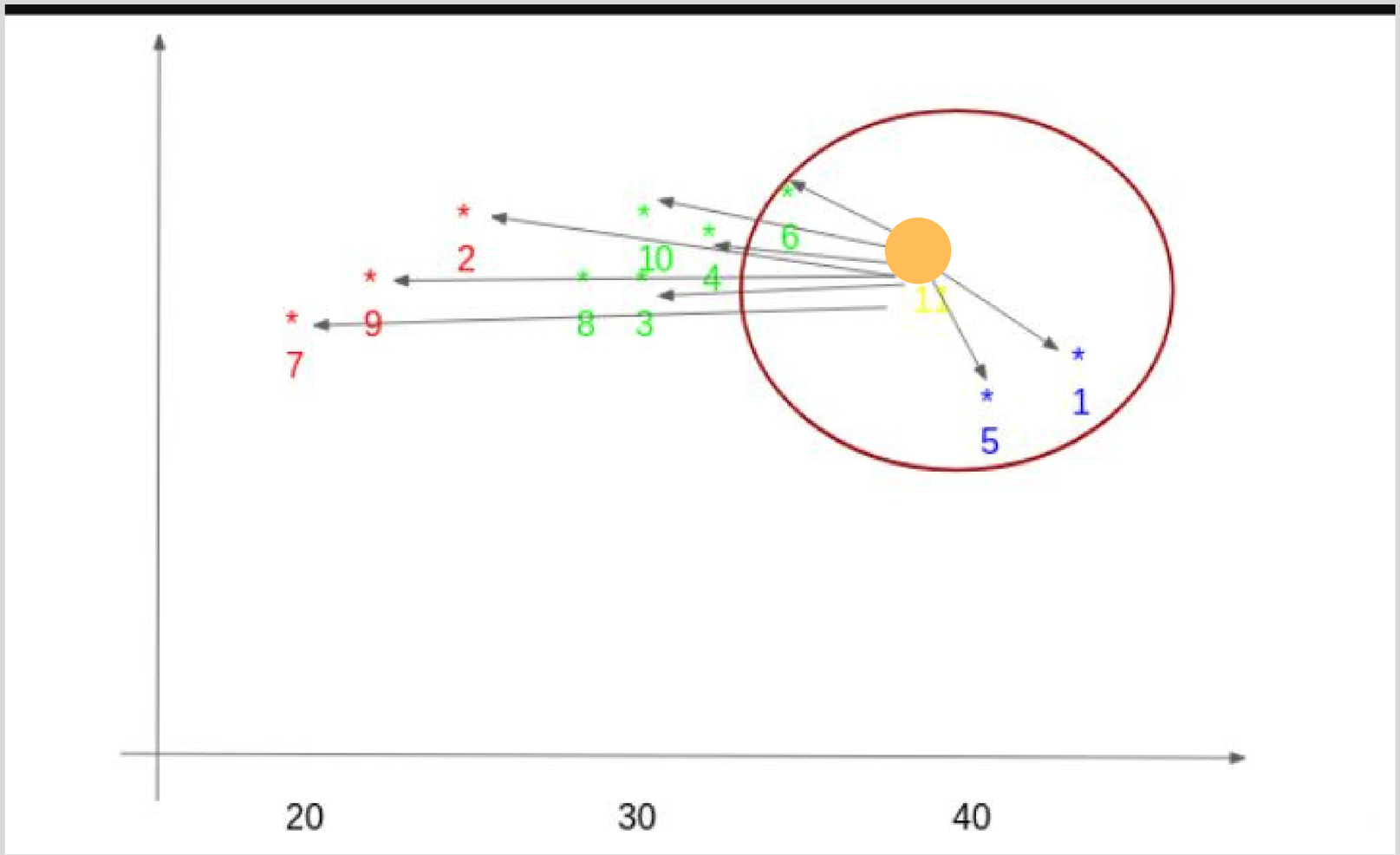
HOW KNN WORKS

1. First, the distance between the new point and each training point is calculated.



HOW KNN WORKS

- The closest k data points are selected (based on the distance). In this example, points 1, 5, 6 will be selected if the value of k is 3.
- The average of these data points is the final prediction for the new point.



