# INDIAN WEATHER ANALYSIS AND FORECAST

# MINI PROJECT REPORT

***Submitted by***

**PRASANNA K (193002073)**

**THARUN ARASU SK (193002111)**

**PADMACHARAN D (193002069)**

**UEC1605**

**MACHINE LEARNING**

****

**Department of Electronics and Communication**

**Engineering**

**Sri Sivasubramaniya Nadar College of Engineering**

(An Autonomous Institution, Affiliated to Anna University)

**Rajiv Gandhi Salai (OMR), Kalavakkam – 603 110**

**EVEN SEM 2021-2022**

**Sri Sivasubramaniya Nadar College of Engineering**

**(An Autonomous Institution, Affiliated to Anna University)**

**BONAFIDE CERTIFICATE**

Certified that this mini project titled “**INDIAN WEATHER ANALYSIS AND FORECAST**” is the bonafide work of “**PRASANNA K (193002073)**, **PADMACHARAN D (193002069)** and **THARUN ARASU SK (193002111) of VI Semester Electronics and Communication Engineering Branch during Even Semester 2021 – 2022 for UEC1605 Machine Learning**

Submitted for examination held on 10/06/2022

**INTERNAL EXAMINER**

**ABSTRACT**

India has a typical weather condition consisting of various seasons and geographical conditions. Country has extreme high temperatures at Rajasthan desert, cold climate at Himalayas and heavy rainfall at Chirapunji. These extreme variations in temperatures make us to feel difficult in inferring / predicting the weather effectively. It requires higher scientific techniques/methods like machine learning algorithms applications for effective study and predictions of weather conditions.

Data cleaning is done to improve the quality of the training data for analytics and enables accurate decision-making and data reshaping is done improve the performance speed.

Various data visualization techniques are used to observe the impact of temperature in India. From the inferences of the visualization techniques temperature throughout the time is plotted, and the seasonal analysis is performed.

In this mini project, we applied K – Means cluster algorithm for grouping similar datasets together and Decision tree regressor for prediction.

Before clustering the dataset using K – Means clustering the better value of K is chosen using elbow criterion, inorder to cluster the dataset in an accurate way.

Using Decision tree algorithm the mean temperature for the every month in the year of 2022 is forecasted or predicted.

This mini project is helpful in providing information to people and organizations inorder to reduce weather-related losses and enhance societal benefits, including protection of life and property, public health and safety, and support of economic prosperity and quality of life.

**TABLE OF CONTENTS**

**CHAPTER No. TITLE PAGE No.**

1 **INTRODUCTION**  1

2 **LITERATURE SURVEY**

3 **MACHINE LEARNING**

**ALGORTIHMS**

3.1 K – Means Cluster Algorithm

3.2 Decision Tree Algorithm

4 **RESULTS AND DISCUSSION**

5 **CONCLUSIONS**

**REFERENCES**

**LIST OF FIGURES**

**Figure no Content Page no.**

1.1 CSV format of dataset 2

1.2 (a) Libraries Used

1.2 (b) Reshaped Data frame

1.2 (c) Updated Data frame

1.3 Flowchart

## CHAPTER 1

## INTRODUCTION

Weather forecasting is the application of scientific techniques and technology to predict the conditions of the atmosphere at a certain location and time.

Weather Forecasting in old times is carried out by hand, using changes in barometric pressure, current weather conditions, and sky conditions or cloud cover, weather forecasting now relies on computer-based models that take many atmospheric factors into accounting now relies on computer-based models that take many atmospheric factors into account.

For a long time, the researcher had attempted to establish a linear relationship between the input weather data attributes and the corresponding target attribute. But the discovery of nonlinearity within different attributes of weather data, the focus has shifted towards the nonlinear prediction of the weather.

Weather forecasts are made by collecting quantitative data about the current state and previous trends of the atmosphere and using a scientific understanding of atmospheric processes to predict how the atmosphere will evolve. The weather warning is important for the protection of life and property. Rain predictions can be used by farmers.

The beginning of the 21st century, with the advent of big data, efficient supercomputers with Graphics Processing Units (GPU), and scientific interest in emerging new methods, turned out to be crucial in the history of machine learning. Although many methods are known from the 1960s and have been examined in detail in many studies since then, recent years, with unprecedented increases in data volume and computer power, are seen as the golden era for artificial intelligence and machine learning.

In order to analyze weather forecasting and learn about it, we have used K-Means clustering and decision tree as machine learning algorithms in our project for training and forecasting.

We have trained the above-mentioned machine learning models for the data collected for the mean monthly temperature for India from 1901 to 2021.

