Data Science Terminology Quick Reference Sheet
Version 1.0.3 by Joff Thyer
Rivergum Security LLC

Term	Description	Term	Description
Accuracy	how often a classification model correctly predicts outcomes	Machine Learning	A subset of artificial intelligence that enables systems to learn and make predictions from data.
Bias	An error in a model that causes it to consistently predict values away from the true values	Mean Absolute Error (MAE)	A measure of the average absolute differences between predicted and actual values.
Binary Classification	Categorizing data into two categories	Mean Squared Error (MSE)	A measure of the average squared difference between predicted and actual values.
Categorical Data	Data that represent categories or groups	Mean	The average value of a set of numbers.
Classification	Categorizing data points into predefined classes or groups.	Median	The middle value in a set of sorted numbers.
Clustering	Grouping similar data points together based on certain criteria.	Metrics	Criteria used to assess the performance of a machine learning model, such as accuracy, precision, and recall
Confidence Interval	A range of values used to estimate the true value of a parameter with a certain level of confidence.	Model Evaluation	Assessing the performance of a machine learning model using various metrics.
Confusion Matrix	A table used to evaluate the performance of a classification algorithm.	Multivariate Analysis	Analyzing data with multiple variables to understand relationships between them.
Correlation	A statistical measure that describes the degree of association between two variables.	Normalization	Scaling numerical variables to a standard range.
Data Preprocessing	Cleaning and transforming raw data into a format suitable for analysis.	One-Hot Encoding	A technique to convert categorical variables into a binary matrix for machine learning models.
Data Visualization	Presenting data in graphical or visual formats to aid understanding.	Outlier	An observation that deviates significantly from other observations in a dataset.
Decision Tree	A tree-like model that makes decisions based on a set of rules.	Overfitting	A model that performs well on the training data but poorly on new, unseen data.
False Positive	Incorrect positive prediction.	Pandas	A standard data manipulation library for Python for working with structured data.
False Negative	Incorrect negative prediction.	Precision	The ratio of true positive predictions to the total number of positive predictions made by a classification model.
Feature	data column that's used as the input for ML models to make predictions.	Random Forest	An ensemble learning method that constructs a multitude of decision trees and merges them together for more accurate and stable predictions.
Feature Engineering	Creating new features from existing ones to improve model performance.	Random Sample	A sample where each member of the population has an equal chance of being selected.
Gaussian Distribution	A type of probability distribution often used in statistical modeling.	Regression Analysis	A statistical method used for modeling the relationship between a dependent variable and one or more independent variables.
Gradient Descent	An optimization algorithm used to minimize the error in a model by adjusting its parameters.	Root Mean Squared Error (RMSE)	A measure of the difference between predicted and actual values.
Imputation	Filling in missing values in a dataset using various techniques.	Sampling Bias	A bias in the selection of participants or data points that may affect the generalizability of results.
Joint Plot	A type of data visualization in Seaborn used for exploring relationships between two variables and their individual distributions.	Sampling	The process of selecting a subset of data points from a larger dataset.
Joint Probability	The probability of two or more events happening at the same time, often used in statistical analysis.	Sigmoid Function	A mathematical function used in binary classification problems.
Jupyter Notebook	An open-source web application for creating and sharing documents containing live code, equations, visualizations, and narrative text.	Standard Deviation	A measure of the amount of variation or dispersion in a set of values.
Loss Function	Measures the difference between predicted and actual values for a single training example.	Supervised Learning	Learning from labeled data where the algorithm is trained on a set of input-output pairs.
Linear Regression	A statistical method for modeling the relationship between a dependent variable and one or more independent variables.	Time Series Analysis	Analyzing data collected over time to identify patterns and trends.
Logistic Function	A sigmoid function used in logistic regression to model the probability of a binary outcome.	Univariate Analysis	Analyzing the variation of a single variable in the dataset.
Logistic Regression	A statistical method for predicting the probability of a binary outcome.	Unsupervised Learning	Learning from unlabeled data where the algorithm identifies patterns and relationships on its own.
		Validation Set	A subset of data used to assess the performance of a model during training.
		Zero-Shot Learning	Training a model to perform a task without explicit examples.