SUPPORT VECTOR MACHINES(SVMs)

- SVM algorithm is considered one of the most robust and accurate strategies among the ML-based approaches.
- The main objective of SVM is to map the input data x into a high-dimensional feature space by means of a nonlinear mapping and generate an optimal hyper-plane $w.x + b = O$ in this new space.
- SVM seeks to minimize an upper bound of the generalization error. In order to obtain the optimal hyper-plane the norm of the vector w must be minimized while the margin defined between the 2 classes 1 is maximized

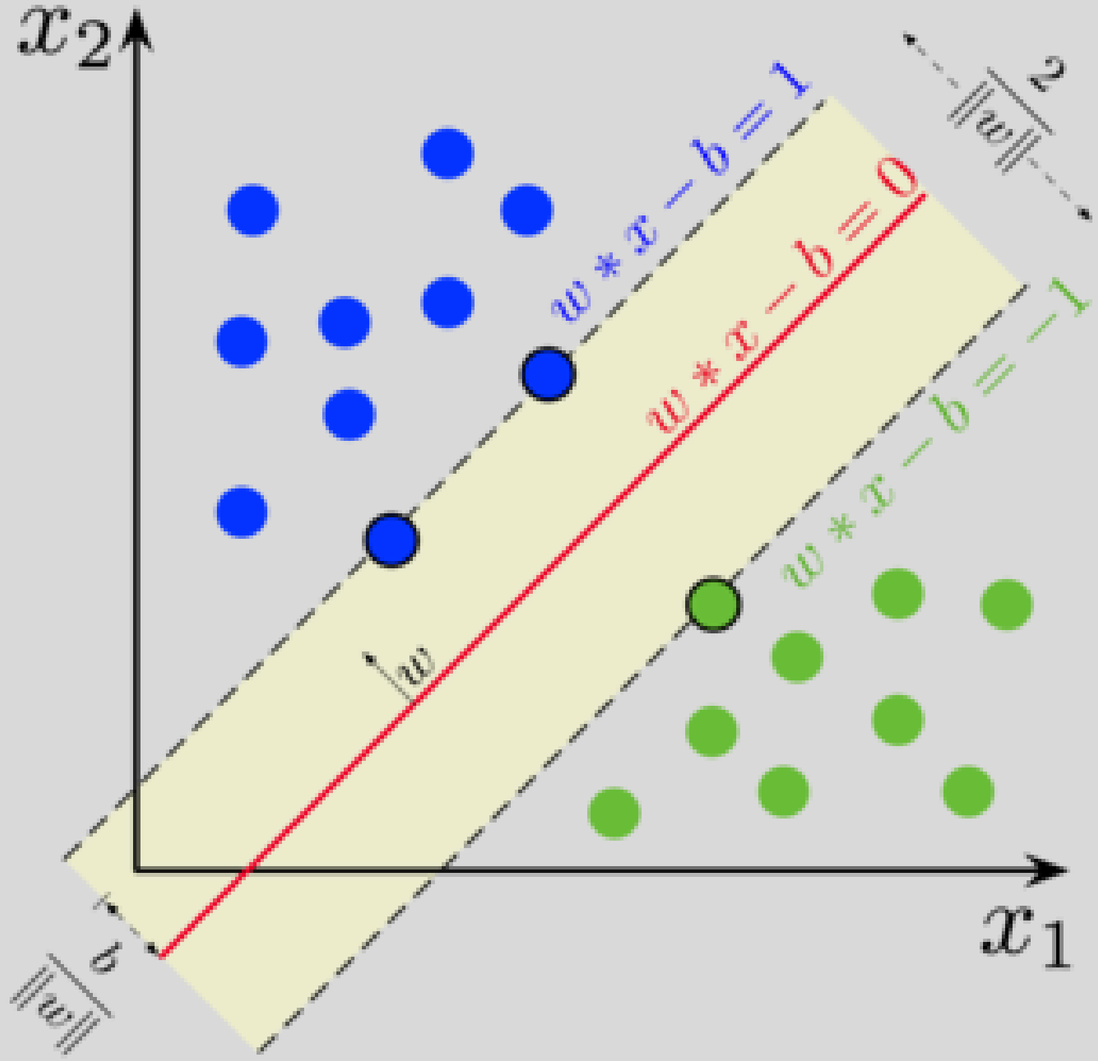


FIGURE REPRESENTING SVM

DECISION TREES

- Decision tree builds regression models in the form of a tree structure.
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.

