CDSP

Certified Data Science Practitioner Professional Certificate

Glossary

A

A/B test

A type of hypothesis test that compares two different values of the same variable in order to determine which value is most effective.

accuracy

A measure of how frequently each prediction is correctly deemed positive or negative.

AI

(artificial intelligence) The ability of machines to exhibit human-like intelligence, as well as the scientific discipline concerned with this idea.

algorithm

A set of rules that defines how a problem-solving operation is performed.

ANN

(artificial neural network) A machine approximation of biological neural networks. Used in deep learning.

Amazon SageMaker

A cloud-based data science and machine learning automation service that is part of the Amazon Web Services (AWS) cloud platform.

ANOVA

(analysis of variance) A type of hypothesis test that compares the mean of three or more distributions.

API

(application programming interface) A service that facilitates interaction between multiple software environments.

area plot

A type of line plot in which the space below the line is filled in with some color or texture.

ARIMA

(autoregressive integrated moving average) A common algorithm for performing univariate time series forecasting.

arithmetic mean

The average of all numbers in a set.

attribute

See *feature*.

attrition bias

A type of bias that occurs when the training data excludes participants that dropped out over time.

AUC

(area under curve) The total space that is under a learning model's ROC curve.

Azure Machine Learning

A cloud-based data science and machine learning automation service offered as part of the Microsoft Azure cloud platform.

B

bag-of-words

An approach to representing textual content as a list of individual words, irrespective of other language components like grammar and punctuation.

bagging

(bootstrap aggregating) An ensemble learning technique for data sampling with replacement.

bar chart

A type of plot that represents the proportion measurement of categorical variables by using either horizontal or vertical bars.

Bayesian optimization

A hyperparameter optimization method that determines the next optimal hyperparameter space to sample from by using past samples to influence where sampling is conducted in subsequent iterations.

BCSS

(between-cluster sum of squares) A clustering model evaluation metric that measures the separation between clusters.

Bessel's correction

A value of 1 subtracted from the number of items in a sample set in order to compensate for the lack of information that would be provided by the larger population.

bias

In machine learning, a type of error that occurs when a model's estimations are different than the ground truth.

big data

Collections of data that are so large and complex that they require advanced tools to process and analyze them.

bimodal

See *multimodal*.

binary classification

A type of classification task that categorizes data as either a 1 or 0 (i.e., there are only two choices).

box plot

A type of plot that represents the distribution of a numeric variable using summary statistics like quartiles and minimum and maximum.

Box–Cox

A transformation function that obtains a normal distribution of data using log and power transformations.

C

CART

(classification and regression tree) A machine learning decision tree algorithm that uses the Gini index for data splitting to solve classification or regression problems.

CCPA

(California Consumer Privacy Act) A law that protects the data privacy and access rights of California citizens.

central tendency

A measure such as the mean, median, and mode intended to identify the typical value in a dataset.

centroid

In a clustering algorithm, the mean (average) of all of the data points that the cluster contains, across all features.

chi-squared test

A type of hypothesis test that compares the effect of categorical variables.

classification

A type of data science task in which a data example is placed into one or more categories.

clustering

A type of data science task that places data examples into groups based on their similarities.

coefficient of determination

A statistical measure that indicates how much of a dependent variable's variance is explainable by a statistical model.

collinearity

See *multicollinearity*.

concept drift

See *model drift*.

confidence interval

A measurement that returns a range of plausible values for some unknown variable, like the population mean.

confusion matrix

A method of visualizing the truth results of a classification problem.

continuous variable

A quantitative variable whose values are uncountable and can extend infinitely.

correlation

A mathematical association between two variables. When variables correlate, it suggests that each variable has some kind of effect on the other.

cost function

A function that attempts to quantify the error between the predicted values and the actual labeled training values.

cross-validation

A set of methods for partitioning data so that a model is able to generalize to new test data.

