**A PROJECT REPORT**

**ON**

**Weather prediction model to predict Rain**

**Using Decision tree Algorithm**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



*Submitted By*

**Anshu(**College of Agricultural Engineering and Technology, Hisar**)**

*Under the guidance of*

**Shivam Bhatia**

# *Introduction*

Rainfall prediction is a crucial task in meteorology and has numerous applications in various fields. This project and I'm using the weather data from multiple cities fromAustralia.The data Consist of 10 years of.Daily weather observations.In this data. The target variables are RainToday and RainTomorrow Both of which are categorical variable.And therefore I decided to use.The decision tree algorithm For this project.

The decision tree algorithm is a popular ML technique that builds a tree-like model of decisions and their possible consequences. By analyzing historical weather data and relevant atmospheric parameters, the decision tree algorithm can identify patterns and make predictions about future rainfall.

The model takes into account various factors such as temperature, humidity, wind speed, wind direction, and cloud cover to make predictions.

The report will provide an overview of the project’s objectives, methodology, dataset used, and the implementation details of the decision tree algorithm.

# Problem Statement

The objective of this project is to that if it is going to rain, be rain today or tomorrow.Based on the atmospheric factors such as temperature.Humidity.The pressure.Wind speed, wind direction. There will be two target variables in this first select.Which are categorical variables.The values of this variable is either a yes or a no.

# Data set

The data set contains about 10 years of daily weather observation from many regulations across Australia. The rain today and rain tomorrow The column is yes.The rainfall is more than 1M.And know if the rainfall is less than 1Mmm.

The dataset has the following 23 columns of data..

Data columns (total 23 columns):

| **Column** | **Non-Null Count** | **Dtype** |
| --- | --- | --- |
| Date | 145460 | object |
| Location | 145460 | object |
| MinTemp | 143975 | float64 |
| MaxTemp | 144199 | float64 |
| Rainfall | 142199 | Float64 |
| Evaporation | 82670 | float64 |
| Sunshine | 75625 | float64 |
| WindGustDir | 135134 | object |
| WindGustSpeed | 135197 | float64 |
| WindDir9am | 134894 | object |
| WindDir3pm | 141232 | object |
| WindSpeed9am | 143693 | float64 |
| WindSpeed3pm | 142398 | float64 |
| Humidity9am | 142806 | float64 |
| Humidity3pm | 140953 | float64 |
| Pressure9am | 130395 | float64 |
| Pressure3pm | 130432 | float64 |
| Cloud9am | 89572 | float64 |
| Cloud3pm | 86102 | float64 |
| Temp9am | 143693 | float64 |
| Temp3pm | 141851 | float64 |
| RainToday | 142199 | object |
| RainTomorrow | 142193 | object |

**Data source:** <https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package?resource=downloa>

# Methodology

## Step: 1 Data exploration

It is the initial steps in data.Look at the data and try.Fine.Characteristics or Patterns.In that data.That.Preparing this data.This steps.Include.Things like.Looking for missing values anImputing, treating them and looking for outliers and treating them, etc.

## Target Variable

Here I am considering both rain today and rain tomorrow as my target variable.These are categorical variables and and has either the value yes or no And therefore there are no outliers in this collapse, but they indeed are.Some missing values.Since this is our target air variable, imputing these values won't be a good idea, so I simply deleted Seems all the columns were That data.Is missing in these columns.And and finally I mapped 1 to “Yes” and 0 to “No”.

## The MinTemp and MaxTemp

These variable represent the minimum temperature throughout the day and the maximum temperature throughout the day respectively. I am building the imputed the missing values in these columns using the escalans K and an imputed library With the nearest Neighbors set to 2. I also tried some other method.But I didn't find much of a difference in the data and I thought.This.Should be.The best way To impute these values.

## Rainfall

This column represents the amount of rain happened during the day. Most of the days are.Generally DRY, therefore the data in this.The column is skewed highly towards this zero. Later I realized it won't be a good idea to use this variable to predict the rain today variable as if its value is more than 1M then.The decision tree would Simply Gives yes to that answer and if it is less than 1M.Then it will give no.

## Temp9am and Temp3pm

As the name suggests, these variables represent that temperature at 9:00 AMIn the morning. And as well as 3:00 PM in the evening. While exploring this data.I found out that there is a correlation between the temperature at 9:00 AM.And the.The minimum temperature column.There is also called relation between temperature at 3:00 PM.The Max temperature Column.So I used this column to fill the missing values in this column.

## Humidity9am and Humidity3pm

To create the missing values.In this column.I simply.Use.The meaning of difference between these values to .fill one.Using the other.

And fill the rest of the missing values using the sklearn’s interpolate function.

## WindGust speed

There were three columns which are as follows. Wind gust speed, Wind gust speed at 9:00 AM.Wind gust at 3:00 PM. In the wind gusts speed column I use the other two columns to fill the missing values assuming that.The gusts speed.Would not.Differ.Throughout the day.

## Other

The evaporation sunshine.Column cloud at 9:00 AM, cloud at 3:00 PM.Has a lot of missing values, so I decided to drop these columns.As it won't be a good idea.To impute these value.

Then I filled the missing values of pressure and wind speed using the scale learns.Library.

## Wind Gust Direction

In this variable as well.I use.The.Other two variables which were wind gust direction at 9:00 AM and wind gust direction at 3:00 PM to fill the missing values in the wind gust direction column and then dropped the.Direction column for 9:00 AM and 3:00 PM.As they are categorical variable and would.Increase.The number of.The dummy variables and which as a result increase.

## Dummy variables

Next I break the date column into three different columns called EAR, month and day and then drop the day and the date column.I then converted the month column.Into.Assume categorical.Column.As the order of the month.Note.Would be.Deciding.Factor to determine.The possibility of rain. I kept the month column For accounting the seasonal change in the weather condition.

Then I create that dummy variables for location.Months and wind gust direction.

Here ends the data exploration part.

# Model building

The decision tree algorithm is a popular machine learning technique that builds a tree-like model of decisions and their possible consequences. It is particularly suitable for classification tasks, making it an ideal choice for predicting rain. The decision tree algorithm works by recursively partitioning the data into subsets based on the values of the input features. The partitioning process continues until the subsets are homogeneous with respect to the target variable, i.e., rainfall in our case.

The first step in building a decision tree mo del is to split the dataset into training and testing sets. The training set is used to build the model, while the testing set is used to evaluate its performance. The decision tree algorithm then recursively partitions the training set into subsets based on the values of the input features. The partitioning process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf.

Once the decision tree model is built, it can be used to predict rainfall for new data points. The model takes as input various atmospheric factors such as temperature, humidity, wind speed, and cloud cover and outputs a binary classification indicating whether it will rain or not.

In our project, we utilized the decision tree algorithm to develop a robust rainfall prediction model. We trained the model on a comprehensive dataset of historical weather records and evaluated its performance using various metrics such as accuracy, precision, recall, and F1 score. The results showed that the model achieved high accuracy and performed well in predicting rainfall.

# 

## Conclusion

We evaluated the performance of the model using various metrics such as accuracy, precision, recall, and F1 score. The results showed that the model achieved above 70% accurecy. The findings of this project have significant implications for weather forecasting and can aid in various applications such as agriculture, disaster management, and urban planning.