The Complete Data Science Dictionary

PART 2

O,P,Q

O

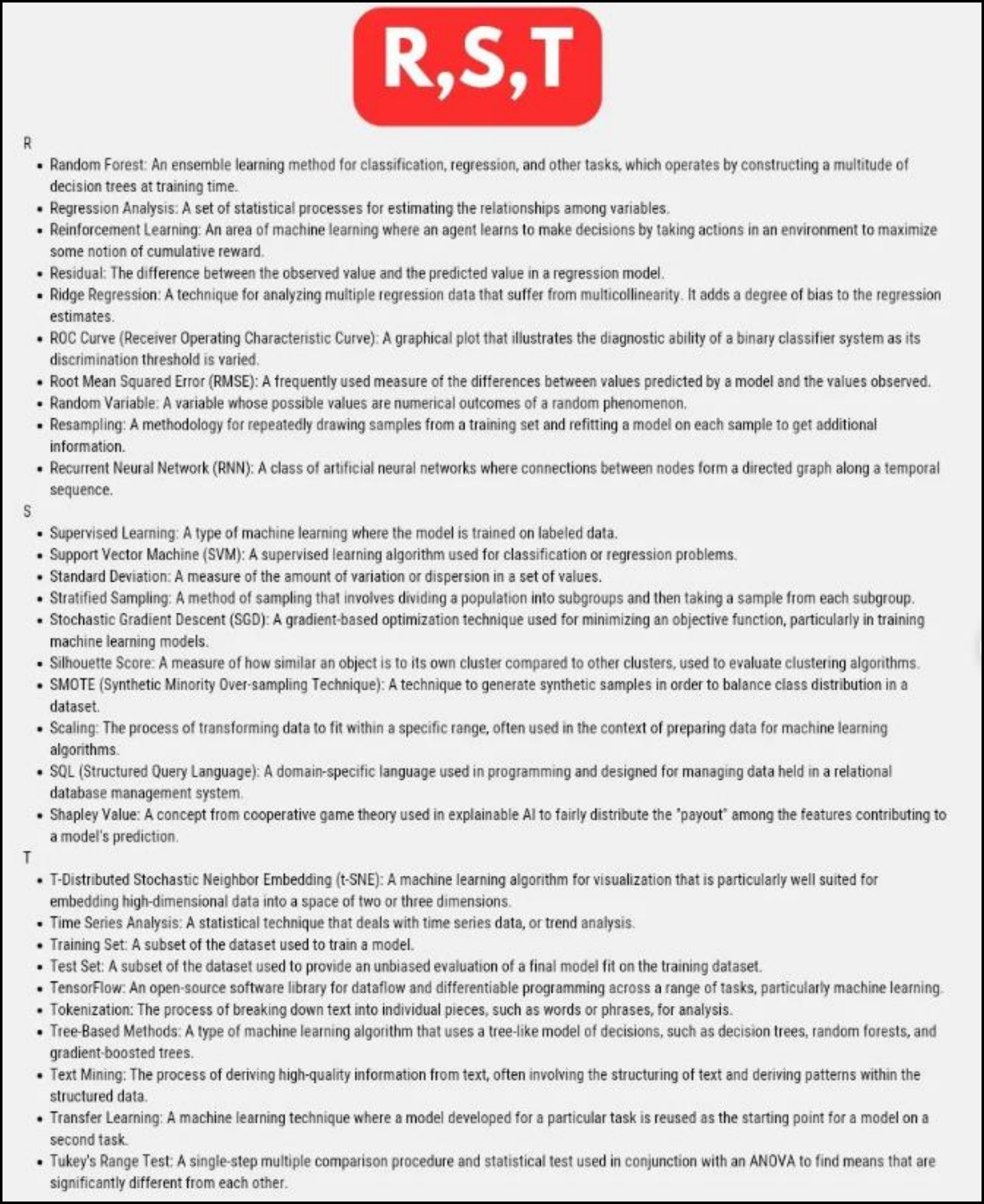
- One-Hot Encoding: A process of converting categorical variables into a binary matrix representation.
- Overfitting: A modeling error in machine learning when a model is too closely aligned to the training data and may perform poorly on new, unseen data.
- Optimization: The process of making a model as effective or functional as possible, often through adjusting parameters to minimize or maximize some objective function.
- Outlier: An observation point that is distant from other observations in data.
- Ordinal Data: Data that can be categorized and ordered but not measured.
- Objective Function: A function that needs to be optimized (maximized or minimized) in a machine learning algorithm.
- Online Learning: A method in machine learning where the model is trained incrementally by processing one observation at a time.
- OverSampling: A technique used to deal with imbalanced datasets by increasing the number of instances in the minority class.
- Optimization Algorithm: A method used to find the optimal parameters for a model, such as gradient descent.
- Ontology: A formal representation of a set of concepts within a domain and the relationships between those concepts.