CSS

(Cascading Style Sheets) A set of rules that define the appearance of web content.

cumulative gains chart

A method of plotting the percentage of some target number of examples in a given class against a percentage of the total number of examples.

D

dashboard

A high-level visualization tool that summarizes the status of an entire project or one or more parts of a project.

data binning

The process of discretizing a continuous variable by placing its values within specific intervals.

data cleaning

The process of locating and addressing errors and inconsistencies in data.

data encoding

The process of converting data of a certain type into a coded value of a different type.

data governance

A concept in which stakeholders ensure that those who govern data are fulfilling objectives and strategies and creating value for the business.

data munging

See *data wrangling*.

data parsing

The process of taking data as input and then representing that data in a certain structure or syntax.

data preparation

The process of altering data so that it more effectively supports tasks like data analysis and modeling.

data preprocessing

The task of applying various transformation and encoding techniques to data so that it can be interpreted and analyzed by a machine learning algorithm.

data science

The discipline that involves accumulating data, analyzing the data, extracting value from the data, and presenting the value of the data in a meaningful way.

data wrangling

The process of transforming data into a usable form.

Datadog

A general-purpose cloud service monitoring platform.

dataset

A collection of data that will be directly used to accomplish the business goals set forth in the project specifications.

decision boundary

The division line that separates negative classes and positive classes in a classification problem.

decision tree

An arrangement of conditional statements and their conclusions in a branch–leaf structure.

deduplication

The process of identifying and removing duplicate entries from a dataset.

deep learning

A type of machine learning that makes complex decisions using multiple layers of information.

deliverable

A tangible, measurable result or outcome required to complete a project or portion of a project.

dendrogram

A diagram that represents a tree-like hierarchy, commonly used to visualize hierarchical clustering tasks.

dependent variable

In an experiment, the variable that is being studied and that is affected by one or more independent variables.

descriptive statistics

A type of data analysis that quantitatively summarizes the patterns and relationships in a dataset using various mathematical calculations and visualizations. May also refer to an individual calculation that is part of this analysis.

design thinking

An approach to generating business ideas that focus on human needs and innovation.

DevOps

An IT approach in which software development practices help automate systems operations.

dimension

The number of features used in a model.

dimensionality reduction

A type of data science task that minimizes irrelevant or unnecessary elements from a dataset in order to improve the data science process.

discrete variable

A quantitative variable whose values are countable and limited, because there is a definite gap between each value in a range of values.

discretization

The process of converting a continuous variable into a discrete variable.

Django

A fully featured web framework for Python that offers robust tools and feature versatility out of the box.

Docker

An open source platform for building and maintaining virtual containers.

DOE

(design of experiments) An approach to identifying, analyzing, and controlling variables used in an experiment. Also referred to as experimental design or DOX.

E

EDA

(exploratory data analysis) A data science approach to closely examining data in order to reveal new information and insights.

elastic net regression

A regularization technique that uses a weighted average of both ridge regression and lasso regression when training a model.

elbow point

In clustering, the point at which the mean distance between each data example and its associated centroid no longer decreases in a significant way.

embedding

The process of condensing a language vocabulary into vectors of relatively small dimensions.

ensemble learning

An application of machine learning in which the estimations of multiple models are considered in combination.

error

Incorrect or missing values in data.

ETL

(extract, transform, load) The process of combining data from multiple sources, preparing the data, and loading the resulting data into a destination format.

evaluation metric

A method of assessing the skill, performance, and characteristics of a model based on a specific measurement.

experimental design

See *DOE*.

F

F_1 score

The weighted average (harmonic mean) of both precision and recall.

feature

In the field of data science, a measurable property of an example in a training set.

feature engineering

The technique of generating and extracting features from data in order to improve the ability for a machine learning model to make estimations.

feature extraction

A type of dimensionality reduction in which you derive new features from the original features.

feature selection

A type of dimensionality reduction in which you select a subset of the original features.

