MARIA KOILALOU *JUSTINE LE BOURG*

Artificial Intelligence

Laboratory 4

Variant 1

Our code implement two differents methods of regression (linear regression and random forest regression) to predict house prices based on the provided dataset. The dataset is provided in csv and opened with the library pandas.

- Importing Libraries: The necessary libraries have been imported at the beginning of the code. It is a good practice to import all the required libraries at the beginning of the code.

- Loading the Dataset: The dataset has been loaded into a pandas DataFrame, which is a widely used library for data manipulation and analysis.

- Encoding Categorical Variables: Categorical variables have been converted to numerical using one-hot encoding or label encoding. This is a crucial step for most machine learning algorithms, as they cannot handle categorical data.

- Splitting the Dataset: The dataset has been split into training, validation, and test sets. Splitting the data into different sets is important to avoid overfitting and to evaluate the model's performance on unseen data.

- Defining Features and Target Variable: The features and target variable have been defined for both models.

- Linear Regression Model: A Linear Regression model has been implemented, trained on the training data, and evaluated using the validation and test sets. The evaluation has been performed using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. These metrics provide insight into the model's performance and help in model selection.

- Random Forest Regression Model: A Random Forest Regression model has been implemented, trained on the training data, and evaluated using the validation and test sets. The evaluation has been performed using the same metrics as the Linear Regression model.

- Comparison of Metrics: The metrics of both models have been compared, and the model with the lower MSE on the test set has been selected.

- Output: The output has been printed, which shows the metrics of both models and which model performed better.

Linear Regression is a statistical method used to establish a linear relationship between a dependent variable and one or more independent variables. It assumes that the relationship between the variables is linear, and it seeks to find the best-fit line that describes this relationship. Linear Regression is a simple and straightforward model that is widely used for prediction and inference.

Random Forest Regression, on the other hand, is a machine learning algorithm that combines multiple decision trees to predict numerical outcomes. It does not assume a linear relationship between the variables, and it can handle nonlinear relationships and interactions between variables. Random Forest Regression is a more complex model that can handle a wide range of data types and is often used for prediction tasks.

The main advantages of Linear Regression are its simplicity and interpretability. It is easy to understand and interpret the coefficients of the linear equation, and it can be used to make predictions and draw inferences about the relationship between variables.

The main advantages of Random Forest Regression are its flexibility and accuracy. It can handle complex relationships between variables, including nonlinear relationships and interactions, and it can be used to predict outcomes with high accuracy. It is also robust to outliers and noise in the data.

In summary, Linear Regression is a simple and interpretable model that is widely used for prediction and inference, while Random Forest Regression is a more complex and flexible model that can handle a wider range of data types and is often used for prediction tasks requiring higher accuracy. The choice of model depends on the specific problem and the nature of the data.

OUTPUT

The output includes the validation set metrics and test set metrics for both the Linear Regression and Random Forest Regression models.

The validation set metrics are used to evaluate the performance of the models during the training phase, where the model is trained on a subset of the data and evaluated on another subset. The test set metrics are used to evaluate the performance of the models on a completely new and unseen dataset.

For the Linear Regression model, the validation set MSE (mean squared error) is 44489199428.204185, indicating that the average squared difference between the predicted values and the actual values in the validation set is quite high. The R-squared value for the validation set is 0.6261790215907379, which indicates that the model explains about 62.6% of the variance in the data.

For the test set, the MSE is even higher at 57147310796.946175, indicating that the model's performance is worse on the unseen data. The RMSE (root mean squared error) is also provided, which is a more interpretable measure of the error in the model's predictions, and is 239055.03717124677. The R-squared value for the test set is 0.6219833733396046, which indicates that the model explains about 62.2% of the variance in the data.

For the Random Forest Regression model, the validation set MSE is much lower at 18053937064.13353, indicating that the average squared difference between the predicted values and the actual values in the validation set is much lower than that of the Linear Regression model. The R-squared value for the validation set is higher at 0.8483015989454936, indicating that the model explains about 84.8% of the variance in the data.

For the test set, the MSE is also lower than that of the Linear Regression model at 28655263317.891224, indicating that the Random Forest Regression model performs better on the unseen data. The RMSE is also provided, which is 169278.65582491853. The R-squared value for the test set is 0.8104518686105269, which indicates that the model explains about 81.0% of the variance in the data.

Overall, the Random Forest Regression model outperforms the Linear Regression model on both the validation set and the test set, as evidenced by its lower MSE and higher R-squared values. The RMSE for the Random Forest Regression model is also lower than that of the Linear Regression model. Therefore, it can be concluded that the Random Forest Regression model is a better choice for predicting numerical outcomes in this particular dataset.