P

- Principal Component Analysis (PCA): A technique used to emphasize variation and bring out strong patterns in a dataset by transforming it into a set of linearly uncorrelated variables called principal components.
- Precision: A measure of a classifier's exactness, calculated as the number of true positive results divided by the number of all positive results.
- Predictive Modeling: The process of using statistical techniques to create a model that can predict future outcomes based on historical data.
- Probability Distribution: A function that describes the likelihood of obtaining the possible values that a random variable can take.
- p-Value: A measure that helps determine the significance of results in hypothesis testing.
- Parameter: A configuration variable used in machine learning algorithms that is set before the learning process begins.
- Permutation Test: A type of statistical significance test in which the distribution of the test statistic is obtained by calculating all possible values under rearrangements of the labels on the observed data points.
- Partial Dependence Plot (PDP): A graphical representation of the relationship between a subset of the input features and the predicted outcome of a machine learning model.
- Poisson Distribution: A probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space.
- Polychoric Correlation: A technique for estimating the correlation between two theorized normally distributed continuous latent variables, from two observed ordinal variables.

Q

- Quantile: Values that divide a sample of data into equal-sized, consecutive subsets. Examples include quartiles (dividing data into four parts) and percentiles (dividing data into 100 parts).
- Quantitative Data: Data that can be quantified and verified, and is amenable to statistical manipulation.
- Q-Learning: A model-free reinforcement learning algorithm that seeks to learn the value of an action in a particular state.
- Quartile: A type of quantile that divides a dataset into four equal parts.
- Query: A request for information or data from a database.
- Queueing Theory: The mathematical study of waiting lines, or queues, used to predict queue lengths and waiting times.
- Qualitative Data: Data that describes qualities or characteristics and is typically non-numeric.
- Quantization: The process of mapping a large set of input values to a smaller set, often used in signal processing.
- Quasi-Newton Method: A category of algorithms used to find local maxima and minima of functions, which are iterative methods for solving unconstrained nonlinear optimization problems.
- Quality Control (QC): The process of ensuring that products and services meet consumer expectations and regulatory requirements.



