Academic Year	Module	Project	Project Type
2	Concepts and Technologies of AI	Final	Individual Report

Predicting Temperature using Solar radiation and Meteorological Data: A Regression Analysis Approach

Student Id : 2408947
Student Name : Rujan Maharjan
Section : L5CG16
Module Leader : Siman Giri
Tutor : Ronit Shrestha
Submitted on : 20/12/2024

Abstract:

Aim: This report aims to predict temperature using regression techniques applied to the Solar Radiation and Meteorological Dataset.

Approach: The analysis makes use of the Solar Radiation and Meteorological Dataset, which includes a comprehensive collection of values for the three primary types of solar radiation and meteorological data. The methodology includes Exploratory Data Analysis, developing regression models with hyperparameter optimization, and selecting key features.

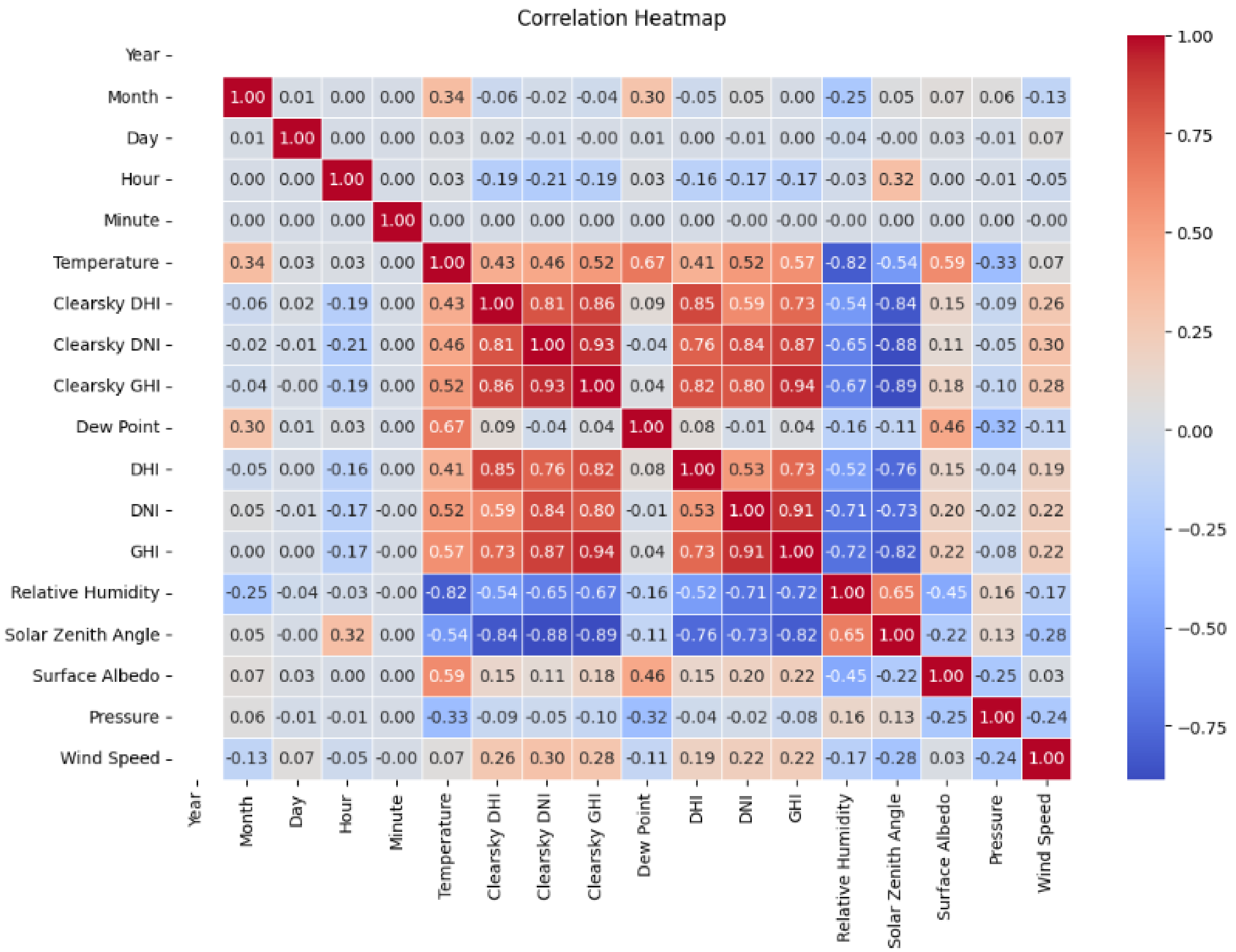
Key Results: The models' performance was evaluated and compared based on R-squared and Mean Squared Error (MSE).

