Data Science Glossary

A - D

1. Accuracy: The ratio of correct predictions to the total number of input samples.
2. Algorithm: A set of rules or processes to be followed in calculations or other problem-solving operations.
3. ANOVA (Analysis of Variance): A statistical method used to test differences between two or more means.
4. AUC-ROC: A performance measurement for classification problems at various threshold settings.
5. Backpropagation: An algorithm used in artificial neural networks to calculate gradients of loss functions.
6. Bias: The error introduced by approximating a real-life problem with a simplified model.
7. Big Data: Large, complex datasets that traditional data processing software cannot manage.
8. Binary Classification: A task of classifying elements into one of two groups.
9. Boosting: An ensemble technique that attempts to create a strong classifier from a number of weak classifiers.
10. Categorical Variable: A variable that can take on one of a limited, and usually fixed, number of possible values.
11. Classification: A supervised learning task where the output is a category or class label.
12. Clustering: An unsupervised learning task of grouping a set of objects such that objects in the same group are more similar to each other than to those in other groups.
13. Confusion Matrix: A table used to describe the performance of a classification model.
14. Correlation: A statistical measure that expresses the extent to which two variables are linearly related.
15. Cross-Validation: A technique for assessing how the results of a statistical analysis will generalize to an independent dataset.
16. Data Augmentation: Techniques used to increase the amount of data by adding slightly modified copies of already existing data.
17. Data Cleaning: The process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset.
18. Data Frame: A two-dimensional, size-mutable, potentially heterogeneous tabular data structure with labeled axes.
19. Data Mining: The process of discovering patterns in large datasets.
20. Data Normalization: The process of scaling individual samples to have a unit norm.
21. Data Visualization: The graphic representation of data.
22. Data Warehousing: The process of collecting and managing data from varied sources to provide meaningful business insights.
23. Deep Learning: A subfield of machine learning based on artificial neural networks with multiple layers.
24. Dimension Reduction: The process of reducing the number of random variables under consideration.
25. Dropout: A regularization technique where randomly selected neurons are ignored during training.

E - H

26. Ensemble Learning: A process where multiple models are combined to solve a particular computational intelligence problem.
27. Epoch: One complete pass of the entire training dataset through the neural network.
28. ETL (Extract, Transform, Load): A process that involves extracting data from various sources, transforming it, and loading it into a data warehouse.
29. Exploratory Data Analysis (EDA): An approach to analyzing datasets to summarize their main characteristics, often with visual methods.
30. F1 Score: The harmonic mean of precision and recall.
31. Feature Engineering: The process of using domain knowledge to extract features from raw data.
32. Feature Selection: The process of selecting a subset of relevant features for use in model construction.
33. Flask: A micro web framework written in Python, often used to deploy ML models.
34. GAN (Generative Adversarial Network): A class of machine learning frameworks where two neural networks contest with each other in a game.
35. Gradient Descent: An optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent.
36. Heuristic: A technique designed for solving a problem more quickly when classic methods are too slow.
37. Hyperparameter: A parameter whose value is set before the learning process begins.
38. Hypothesis Testing: A statistical method used to make inferences or draw conclusions about a population based on sample data.

I - L

39. Imputation: The process of replacing missing data with substituted values.
40. Inferential Statistics: The process of using data analysis to deduce properties of an underlying distribution of probability.
41. Information Retrieval: The activity of obtaining information system resources that are relevant to an information need.
42. Interquartile Range (IQR): A measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles.
43. K-Means Clustering: A method of vector quantization that aims to partition n observations into k clusters.
44. K-Nearest Neighbors (KNN): A non-parametric supervised learning method used for classification and regression.
45. Kurtosis: A measure of the “tailedness” of the probability distribution of a real-valued random variable.
46. Label Encoding: Converting each value in a column to a number.
47. Lasso Regression: A regression analysis method that performs both variable selection and regularization.
48. Linear Regression: A linear approach to modeling the relationship between a scalar response and one or more explanatory variables.
49. Logistic Regression: A statistical model used to model binary dependent variables.
50. Long Short-Term Memory (LSTM): An artificial recurrent neural network architecture used in deep learning.