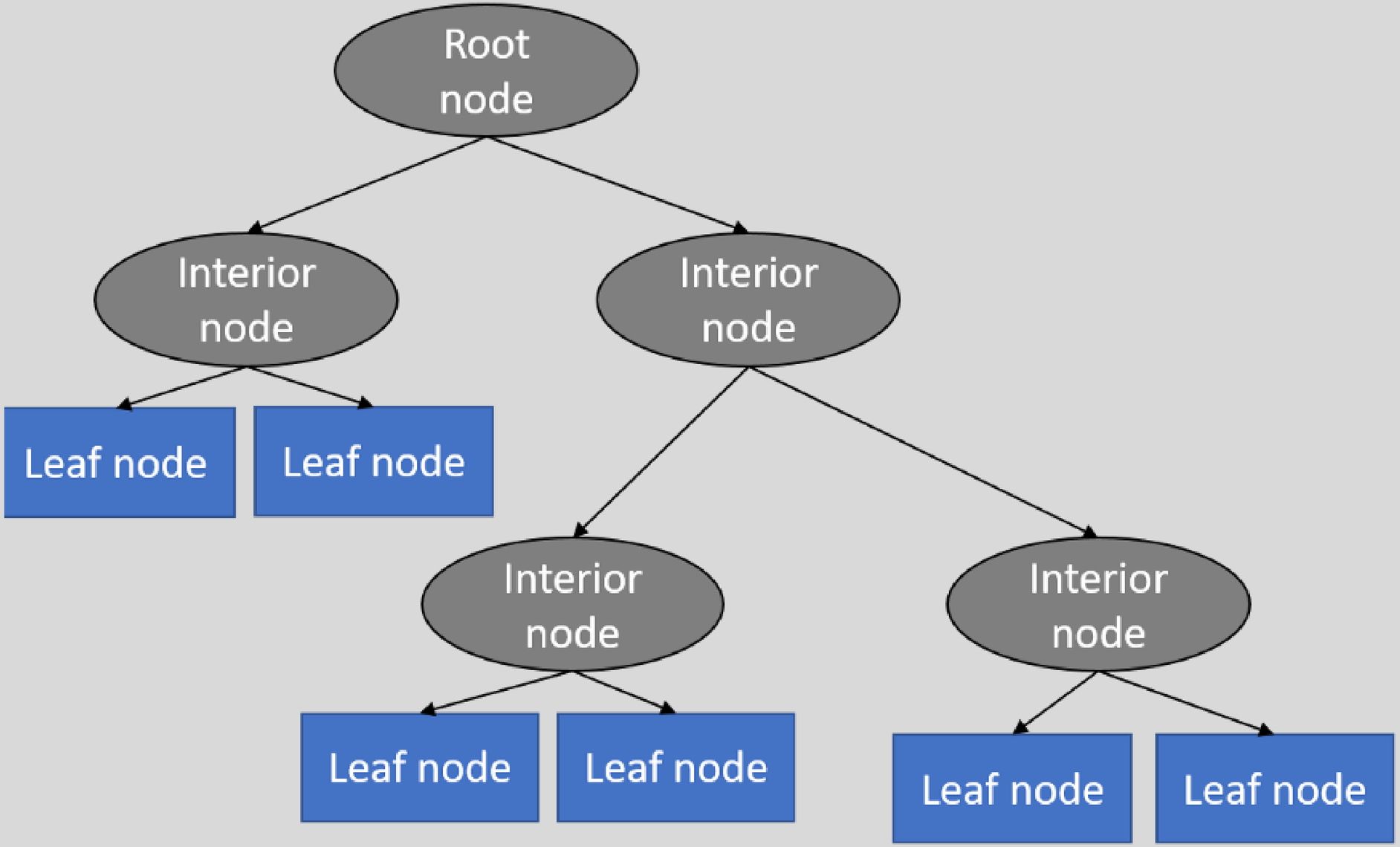