R,S,T

R

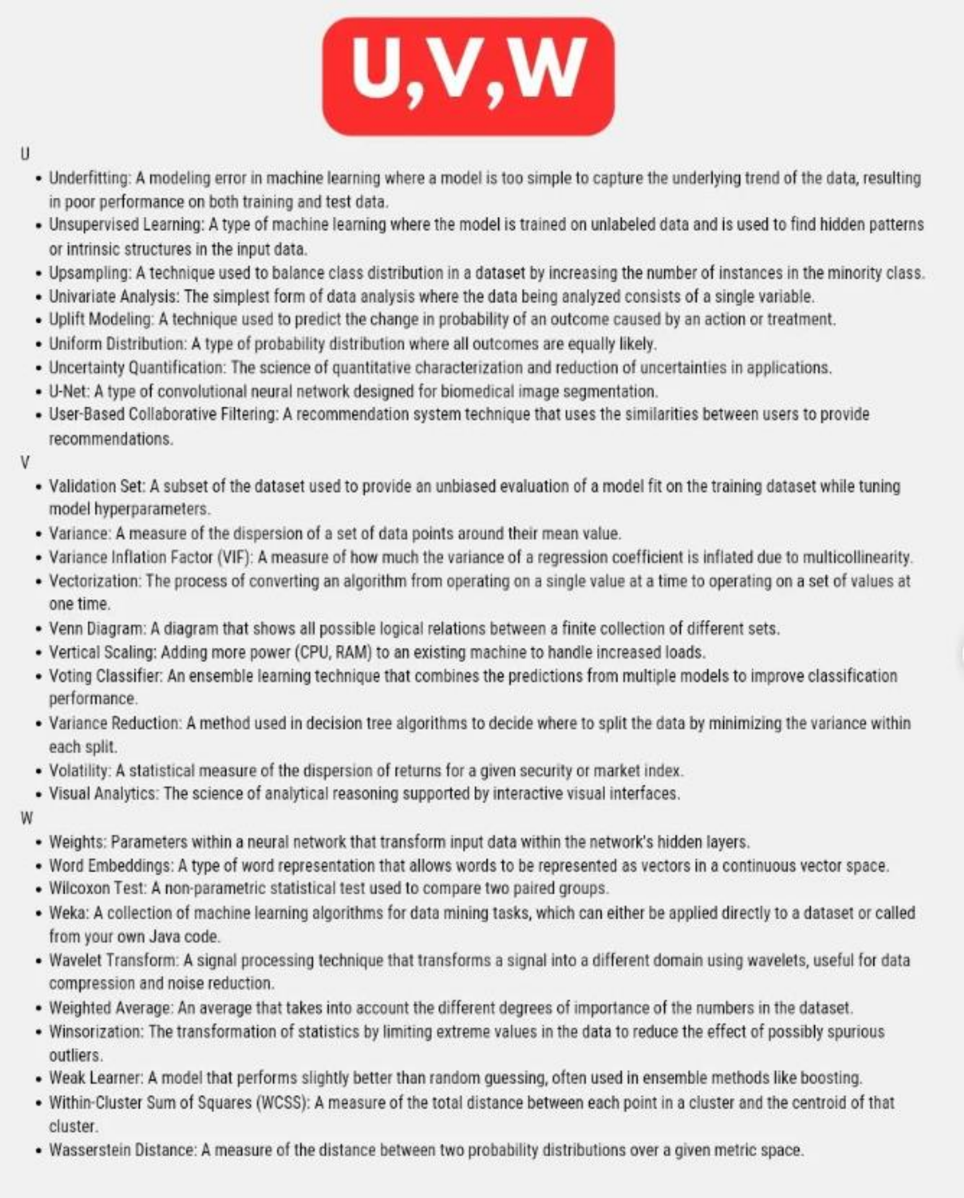
- Random Forest: An ensemble learning method for classification, regression, and other tasks, which operates by constructing a multitude of decision trees at training time.
- Regression Analysis: A set of statistical processes for estimating the relationships among variables.
- Reinforcement Learning: An area of machine learning where an agent learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward.
- Residual: The difference between the observed value and the predicted value in a regression model.
- Ridge Regression: A technique for analyzing multiple regression data that suffer from multicollinearity. It adds a degree of bias to the regression estimates.
- ROC Curve (Receiver Operating Characteristic Curve): A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- Root Mean Squared Error (RMSE): A frequently used measure of the differences between values predicted by a model and the values observed.
- Random Variable: A variable whose possible values are numerical outcomes of a random phenomenon.
- Resampling: A methodology for repeatedly drawing samples from a training set and refitting a model on each sample to get additional information.
- Recurrent Neural Network (RNN): A class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence.

S

- Supervised Learning: A type of machine learning where the model is trained on labeled data.
- Support Vector Machine (SVM): A supervised learning algorithm used for classification or regression problems.
- Standard Deviation: A measure of the amount of variation or dispersion in a set of values.
- Stratified Sampling: A method of sampling that involves dividing a population into subgroups and then taking a sample from each subgroup.
- Stochastic Gradient Descent (SGD): A gradient-based optimization technique used for minimizing an objective function, particularly in training machine learning models.
- Silhouette Score: A measure of how similar an object is to its own cluster compared to other clusters, used to evaluate clustering algorithms.
- SMOTE (Synthetic Minority Over-sampling Technique): A technique to generate synthetic samples in order to balance class distribution in a dataset.
- Scaling: The process of transforming data to fit within a specific range, often used in the context of preparing data for machine learning algorithms.
- SQL (Structured Query Language): A domain-specific language used in programming and designed for managing data held in a relational database management system.
- Shapley Value: A concept from cooperative game theory used in explainable AI to fairly distribute the "payout" among the features contributing to a model's prediction.

T

- T-Distributed Stochastic Neighbor Embedding (t-SNE): A machine learning algorithm for visualization that is particularly well suited for embedding high-dimensional data into a space of two or three dimensions.
- Time Series Analysis: A statistical technique that deals with time series data, or trend analysis.
- Training Set: A subset of the dataset used to train a model.
- Test Set: A subset of the dataset used to provide an unbiased evaluation of a final model fit on the training dataset.
- TensorFlow: An open-source software library for dataflow and differentiable programming across a range of tasks, particularly machine learning.
- Tokenization: The process of breaking down text into individual pieces, such as words or phrases, for analysis.
- Tree-Based Methods: A type of machine learning algorithm that uses a tree-like model of decisions, such as decision trees, random forests, and gradient-boosted trees.
- Text Mining: The process of deriving high-quality information from text, often involving the structuring of text and deriving patterns within the structured data.
- Transfer Learning: A machine learning technique where a model developed for a particular task is reused as the starting point for a model on a second task.
- Tukey's Range Test: A single-step multiple comparison procedure and statistical test used in conjunction with an ANOVA to find means that are significantly different from each other.