Flask

A lightweight web framework for Python that emphasizes simplicity and modularity.

forecasting

A task that involves making predictions about future events based on the analysis of relevant past events.

FPR

(false positive rate) A measure of how frequently the learning model incorrectly predicted positive values.

frequency distribution

A type of distribution that demonstrates the frequency of outcomes for a particular sample of a random variable.

full outer join

A method of merging two relational database tables in which all rows from a combination of both tables are returned.

G

Gaussian

Having the shape of a normal curve or a normal distribution.

GDPR

(General Data Protection Regulation) A European Union regulation that regulates the export of EU citizens' personal data for entities that collect or process this data, even if said entities are not based in the EU.

generalization

A model's ability to adapt properly to new, previously unseen data.

geographical map

A type of plot that visually represents data points as they relate to a location.

Gini index

A decision tree splitting metric that splits trees based on the "purity" of decision nodes by squaring each feature's class probability.

global interpretability

A method of measuring the overall decision-making processes of a model.

Goodhart's law

A principle that states: "When a measure becomes a target, it ceases to be a good measure." Used as a reminder not to rely too heavily on one or a small number of metrics when evaluating machine learning model performance.

gradient boosting

An iterative ensemble learning method that builds multiple decision trees in succession, where each tree attempts to reduce the errors of the previous tree.

gradient descent

A method of minimizing a cost function in which a model's parameters are tuned over several iterations by taking gradual "steps" down a slope, toward a minimum error value.

grid search

A hyperparameter optimization method that takes a set (or grid) of parameter combinations, trains a model using each of those combinations, and then returns the combination that best optimizes a specified evaluation metric.

H

HAC

(hierarchical agglomerative clustering) A type of clustering algorithm that initializes each data example in its own cluster, then gradually merges the closest examples and clusters.

hard-margin classification

A type of classification in SVMs where all data examples are outside of the margins, and each example is on the "correct" side of the margins.

HDC

(hierarchical derivative clustering) A type of clustering algorithm that initializes all data examples in a single cluster, then gradually splits the data into more and more clusters.

heatmap

A type of plot that shows different shades or intensities of color on a matrix based on data values in that location of the matrix.

HIPAA

(Health Insurance Portability and Accountability Act) A law enacted in 1996 to establish several rules and regulations regarding healthcare in the United States.

histogram

A type of plot that represents the probability distribution of a given variable using bins.

holdout

A cross-validation method in which the dataset is split into two: the training dataset and the test dataset.

hyperparameter

A parameter that is external to a machine learning model; i.e., set on the algorithm itself and not the learning model.

hyperparameter optimization

The process of repeatedly altering the hyperparameters that an algorithm uses to train a model in order to determine the set of hyperparameters that lead to the best or the desired level of model performance.

hyperplane

In SVMs, a decision boundary that has parallel and equidistant lines or curves on either side of the boundary.

hypothesis

A candidate machine learning model that you create to test its performance, particularly whether it is able to produce the outcome that you require.

I

identity matrix

A matrix of all zeros except for the main diagonal, which consists of all 1s.

imputation

The process of filling in missing data values that consists of using statistical calculations to determine what the missing values should be.

independent variable

In an experiment, a variable that can have an effect on the dependent variable.

inner join

A method of merging two relational database tables in which only rows that match from both tables are returned.

IQR

(interquartile range) The middle half of data values in a distribution.

irreducible error

Errors that cannot be reduced any further when fitting a machine learning model, due to the way the problem was framed, and caused by factors such as unused or unknown features that would have an effect on the output had they been used.

J

JavaScript

A scripting language that provides interactive functionality to web apps.

K

k-fold cross-validation

A cross-validation method in which the dataset is split into *k* groups (folds). One group is the test set, and the remaining groups form the training set.

k-means clustering

A type of clustering algorithm that iteratively updates cluster centroids based on the mean value of each data example in the centroid's cluster.

k-NN

(*k*-nearest neighbor) An algorithm commonly used to classify data examples based on their similarities to other data examples within the feature space.

