Machine Learning Model on Rain Prediction

By Sidharth Khurana

Objective

Problem Statement:

- Create a Machine Learning model using various Classification Models to predict rainfall in Sydney.
- Project is a classification problem that has the following input features:
  - Date
  - Location
  - Temperature(Min, Max, 9am, 3pm)
  - Rainfall
  - Evaporation
  - Sunshine
  - Humidity(9am, 3pm), Pressure(9am, 3pm), Cloud(9am, 3pm)
  - RainToday

Output Feature is RainTomorrow, which we have to predict.

Objective

Stakeholders:

Agriculture & Farmers

- Planning irrigation, sowing, and harvesting schedules.
- Reducing crop loss from unexpected rain.

Water Management Authorities

- Reservoir and dam management.
- Flood risk assessment and early warning systems.

Logistics & Transport Companies

- Optimizing shipping and delivery schedules.
- Preventing delays due to bad weather.

Insurance Companies

- Crop or property insurance planning.
- Estimating risk for weather-related claims.

Approach

Data Preparation:

Cleaned the data by dropping irrelevant columns, one-hot encoding categorical columns, and imputing the missing values with the mean.

Data is prepared for machine learning by splitting it into train and test data.

Approach

Train Baseline Model

Build a baseline classifier decision tree model to check for the performance.

Use bagging with classifier decision tree to improve performance.

Approach

Further models and evaluation

Build random forest model and evaluate the same

Use Grid Search CV to identify the best hyperparameters for random forest

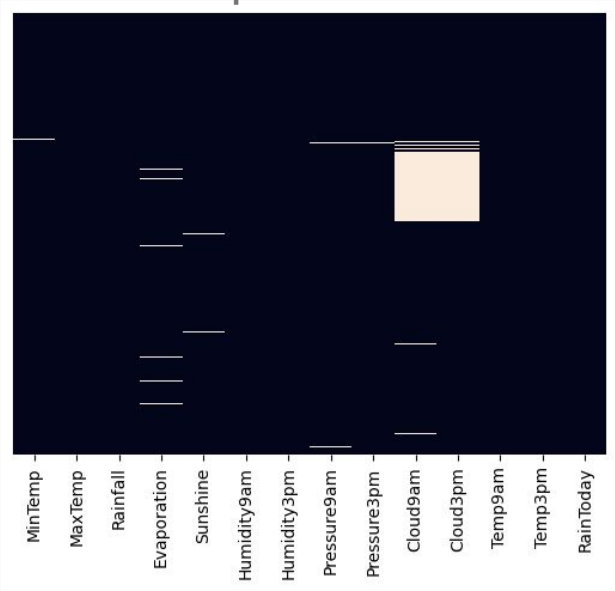
Inference based on the best parameters

Model Summary

1. Datapoints

Description	Count
Data	3,337
% Missing values	16.9%
Train data	2235
Test Data	1102

Heat Map for Null values



Model Summary

Data Cleaning

- Removed Date and Location columns
- One-hot encoding of RainToday and RainTomorrow columns
- Split data into X and y
- Impute the missing values with mean

Model Summary

Models and metrics:

Models used:

- Decision Tree
- Random Forests

Metrics used:

- Mean squared error
- R2 score
- Accuracy score

Model Summary

Parameters of Grid Search CV:

- CV=5
- Scoring: Accuracy

Details for Grid Search:

Model name: Random Forest

Max_features: [4,5,6,7,8,9,10]

Min_samples_split: [2,3,10]

Results(Baseline Model)

Classification Tree:

Model Performance:

Mean Squared Error = 0.194 (test data)

Accuracy score = 0.828 (train data)

Accuracy score = 0.806 (test data)

Results(with Bagging)

Classification Tree + Bagging

Model Performance:

Accuracy score = 1.0 (train)

Accuracy score = 0.831 (test)

Results(Random Forest)

Model Performance:

Accuracy score = 1.0 (train)

Accuracy score = 0.823 (test)

Results(RF+Grid Search CV)

Model Performance:

Accuracy Score = 1.0 (train)

Accuracy Score = 0.8357 (test)

Reference

Code link:

https://github.com/oxerz8/IIT-Roorkee-Capstone-Project-Sydney-Rain-Prediction

Dataset link:

https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package