Course Project: Interview Preparation (Data Science- PGC, Internshala Trainings)

Problem Statement: Create a Machine Learning model using various Classification Models to predict rainfall.

Description:

Once upon a time in Sydney, there was a small newspaper company called "The Daily Buzz."
The Daily Buzz was founded many years ago by a group of journalists and entrepreneurs who were passionate about providing accurate and trustworthy news to the people of Sydney. They believed that a strong, independent newspaper was vital to the health and well-being of any community, and so they set about creating one.
The Daily Buzz was a small operation at first, with a limited budget and a small team of dedicated journalists. Despite these challenges, the newspaper quickly gained a reputation for its high-quality journalism and its commitment to the truth.
Over the years, The Daily Buzz grew and evolved. It added new sections, hired more journalists, and expanded its reach. Eventually, the newspaper became one of the most widely read and respected sources of news and information in Sydney.
They were well-known for their interesting articles and local news coverage. However, they were struggling to attract more readers and stay ahead of the competition.
One day, the editor-in-chief had a brilliant idea. They would start a new column called "The Weather Oracle," where they would predict the weather for the coming days. This was an exciting opportunity for the newspaper, as it would provide a unique and useful service to the community.
The editor handpicked a team of experienced meteorologists to work on the new column. They were equipped with state-of-the-art technology and the latest weather forecasting tools. The team worked tirelessly to provide accurate and reliable predictions.
The first edition of "The Weather Oracle" was published on a rainy Monday. The column included detailed predictions for the week ahead, including the temperature, humidity, and chances of precipitation. The response was overwhelmingly positive, with many readers saying they relied on the column to plan their activities for the week.
As the popularity of "The Weather Oracle" continued to grow, the editor-in-chief of The Daily Buzz realized that they could further improve the accuracy of their weather predictions. With this in mind, they decided to hire a machine learning expert to create a machine learning model for rainfall prediction.

The editor-in-chief of "The Daily Buzz.” has hired you as you are an ML expert and he wants you to create an ML model to accurately predict the rainfall in Sydney.
The Editor wants you to use Ensemble methods to get the best accuracy. So, your task is to try various methods and use the one with the best accuracy for prediction.
You have given the weather information of Sydney from 2008 to 2017, you can download the data from Here:

The dataset contains 18 columns:
Date:The date of observation
Location:The common name of the location of the weather station
MinTemp:The minimum temperature in degrees celsius
MaxTemp:The maximum temperature in degrees celsius
Rainfall:The amount of rainfall recorded for the day in mm
Evaporation:The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine:The number of hours of bright sunshine in the day
Humidity 9am:Humidity (percent) at 9am
Humidity3pm:Humidity (percent) at 3pm
Pressure 9am:Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure 3pm:Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud 9 Am:Fraction of sky obscured by cloud at 9am. This is measured in “oktas”, which are a unit of eighths. It records how many eighths of the sky are obscured by clouds. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
Cloud3pm:Fraction of sky obscured by clouds (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
Temp 9am:Temperature (degrees C) at 9am
Temp3pm:Temperature (degrees C) at 3pm
RainToday:Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow:Boolean: next day rain

Classification

Classification techniques such as decision tree classifiers and ensemble methods can be useful in predicting the weather. Weather prediction involves analyzing various factors such as temperature, humidity, wind speed, precipitation, and pressure, among others. These factors can be used as input features for building classification models.

A decision tree classifier is a type of supervised learning algorithm that can be used to predict the weather. It works by recursively splitting the data into subsets based on the most significant input feature, which helps to create a tree-like structure. Each internal node of the tree represents a test on an input feature, and each leaf node represents a prediction. The decision tree classifier can be trained on historical weather data, and once the tree is constructed, it can be used to predict future weather conditions.

Ensemble methods, on the other hand, combine multiple classification models to improve the accuracy of predictions. Random Forest is an example of an ensemble method that can be used to predict the weather. It combines multiple decision trees and uses bagging (bootstrap aggregating) to improve the accuracy of the predictions. In other words, it creates multiple decision trees using different subsets of the data, and then averages the predictions to produce a final prediction.

Another ensemble method is the Gradient Boosting algorithm, which combines multiple weak learners to create a strong learner. It works by sequentially adding decision trees to the model,

with each tree trying to correct the errors of the previous tree. This approach can be used to predict the weather by combining multiple decision trees to create a more accurate model.
Overall, decision tree classifiers and ensemble methods can be useful in predicting the weather. By analyzing historical weather data and using various input features, these techniques can help to create accurate predictions of future weather conditions.

Task:

As an ML expert at The Daily Buzz, you are given the task to create a ML model to predict the rainfall. So, you have to create a Machine Learning Model using various Classification Models including Decision Trees and Ensemble methods, and compare the accuracy of each model.
First Load the data and perform data preprocessing and after data cleaning use decision tree classification and then use Bagging And Boosting techniques along with the Random Forest Classifier then find out the accuracy score of each and create a confusion matrix to evaluate the performance. After completing this, take your best model and write why this model performed better than other models and in what ways you can further improve the accuracy of the selected model.

You have to share the .ipynb file in which you will perform all of the required steps this file should also contain the answer of following question (you can use markdown option to answer these questions in same notebook)

1. Your views about the problem statement?
2. What will be your approach to solving this task?
3. What were the available ML model options you had to perform this task?
4. Which model's performance is best and what could be the possible reason for that?
5. What steps can you take to improve this selected model's performance even further?

Note:Your final deliverable is your Jupyter Notebook, which should contain all the preprocessing steps, visualizations, and the trained model. Additionally, it should also include the answers to the above 5 questions. You can use the markdown option of Jupyter Notebook to accomplish this