DATA SCIENCE : Rainfall Prediction UsingMachine Learning

Date :31 th December, 2022

Author : Yuvraj dutta

1. *Problem Definition.*

Climate change is the biggest problem in the whole world. It is an important aspect of human life, so accurate prediction should be made as much as possible. Climate change is the biggest problem all over the world. Scientists are working to identify the patterns of climate change as it affects the economy from production to infrastructure. Predicting rainfall cannot be done in a traditional way, so scientists and analysts are using machine learning and deep learning concepts to figure out the patterns for predicting rainfall. Heavy and erratic rainfall can have many impacts, such as damage to property, crops, and farms - so a better predictive model is essential for early warning that can better minimize risks to lives, property, and farms. The main goal is to help farmers and also use water resources efficiently.

2. *Data Analysis.*

Context - Predicting rain the next day by training classification models on the target variable rain tomorrow.

Contents - This dataset contains about 10 years of daily weather observations from many locations in Australia.
Rain Tomorrow is the target variable to be predicted. It means - did it rain the next day, yes or no? This column is Yes if the rain that day was 1 mm or more.
Source & Acknowledgements -

Observations were obtained from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data.

An example of latest weather observations in Canberra: http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

Datasource: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.

Dataset - https://raw.githubusercontent.com/dsrscientist/dataset3/main/weatherAUS.csv

https://github.com/dsrscientist/dataset3

IMPORTING PACKAGES -

1. NumPy:

It is the basic scientific computing package in Python. A Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), including mathematical, logical, shape manipulation, sorting, selection, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation, and more.

2. Pandas:

Pandas is a Python package that provides fast, flexible, and expressive data structures that make working with "relational" or "labeled" data easy and intuitive. A fundamental high-level building block for performing practical, real-world data analysis in Python. The most powerful and flexible open-source data analysisand manipulation tool available in any language.

3. Matplotlib:

It is a comprehensive library for creating static, animated and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of print formats and interactive environments on multiple platforms. Matplotlib can be used in Python scripts, web application servers, and various graphical user interface toolkits.

4. Seaborn :

It is a library for creating statistical graphs in Python. It is built on matplotlib and is tightly integrated with pandas data structures. Seaborn helps you explore and understand your data. The plot functions work with data frames and arrays containing entire data sets, and internally perform the necessary semantic mapping and statistical aggregation to create informative graphs.

5. SKlearn :

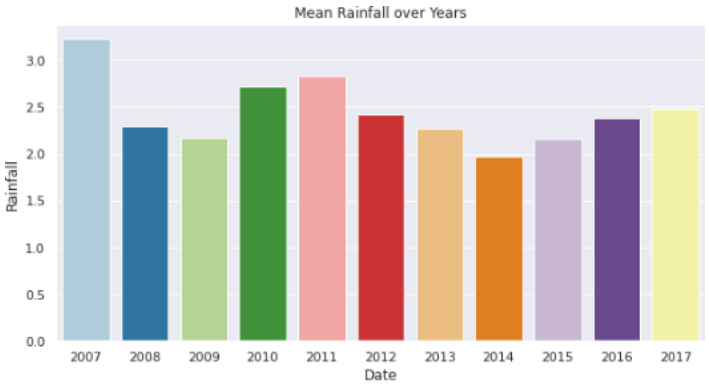
Scikit-learn (Sklearn) is the most useful and robust machine learning library in Python. It provides a range of efficient machine learning and statistical modeling tools, including classification, regression, clustering, and dimensionality reduction through a unified interface in Python.

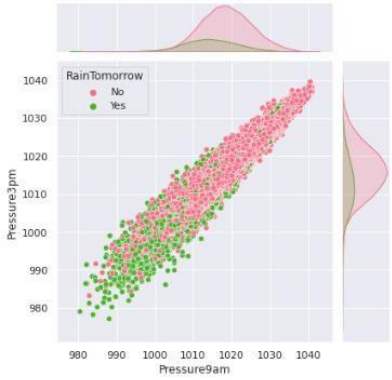
LOAD DATASET -

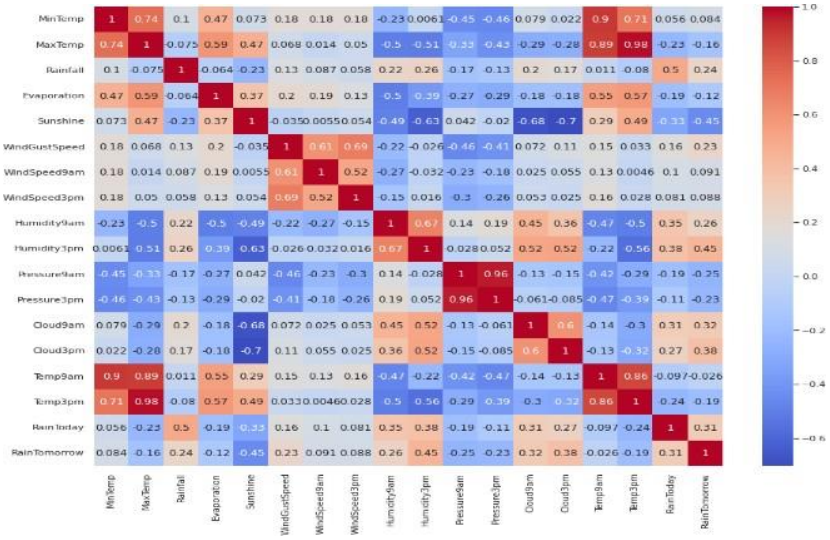
1. Selecting the appropriate dataset from Kaggle or GitHub 2. Downloading the API credentials 3. Setting up the Collab Notebook 4. Downloading the dataset 5. Reading the csv file.