* 1. **Dataset Description**

The Dataset is derived from Kaggle. The dataset contains the monthly mean temperature of India from 1901 to 2021 in CSV format. It contains only numerical input values with no missing or undefined values.

The dataset has a total of 1680 values in 120 rows and 14 columns. The figure1.1 shows the head of the dataset in CSV format.

Graphical user interface, application

Description automatically generated

**Figure 1.1**

* 1. **Libraries Used**

In this project, we have imported numpy, pandas, plotly express, and datetime libraries.

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms basic linear algebra, basic statistical operations, random simulation and much more.

At the core of the NumPy package, is the array object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

Plotly express is a built-in part of the plotly library and is recommended starting point for creating most common figures. Instead of matplotlib, we have used plotly as it is very helpful for data visualization and understanding the data easily with its easily built API. It is easy to switch from scatter plot to histogram to sunburst with plotly.

DateTime is a class and an object in Python's DateTime module. The arguments are a combination of date and time attributes, starting from the year and ending in microseconds.

* 1. **Data Pre – processing**

The dataset we have taken has an unnamed column that is from the data as it is of no use. So, the column must be removed from the dataset.

So, we define the index column as zero and the modified dataset can be seen in figure1.2 (a).

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 1.2 (a)**

As we must improve the performance of our analysis the data can be melted. To do that the YEAR column is made an identifier column and the other two columns are melted into two non-identifier columns namely variable and value. The melted data frame can be seen in figure 1.2 (b).

Text

Description automatically generated with medium confidence

**Figure 1.2 (b)**

Then a new column named ‘Date’ is added to the dataframe as an attribute for analysis. Using the DateTime library for time series analysis lambda function and loc[] method are used and strptime() is used to format and return a string representation of date and time.

The updated data frame can be seen in figure 1.2 (c).

Text

Description automatically generated

**Figure 1.2 (c)**

From the above figure we can see that column has a name variable which doesn’t convey the meaning so we rename the column to ‘Temperature’ and sort the values sort\_value() function. This will return a dataframe with sorted values in ascending order.

This can be seen in figure 1.2 (d) and this modified or pre-processed data is used in the whole project.

Text

Description automatically generated

**Figure 1.2 (d)**

figure 1.3 shows the flowchart of mini project

Diagram

Description automatically generated

**Figure 1.3**

## CHAPTER 2

## LITERATURE SURVEY

This section presents previous work related to our proposed method.

For this project, we have referred to a few journals and papers to understand which procedure or algorithm to follow to successfully complete the project with greater accuracy and the best-suited approach for the problem statement we have taken. Some of the reviews we have taken and our understanding from the papers is presented below.

**Mark Holmstrom, Dylan Liu, and Christopher Vo (2016)** concluded that both linear and functional regression didnot perform as well as professional weather forecastingmethods but in the longer run differences in theirperformances decreased, suggesting that over a longerperiod of time, Machine learning can indeed outperformprofessional and traditional methods. Linear regression isa low bias and high variance algorithm and hence its accuracy can be improved by collecting further data.

**PiyushKapoor and Sarabjeet Singh Bedi (2013)** concluded that if we perform the comparison of weathercondition variation by sliding window algorithm, theresults are highly accurate except for the months ofseasonal change. The results can be altered by changingthe size of the window. The accuracy of the unpredictablemonths can be increased by increasing the window sizeto one month.

**DivyaChauhan and Jawahar Thakur (2013)** made a comparison in their paper, which shows that the algorithms such as k-mean clustering and decision trees are well suited for mining data to predict future weather conditions.

**Siddharth S. Bhatkande, Roopa G. Hubballi(2016)** In their work, the authors have used data mining techniques and Decision tree algorithm to classify weather parameters like maximum temperature and minimum temperature in terms of the day, month, and year.

From the above journals, we concluded from DivyaChauhan and Jawahar Thakur that for weather forecasting project k-means and decision tree algorithms yield better results compared to other methods, as our data do not have specific labels and come under unsupervised learning these algorithms were a go to approach.

As we have decided which algorithms to use, we selected and studied a paper each for k-means clustering and decision tree algorithm. The papers we referred to are given below.

A review on K-Means data clustering approach by Shraddha Shukla and Naganna S inferred that amount of information in the world doubles every 20 months. Data mining is the future of computing science as it extracts data for visualization and extracts different patterns for classification, clustering, regression, etc.

K-Means clustering is used in computer vision and pattern recognition. It assigns n data points into k-clusters so that similar data can be grouped together. Outliers, empty clusters are some limitations of clustering. Setting an appropriate initial number of clusters is always a challenging task.

