Task 2 : Regression

For this assignment task, you'll work with a dataset that records daily bike rental activity, along with weather conditions, seasonal information, and calendar details. Your goal is to develop a regression model that predicts the `total_users`.

While the data is collected daily, this is not a time series forecasting task. Instead, you'll use time-related features (such as season, weekday, and holiday indicators) as inputs to help explain patterns in bike rental behavior.

Begin by exploring the dataset to identify how features such as season, temperature, humidity, and day type affect rental activity. Apply appropriate preprocessing techniques, including encoding categorical variables and scaling numerical features.

Since the number of features in this task is relatively high, selecting the most relevant features that contribute meaningfully to the regression performance is important. It is recommended to apply statistical methods for selecting high-quality features to improve the model's accuracy and robustness.

Next, experiment with a variety of traditional regression algorithms and select the one that performs best on the training data. Finally, train your model and use it to predict the number of users in the test dataset.

Note: The `total_users` column is the target variable and is missing in the test set. Your model should generate accurate numeric predictions for each day and output them in a CSV file with two columns: `id` and `label`.

Column Name	Description
id	Unique identifier for each record (starting from 1)
date	The specific calendar date (DD-MM-YYYY) of the record
season_id	Categorical identifier for the quarter of the year based on the Gregorian calendar
year	Year of the record (e.g., 0 for 2018, 1 for 2019)
month	Numeric month of the year (1–12)
is_holiday	Indicates if the day is a holiday (1 = Yes, 0 = No)
weekday	Day of the week (0 = Tuesday, 6 = Monday)
is_workingday	Indicates if the day is a working day (1 = Yes, 0 = No)

weather_condition	Categorical weather rating
temperature	temperature in Celsius (continuous variable)
feels_like_temp	apparent temperature (what it feels like)
humidity	humidity level (%)
wind_speed	wind speed (km/h)
total_users	Total number of bike rentals (casual_users + registered_users). This is the target variable

Notes:

1. **Evaluation Metrics**: To assess how well your regression model is performing, consider using the following metrics (you will show these metrics in an in-person session).
    - **Mean Squared Error (MSE)**: Measures the average of the squares of errors — the average squared difference between predicted and actual values.
    - **Root Mean Squared Error (RMSE)**: The square root of MSE, providing error in the same unit as the target variable.
    - **R-Squared (R2 Score)**: Indicates the proportion of the variance in the target variable that is predictable from the input features. A higher R2 value means better model performance.
    - **Mean Absolute Percentage Error (MAPE)**: Measures the accuracy as a percentage of the error in predictions.
    - **Mean Absolute Error (MAE)**: Averages the absolute differences between predicted and actual values.
    -
2. **Data Preparation**: In the Kaggle competition, you will need to submit your predictions in a CSV file that includes two columns:
    - `id`: The unique identifier for each record (starting from 1).
    - `label`: The predicted value of `total_users` for each record.

    Make sure your predictions match the required format before submission.

3. **Feature Selection**: Besides correlation analysis, it is highly recommended to use **p-values** for feature selection. This statistical test helps determine which features are significantly related to the target variable. By doing so, you can filter out irrelevant or weakly correlated features that may negatively affect model performance.

4. **Data Augmentation:** Since the dataset contains real-world data, you may also consider using additional data sources to enhance your model. For example, you could retrieve external information about specific days (e.g., holidays, events, or special weather conditions) that might influence bike rental patterns. Although it's not required, leveraging external data could potentially improve prediction accuracy.

Make sure to experiment with different regression algorithms (e.g., Linear Regression, Decision Trees, Random Forest, or Gradient Boosting) and fine-tune the hyperparameters to achieve the best performance on the training data.

To participate in the competition related to this task Click on the competition link: Kaggle Competition.

Good luck with the competition!