DIY PROJECTS
WINTER 2021

Coding Club
IIT Guwahati

Precipitation Prediction using ML

Expected Time To Finish: 7 Days

By Azal Ahmad Khan

INTRODUCTION

This project aims to create models that can predict whether precipitation will occur or not in LA using common machine learning techniques.

TECHNOLOGIES USED

- **Python, Pandas, Matplotlib, Scikit-learn & Seaborn**

RESOURCES

Part 1: Setting up the project

- **Download Anaconda and Jupyter on your device.**
- **Introduction to Jupyter Notebook**
- **Learn basics of pandas**
- **Learn basics of Matplotlib**
- **Learn basics of Seaborn**
- **Learn basics of Scikit-learn**

Part 2: Data importing and exploration

- **We will use dataset from here.**
- **We will use pandas framework to import the data and perform further analysis on it.**
- **PRCP column in the dataframe will be our target feature in this model. We have to replace all values greater than 0 as 1 (representing precipitation will occur), and values that are equal to 0 representing precipitation will not occur.**

Part 3: Handling class imbalance and missing values

- **In our dataset, there is an imbalance between examples where precipitation occurs or not. Use matplotlib to visualize it.**
- **Most of the ML algos used for classification were designed with the assumption of an equal no. of examples in each case. Therefore we need to balance it.**
- **We will now overbalance the minority class using sklearn.utils.resample.**
  Use this.
- **We will now check for null value**s.
- **If any feature contains many null values, we will drop it.**
- **Now, we will convert the rest of the null values with mode.**

Part 4: Standardizing data and feature selection

- **We will now normalize our data.**
- **Feature selection will be made using the chi-square test.**
  What is the chi-square test for feature selection?
  Read this.
  How will we do this?
  Use SelectKBest and chi2.
- **We will now normalize our data.**

Part 5: Training model using different techniques

- **Split data into test and train datasets.**
- **We can use logistic regression classifier, decision tree classifier, neural networks, etc on training dataset.**
- **Calculate accuracy, precision, recall, F-1 score, and ROC_AUC on the test dataset and visualize it.**
- **Plot confusion matrix using sklearn.**

Part 6: Model Comparison

- **Compare models based on accuracy and ROC_AUC score and visualize it using seaborn.**

Congrats! for Completing this project. Happy Coding !

SUBMISSION

The best projects will get featured in the next edition of **Debugged (Issue 3)**, our highly acclaimed club magazine, so don't forget to submit your projects upon completion!
After completing the project, submit it at:
https://forms.gle/fJjv7TPwQDy5ttPG6