|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Generative AI Consortium (Ltd)**  **AI/ML Internship: Assignment 1 (Simple Machine Learning Problem)**  **Name: KARTHYAYENI P** | | **Email: kavikrithi0605@gmail.com**  **Fraud detection dataset** | | | | | | |  |
| **Transaction ID** | | **Transaction Amount (INR)** | **Merchant**  **Type** | **Location** | **Time of Transaction** | **Fraudulent** |
| 1 | | 86400 | Online Retail | USA | Morning | No |
| 2 | | 36000 | Gas Station | Canada | Evening | Yes |
| 3 | | 216000 | Store | UK | Afternoon | Yes |
| 4 | | 10800 | Coffee Shop | Australia | Morning | No |
| 5 | | 57600 | Online Retail | Germany | Night | Yes |
| 6 | | 50000 | Electronics | USA | Afternoon | No |
| 7 | | 25000 | Pharmacy | Canada | Evening | Yes |

FFfffff

**Feature**: Individual measurable properties **(e.g., Transaction Amount, Merchant Type, Location, Time of Transaction)** used as inputs to the model.

**Label**: The output variable that the model aims to predict **(e.g., Fraudulent or Not Fraudulent)**.

**Prediction**: The classification (Yes/No) indicating whether a transaction is likely to be fraudulent based on input features.

**Outlier**: A data point that significantly deviates from the rest of the data **(e.g., a transaction with an unusually high amount compared to others)**.

**Training Data**: The initial dataset used to train the machine learning model.

**Test Data**: Data used to evaluate the model's performance after training.

**Model**: A mathematical representation of patterns found in data, used to make predictions **(e.g., a classification model using logistic regression or random forest)**.

**Validation Data**: A subset of the training data used to tune the model's hyperparameters and evaluate its performance.

**Hyperparameter**: Parameters that are set before the learning process begins, affecting the learning algorithm's behaviour **(e.g., regularization parameter in logistic regression)**.

**Epoch**: One complete pass through the entire training dataset during the learning process (relevant in iterative learning algorithms).

**Loss Function**: A measure of how well the model's predictions match the actual results **(e.g., binary cross-entropy loss in classification tasks).**

**Learning Rate**: A tuning parameter that determines the step size at each iteration while moving toward a minimum of the loss function **(e.g., 0.01 in gradient descent).**

**Overfitting**: When a model learns the detail and noise in the training data to the extent that it negatively impacts the model's performance on new data **(e.g., capturing noise instead of generalizing patterns)**.

**Underfitting**: When a model is too simple to capture the underlying patterns in the training data, resulting in poor performance on new data **(e.g., using a linear model for highly non-linear fraud patterns).**

**Regularization**: Techniques used to prevent overfitting by penalizing large coefficients in the model **(e.g., L2 regularization)**.

**Cross-Validation**: A technique used to evaluate the model's performance by splitting the data into multiple subsets and training the model on different combinations of these subsets.

**Feature Engineering**: The process of selecting and transforming variables when creating a machine learning model to improve its performance **(e.g., creating new features like transaction frequency or amount per time period)**.

**Dimensionality Reduction**: Techniques used to reduce the number of input variables in a dataset while retaining as much information as possible **(e.g., using feature selection based on importance)**.

**Bias**: Error that occurs when a model's predictions are consistently off-target due to erroneous assumptions in the learning algorithm **(e.g., assuming all transactions from certain merchant types are non-fraudulent)**.

**Variance**: Error that occurs when a model is overly sensitive to small fluctuations in the training data, leading to poor generalization to new data **(e.g., a model that performs well on training transactions but poorly on unseen fraudulent patterns due to high variance)**.