KPI

(key performance indicator) A metric used to evaluate the success of a project or its activities.

Kubeflow

A component of Kubernetes that orchestrates data science and machine learning pipelines.

Kubernetes

An open source platform for orchestrating the deployment and management of virtual containers.

kurtosis

A measure of the shape of the tails of a distribution, representing the combined weight of the tails relative to the center of the distribution.

L

label

In the field of machine learning, the variable in a training set that you are trying to predict for new samples of data.

lasso regression

A regularization technique that uses an ℓ_1 norm to reduce irrelevant features to 0 when training a model.

LCA

(latent class analysis) A form of unsupervised learning that groups data examples together into unobservable groups called latent classes.

learning curve

A method of visually comparing the change in a model's score or error to the number of data examples used as input.

learning rate

In gradient descent, the size of each "step" down the slope.

left outer join

A method of merging two relational database tables in which all rows from the first table are returned, as well as column data from the second table for any matching rows.

lemmatization

The process of using language morphology to determine the base dictionary form of an inflected word.

leptokurtic

Used to describe a distribution curve that is bunched toward the center, with heavy tails on the right and left sides.

lift chart

The ratio of a percentage of examples in a given class to a baseline, plotted against a percentage of the total number of examples.

line plot

A variant of a scatter plot in which a series of lines connects data points in order.

linear regression

A type of regression analysis in which there is a linear relationship between one independent variable and one dependent variable.

local interpretability

A measure of the decision-making processes in a model as applied to specific data examples.

logistic function

The value between 0 and 1 that a logistic regression algorithm outputs, taking an *S* shape.

logistic regression

A type of linear regression in which the output is a classification probability between 0 and 1.

LOOCV

(leave-one-out cross-validation) A leave-*p*-out cross-validation method in which *p* is set to 1 to minimize performance issues.

LPOCV

(leave-*p*-out cross-validation) A *k*-fold cross-validation method in which *k* (folds) is equal to all data points in the dataset (*n*), with $n - p$ being the training set and *p* being the test set.

M

machine learning

An AI discipline in which a machine is able to gradually improve its estimative capabilities without being given explicit instructions.

machine learning model

A specific implementation of an algorithm that is used to generate predictions and other decision-making outcomes based on some training data.

MAE

(mean absolute error) A cost function that calculates the average difference between estimated and actual values without considering the sign of those values.

mesokurtic

A distribution with average or normal shaped tails on the right and left.

milestone

An event during a project that triggers a reporting requirement or that requires approval from stakeholders before proceeding with the project.

MLflow

An open source data automation platform that is part of the Linux Foundation.

model drift

A process through which the patterns initially used to train a machine learning model change over time such that the model no longer performs well with new data.

model parameter

A parameter that is internal to a machine learning model; i.e., derived from the model as it undergoes the training process.

moment

A set of four statistical parameters commonly used to measure a distribution, including mean, variance, skewness, and kurtosis.

MSE

(mean squared error) A cost function that squares the error between estimated and actual values, then calculates the average of all squares.

multi-class classification

A classification problem in which a data example can be placed into one of three or more classes.

multi-label classification

A classification problem in which a data example can be given multiple labels.

multicollinearity

The property that describes multiple variables as exhibiting a linear relationship.

multimodal

A distribution with more than one peak, or mode.

multinomial logistic regression

An algorithm commonly used to solve multi-class classification problems.

MVP

(minimum viable product) A version of the product or service that is, at the bare minimum, usable by early adopters.

