|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | A logo of a company  Description automatically generated A close-up of a logo  Description automatically generated | | **Joint Tech Internship Community Program**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)**  **Name: JAYAVIJAAY CJ**  Let's create an example dataset using household phone products. Here's a dataset:   | **Product ID** | **Brand** | **Model** | **Price ($)** | **Rating** | **Units Sold** | **Discount (%)** | | --- | --- | --- | --- | --- | --- | --- | | 1 | Samsung | Galaxy S21 | 799 | 4.6 | 500 | 10 | | 2 | Apple | iPhone 12 | 999 | 4.8 | 450 | 5 | | 3 | Google | Pixel 5 | 699 | 4.5 | 300 | 15 | | 4 | OnePlus | 9 Pro | 969 | 4.4 | 250 | 20 | | 5 | Motorola | Edge | 699 | 4.3 | 150 | 25 |   **Terminologies Explained:**   1. **Feature**: Individual measurable properties or characteristics used as inputs to the model.    * Example: Product ID, Brand, Model, Price, Rating, Units Sold, Discount. 2. **Label**: The output variable that the model aims to predict.    * Example: Units Sold (if we are trying to predict how many units will be sold based on the other features). 3. **Prediction**: The output of the model after it has been trained, given new input data.    * Example: Predicting the units sold for a new phone model based on its price, rating, etc. 4. **Outlier**: A data point that deviates significantly from the rest of the data.    * Example: If there was a phone with a price of $5000 in this dataset. 5. **Test Data**: A subset of the dataset used to assess the performance of the model.    * Example: The last two records in the table (Product ID 4 and 5) can be used as test data. 6. **Training Data**: The subset of the dataset used to train the model.    * Example: The first three records in the table (Product ID 1, 2, and 3) can be used as training data. 7. **Model**: The mathematical representation of the relationship between features and labels.    * Example: A linear regression model predicting units sold based on price and rating. 8. **Validation Data**: A subset of the dataset used to tune the hyperparameters of the model.    * Example: A separate set of records not in the training or test set. 9. **Hyperparameter**: Parameters whose values are set before the learning process begins.    * Example: The learning rate or the number of epochs in a neural network. 10. **Epoch**: One complete pass through the entire training dataset.     * Example: In a neural network, an epoch would mean the model has seen each record in the training data once. 11. **Loss Function**: A function that measures the discrepancy between the predicted and actual labels.     * Example: Mean Squared Error (MSE) used in regression tasks. 12. **Learning Rate**: A hyperparameter that controls how much the model's parameters are adjusted with respect to the loss gradient.     * Example: A learning rate of 0.01. 13. **Overfitting**: When a model performs well on training data but poorly on unseen data.     * Example: If our model predicts the units sold perfectly on training data but fails on test data. 14. **Underfitting**: When a model is too simple to capture the underlying pattern in the data.     * Example: If our model has a low prediction accuracy on both training and test data. 15. **Regularization**: Techniques to prevent overfitting by penalizing large coefficients.     * Example: L1 or L2 regularization. 16. **Cross-Validation**: A technique for assessing how the results of a statistical analysis will generalize to an independent dataset.     * Example: k-fold cross-validation. 17. **Feature Engineering**: The process of using domain knowledge to create features that make machine learning algorithms work.     * Example: Creating a new feature 'Price per Unit' from the price and units sold. 18. **Dimensionality Reduction**: Techniques to reduce the number of input variables.     * Example: Principal Component Analysis (PCA). 19. **Bias**: Error due to overly simplistic assumptions in the learning algorithm.     * Example: High bias could mean the model is too simple (underfitting). 20. **Variance**: Error due to too much complexity in the learning algorithm.     * Example: High variance could mean the model is too complex (overfitting). | |  | |  |

|  |
| --- |
| FFfffff |