The Complete Data Science Dictionary

A

- Accuracy: The ratio of correctly predicted observations to the total observations.
- Active Learning: A type of learning in which the algorithm can query the user to obtain the desired outputs at new data points.
- AdaBoost (Adaptive Boosting): An ensemble learning technique that combines the outputs of several weak classifiers to create a strong classifier.
- Adversarial Networks: Neural networks used to generate data, typically used in the context of Generative Adversarial Networks (GANs).
- AIC (Akaike Information Criterion): A measure used in statistics to compare different possible models and determine which one is the best fit for the data.
- Algorithm: A set of rules or instructions given to an AI, computer, or machine to help it learn on its own. Algorithms are used to solve problems or perform tasks.
- Alpha: A parameter in Ridge and Lasso regression that controls the amount of regularization applied to the model.
- ANOVA (Analysis of Variance): A statistical method used to test differences between two or more means.
- Artificial Intelligence (AI): The simulation of human intelligence in machines that are programmed to think like humans and mimic their actions.
- Association Rule Learning: A rule-based machine learning method for discovering interesting relations between variables in large databases.
- Autoencoder: A type of neural network used to learn efficient codings of unlabeled data. An autoencoder learns to compress data (encoding) and then reconstruct it back (decoding).
- Autoregressive Model: A type of random process that is often used to model and predict various types of natural phenomena.
- A/B Testing: A statistical method used to compare two versions of a webpage or app against each other to determine which one performs better.
- Attribute: A variable that represents a feature or characteristic of the data.
- Augmented Reality (AR): An interactive experience where objects in the real world are enhanced by computer-generated perceptual information.
- AutoML (Automated Machine Learning): The process of automating the end-to-end process of applying machine learning to real-world problems.

B

- Bagging (Bootstrap Aggregating): An ensemble technique that improves the stability and accuracy of machine learning algorithms by training multiple models on different subsets of the data and averaging their predictions.
- Bayesian Inference: A method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.
- Bayesian Networks: Probabilistic graphical models that represent a set of variables and their conditional dependencies via a directed acyclic graph.
- Bias: A systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others.
- Big Data: Extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
- Binary Classification: Classification tasks with two possible outcomes, such as spam vs. non-spam or positive vs. negative.
- Bias-Variance Tradeoff: The property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.
- Bayes' Theorem: A mathematical formula used for calculating conditional probabilities.
- Boosting: An ensemble technique that combines the outputs of several weak learners to create a strong learner.
- Bootstrap: A resampling method used to estimate statistics on a population by sampling a dataset with replacement.
- Box Plot: A graphical representation of data that shows the distribution of a dataset based on a five-number summary: minimum, first quartile, median, third quartile, and maximum.
- Bag-of-Words (BoW): A method of text representation in natural language processing where a text is represented as an unordered collection of words, disregarding grammar and word order but keeping multiplicity.
- Bayesian Optimization: A probabilistic model-based optimization technique used to find the maximum or minimum of an objective function that is expensive to evaluate.
- Bias (in Machine Learning): The error introduced by approximating a real-world problem, which may be extremely complicated, by a much simpler model.
- BIC (Bayesian Information Criterion): A criterion for model selection among a finite set of models; it is based on the likelihood function and is closely related to AIC.
- Batch Processing: The processing of data in large groups, or batches, at a specific time.
- Benchmarking: The process of comparing a system's performance against a standard or the performance of other systems.
- Bayesian Nonparametrics: A class of statistical methods that provide flexible models and are particularly useful when the number of parameters is unknown or infinite.
- Backpropagation: A method used in artificial neural networks to calculate the gradient of the loss function with respect to the weights by the chain rule, often used for training deep neural networks.
- Behavioral Analytics: The use of data to understand how users interact with a system or product, often used to optimize user experience and increase engagement.