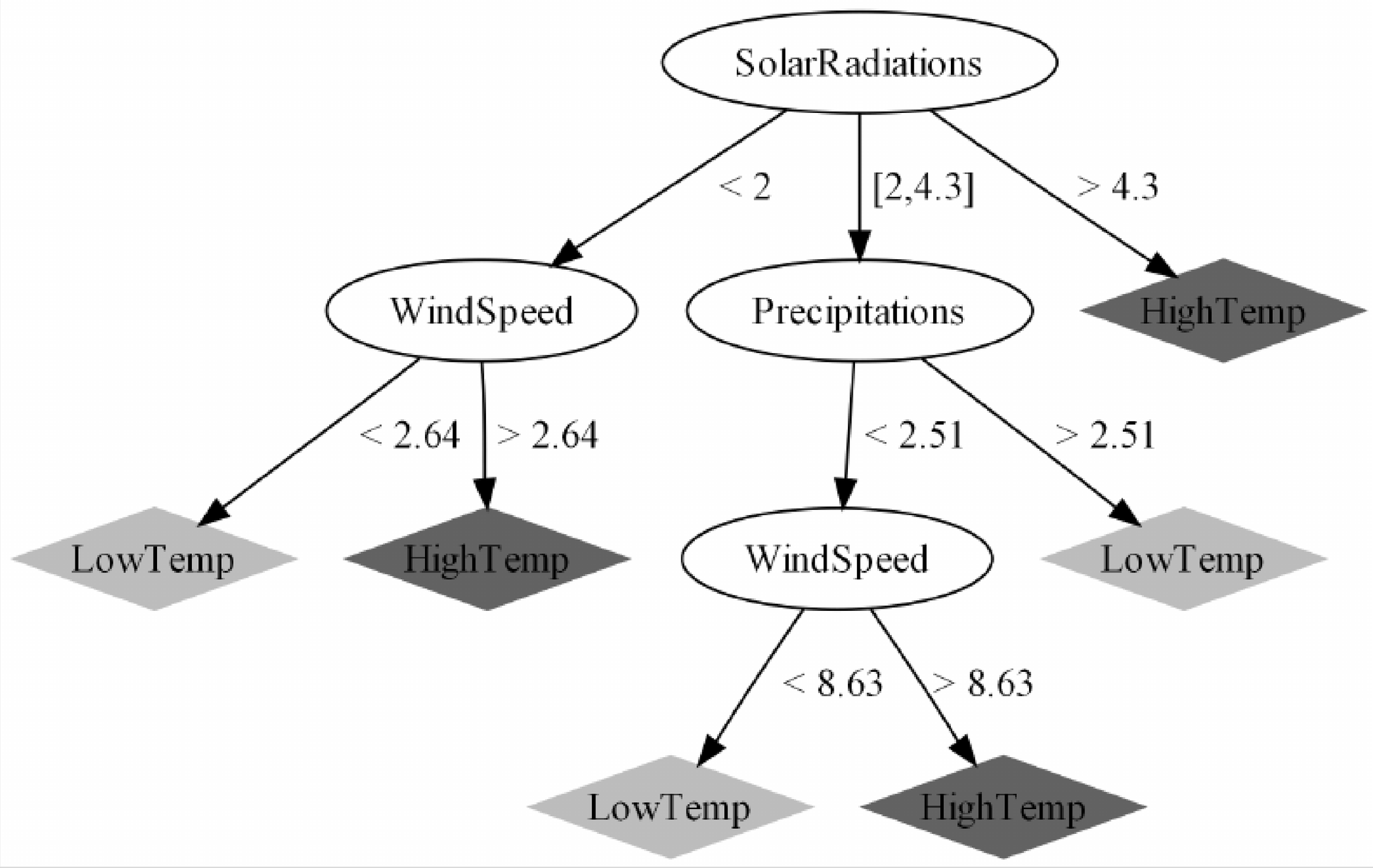