3. *EDA Concluding Remark.*

Exploratory data analysis is an approach to analyzing data sets to summarize their key characteristics, often using statistical graphics and other data visualization methods. EDA is not about whether or not statistical models are used, but primarily about finding out what the data can tell us beyond formal modeling or hypothesis testing. In EDA, we are primarily concerned with data exploration and visualization.







Data visualization is defined as a graphical representation that contains the information and data. Visual elements such as charts, diagrams, maps, etc. are used.

Data visualization techniques provide an accessible way to identify and understand trends, outliers, and patterns in data. Some of the data visualization methods are: Line chart, bar chart, distribution chart, common chart, scatter plot, violin plot, count chart, pair chart, etc.
Data exploration refers to the first step of data analysis, where data analysts use data visualization and statistical techniques to describe the characteristics of data sets, such as size, quantity, and precision, to better understand the nature of the data. The steps of data exploration are

1. Identifying variables and data types.
2. Analyzing basic metrics.
3. Non-graphical univariate analysis.
4. Graphical univariate analysis.
5. Bivariate analysis.
6. Variable transformations.
7. Treatment of missing values.
8. Treatment of outliers.

4. Pre-Processing Pipeline.

Data preprocessing in machine learning refers to the technique of preparing (cleaning and organizing) raw data to make it suitable for building and training machine learning models.
Data processing is divided into four stages: Data cleaning, data integration, data reduction, and data transformation.
The different types of data preprocessing are -
a. Aggregation b. Sampling c. Dimensionality reduction d. Feature subset selection e. Feature generation f. Discretization and binarization g. Variable transformation.
In the preprocessing pipeline, the EDA dataset is highly imbalanced. This data leads to biased results. Therefore, the dataset is adjusted with respect to the RainTomorrow attribute using a resample from Sklearn. Thus, we need to deal with the class imbalance.
Feature Selection - A process to reduce the number of input variables when developing a predictive model. We considered the attribute "RainTomorrow" as our dependent variable (Y) since it is what we are predicting, and considered all other attributes except "Date", "Evaporation", and "Sunshine" as independent variables (X) since Date has no effect on our model and Evaporation and Sunshine have a very high percentage of missing values.

Dealing with missing values - EDA has many attributes with a high percentage of missing values, which could lead to poor accuracy of our model. We used Simple Imputer from ski-kit learn to fill the missing values with the most common values in each column.

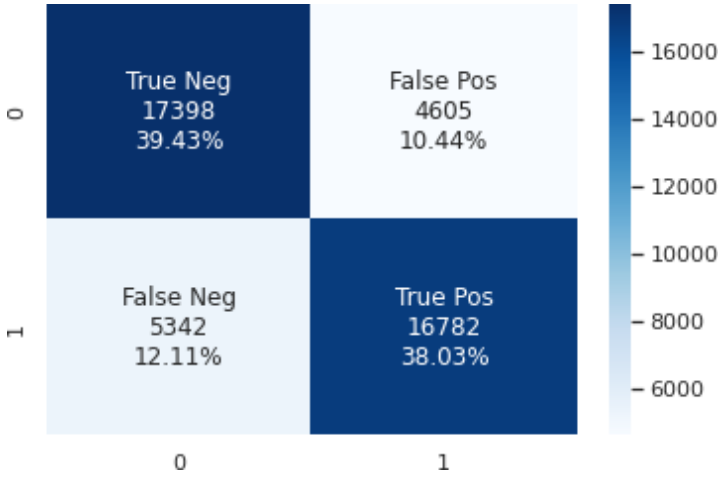
Categorical Data Coding - This is data with two or more categories, but no order of its

own. We have a few categorical features - location, WindGustDir, WindDir9am, WindDir3pm, RainToday. Now it becomes more complicated for machines to understand and process texts than numbers, because the models are based on mathematical equations and calculations. Therefore, the categorical data should be encoded using the label encoder of ski-kit learn.
Feature Scaling - The dataset contains features with widely varying magnitudes and ranges. In general, most machine learning algorithms use Euclidean distance between two data points in their calculations, which is a problem. The features with high values will have much more weight in the distance calculations than features with low values. To reduce this effect, we need to bring all features to the same scale. This can be achieved by scaling. Use the default sk-learn scaler to scale all data points in a given range.