Introduction:

The objective of this project is to predict temperature using meteorological data and serial measurements of solar radiation. The dataset used in this analysis is the Solar Radiation and Meteorological Dataset, obtained from Kaggle (https://www.kaggle.com/datasets/ibrahimkiziloklu/solar-radiation-dataset). It includes data such as Dew Point, Pressure, Month, DHI, GHI, and more. This dataset supports the UNSDG 13: Climate Action (https://sdgs.un.org/goals/goal13), as it provides insights into solar radiation variations and meteorological data, highlighting their impact on climate. The goal of this analysis is to develop a predictive regression model to estimate temperature based on the dataset's features.

Methodology:

Before contructing the model, the data was preprocessed by addressing missing values using isnull().sum(). A column with no values was removed. Next, a correlation matrix was plotted to examine the relationships between temperature and the other features. Based on the correlations, the year, minute, hour, and wind speed were excluded from the dataset. Histograms and box plots were then created for the remaining features. Outliers were handled, followed by re-plotting the box plots and histograms. Two regression models—Linear Regression and Random Forest Regression—were used for this analysis, with an 80-20 train-test split.



	Year	Month	Day	Hour	Minute	Temperature	Clearsky DHI	Clearsky DNI	Clearsky GHI	Dew Point	DHI	DNI	GHI	Relative Humidity	Solar Zenith Angle	Surface Albedo	Pressure	Wind Speed
Year																		
Month		1.00	0.01	0.00	0.00	0.34	-0.06	-0.02	-0.04	0.30	-0.05	0.05	0.00	-0.25	0.05	0.07	0.06	-0.13
Day		0.01	1.00	0.00	0.00	0.03	0.02	-0.01	-0.00	0.01	0.00	-0.01	0.00	-0.04	-0.00	0.03	-0.01	0.07
Hour		0.00	0.00	1.00	0.00	0.03	-0.19	-0.21	-0.19	0.03	-0.16	-0.17	-0.17	-0.03	0.32	0.00	-0.01	-0.05
Minute		0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00
Temperature		0.34	0.03	0.03	0.00	1.00	0.43	0.46	0.52	0.67	0.41	0.52	0.57	-0.82	-0.54	0.59	-0.33	0.07
Clearsky DHI		-0.06	0.02	-0.19	0.00	0.43	1.00	0.81	0.86	0.09	0.85	0.59	0.73	-0.54	-0.84	0.15	-0.09	0.26
Clearsky DNI		-0.02	-0.01	-0.21	0.00	0.46	0.81	1.00	0.93	-0.04	0.76	0.84	0.87	-0.65	-0.88	0.11	-0.05	0.30
Clearsky GHI		-0.04	-0.00	-0.19	0.00	0.52	0.86	0.93	1.00	0.04	0.82	0.80	0.94	-0.67	-0.89	0.18	-0.10	0.28
Dew Point		0.30	0.01	0.03	0.00	0.67	0.09	-0.04	0.04	1.00	0.08	-0.01	0.04	-0.16	-0.11	0.46	-0.32	-0.11
DHI		-0.05	0.00	-0.16	0.00	0.41	0.85	0.76	0.82	0.08	1.00	0.53	0.73	-0.52	-0.76	0.15	-0.04	0.19
DNI		0.05	-0.01	-0.17	-0.00	0.52	0.59	0.84	0.80	-0.01	0.53	1.00	0.91	-0.71	-0.73	0.20	-0.02	0.22
GHI		0.00	0.00	-0.17	-0.00	0.57	0.73	0.87	0.94	0.04	0.73	0.91	1.00	-0.72	-0.82	0.22	-0.08	0.22
Relative Humidity		-0.25	-0.04	-0.03	-0.00	-0.82	-0.54	-0.65	-0.67	-0.16	-0.52	-0.71	-0.72	1.00	0.65	-0.45	0.16	-0.17
Solar Zenith Angle		0.05	-0.00	0.32	0.00	-0.54	-0.84	-0.88	-0.89	-0.11	-0.76	-0.73	-0.82	0.65	1.00	-0.22	0.13	-0.28
Surface Albedo		0.07	0.03	0.00	0.00	0.59	0.15	0.11	0.18	0.46	0.15	0.20	0.22	-0.45	-0.22	1.00	-0.25	0.03
Pressure		0.06	-0.01	-0.01	0.00	-0.33	-0.09	-0.05	-0.10	-0.32	-0.04	-0.02	-0.08	0.16	0.13	-0.25	1.00	-0.24
Wind Speed		-0.13	0.07	-0.05	-0.00	0.07	0.26	0.30	0.28	-0.11	0.19	0.22	0.22	-0.17	-0.28	0.03	-0.24	1.00