FIGURE REPRESENTING THE TYPES OF NODES IN A DECISION TREE



HOW DECISION TREE CAN BE USED FOR AIR TEMP. PREDICITION

Extreme Gradient Boosting (Xgboost)

- Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for regression predictive modeling problems.
- Extreme Gradient Boosting, or XGBoost for short, is an efficient open-source implementation of the gradient boosting algorithm.
- It is designed to be both computationally efficient and highly effective, perhaps more effective than other implementations.
- Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoost is one of the ensemble learning methods. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancels out and better one sums up to form final good predictions.

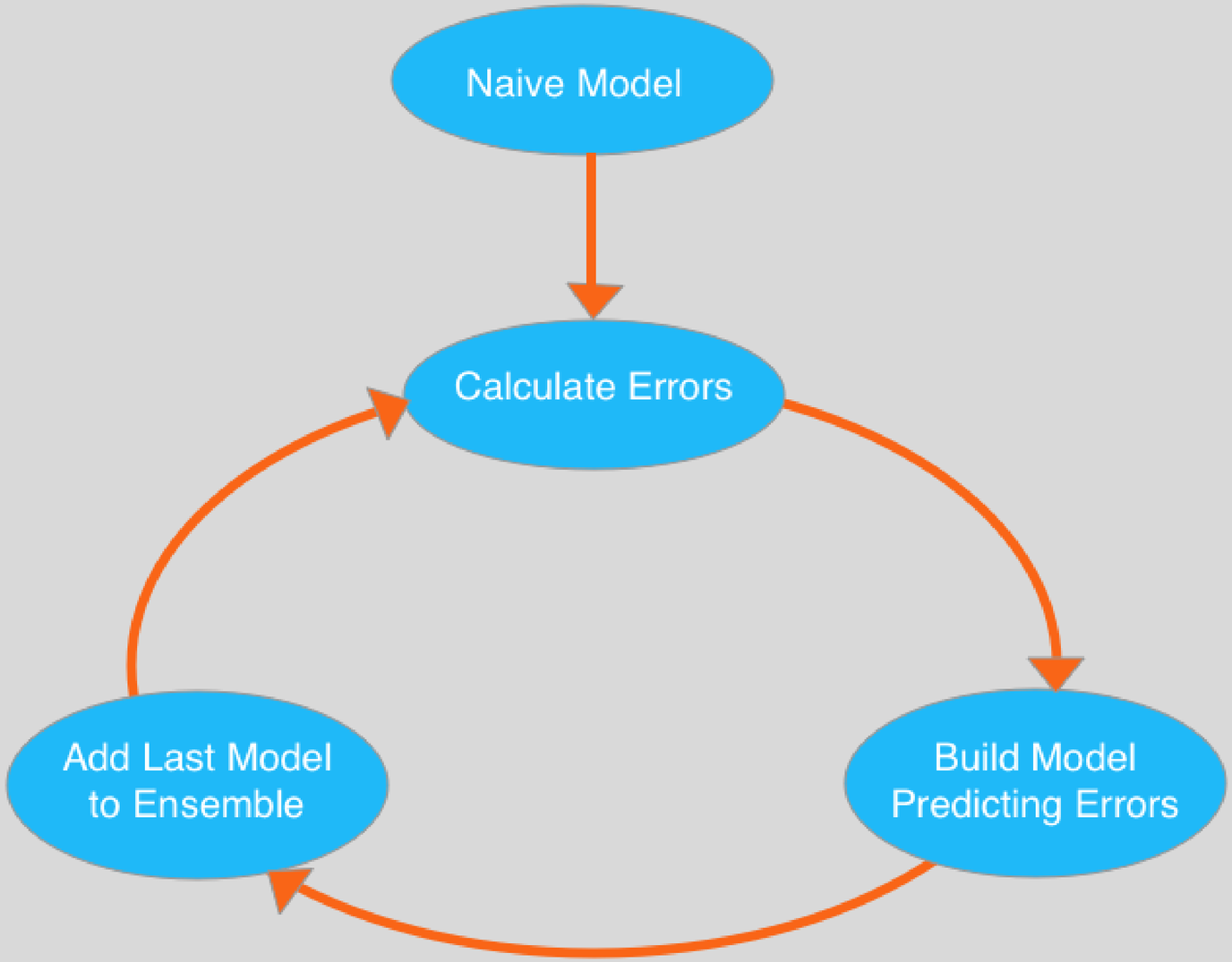


FIGURE REPRESENTING XGBOOST FOR REGRESSION

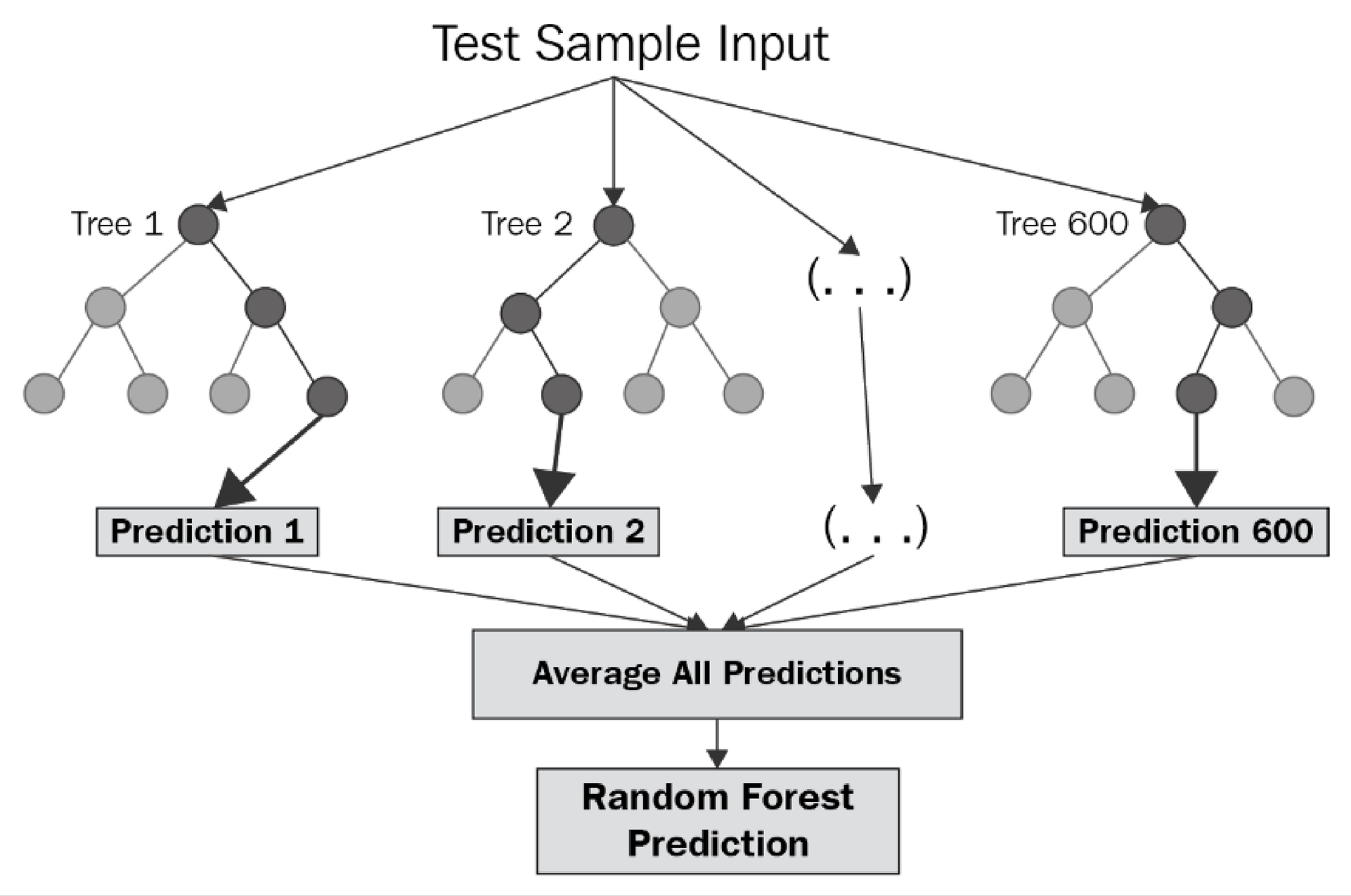


FIGURE REPRESENTING RANDOM FOREST REGRESSOR

Evaluation Measures

- The measures that we used to evaluate the ML strategies implemented for forecasting tasks were:

1. **Mean Absolute Error (MAE)**: This measure is an error statistic that averages the distances between the estimated and the observed data for N samples.
2. **Mean Square Error (MSE)**: This measure is defined as the average squared difference between the predicted and the observed temperature data, for N samples.
3. **Root Mean Square Error (RMSE)**: This measure is the standard deviation of the difference between the estimation and the true observed data.

RESULT

- The best result was obtained after applying random forest regressor through which we achieve the following scores:

  Mean Absolute Error : 19.028.

  Root Mean Squared Error : 36.8199417142395

- The second best result was obtained using the Xgboost algorithm for regression through which we obtained the following results:

  Mean Absolute Error : 24.69975078658646

  Root Mean Squared Error : 40.769059236043134

FURTHER SCOPE OF IMPROVEMENTS

- Most of our work presented is focused on the local analysis of the air temperature. However, there is not an extensive study about the anomalies prediction of temperature at a global level by means of these ML-based approaches.
- A further study about the feature selection, based on their relevance, should be performed.
- In-depth analysis using statistical significance tests is required in order to assess the forecasting model's performance in terms of its ability to generate both unbiased and accurate forecasts.

THANK YOU

THE END OF THE PRESENTATION