N

naïve Bayes

A type of classification algorithm that computes classification probabilities using Bayes' theorem.

noise

Irrelevant or irregular data values, examples, or features that make it difficult to "hear" patterns revealed by other data that is actually relevant.

non-parametric

A description of a machine learning algorithm that indicates the algorithm can generate a potentially infinite number of model parameters.

normal distribution

A function that represents the distribution of a random variable as a symmetrical bell-shaped graph.

normal equation

A closed-form solution to linear regression problems.

normalization

A technique in which features are scaled so that the lowest value is 0 and the highest value is 1.

NoSQL

Any database technology that does not represent data as relational tables.

null hypothesis

The assumption that there is no statistically significant difference between models under comparison.

O

ordinal data

Data that can be placed in an order.

outlier

A value outside the normal distribution, deviating significantly from the rest of the values in the dataset.

overfitting

A problem in machine learning in which a model's estimations fit well to the training data, but fail to generalize well to other data. An overfit model exhibits high variance and low bias.

P

p-value

The probability of obtaining a result from the test given that the null hypothesis is true.

parametric

A description of a machine learning algorithm that indicates the algorithm generates a fixed number of model parameters.

PCC

(Pearson correlation coefficient) provides a measure of the linear correlation between two variables commonly called *x* and *y*. It produces a value between +1 and −1 that shows the strength of their dependence on each other.

PCI DSS

(Payment Card Industry Data Security Standard) A proprietary standard that specifies how organizations should handle information security for major card brands to increase controls on cardholder data and reduce fraudulent use of accounts.

PII

(personally identifiable information) Data that must be protected to ensure the privacy of the people described by that data.

pipeline

A sequential set of processes that automate the data science process by feeding the output of one process into the input of the next process.

platykurtic

Used to describe a distribution curve that is flat, with light tails on the right and left sides.

POC

(proof of concept) Evidence that supports the feasibility of the product or service that the project is meant to create.

precision

A measure of how often the positives identified by the learning model are true positives.

probability distribution

A type of distribution that demonstrates the probability of outcomes for a random variable.

problem formulation

The process of identifying an issue that should be addressed, and putting that issue in terms that are understandable and actionable.

pruning

The process of reducing the overall size of a decision tree by eliminating nodes, branches, and leaves that provide little value for the classification or regression problem at hand.

Q

qualitative data

Data that holds categorical values.

quantitative data

Data that holds numerical values that represent magnitude.

R

R^2

See *coefficient of determination*.

random forest

An ensemble learning method that aggregates multiple decision tree models together and selects the optimal classifier or predictor.

randomized search

A hyperparameter optimization method that takes a distribution of parameter combinations, trains a model using a random sampling of those combinations, and then returns the combination that best optimizes a specified evaluation metric.

recall

A measure of the percentage of positive instances that are found by a machine learning model as compared to all relevant instances.

regression

A type of data science task that measures the relationship between variables and outputs an estimation for a numeric variable.

regular expression

A group of characters that describe how to execute a specific search pattern on a given text.

regularization

The technique of simplifying a machine learning model by constraining the model parameters, which helps the model avoid overfitting to the training data.

reporting bias

A type of bias that occurs when the training data is missing observations that were not reported.

ridge regression

A regularization technique that uses an ℓ_2 norm to constrain features used to train a model.

right outer join

A method of merging two relational database tables in which all rows from the second table are returned, as well as column data from the first table for any matching rows.

RMSE

(root mean squared error) The square root of the *MSE*.

ROC curve

(receiver operating characteristic curve) A method of plotting the relationship between estimated hits (true positive rate) versus false alarms (false positive rate).