M - P

51. Machine Learning: The study of computer algorithms that improve automatically through experience.
52. Mean Absolute Error (MAE): A measure of errors between paired observations expressing the same phenomenon.
53. Mean Squared Error (MSE): The average of the squares of the errors.
54. Multinomial Distribution: A generalization of the binomial distribution.
55. Multivariate Analysis: The physiological and statistical analysis of more than one statistical variable at a time.
56. Naive Bayes: A family of simple “probabilistic classifiers” based on applying Bayes’ theorem.
57. Natural Language Processing (NLP): A subfield of AI concerned with the interactions between computers and human languages.
58. Neural Network: A series of algorithms that endeavors to recognize underlying relationships in a set of data.
29. Normal Distribution: A probability distribution that is symmetric about the mean.
60. Null Hypothesis: A general statement or default position that there is no relationship between two measured phenomena.
61. One-Hot Encoding: A process of converting categorical variables into a form that could be provided to ML algorithms to do a better job in prediction.

62. Outlier: A data point that differs significantly from other observations.
63. Overfitting: The production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.
64. P-Value: The probability of obtaining test results at least as extreme as the results actually observed.
65. Pandas: A fast, powerful, flexible and easy-to-use open-source data analysis and manipulation tool for Python.
66. PCA (Principal Component Analysis): A technique used to emphasize variation and bring out strong patterns in a dataset.
67. Pearson Correlation: A measure of linear correlation between two sets of data.
68. Precision: The ratio of correctly predicted positive observations to the total predicted positives.
69. Predictive Analytics: The use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes.
70. Probability Distribution: A mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.
71. PyTorch: An open-source machine learning library for Python, based on the Torch library.

Q - T

72. Quantile: Cut points dividing the range of a probability distribution into continuous intervals with equal probabilities.
73. Random Forest: An ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees.
74. Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in actual class.
75. Regression: A set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.
76. Regularization: The process of adding information in order to prevent overfitting.
77. Reinforcement Learning: An area of machine learning concerned with how intelligent agents ought to take actions in an environment to maximize cumulative reward.
78. ReLU (Rectified Linear Unit): An activation function used in artificial neural networks.
79. Residuals: The difference between the observed value and the estimated value of the quantity of interest.
80. Ridge Regression: A method of estimating the coefficients of multiple-regression models in scenarios where independent variables are highly correlated.
81. Root Mean Squared Error (RMSE): The square root of the mean squared error.
82. Sample: A set of individuals or objects collected or selected from a statistical population by a defined procedure.
83. Scikit-Learn: A free software machine learning library for Python.
84. Semi-Supervised Learning: A class of machine learning tasks and techniques that also make use of unlabeled data for training.
85. Sigmoid Function: A mathematical function having a characteristic “S”-shaped curve.
86. Silhouette Score: A measure of how similar an object is to its own cluster compared to other clusters.
87. SVD (Singular Value Decomposition): A factorization of a real or complex matrix.
88. Spearman Rank Correlation: A non-parametric measure of rank correlation.
89. Standard Deviation: A measure of the amount of variation or dispersion of a set of values.
90. Stationarity: A property of time series data where statistical properties such as mean and variance are constant over time.
91. Stratified Sampling: A method of sampling from a population which can be partitioned into subpopulations.
92. Supervised Learning: The machine learning task of learning a function that maps an input to an output based on example input-output pairs.
93. SVM (Support Vector Machine): Supervised learning models with associated learning algorithms that analyze data for classification and regression.
94. T-Distribution: A probability distribution that is used when estimating the mean of a normally distributed population in situations where the sample size is small.
95. Time Series Analysis: Methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.
96. T-Test: A type of inferential statistic used to determine if there is a significant difference between the means of two groups.

U - Z

97. Underfitting: A scenario where a machine learning model cannot capture the underlying trend of the data.
98. Univariate Analysis: The simplest form of analyzing data where the data being analyzed contains only one variable.
99. Unsupervised Learning: A type of machine learning that looks for previously undetected patterns in a dataset with no pre-existing labels.
100. Validation Set: A set of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.
101. Variance: The expectation of the squared deviation of a random variable from its population mean.
102. XGBoost: An open-source software library which provides a regularized gradient boosting framework.
103. Zero-Shot Learning: A setup in machine learning where, at test time, a learner observes samples from classes that were not observed during training.
104. Z-Score: A numerical measurement that describes a value’s relationship to the mean of a group of values.