C

- Categorical Data: Data that can be divided into specific groups or categories.
- Classification: The process of predicting the class or category of a given data point.
- Clustering: A type of unsupervised learning that involves grouping similar data points together.
- Confusion Matrix: A table used to describe the performance of a classification algorithm.
- Correlation: A measure of the relationship between two variables and how they move together.
- Cross-Validation: A statistical method used to estimate the skill of machine learning models.
- Curse of Dimensionality: The phenomenon where the feature space becomes exponentially larger as more features are added to the dataset.
- Customer Segmentation: The process of dividing customers into groups based on common characteristics.
- CART (Classification and Regression Trees): A predictive algorithm used for both classification and regression tasks.
- Content-Based Filtering: A recommendation system technique that uses the features of items to recommend additional items similar to what the user likes.
- Collaborative Filtering: A recommendation system technique that uses user-item interactions to recommend items.
- Chi-Square Test: A statistical test used to determine whether there is a significant association between categorical variables.
- Cost Function: A function that measures the performance of a machine learning model for given data. It quantifies the error between predicted and actual values.
- Correlation Coefficient: A measure that describes the direction and strength of a relationship between two variables.
- Cross-Entropy: A loss function often used in classification tasks to measure the performance of a model whose output is a probability value between 0 and 1.
- Causal Inference: The process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
- Conditional Probability: The probability of an event occurring given that another event has already occurred.
- Covariance: A measure of how much two random variables vary together.
- Cluster Analysis: A set of techniques used to classify objects or cases into relative groups called clusters.
- Cumulative Gain: A metric used to evaluate the performance of classification models based on the gain of true positives.

D

- Data Cleaning: The process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset.
- Data Mining: The practice of examining large databases to generate new information.
- Data Warehousing: The process of collecting, storing, and managing large volumes of data.
- Decision Tree: A decision support tool that uses a tree-like model of decisions and their possible consequences.
- Deep Learning: A subset of machine learning that uses neural networks with many layers (deep networks).
- Dimensionality Reduction: The process of reducing the number of random variables under consideration.
- Domain Knowledge: Expertise and understanding in a specific field or industry relevant to the data being analyzed.
- Descriptive Statistics: Statistical methods that summarize and describe the characteristics of a dataset.
- Dependent Variable: The variable being tested and measured in an experiment.
- Deployment: The process of making a machine learning model available for use in a production environment.
- Density Estimation: A technique used to estimate the probability density function of a random variable.
- Data Visualization: The graphical representation of information and data to understand and communicate insights.
- Dropout: A regularization technique used in neural networks to prevent overfitting by randomly dropping units during training.
- Decision Boundary: A hypersurface that partitions the underlying vector space into two sets, one for each class.
- Dummy Variable: A binary variable created to represent an attribute with two or more categories in regression models.
- Dynamic Time Warping (DTW): An algorithm used to measure similarity between two time series that may vary in speed.
- Dependent Variable: The variable that is being predicted or explained in a study or model.
- Deterministic Model: A model in which the output is fully determined by the parameter values and initial conditions.
- Density-Based Clustering: A clustering method that identifies clusters as dense regions in the data space, separated by sparser regions.

E

- EDA (Exploratory Data Analysis): An approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- ETL (Extract, Transform, Load): A process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse.
- Ensemble Learning: Techniques that create multiple models and then combine them to produce improved results.
- Entropy: A measure of the uncertainty or randomness in a data set.
- Error Rate: The proportion of incorrect predictions made by a model.
- Epoch: One complete pass through the entire training dataset in machine learning.
- Euclidean Distance: The straight-line distance between two points in Euclidean space.
- Exponential Smoothing: A time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component.
- Embedded Methods: Feature selection methods that perform feature selection during the model training process.
- Ensemble Methods: Methods that combine the predictions of multiple models to produce a final prediction.
- Evaluation Metrics: Metrics used to measure the performance of a model, such as accuracy, precision, recall, and F1 score.
- Empirical Risk Minimization (ERM): A principle in statistical learning that aims to minimize the average of the loss function over the training set.
- Evolutionary Algorithms: Optimization algorithms based on the principles of natural selection and genetics.
- Expectation-Maximization (EM) Algorithm: An iterative method to find the maximum likelihood or maximum a posteriori estimates of parameters in statistical models.
- Explanatory Variable: A type of independent variable that is used to explain variations in the dependent variable.

F

- Feature Engineering: The process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data.
- Feature Selection: The process of selecting a subset of relevant features for use in model construction.
- False Positive Rate: The proportion of negative cases that were incorrectly classified as positive.
- False Negative Rate: The proportion of positive cases that were incorrectly classified as negative.
- F1 Score: A measure of a test's accuracy, calculated as the harmonic mean of precision and recall.
- Forecasting: The process of making predictions about the future based on past and present data.
- Fourier Transform: A mathematical transform that decomposes a function (often a time signal) into its constituent frequencies.
- Factor Analysis: A statistical method used to describe variability among observed variables in terms of fewer unobserved variables called factors.
- Feature Importance: A technique used to rank the input features of a predictive model by their importance.
- Filter Methods: Feature selection methods that use statistical techniques to evaluate the importance of features.
- Fine-Tuning: The process of making small adjustments to a pre-trained model to adapt it to a specific task.
- Forward Selection: A stepwise regression method that starts with no variables in the model and adds variables one by one.
- Frequent Pattern Mining: The process of finding regular patterns in large datasets, often used in market basket analysis.
- Fuzzy Logic: A form of many-valued logic in which the truth values of variables may be any real number between 0 and 1, used to handle the concept of partial truth.
- Finite State Machine: A computational model used to design both computer programs and sequential logic circuits.
- Fisher's Linear Discriminant: A method used in statistics and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.
- Feature Map: In the context of convolutional neural networks, a feature map refers to the output of one filter applied to the previous layer.

