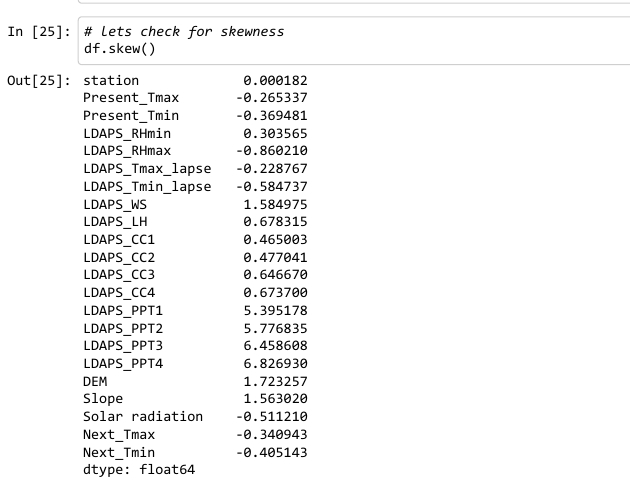
|  |
| --- |
| Problem definition  There are various factors that influence the temperature. In order to forecast temperature we need the historical data. Based on this historical data we need to forecast the temperature. In this project we are going to forecast next day max and min temperature using various machine learning techniques  Data Analysis  Station:  According to the dataset station does not have strong relationship with the target variables. Lets check it out with a plot    date:  **According to the data set Date vs Next\_Tmax and Next\_Tmin, it seems like there is a small drop in temperature in both the cases between 2016 to 2017 and 2017 to 2018 is observed. Lets again check this out with a plot**    Present\_Tmax: there is a strong relationship between Present day max temperature with Next day min temperature compared to with Next day temperature max. lets check this also through a plot   Present\_Tmin: Present day min temp is strongly related with Next day min temp than with Next day max temp. lets this also with a plot   Checking if there are any outlier:  Lets see if there are any outliers using box plot    *from the plots we can observe many columns have outliers, we will remove these outliers using zscore method* |
| EDA (Exploratory data analysis) |
| Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. we can say that EDA is statisticians’ way of storytelling where you explore data, find patterns and tell insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like: - main features of data variables and relationships that hold between them identifying which variables are important for our problem  **1. variable identification:**  Looking at the input data, what will be the output variable  **2. checking and clearing null values:**  It is possible that there may be null values in the dataset. missing values usually occurs when an entry is left empty. It can be filled with the mean, median or mode of the column through fillna () function. Or sometimes SimpleImputer () function is used to replace the null values along with label encoding them, we have null values in all the columns other than lat, lon, Dem  **3. Handling Categorical variables**  Most of the statistical models cannot take Objects / Strings as input they only takes numbers as inputs, with LabelEncoder () it is possible to categorize the string into Numbers as 1,2,3 ,they all are converted through LabelEncoder().fit\_transform function  **4. Descriptive Statistics** –  It is a way of giving a brief overview of the dataset we are dealing with, including some measures and features of the sample. For this we can use the describe() function the describe() function tells the mean(), min(), max(), 25th percentile, median 50th percentile and 75th percentile of the data.   1. **Correlation**   Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr () function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure |

1. **Skewness**

The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side.



1. **Pre-processing Pipeline**

Pipelines are the special way to simplify the code, Pipeline is generally used if we have to perform the code repeatedly usually when there is different train and Test data

Here Our EDA process is completed now moving towards next step

**Building Machine Learning Models**

1. **Separating Features and Target column** –

It is necessary to separate the independent/Features column into a variable (x) and target column into a variable (y). here we have to separate all columns in x Data Frame (variable) and income variable in y Data Frame (variable)

1. **Splitting the Data for Training and Testing**

In ML the separated data is split into 4 parts for Training and Testing of features (x) and for Training and Testing of Target (y) like x\_train, x\_test, y\_train, y\_test. It is possible through a inbuilt library of sklearn’s train\_test\_model

1. **Training the Models**

To find the best model it is necessary to train 3-4 models, In the same way I have trained LinearRegressionModel, KNeighborsRegressor, ExtraTreesRegressor, DecisionTreeRegressor, lets find out which is our best algorithm using the r2 score

Here the above models are giving quite good scores, but ExtraTreesRegressor gave the best result so lets choose ExtraTressRegressor

7) **Exporting the Model**:

Let’s export our model using joblib library in the form of obj or .pkl