From the above journal, we got to know about k-means clustering workflow.

Study and Analysis of decision Tree Based Classification Algorithm by Harsh H. Patel, Purvi Prajapti inferred that A normal tree includes root, branches and leaves. The same structure is followed in Decision Tree. It contains root node, branches, and leaf nodes. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continuous value).

Decision Tree is similar to the human decision-making process and so that it is easy to understand. It can solve in both situations whether one has discrete or continuous data as input. Decision tree algorithms are used to split the attributes to test at any node to determine whether splitting is “Best” in individual classes. The resulting partitioned at each branch is PURE as possible, for that splitting criteria must be identical. If the values are close to each other, the set can be said to be precise. If their average is close to the true value of the quantity being measured, the set can be said to be accurate. Only if given a set of data points from repeated measurements of the same quantity then one can measure the above two terms. Among ID3 CART and C4.5 they conclude that CART algorithm in decision tree gives the best accuracy.

From the above papers we learned about how to approach the project, from the above readings we got a clear idea to complete the project with a good understanding on algorithms.

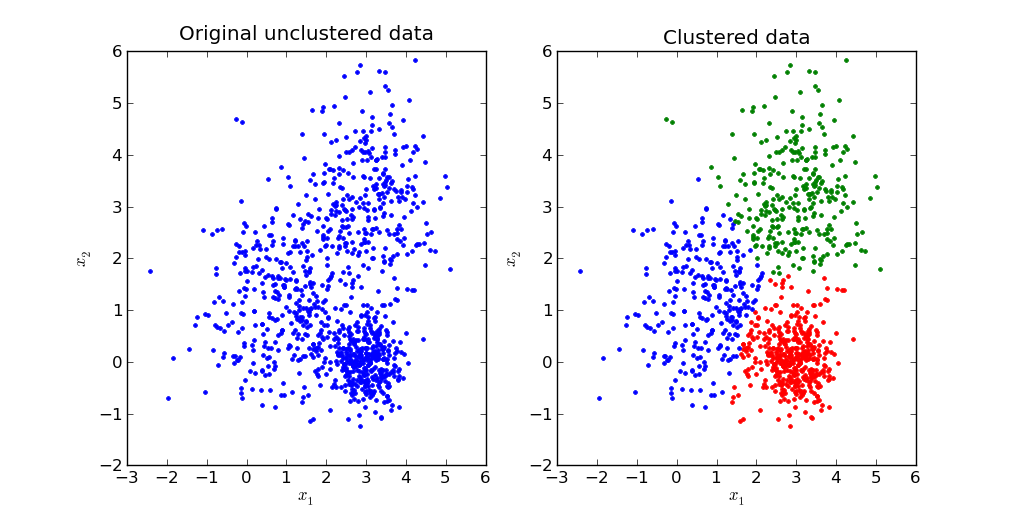
**CHAPTER 3**

**MACHINE LEARNING ALGORITHMS**

In this chapter, the basic concept of two important algorithms namely K-Means cluster algorithm and decision tree algorithm are discussed.

**3.1 K – MEANS CLUSTER ALGORITHM**

According to The IEEE Standard Definitions, k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. Figure 3.1 shows the representation of clustered and unclustered data.



**Figure 3.1 K – Means Data Clustering**

**3.1.1 K – Means Algorithm**

The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, k < n. It is similar to the expectation maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance or the squared error function.

A picture containing text, watch

Description automatically generated

where there are k clusters Si, i = 1, 2, …, k, and µi is the centroid or mean point of all the points xj Si.

**The algorithm steps are:**

1. Initialize k cluster centers to be seed points (These centers can be randomly produced or use other ways to generate).
2. For each sample, find the nearest cluster center, put the sample in this cluster.
3. Recompute centers of the altered cluster (Repeat n times).
4. Examine all samples again and put each one in the cluster identified with the nearest center (don't recompute any cluster centers). If members of each cluster haven't been changed, stop. If changed, go to step 2.

## A picture containing graphical user interface Description automatically generated Shape, rectangle Description automatically generated

## Step – 1 Step – 2

## A picture containing graphical user interface Description automatically generated Shape Description automatically generated

**Step – 3 Step – 4**

**Figure 3.2 Demonstration Standard K – Means Algorithm**

**3.1.3 Implementation of K – Means cluster on temperature dataset**

In this mini project, K – Means clustering is used to classify the months into three clusters based on the temperature. The three clusters are coldest, hottest and neither too hot nor too cold.