G

- Gaussian Distribution: Also known as the normal distribution, it is a bell-shaped distribution that is defined by its mean and standard deviation.
- Gradient Descent: An optimization algorithm used to minimize the cost function in machine learning models by iteratively moving in the direction of steepest descent.
- Grid Search: A hyperparameter tuning technique that is used to find the optimal hyperparameters of a model by performing an exhaustive search over a specified parameter grid.
- Gaussian Process: A collection of random variables, any finite number of which have a joint Gaussian distribution. Used in regression and classification tasks.
- Gini Impurity: A measure of impurity used in decision trees to determine the best split. It represents the probability of a randomly chosen element being misclassified.
- Generative Adversarial Networks (GANs): A class of machine learning frameworks where two neural networks, a generator and a discriminator, contest with each other to produce data that is indistinguishable from real data.
- Generalization: The ability of a model to perform well on new, previously unseen data, as opposed to the training data.
- Geospatial Analysis: The process of collecting, displaying, and manipulating imagery, GPS, satellite photography, and historical data, primarily through the use of GIS (Geographic Information System).
- Gradient Boosting: A machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- Gaussian Mixture Model (GMM): A probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions with unknown parameters.
- Graph Theory: A branch of mathematics concerned with the properties of graphs, which are abstract representations of a set of objects where some pairs of the objects are connected by links.
- GroupBy: A method used in data manipulation to split data into groups based on some criteria and then apply a function to each group.
- Gradient Checking: A technique used to verify the correctness of the implementation of backpropagation in a neural network.
- Greedy Algorithm: An algorithm that makes the locally optimal choice at each stage with the hope of finding the global optimum.
- Gaussian Naive Bayes: A variant of the Naive Bayes algorithm that assumes that the features follow a normal (Gaussian) distribution.
- Granger Causality: A statistical hypothesis test for determining whether one time series can predict another.
- Grid Sampling: A technique used in hyperparameter tuning where the parameter space is explored in a structured manner.
- Genetic Algorithms: A class of optimization algorithms inspired by the process of natural selection, often used to find approximate solutions to difficult problems through techniques such as selection, crossover, and mutation.
- Gated Recurrent Unit (GRU): A type of recurrent neural network (RNN) architecture used in deep learning, which is similar to a long short-term memory (LSTM) unit but has fewer parameters.
- Goodness of Fit: A statistical measure that describes how well a model fits the observed data.
- Gaussian Kernel: A function used in various machine learning algorithms, particularly in support vector machines (SVM) and kernelized algorithms, to transform data into a higher-dimensional space.

H,I,J,K

H

- Hyperparameter: A parameter whose value is set before the learning process begins. It is used to control the learning process.
- Hierarchical Clustering: A method of cluster analysis which seeks to build a hierarchy of clusters.
- Heteroscedasticity: The property of a series of random variables where the variability of the variable is unequal across the range of values of a second variable that predicts it.
- Hidden Markov Model (HMM): A statistical model which assumes that the system being modeled is a Markov process with unobserved (hidden) states.
- Holdout Method: A simple cross-validation technique where a dataset is randomly divided into two separate sets, typically a training set and a test set.
- Histogram: A graphical representation of the distribution of numerical data, often used to visualize the frequency distribution of a dataset.
- Heuristics: Techniques designed to solve a problem faster when classic methods are too slow or to find an approximate solution when classic methods fail to find any exact solution.
- Hinge Loss: A loss function used for training classifiers, especially in support vector machines.
- Harmonic Mean: A type of average often used in F1 scores and other statistical measures that require the average of rates or ratios.
- Heatmap: A data visualization technique that shows the magnitude of a phenomenon as color in two dimensions.
- Hyperplane: A flat affine subspace of one dimension less than its ambient space, used in SVMs to separate data into classes.
- Hybrid Model: A model that combines two or more different techniques to improve predictive performance.