U,V,W

U

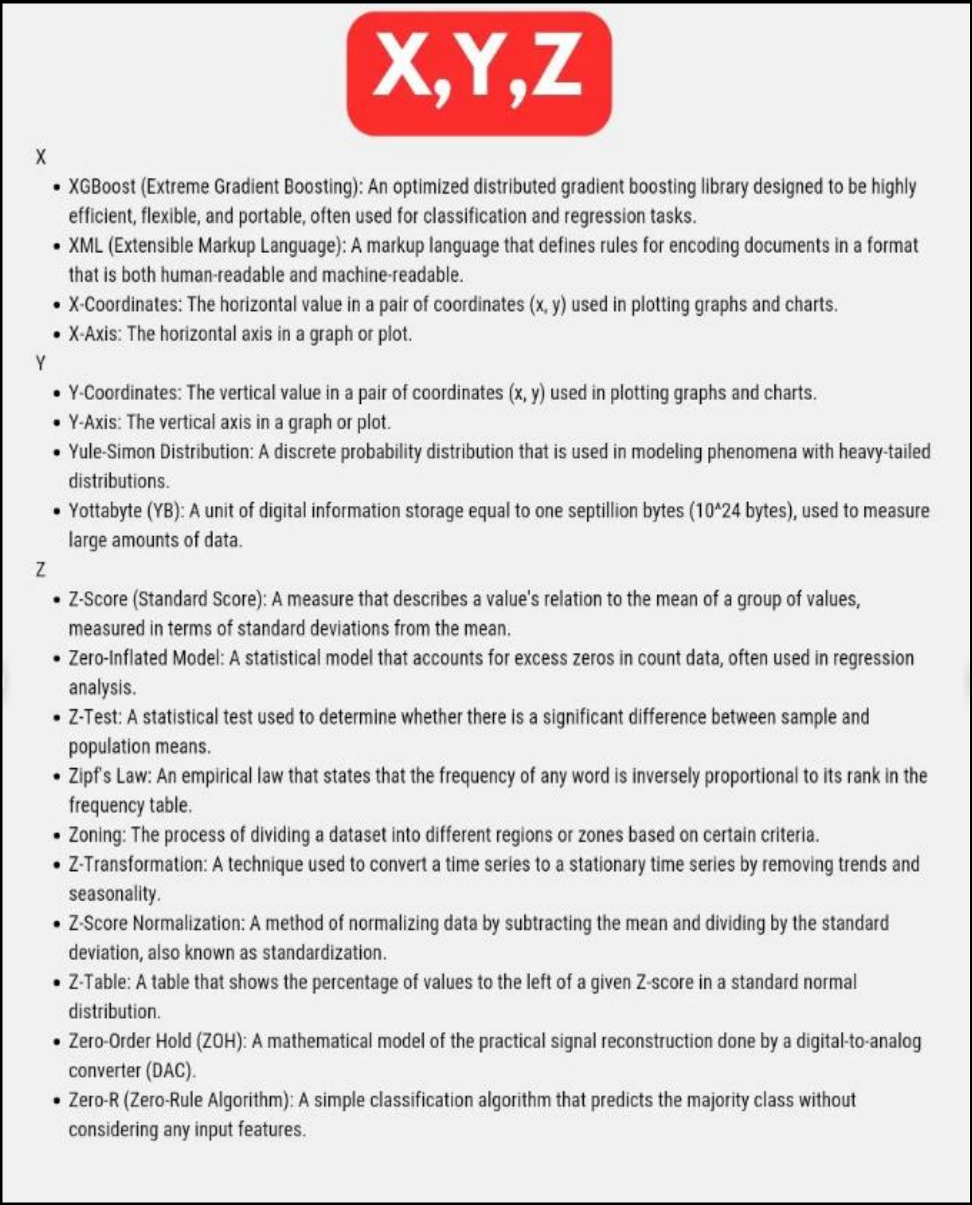
- Underfitting: A modeling error in machine learning where a model is too simple to capture the underlying trend of the data, resulting in poor performance on both training and test data.
- Unsupervised Learning: A type of machine learning where the model is trained on unlabeled data and is used to find hidden patterns or intrinsic structures in the input data.
- Upsampling: A technique used to balance class distribution in a dataset by increasing the number of instances in the minority class.
- Univariate Analysis: The simplest form of data analysis where the data being analyzed consists of a single variable.
- Uplift Modeling: A technique used to predict the change in probability of an outcome caused by an action or treatment.
- Uniform Distribution: A type of probability distribution where all outcomes are equally likely.
- Uncertainty Quantification: The science of quantitative characterization and reduction of uncertainties in applications.
- U-Net: A type of convolutional neural network designed for biomedical image segmentation.
- User-Based Collaborative Filtering: A recommendation system technique that uses the similarities between users to provide recommendations.

