Project: Weather Prediction Models

Overview

This report evaluates two predictive models using weather data from Sylhet. The models are designed to predict whether it will rain today (RainToday) and to predict the Relative Humidity. The dataset includes features such as Wind Speed, Specific Humidity, Relative Humidity, Temperature, and Precipitation.

Dataset:

Secondary.

Kaggle_link=
https://www.kaggle.com/datasets/farukalam/weather-timeseries-data-in-sylhet-bangladesh

Models Used

Random Forest Classifier

Purpose : Predicts whether it will rain today (RainToday).
Type : Ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.

Gradient Boosting Regressor

Purpose : Predicts Relative Humidity.
Type : Ensemble method that builds an ensemble of decision trees where each tree tries to correct the errors of the previous ones.

Rain Prediction Model-Random Forest Classifier

Model Parameters
Best Parameters :

max_depth: None (no maximum depth)
min_samples_split: 2 (minimum number of samples required to split an internal node)
n_estimators: 200 (number of trees in the forest)

Cross-Validation Score
Best Cross-Validation Score : 0.8968720628618507

This score indicates that the model has a high generalization ability on unseen data during cross-validation.

Classification Report

Precision :
Class 0.0 (No Rain): 0.88
Class 1.0 (Rain): 0.92
Precision is the ratio of true positive predictions to the total predicted positives.

Recall :
Class 0.0 (No Rain): 0.93
Class 1.0 (Rain): 0.87
Recall is the ratio of true positive predictions to the total actual positives.

F1-Score :
Class 0.0 (No Rain): 0.91
Class 1.0 (Rain): 0.90
F1-Score is the harmonic mean of precision and recall.

Accuracy : 0.9005053695514845
Overall accuracy of the model.

Confusion Matrix :
True Positives (TP) for No Rain (0.0): 1505
False Negatives (FN) for No Rain (0.0): 112
False Positives (FP) for Rain (1.0): 203
True Positives (TP) for Rain (1.0): 1346

Interpretation
The model has a high accuracy of approximately 90%, indicating that it correctly predicts whether it will rain or not in most cases.

The precision and recall scores are also quite high, suggesting that the model is reliable in identifying both classes (rain and no rain).

The confusion matrix shows that the model correctly identifies a large majority of instances, with only a small number of false positives and false negatives.

Humidity Prediction Model-Gradient Boosting Regressor

Model Parameters
Best Parameters :
learning_rate: 0.1
max_depth: 5
n_estimators: 200
Cross-Validation Score

Best Cross-Validation Score : 0.5987156045520742

This score indicates the model's performance during cross-validation. However, it is important to note that for regression tasks, the cross-validation score is often negative due to the metric used (e.g., negative MSE).

Evaluation Metrics

Mean Squared Error (MSE) : 0.6502789862474931
MSE measures the average squared difference between the actual and predicted values.

Root Mean Squared Error (RMSE) : 0.8063987761941936
RMSE is the square root of MSE, providing an error metric in the same units as the target variable.

Mean Absolute Error (MAE) : 0.5748630024776912
MAE measures the average absolute difference between the actual and predicted values.

R-squared : 0.9972945908040782

R-squared indicates the proportion of variance in the dependent variable that is predictable from the independent variables. An R-squared value close to 1 suggests a very good fit.
Interpretation

The R-squared value of 0.9972945908040782 is extremely high, indicating that the model explains a very large proportion of the variance in the Relative Humidity.

However, the MSE, RMSE, and MAE values suggest that while the model fits the data very well, the absolute prediction errors are relatively high.

Conclusion

Rain Prediction Model :

Performance: The Random Forest Classifier performs very well with high accuracy, precision, recall, and F1-scores.

Implications : The model is reliable for predicting whether it will rain, which can be useful for weather forecasting and planning.

Humidity Prediction Model :

Performance : The Gradient Boosting Regressor has a very high R-squared value but relatively high MSE, RMSE, and MAE.

Implications : While the model fits the training data very well, the high prediction errors suggest potential overfitting or other issues. Further investigation and possibly more sophisticated models or feature engineering might be needed to improve generalization.

Improvements

Rain Prediction Model :

The model is already performing well, but further exploration of feature importance and interaction effects could provide deeper insights.
Consider using other ensemble methods like XGBoost or LightGBM to see if they yield better performance.

Humidity Prediction Model :

Investigate potential overfitting by using techniques like regularization, pruning, or reducing the complexity of the model.
Explore additional features or transformations that might capture more variability in the data.
Consider using more advanced regression models like XGBoost Regressor or neural networks for potentially better performance.