Before proceeding with K – Means Clustering we need to select the better K value. To achieve this elbow criterion is used. Figure 3.3 and 3.4 shows the evaluation on number of clusters and clustered dataset respectively.

Chart, line chart

Description automatically generated

**Figure 3.3 Evaluation on Number of clusters**

**Chart, scatter chart

Description automatically generated**

**Figure 3.4 Temperature Clusters**

**3.2 DECISION TREE ALGORITHM**

According to The IEEE Standard Definitions, A decision tree is a tree whose internal nodes can be taken as tests (on input data patterns) and whose leaf nodes can be taken as categories (of these patterns). These tests are filtered down through the tree to get the right output to the input pattern. Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Any Boolean expression can be represented on discrete attributes using the decision tree.



## Figure 3.5 Decision Tree Representation

**3.2.1 Tool for Decision Making**

A decision tree is the most widely used tool for decision making. To accomplish this, one should draw a decision tree with different branches and leaves. These branches and leafs should point to all the various factors concerning a particular situation. A decision tree is almost like a decision support tool. Its uses a tree-like graph of decisions and their possible outcomes which include resource costs, event outcomes, and utility.

Decision Tree algorithms can be applied and used in various fields. It can be used as a replacement for statistical procedures to find data, to extract text, to find missing data in a class, to improve search engines and it also finds various applications in medical fields.

## 3.2.2 Assumptions while creating Decision Tree

## In the beginning, the whole training set is considered as the root.

## Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.

## Records are distributed recursively on the basis of attribute values.

## Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

## Decision Trees follow Sum of Product (SOP) representation. The Sum of product (SOP) is also known as Disjunctive Normal Form. For a class, every branch from the root of the tree to a leaf node having the same class is conjunction (product) of values, different branches ending in that class form a disjunction (sum).

## 3.2.3 Working of Decision Tree

## The decision of making strategic splits heavily affects a tree’s accuracy. The decision criteria are different for classification and regression trees.

## Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. The algorithm selection is also based on the type of target variables.

## Algorithms Used in decision tree:

## ID3

## C4.5

## CART

## CHAID

## MARS

## Steps in ID3 Algorithm:

## It begins with the original set S as the root node.

## On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.

## It then selects the attribute which has the smallest Entropy or Largest Information gain.

## The set S is then split by the selected attribute to produce a subset of the data.

## The algorithm continues to recur on each subset, considering only attributes never selected before.

## Attribute Selection Measures:

## If the dataset consists of N attributes, then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can’t solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

## Solutions for attribute selection problem:

## Entropy

## Information gain

## Gini index

## Gain Ratio

## Reduction in Variance

## Chi-Square

## 3.2.4 Pruning

## The common problem with Decision trees, especially having a table full of columns, they fit a lot. Sometimes it looks like the tree memorized the training data set. If there is no limit set on a decision tree, it will give you 100% accuracy on the training data set because in the worst case it will end up making 1 leaf for each observation. Thus, this affects the accuracy when predicting samples that are not part of the training set.

## In pruning, the branches of the tree are trimmed off, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

## Decision tree pruning - Wikipedia

## Figure 3.6 Decision Tree Pruning

## CHAPTER 4

## RESULTS AND DISCUSSION

## Timeline Description automatically generated

**Figure 4.1**

**Inference from the above plot:**

* May 1921 has been the hottest month in India in the history.
* Dec, Jan and Feb are the coldest months. We can group them together as "Winter".
* Apr, May, Jun, July and Aug are the hottest months. We could group them together as "Summer".
* But since this is not how seasons work. We have four main seasons in India, and this is how they are grouped:
* Winter : December, January and February.
* Summer(Also called, "Pre-Monsoon Season") : March, April and May.
* Monsoon : June, July, August and September.
* Autumn(Also called "Post Monsoon Season) : October and November.
* We will also stick to these seasons for further our analysis.

## Chart, box and whisker chart Description automatically generated

## Figure 4.2

**Inference from the above box plot:**

* January has the coldest Days in a Year.
* May has the hottest days in a Year.
* July is the month with least Standard Deviation which means, temperature in July vary least. We can expect any day in July to be a warm day.

## Chart Description automatically generated

## Figure 4.3

**Inference from the above plot:**

* Sum of Squared Errors are calculated.
* Using Elbow Criterion, the optimised K value is chosen.
* Here cluster size of 3 seems to be a good choice.

## Chart, scatter chart Description automatically generated

## Figure 4.4

**Inference from the above Scatter plot:**

* Despite having 4 seasons we can see 3 main clusters based on temperatures.
* Jan, Feb and Dec are the coldest months.
* Apr, May, Jun, Jul, Aug and Sep; all have hotter temperatures.
* Mar, Oct and Nov are the months that have temperatures neither too hot nor too cold.