V

- Validation Set: A subset of the dataset used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
- Variance: A measure of the dispersion of a set of data points around their mean value.
- Variance Inflation Factor (VIF): A measure of how much the variance of a regression coefficient is inflated due to multicollinearity.
- Vectorization: The process of converting an algorithm from operating on a single value at a time to operating on a set of values at one time.
- Venn Diagram: A diagram that shows all possible logical relations between a finite collection of different sets.
- Vertical Scaling: Adding more power (CPU, RAM) to an existing machine to handle increased loads.
- Voting Classifier: An ensemble learning technique that combines the predictions from multiple models to improve classification performance.
- Variance Reduction: A method used in decision tree algorithms to decide where to split the data by minimizing the variance within each split.
- Volatility: A statistical measure of the dispersion of returns for a given security or market index.
- Visual Analytics: The science of analytical reasoning supported by interactive visual interfaces.

W

- Weights: Parameters within a neural network that transform input data within the network's hidden layers.
- Word Embeddings: A type of word representation that allows words to be represented as vectors in a continuous vector space.
- Wilcoxon Test: A non-parametric statistical test used to compare two paired groups.
- Weka: A collection of machine learning algorithms for data mining tasks, which can either be applied directly to a dataset or called from your own Java code.
- Wavelet Transform: A signal processing technique that transforms a signal into a different domain using wavelets, useful for data compression and noise reduction.
- Weighted Average: An average that takes into account the different degrees of importance of the numbers in the dataset.
- Winsorization: The transformation of statistics by limiting extreme values in the data to reduce the effect of possibly spurious outliers.
- Weak Learner: A model that performs slightly better than random guessing, often used in ensemble methods like boosting.
- Within-Cluster Sum of Squares (WCSS): A measure of the total distance between each point in a cluster and the centroid of that cluster.
- Wasserstein Distance: A measure of the distance between two probability distributions over a given metric space.



X,Y,Z

X

- XGBoost (Extreme Gradient Boosting): An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable, often used for classification and regression tasks.
- XML (Extensible Markup Language): A markup language that defines rules for encoding documents in a format that is both human-readable and machine-readable.
- X-Coordinates: The horizontal value in a pair of coordinates (x, y) used in plotting graphs and charts.
- X-Axis: The horizontal axis in a graph or plot.

Y

- Y-Coordinates: The vertical value in a pair of coordinates (x, y) used in plotting graphs and charts.
- Y-Axis: The vertical axis in a graph or plot.
- Yule-Simon Distribution: A discrete probability distribution that is used in modeling phenomena with heavy-tailed distributions.
- Yottabyte (YB): A unit of digital information storage equal to one septillion bytes (10^24 bytes), used to measure large amounts of data.

Z

- Z-Score (Standard Score): A measure that describes a value's relation to the mean of a group of values, measured in terms of standard deviations from the mean.
- Zero-Inflated Model: A statistical model that accounts for excess zeros in count data, often used in regression analysis.
- Z-Test: A statistical test used to determine whether there is a significant difference between sample and population means.
- Zipf's Law: An empirical law that states that the frequency of any word is inversely proportional to its rank in the frequency table.
- Zoning: The process of dividing a dataset into different regions or zones based on certain criteria.
- Z-Transformation: A technique used to convert a time series to a stationary time series by removing trends and seasonality.
- Z-Score Normalization: A method of normalizing data by subtracting the mean and dividing by the standard deviation, also known as standardization.
- Z-Table: A table that shows the percentage of values to the left of a given Z-score in a standard normal distribution.
- Zero-Order Hold (ZOH): A mathematical model of the practical signal reconstruction done by a digital-to-analog converter (DAC).
- Zero-R (Zero-Rule Algorithm): A simple classification algorithm that predicts the majority class without considering any input features.