I

- Independent Variable: A variable that is manipulated to determine its effects on a dependent variable.
- Imputation: The process of replacing missing data with substituted values.
- Instance-Based Learning: Learning that involves storing and using specific instances rather than inferring a general model.
- Interactive Data Visualization: Visualization methods that allow users to interact with data representations, such as zooming, filtering, and selecting subsets.
- Interval Data: Data measured along a scale in which each point is placed at equal intervals from one another.
- Interpolation: A method of constructing new data points within the range of a discrete set of known data points.
- IQR (Interquartile Range): A measure of statistical dispersion, or how spread out the data is, which is the difference between the third quartile (Q3) and the first quartile (Q1).
- Image Recognition: The process of identifying and detecting an object or a feature in a digital image or video.
- Incremental Learning: A type of learning where the model is capable of learning continuously as new data becomes available.
- Independent Component Analysis (ICA): A computational method for separating a multivariate signal into additive, independent components.
- Inertia: A measure used in clustering to quantify how tightly the clusters are packed.
- Information Gain: A measure of the reduction in entropy or surprise by transforming a dataset and used to build decision trees.
- Instance Segmentation: A type of object detection that identifies each distinct object in an image and segments it.

J

- Jackknife Resampling: A technique for estimating the bias and variance of a statistical estimate.
- Jaccard Index: A statistic used for comparing the similarity and diversity of sample sets, defined as the size of the intersection divided by the size of the union of the sample sets.
- Jittering: The process of adding random noise to data, often used to improve the robustness of machine learning models.
- Joint Probability Distribution: A probability distribution that gives the probability that each of two or more random variables takes at a particular value.
- Jupyter Notebook: An open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text.
- JSON (JavaScript Object Notation): A lightweight data-interchange format that is easy for humans to read and write, and easy for machines to parse and generate, often used in data exchange.
- Jensen-Shannon Divergence: A method of measuring the similarity between two probability distributions, often used in machine learning and statistics.
- Johnson-Lindenstrauss Lemma: A result in mathematics that states high-dimensional data can be projected into a lower-dimensional space while approximately preserving pairwise distances.

K

- K-Means Clustering: A type of unsupervised learning used when you have unlabeled data. The algorithm partitions the data into K clusters, each represented by the mean of the points in the cluster.
- K-Nearest Neighbors (K-NN): A simple, instance-based learning algorithm used for classification and regression, where the output is based on the closest K training examples in the feature space.
- Kernel Trick: A technique used in machine learning algorithms to transform data into a higher-dimensional space, making it easier to classify with a linear separator.
- Kernel Density Estimation (KDE): A non-parametric way to estimate the probability density function of a random variable.
- Kurtosis: A statistical measure used to describe the distribution of observed data around the mean, particularly the "tailedness" of the distribution.
- K-Fold Cross-Validation: A resampling procedure used to evaluate machine learning models on a limited data sample by dividing the data into K subsets and using each subset as test set while the remaining K-1 subsets are used for training.
- K-S Test (Kolmogorov-Smirnov Test): A non-parametric test used to determine if two samples come from the same distribution or if a sample comes from a reference probability distribution.
- Knowledge Discovery in Databases (KDD): The process of discovering useful knowledge from a collection of data, which involves data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.
- Kappa Statistic: A statistic that measures inter-rater agreement for categorical items, correcting for the agreement that could happen by chance.
- Kalman Filter: An algorithm that uses a series of measurements observed over time to estimate unknown variables by minimizing the mean of the squared error.

L,M,N

L

- Label Encoding: The process of converting categorical data into numerical data using a mapping of each category to a unique number.
- Latent Variable: A variable that is not directly observed but is inferred from other variables that are observed.
- Lasso Regression: A type of linear regression that uses shrinkage, where data values are shrunk towards a central point as the mean. Lasso stands for Least Absolute Shrinkage and Selection Operator.
- Learning Rate: A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
- Levenshtein Distance: A string metric for measuring the difference between two sequences, also known as edit distance.
- Linear Discriminant Analysis (LDA): A method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes.
- Logistic Regression: A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome, used for binary classification.
- Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) architecture used in deep learning for tasks that require learning sequences, such as speech recognition and time series forecasting.
- Lift Chart: A graphical representation of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- Likelihood: A function of the parameters of a statistical model that measures the probability of the observed data under specific parameter values.

