



## Reading: Glossary terms module 2

---

### Terms and definitions from Course 6, Module 2

**Accuracy:** Refers to the proportion of data points that were correctly categorized

**Bayes' Theorem:** An equation that can be used to calculate the probability of an outcome or class, given the values of predictor variables

**Class imbalance:** When a dataset has a predictor variable that contains more instances of one outcome than another

**Continuous variables:** Variables that can take on an infinite and uncountable set of values

**Customer churn:** The business term that describes how many and at what rate customers stop using a product or service, or stop doing business with a company

**Downsampling:** The process of removing some observations from the majority class, making it so they make up a smaller percentage of the dataset than before

**F1-Score:** The harmonic mean of precision and recall

**Feature engineering:** The process of using practical, statistical, and data science knowledge to select, transform, or extract characteristics, properties, and attributes from raw data

**Feature extraction:** A type of feature engineering that involves taking multiple features to create a new one that would improve the accuracy of the algorithm

**Feature selection:** A type of feature engineering that involves select the features in the data that contribute the most to predicting the response variable

**Feature transformation:** A type of feature engineering that involves modify existing features in a way that improves accuracy when training the model

**Naive Bayes:** A supervised classification technique that is based on Bayes's Theorem with an assumption of independence among predictors

**Plan stage:** Stage of the PACE workflow where the scope of a project is defined and the informational needs of the organization are identified

**Posterior probability:** The probability of an event occurring after taking into consideration new information

**Precision:** The proportion of positive predictions that were correct to all positive predictions

**Recall:** The proportion of actual positives that were identified correctly to all actual positives

**Supervised model:** A machine learning model that is used to make predictions about unseen events

**Unsupervised model:** A machine learning model that is used to discover the natural structure of the data, finding relationships within unlabeled data

**Upsampling:** The process of taking observations from the minority class and either adding copies of those observations to the dataset or generating new observations to add to the dataset

**“Zero Frequency” problem:** Occurs when the dataset has no occurrences of a class label and some value of a predictor variable together

## **Terms and their definitions from the previous module**

### **C**

**Categorical variables:** Variables that contain a finite number of groups or categories

**Collaborative filtering:** A technique used by recommendation systems to make comparisons based on who else liked the content

**Content-based filtering:** A technique used by recommendation systems to make comparisons based on attributes of content

**Continuous variables:** Variables that can take on an infinite and uncountable set of values

### **D**

**Decision tree:** A flowchart-like structure that uses branching paths to predict the outcomes of events, or the probability of certain outcomes

**Discrete features:** Features with a countable number of values between any two values

**Documentation:** An in-depth guide that is written by the developers who created a package that features very specific information on various functions and features

**I**

**Integrated Development Environment (IDE):** A piece of software that has an interface to write, run, and test a piece of code

**M**

**Machine learning:** The use and development of algorithms and statistical models to teach computer systems to analyze and discover patterns in data

**P**

**Popularity bias:** The phenomenon of more popular items being recommended too frequently

**R**

**Recommendation systems:** Unsupervised learning techniques that use unlabeled data to offer relevant suggestions to users

**S**

**Supervised machine learning:** A category of machine learning that uses labeled datasets to train algorithms to classify or predict outcomes

---