**Joint Tech Internship Community Program**

A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Shree Varshana R**

**22CSR190**

**Assignment - 1**

**Dataset Example:**

| **Day** | **Temperature (°C)** | **Humidity (%)** | **Wind Speed (km/h)** | **Rainfall (mm)** | **Weather Condition** |
| --- | --- | --- | --- | --- | --- |
| 1 | 25 | 60 | 15 | 0 | Sunny |
| 2 | 30 | 70 | 10 | 0 | Sunny |
| 3 | 22 | 80 | 20 | 5 | Rainy |
| 4 | 28 | 65 | 25 | 0 | Cloudy |
| 5 | 24 | 75 | 15 | 10 | Rainy |

**Terminologies :**

**1. Feature:**

* **Definition: The inputs to the model that are measurable properties or characteristics to make some predictions.**
* **Example: In this dataset, features include Temperature (°C), Humidity (%), Wind Speed (km/h), and Rainfall (mm).**

**2. Label:**

* **Definition: The prediction that the model makes that is defined as the output variable.**
* **Example: Here, the label is the Weather Condition.**

**3. Prediction:**

* **Definition: The output generated by the model after that is being trained, using some of the features the model predicts the label.**
* **Example: Predicting the weather condition (Sunny, Rainy, Cloudy) based on temperature, humidity, wind speed, and rainfall.**

**4. Outlier:**

* **Definition: It is the data point that significantly deviates from the other data, it doesn’t fit the pattern that rest of the data follow’s.**
* **Example: If there was a day with an exceptionally high temperature or rainfall compared to others in this dataset, that would be an outlier.**

**5. Test Data:**

* **Definition: This dataset is not used during the training period of the model. It is the subset of data used to evaluate the model's performance.**
* **Example: Using additional weather data that the model has not seen during training to test it.**

**6. Training Data:**

* **Definition: The dataset used to train the model.**
* **Example: The provided data of days' temperature, humidity, wind speed, rainfall, and weather conditions can be used as training data.**

**7. Model:**

* **Definition: The mathematical representation created during training that makes predictions based on input features. Simply, it shows the relationship between the features and the label.**
* **Example: A classification model predicting weather conditions.**

**8. Validation Data:**

* **Definition: A subset of data used to fine-tune hyperparameters and prevent overfitting.**
* **Example: A separate set of weather data used to validate the model during training.**

**9. Hyperparameter:**

* **Definition: These are the parameters set before the learning process begins, like the learning rate or the number of epochs.**
* **Example: Learning rate and the number of epochs.**

**10. Epoch:**

* **Definition: It is the one complete pass through the entire training dataset.**
* **Example: Running the model through all the weather records once.**

**11. Loss Function:**

* **Definition: A function that measures how well the model's predictions match the actual data.**
* **Example: Cross-Entropy Loss could be used to measure the difference between predicted and actual weather conditions.**

**12. Learning Rate:**

* **Definition: A hyperparameter that controls the step size during the optimization process.**
* **Example: How much the model's weights are adjusted with each update.**

**13. Overfitting:**

* **Definition: When the model learns the training data too well, including its noise and outliers, and so performs poorly on new data.**
* **Example: If our weather condition prediction model only works well on the given dataset but not on new data.**

**14. Underfitting:**

* **Definition: When the model is too simple to capture the underlying pattern in the data.**
* **Example: A model that fails to capture the relationship between temperature, humidity, wind speed, rainfall, and the weather condition.**

**15. Regularization:**

* **Definition: Techniques prevent overfitting by adding a penalty to the loss function.**
* **Example: Adding a regularization term to the loss function.**

**16. Cross-Validation:**

* **Definition: A technique to evaluate the model by splitting the data into multiple subsets and training/testing the model on different combinations of these subsets.**
* **Example: k-fold cross-validation on the weather data.**

**17. Feature Engineering:**

* **Definition: This involves creating new features from existing ones to improve model performance.**
* **Example: Creating a new feature "Dew Point" from the given data.**

**18. Dimensionality Reduction:**

* **Definition: Methods employed to decrease the number of variables in the dataset.**
* **Example: Reducing the features in the weather dataset to only the most impactful ones.**

**19. Bias:**

* **Definition: The error that arises from using an overly simplified model to represent a complex real-world problem.**
* **Example: Assuming a linear relationship between temperature and weather conditions when it might be more complex.**

**20. Variance:**

* **Definition: The error that occurs because the model is overly responsive to minor variations in the training data.**
* **Example: A model that performs well on the training data but poorly on test data due to overfitting.**