M

- Machine Learning: A branch of artificial intelligence that involves teaching computers to learn from data.
- Mean Absolute Error (MAE): A measure of errors between paired observations expressing the same phenomenon.
- Mean Squared Error (MSE): A measure of the average of the squares of the errors, which is used to evaluate the performance of a regression model.
- Median Absolute Deviation (MAD): A robust measure of the variability of a univariate sample of quantitative data.
- Monte Carlo Simulation: A statistical technique that allows for the modeling of complex situations by random sampling.
- Multicollinearity: A situation in which several independent variables in a multiple regression model are highly correlated.
- Markov Chain: A mathematical system that undergoes transitions from one state to another according to certain probabilistic rules.
- Multivariate Analysis: The analysis of more than two variables to understand the effect of variables on responses.
- Missing Data: Instances in a dataset where values are not stored (missing values) for certain variables.
- Mutual Information: A measure of the mutual dependence between two variables.

N

- Naive Bayes: A simple yet effective classification algorithm based on Bayes' theorem with an assumption of independence among predictors.
- Normalization: The process of scaling individual samples to have zero mean and unit variance.
- Natural Language Processing (NLP): A field of AI that gives machines the ability to read, understand, and derive meaning from human languages.
- Neural Network: A series of algorithms that mimic the operations of a human brain to recognize relationships in a set of data.
- N-grams: Contiguous sequences of n items from a given sample of text or speech used in text mining and natural language processing.
- Nominal Data: Data that can be categorized but not ordered or ranked.
- Numerical Data: Data that is expressed in numbers and can be used in arithmetic operations.
- Noise: Random variations in data that do not represent true signal or information, often referred to as random error.
- Neural Architecture Search (NAS): The process of automating the design of artificial neural networks, where the goal is to find optimal network architectures.
- Network Analysis: The process of investigating social structures through the use of networks and graph theory.

O,P,Q

L

- Label Encoding: The process of converting categorical data into numerical data using a mapping of each category to a unique number.
- Latent Variable: A variable that is not directly observed but is inferred from other variables that are observed.
- Lasso Regression: A type of linear regression that uses shrinkage, where data values are shrunk towards a central point as the mean. Lasso stands for Least Absolute Shrinkage and Selection Operator.
- Learning Rate: A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.
- Levenshtein Distance: A string metric for measuring the difference between two sequences, also known as edit distance.
- Linear Discriminant Analysis (LDA): A method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes.
- Logistic Regression: A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome, used for binary classification.
- Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) architecture used in deep learning for tasks that require learning sequences, such as speech recognition and time series forecasting.
- Lift Chart: A graphical representation of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- Likelihood: A function of the parameters of a statistical model that measures the probability of the observed data under specific parameter values.

M

- Machine Learning: A branch of artificial intelligence that involves teaching computers to learn from data.
- Mean Absolute Error (MAE): A measure of errors between paired observations expressing the same phenomenon.
- Mean Squared Error (MSE): A measure of the average of the squares of the errors, which is used to evaluate the performance of a regression model
- Median Absolute Deviation (MAD): A robust measure of the variability of a univariate sample of quantitative data.
- Monte Carlo Simulation: A statistical technique that allows for the modeling of complex situations by random sampling.
- Multicollinearity: A situation in which several independent variables in a multiple regression model are highly correlated.
- Markov Chain: A mathematical system that undergoes transitions from one state to another according to certain probabilistic rules.
- Multivariate Analysis: The analysis of more than two variables to understand the effect of variables on responses.
- Missing Data: Instances in a dataset where values are not stored (missing values) for certain variables.
- Mutual Information: A measure of the mutual dependence between two variables.

N

- Naive Bayes: A simple yet effective classification algorithm based on Bayes' theorem with an assumption of independence among predictors.
- Normalization: The process of scaling individual samples to have zero mean and unit variance.
- Natural Language Processing (NLP): A field of AI that gives machines the ability to read, understand, and derive meaning from human languages.
- Neural Network: A series of algorithms that mimic the operations of a human brain to recognize relationships in a set of data.
- N-grams: Contiguous sequences of n items from a given sample of text or speech used in text mining and natural language processing.
- Nominal Data: Data that can be categorized but not ordered or ranked.
- Numerical Data: Data that is expressed in numbers and can be used in arithmetic operations.
- Noise: Random variations in data that do not represent true signal or information, often referred to as random error.
- Neural Architecture Search (NAS): The process of automating the design of artificial neural networks, where the goal is to find optimal network architectures.
- Network Analysis: The process of investigating social structures through the use of networks and graph theory.