DATA SCIENCE : Rainfall Prediction Using Machine Learning

Date : 16th October, 2022

Author : Rajesh K (1843)

1. *Problem Definition.*

Climate change is the biggest issue all over the world. It is an important aspect of human life so, accurate prediction should be done as much as possible. Climatic changes is the biggest issue happening all over the world. Scientists are working on, to detect the patterns in climate change as it affects the economy in production to infrastructure. Making prediction on rainfall cannot be done by the traditional way, so scientists and analysist are using machine learning and deep learning concepts to find out the pattern for rainfall prediction. Heavy and irregular rainfall can have many impacts like damage of property, crops and farms - so a better forecasting model is essential for an early warning that can minimize risks to life and property and agricultural farms in a better way. The main objective is to help farmers and also water resources that can be utilized efficiently.

2. *Data Analysis.*

Context – Predict next day rain by training classification models on the target variable Rain tomorrow.

Content -This dataset contains about 10 years of daily weather observations from many locations across Australia.

Rain Tomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.
Source & Acknowledgements -

Observations were drawn 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 –

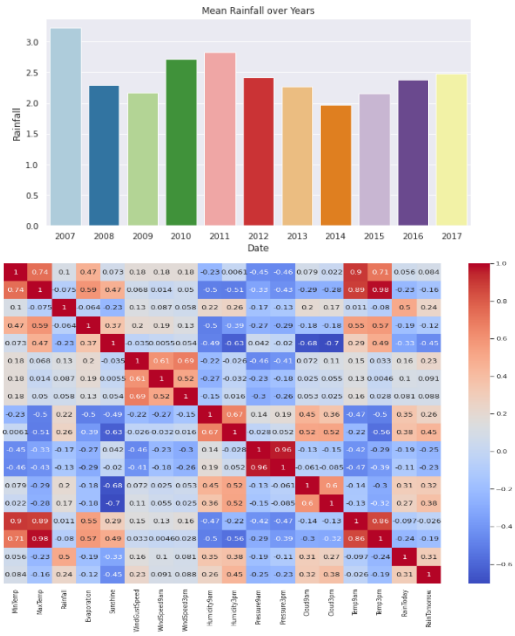
1. NumPy:
   It is the fundamental package for scientific computing in Python. A Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
2. Pandas:
   Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. A fundamental high-level building block for doing practical, real-world data analysis in Python. The most powerful and flexible open-source data analysis/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 hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, web application servers, and various graphical user interface toolkits.
4. Seaborn :
   It is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.
5. SKlearn
   Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

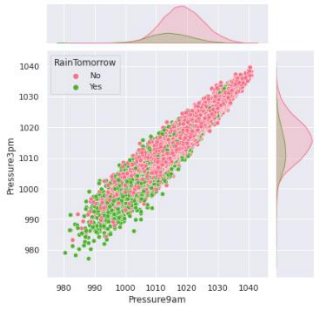
LOADING DATASET –

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

3. *EDA Concluding Remark.*

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Rather than statistical models used or not, primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. In EDA we primarily do Data Exploration and Visualization.





Data visualization is defined as a graphical representation that contains the information and the data. By using visual elements like charts, graphs, maps. etc. Data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data. Some of the Data Visualization methods are: Line plot, Bar plot, Distribution plot, Joint plot, Scatter plot, Violin Plot, Count plot, Pair Plot etc.

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data. The steps in Data Exploration are–

1. Identification of variables and data types.
2. Analyzing the basic metrics.
3. Non-Graphical Univariate Analysis.
4. Graphical Univariate Analysis.
5. Bivariate Analysis.
6. Variable transformations.
7. Missing value treatment.
8. Outlier treatment.

4. *Pre-Processing Pipeline.*

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.
The Data Processing is divided into four stages – Data cleaning, Data integration, Data reduction and Data transformation.
The various types of Data preprocessing are –

a. Aggregation
b. Sampling
c. Dimensionality Reduction
d. Feature Subset Selection
e. Feature Creation
f. Discretization and Binarization
g. Variable Transformation.

In Pre-processing pipeline- the EDA dataset is highly imbalanced. These data results in biased results. So, the dataset with respect to “RainTomorrow” attribute with the help of resample from sklearn. So, we have to deal with Class Imbalance.

Feature selection- A process of reducing the number of input variables when developing a predictive model. Considered the attribute “RainTomorrow” as our dependent variable (Y) as it is what we predict and have considered all the remaining attributes except “Date”, “Evaporation”, “Sunshine” as Independent Variables (X) because Date doesn’t affect our model and Evaporation and Sunshine have very high percentage of missing values.

Dealing with Missing values - EDA has many attributes contain high percentage of missing values which could result in bad accuracy of our model. Used Simple Imputer from ski-kit learn to fill the missing values with most frequent values in respective columns.

Encoding Categorical data - It is one which has two or more categories, but there is no intrinsic ordering to the categories. We have a few categorical features – Location, WindGustDir, WindDir9am, WindDir3pm, RainToday. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, it is to be encoded the categorical data with Label Encoder from ski-kit learn.

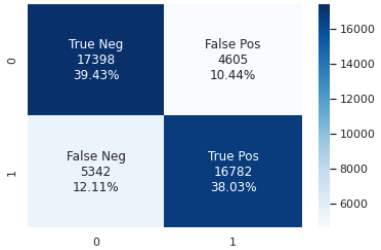
Feature Scaling- The data set contains features with highly varying magnitudes and range. Generally, most of the machine learning algorithms use Euclidean distance between two data points in their computations, which is an issue. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To reduce this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling. Used sk-learn's Standard Scaler to scale all the data points in a certain range.

5. *Building Machine Learning Models.*

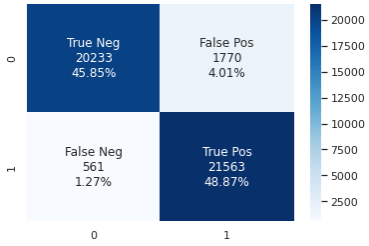
Python implementations do some of the fundamental Machine Learning models and algorithms from scratch. The purpose of this project is not to produce as optimized and computationally efficient algorithms as possible but rather to present the inner workings of them in a transparent and accessible way. What are models of machine learning. A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

The steps to be followed –

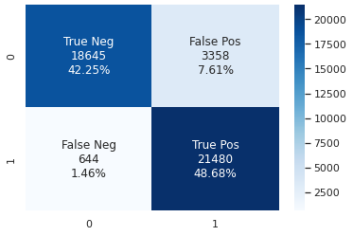
1. Splitting Dataset into Training set and Testing set.
   We have split our dataset into training set (80%) and testing set (20%) to train and test rainfall prediction models.
2. Training and Testing
   Different classifiers to predict rainfall with the dataset.
   a. Logistic Regression – which comes under 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 in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).



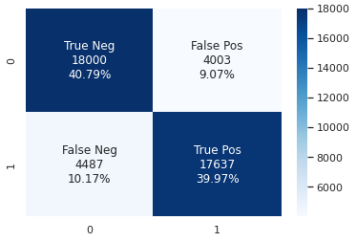
b. Random-Forest - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



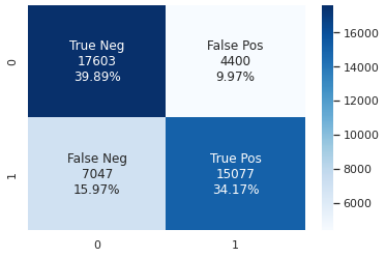
c. Decision Trees - It falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal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