Figure-1: Correlation heatmap of the columns in dataset

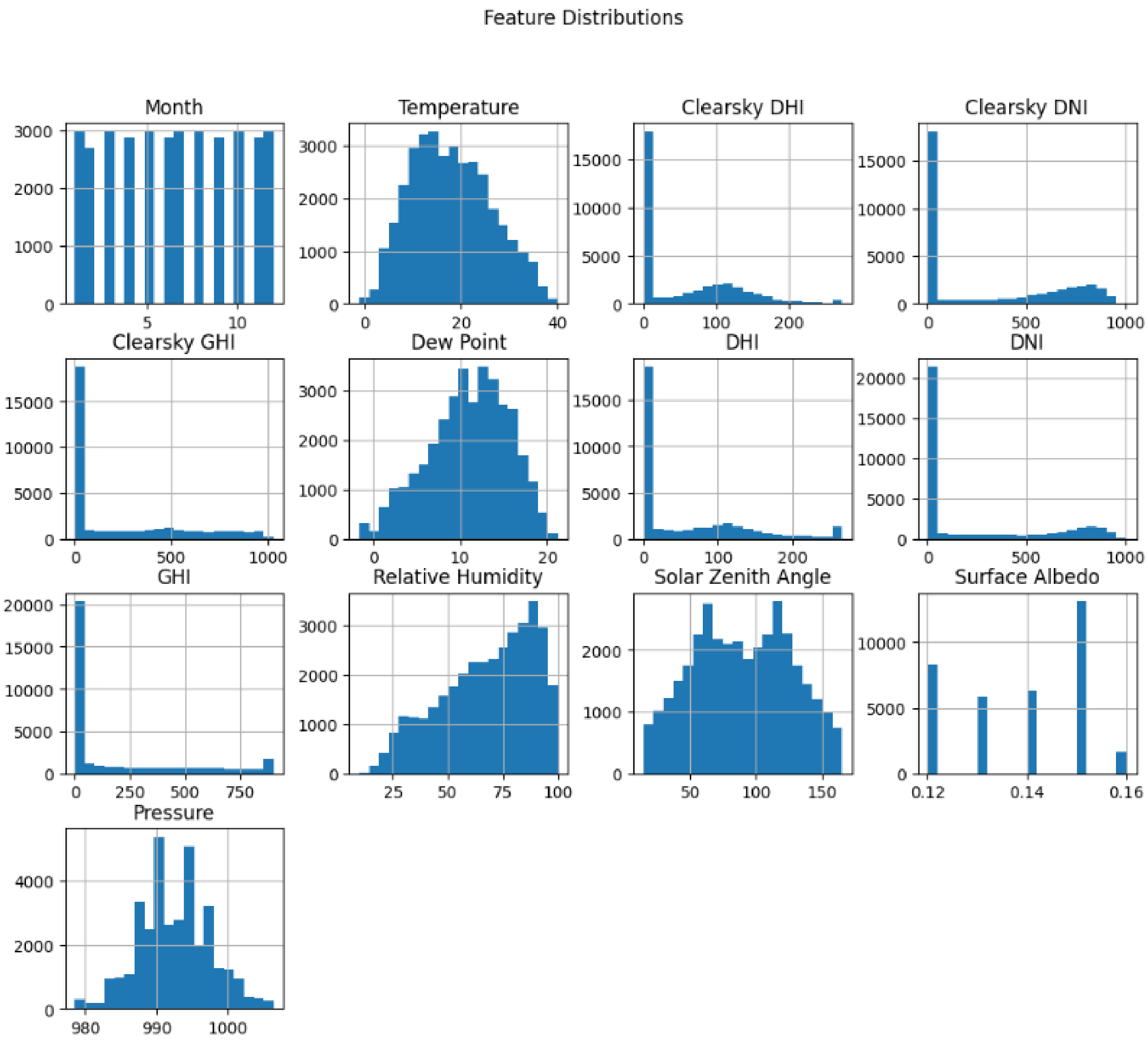


Figure-2: Histogram of the features after handling outliers

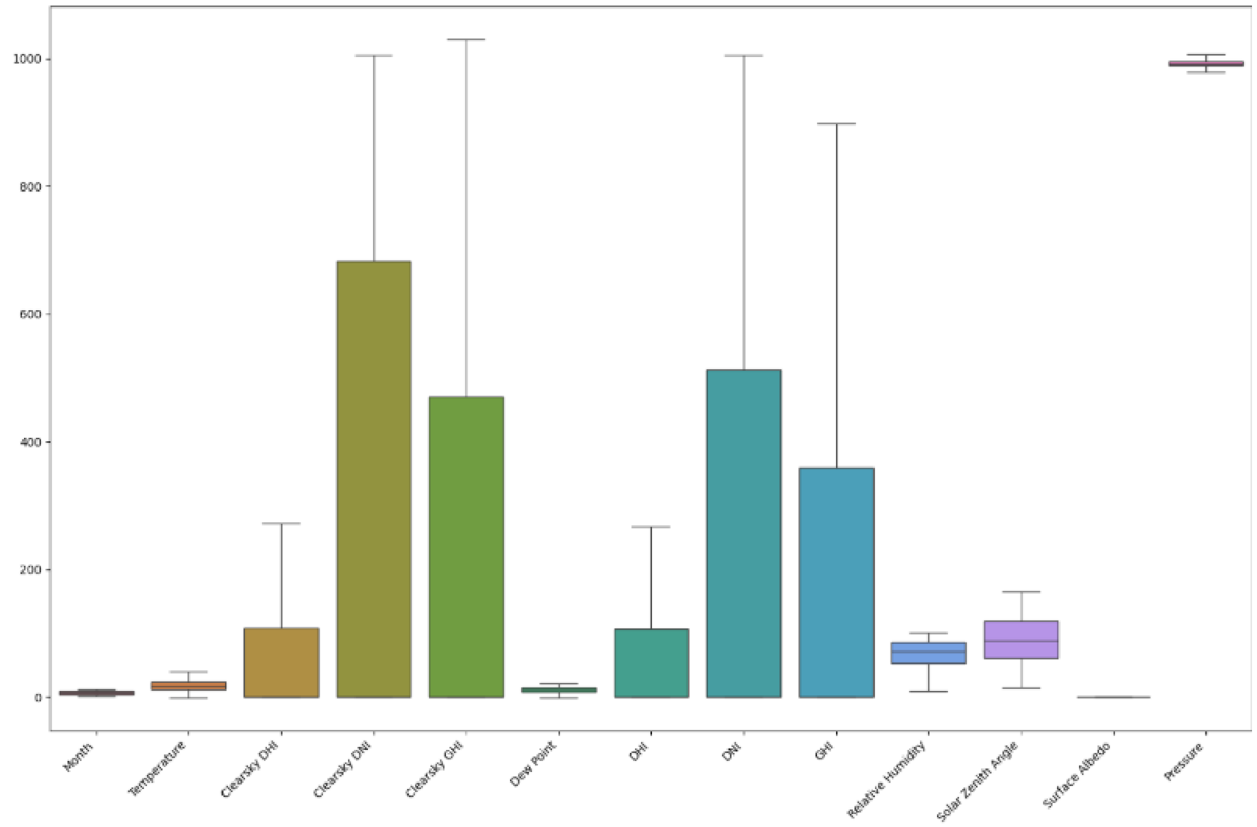


Figure-3: Box plot of the features after handling outliers

Model Evaluation:

The model's accuracy was evaluated using R-squared and Mean Squared Error (MSE), which are standard metrics for assessing regression models. The Linear Regression model resulted in an MSE of 1.5 and an R-squared value of 0.9, while the Random Forest Regression model yielded an MSE of 0.0009 and an R-squared value of 0.9.

Hyper-parameter Optimization:

To improve the performance of model parameter optimization was conducted using GridsearchCV for linear regression and random search CV for random forest regression.

And the list of optimal parameters was Month, clear-sky DHI, clear-sky DNI, clear-sky GHI, DHI, GHI, DNI, and Surface Albedo.

Feature Selection:

There was no use of Recursive Feature Elimination (RFE) this time as there were clear signs of which features were the superior ones. The selected features were Month, clear-sky DHI, clear-sky DNI, clear-sky GHI, DHI, GHI, DNI, and Surface Albedo.

Conclusion:

The model's evaluation on the test dataset was assessed using Mean Squared Error (MSE) and R-squared. The results indicated that Linear Regression performed better with an MSE of 1.5, while the Random Forest model had an MSE of 0.009, although both models showed similar R-squared values, each exceeding 0.9. There were no significant challenges with this dataset; the process went smoothly compared to the classification dataset, and data handling and visualization were more efficient. To improve this model, it would be advisable to avoid completing the work the day before the deadline, as the effort required may not justify the time crunch. Additionally, using methods other than Randomized Search CV could be beneficial, as the time it took to get results was over an hour.

Discussion:

The model that performed best according to the evaluation metrics was the linear regression model. The results suggest that the Random Forest Regression model doesn't outperform the linear regression model by a significant margin for this task. After hyper-parameter tuning, the optimal value for the model's parameter was set to 11.8, but this caused the R-squared to decrease to 0.8, down from over 0.9. The outcome was as anticipated, and the linear regression model performed as expected. As for the limitations of the model, I am not yet in a position to identify them, as I haven't explored the full potential of the model. Regarding future suggestions, I will focus on further refining the model's performance.