## Chart, histogram Description automatically generated

## Figure 4.5

**Inference from the above histogram plot:**

* The mean temprature for most months during history has been between 26.8-26.9.

## Graphical user interface, application Description automatically generated

## Figure 4.6

## Chart, scatter chart Description automatically generated

**Figure 4.7**

**Inference from the above scatter plots:**

We can see that the issue of global warning is true.

* The yearly mean temprature was not incresing till 1980. It was only after 1979 that we can see the gradual increse in yearly mean temprature.
* After 2015, yearly temprature has incresed drastically.
* But, There are some problems in this figure.
* We are seeing a monthly like up-down pattern in yearly tempratures as well.
* This is not understandable. Because with months, we have a phenominan of seasons and the earth the revolving around sun in a eliptic path. But this pattern is not expected in yearly temprature.
* Even from the plot we can infer that the sample points are not linearly separable.
* Hence decision tree regression algorithm is implemented in our project.

## Calendar Description automatically generated with low confidence

## Figure 4.8

**Inference from the above plot:**

* From this plot we can predict the temperature range for a month in a year.
* We can also see clear positive trendlines.

## Graphical user interface, application, Teams Description automatically generated

## Figure 4.9

**Inference from the above plot:**

* From the above visualization, we get an idea about the seasonal mean temperature.
* The trendline does not have a very high positive correlation with years, still it is not negligible.

## Chart, scatter chart Description automatically generated

## Figure 4.10

**Inference from the Animation:**

* At first, we can see some fluctuations but it doesn't give much of insights for us.
* If we analyse by adjusting the bar below to early years and late years we can notice the change.

## Chart, scatter chart Description automatically generated

## Figure 4.11

**Inference from the above plot:**

* Thus, the yearly mean temperature of India in 2022 is 26.19.

## CHAPTER 5

## CONCLUSION

## Data cleaning and data reshaping has been done.

## K-means clustering algorithm has been used to classify the data samples and decision tree regression algorithm has been used to forecast the weather.

## Decision trees prove as an effective method of decision making in Weather prediction. As, decision trees are ideal for multiple variable analyses, it is particularly important in current problem-solving task,like weather forecasting.

## The dataset has been classified according to the mean temperature of every month in each year and the weather prediction has been done.

## Professional weather forecasters are not perfect, but their predictions are typically more accurate than those of this machine learning model. This implies that weather is a non-linear system. Additionally, our predictions were all based on data from one location as opposed to multiple locations that most forecasters use. Even in our model, we have described the limitation of using linear regression on predicting the weather.

## This mini project is helpful in providing information to people and organizations inorder to reduce weather-related losses and enhance societal benefits, including protection of life and property, public health and safety, and support of economic prosperity and quality of life.

## REFERENCES

[1] Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting, Stanford University, 2016.

[2] PiyushKapoor and Sarabjeet Singh Bedi “ Weather Forecasting Using Sliding Window Algorithm”, Kvantum Inc., Gurgaon 122001, India MJP Rohilkhand University, Bareilly 243006, India,2013.

[3] DivyaChauhan, Jawahar Thakur “ Data Mining Techniques for Weather Prediction: A Review Shimla 5, India: ISSN, 2013.

[4] Siddharth S. Bhatkande1 ,Roopa G. Hubballi2 “Weather Prediction Based on Decision Tree Algorithm Using Data Mining Techniques”, Belgaum India:

International Journal of Advanced Research in Computer and Communication Engineering, 2016.

[5] Shraddha Shukla and Naganna S, “A Review on K-Means data Clustering Approach”, School of Computing Sciences and Engineering, Galgotias University, Greater Noida, India, International Journal of Information and Computation Technology (IJICT),2014.

[6] Harsh H. Patel ,Purvi Prajapti, “Study and Analysis of Decision tree Classification Algorithm”, Charotar University Gujarat,India,International Journal of Computer Science Engineering,2018.

[7] A. Kumar, R. Sinha, V. Bhattacherjee, D. S. Verma, S. Singh, “modeling using K-means clustering algorithm”, IEEE 2012, 1 st international conference on recent advances in information technology(RAIT).

[8] D. T. Pham, S. S. Dimov, and C. D. Nguyen, “Selection of K in K-means clustering”, Proc. IMechE Vol. 219 Part C: J. Mechanical Engineering Science, IMechE 2005.

[9] J. Zhu; H. Wang, "An improved K-means clustering algorithm, " 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME), vol., no., pp.190, 192, 16-18 April 2010.