5. *Building Machine Learning Models.*

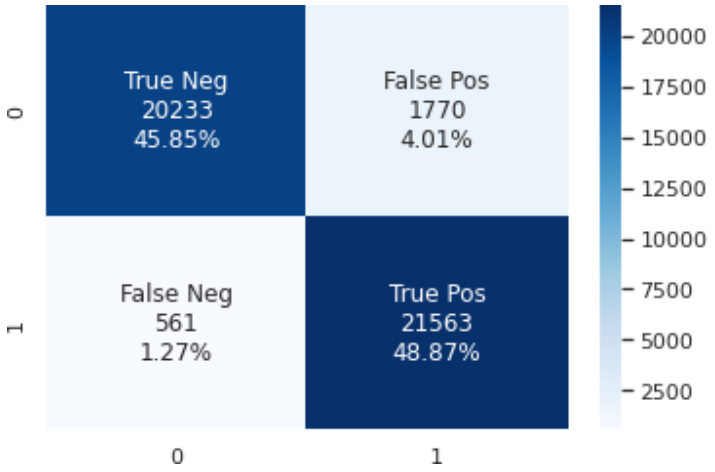
Some of the basic machine learning models and algorithms are implemented from scratch in Python. The goal of this project is not to create the most optimized and computationally efficient algorithms possible, but rather to make the inner workings of these algorithms transparent and accessible. What are machine learning models? A machine learning model is a data file that has been trained to recognize certain types of patterns. You train a model on a set of data and give it an algorithm to use to infer and learn from that data.
The steps to follow -
1. Splitting the dataset into a training dataset and a test dataset.
We split our dataset into a training set (80%) and a test set (20%) to train and test the rainfall prediction models.
2. Training and testing
Different classifiers to predict rainfall using the dataset.

a. Logistic regression - falls under the supervised learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. In simple words, the dependent variable is binary and the data is coded as either 1 (representing success/yes) or 0 (representing failure/no).

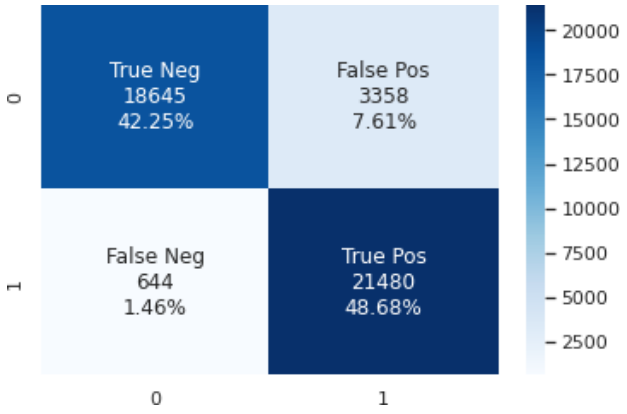


b. Random-Forest - A random forest is a meta estimator that fits a number of decision tree

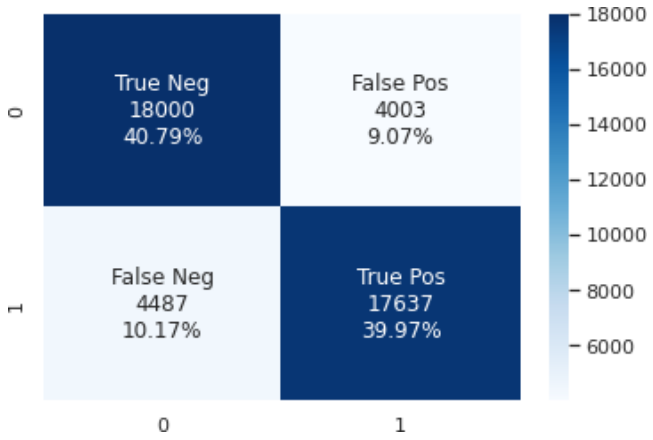
classifiers on various sub-samples of the dataset and usesaveraging to improve the predictive accuracy and control over-fitting.



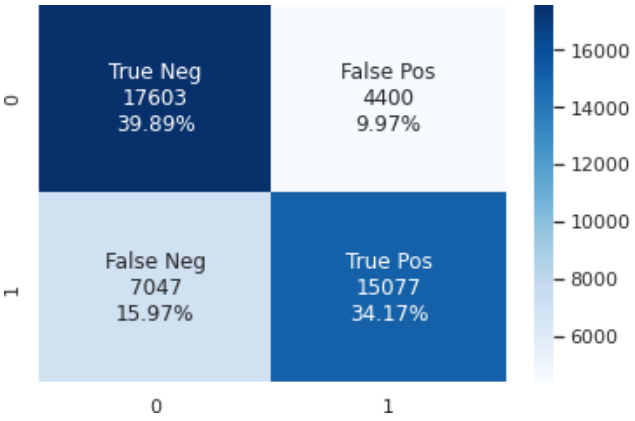
c. Decision Trees - It falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. It can represent Boolean function on discrete attributes using the decision tree.



d. LightGBM - is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage. It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfils the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks.



e. Naive Bayes - classifier assumes all the features are independent to each other. A Gaussian Naive Bayes algorithm is a special type of NB algorithm.It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e., normal distribution.



6. *Concluding Remarks.*

Accuracy Comparison - Accuracies of all the models built and plotted the same using a bar plot.