S

sample set

A smaller group than the entire population.

scatter plot

A type of plot that represents the relationship between two variables through the use of points on a graph.

scope

In project management, an outline of all aspects of a project, including any constraints.

scope creep

The condition by which a project continues to grow beyond its ability to be sustained or meet expectations.

selection bias

A type of bias that occurs when the training dataset doesn't truly represent the population the model will ultimately be applied to.

semi-structured data

Data that is in a format that facilitates searching, filtering, or extracting some elements of that data, whereas other elements are not so easy to work with.

sensitivity

See *recall*.

silhouette analysis

A method of calculating how well a particular data example fits within a cluster as compared to its neighboring clusters.

skewness

The property of a distribution that has a high density of values distributed toward the lower or higher end of the x-axis.

skillful

Used to describe a model that is useful for its intended task. There are degrees of skill; some models are more useful than others. Improving a model's skill is the ultimate goal of the iterative tuning process.

soft-margin classification

An approach to classification with SVMs that keeps the distance between the margins as large as possible while minimizing the number of examples that end up inside the margins.

specificity

A measure of how frequently a machine learning model correctly identifies all actual negative instances.

SQL

(Structured Query Language) A language for programmatically creating, retrieving, modifying, and deleting data in a relational database.

stakeholder

A person who has a vested interest in the outcome of a project or who is actively involved in its work.

standard deviation

A measure of variability; the square root of variance.

standardization

A technique in which features are scaled so that the mean value is 0 and the standard deviation is 1.

statistical model

A mathematical system that generates assumptions about data using statistical methods and probability.

stemming

The process of removing the affix of a word in order to retrieve the word stem.

stochastic

The property by which a randomly determined process cannot perfectly estimate individual events or data points, but can demonstrate a general pattern common to the entire set of data.

stop word

A word in a text document that is so common it is typically removed when the text is processed.

stratified *k*-fold cross-validation

A *k*-fold cross-validation method in which each fold has a representative sample of data in datasets that exhibit class imbalance.

stratified random sampling

A statistical sampling method in which the population is divided into groups (strata) according to chosen characteristics (e.g., demographics), then members of each group are randomly sampled, and those samples are combined to form the ultimate sample.

structured data

Data that is in a format that facilitates searching, filtering, or extracting that data.

summary statistics

See *descriptive statistics*.

supervised learning

A type of machine learning in which known label values are provided as input so that a model can estimate these values in future datasets.

SVMs

(support-vector machines) Supervised learning algorithms that can be used to solve classification and regression problems by separating data values using a hyperplane.

T

t-test

A type of hypothesis test that compares the mean of two distributions in which the population standard deviation is not known.

target feature

A variable that the data science practitioner is interested in learning more about.

target function

A representation of the mapping between input variables and output variables that best approximates some desired outcome from a machine learning model.

threshold

A value used by a classification model to classify anything higher than the threshold as positive, and anything lower than the threshold as negative.

time series

A representation of data in which observations are ordered according to a sequential change in time.

TNR

(true negative rate) See *specificity*.

tokenization

The process of partitioning text into smaller units.

TPR

(true positive rate) See *recall*.

training

In the field of machine learning, the process by which a model learns from input data.

training sample

A dataset of examples used to generate a machine learning model.

U

underfitting

A problem in machine learning in which a model cannot make effective estimations due to an inability to identify the underlying patterns in the data. An underfit model exhibits low variance and high bias.

unimodal

A distribution with one peak, or mode.

unstructured data

Data that is in a format that makes it difficult to search, filter, or extract that data.

unsupervised learning

A type of machine learning in which label values are not provided as input, so the model does not have an explicit variable that it is estimating.

V

variability

The property that indicates the extent to which data varies across all values in a dataset.

variance

A measurement of the spread between numbers in a dataset, or the variation of a model's estimations across datasets.

violin plot

A type of plot that shows the distribution of a numerical value through probability density.

W

WCSS

(within-cluster sum of squares) A clustering model evaluation metric that measures the compactness of clusters.

web app

An application that runs on a server using a web-based protocol, particularly Hypertext Transfer Protocol (HTTP) and its secure equivalent, HTTPS.

web framework

A programming library that supports the development of web apps and other web-based services through a standardized interface.

Z

z-score

The number of standard deviations that a sample is above or below the mean of all values in the sample.

z-test

A type of hypothesis test that compares the mean of two distributions when the standard deviation of a population is known.