Introduction to Machine Learning with Apache Spark

Module 1: Get Started with Machine Learning

Welcome! This alphabetized glossary contains many terms you will find in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and participating in other certificate programs.

Terms	Definition	Video
AI (Artificial Intelligence)	The field of computer science aims to create intelligent machines that can mimic human cognitive functions.	Introduction to Machine Learning for Everyone
Anomaly Detection	An application of clustering that focuses on identifying data points that are unusual, abnormal, or deviate significantly from the established patterns or clusters.	Clustering
Augmented Intelligence	The concept of using AI technologies to enhance and augment human capabilities allows experts to scale their abilities while machines manage time-consuming tasks.	Generative AI Overview and Use Cases
Categorical data	Non-numeric data that represent categories or labels.	Supervised vs Unsupervised Learning
Classification	A supervised learning technique that predicts the class or category of a case, such as classifying a cell as benign or malignant.	Introduction to Machine Learning for Everyone
Classifier	A machine learning algorithm or model is used to solve classification problems by learning patterns and making predictions about the class of new, unseen data.	Classification
Cluster Centroid	Cluster centroid refers to a cluster's representative or central point in a clustering algorithm. It is calculated as the mean or median of the data points assigned to that cluster.	Clustering
Clustering	An unsupervised learning technique that groups similar cases together based on their features, aiming to identify patterns or clusters within the data.	Introduction to Machine Learning for Everyone
Confusion Matrix	A table that summarizes a classification model's performance by showing the counts of true positives, true negatives, false positives, and false negatives.	Evaluating Machine Learning Models

Decision Tree	A predictive model that uses a tree-like structure to make decisions or predictions based on input features.	Regression
Deep learning	An exceptional field of machine learning where computers can learn and make intelligent decisions independently.	Introduction to Machine Learning for Everyone
Density Estimation	An unsupervised learning technique that focuses on estimating the underlying probability density function of a dataset.	Supervised vs Unsupervised Learning
Dependent variable	The continuous variable that is being predicted, explained, or estimated based on the input or independent variables	Supervised vs Unsupervised Learning
Dimensionality Reduction	An unsupervised learning technique is used to reduce the number of input features while preserving valuable information.	Supervised vs Unsupervised Learning
Eager Learner	A type of classification algorithm that spends time training and generalizing the model, making it faster in predicting test data. Examples include decision trees and logistic regression.	Classification
Ethical Concerns	Issues and considerations related to the responsible and ethical use of AI, including potential misuse of AI-generated content and implications for intellectual property and copyright laws.	Generative AI Application and Examples
Euclidean Distance	Euclidean distance is a measure of distance or similarity between two data points in a multidimensional space.	Clustering
Extract, Transform, and Load (ETL)	The process within the machine learning model lifecycle refers to the data collection and preparation stage.	Machine Learning Model Lifecycle
F1-Score	A metric that combines precision and recalls into a single value to assess a classification model's overall performance. It is calculated as the harmonic mean of precision and recall, providing a balanced measure when both metrics are equally important.	Evaluating Machine Learning Models
Feature Engineering	The process of creating new features or representations from existing data to enhance the performance and predictive capabilities of machine learning models.	Role of data Engineering in Machine learning
Feature extraction	The process in which relevant information or characteristics are extracted from raw data and transformed into a reduced and more informative representation, known as features	Role of data Engineering in Machine learning
Generative AI	A technology that uses machine learning and deep learning techniques to generate original content based on patterns learned during training, enabling	Generative AI Application and Examples

	software applications to create and simulate new content.	
Gradient Boosting	A machine learning technique that builds an ensemble of weak models like decision trees sequentially, where each subsequent model focuses on correcting the errors made by the previous models.	Regression
Image Segmentation	Image segmentation is an application of clustering that involves dividing images into categories based on color, content, or other features.	Clustering
Independent variable	A variable that is used to explain, predict, or estimate the value of the dependent variable.	Supervised vs Unsupervised Learning
K-means Algorithm	The K-means algorithm is a popular clustering algorithm that aims to divide a dataset into K clusters, where K is a user-specified parameter.	Clustering
k-nearest neighbor (KNN)	A lazy learner algorithm is used for classification. It classifies unknown data points by finding the k most similar examples in the training set and assigning the majority class among those neighbors to the test data point.	Classification
Large Language Model (LLM)	A type of artificial intelligence model based on deep learning techniques designed to process and generate natural language, which can be incorporated into Generative AI systems.	Generative AI Overview and Use Cases
Lazy Learner	A type of classification algorithm that does not have a specific training phase. It waits until it receives test data before making predictions, often resulting in longer prediction times.	Classification
Line of Best Fit	A straight line represents the best approximation of the relationship between two variables in a scatter plot.	Regression
Machine learning	The subfield of computer science gives computers the ability to learn from data without being explicitly programmed.	Introduction to Machine Learning for Everyone
Machine Learning Model Lifecycle	The end-to-end process involved developing, deploying, and maintaining a machine learning model.	Machine Learning Model Lifecycle
Market Basket Analysis	An unsupervised learning technique used to identify associations or relationships between items in a dataset.	Supervised vs Unsupervised Learning
Mean Absolute Error (MAE)	A metric that uses the absolute differences between the predicted and actual values. It calculates the average of the absolute values of the errors.	Evaluating Machine Learning Models

Model Deployment	The process of making the trained machine learning model available for use in a production environment or real-world application.	Machine Learning Model Lifecycle
Natural language processing	The field of study that focuses on enabling computers to understand and process human language, both written and spoken.	Introduction to Machine Learning for Everyone
Neural Networks	A class of machine learning models inspired by the structure and functioning of biological neural networks. Neural networks consist of interconnected nodes (neurons) organized in layers and are capable of learning complex patterns from data. They are used for regression tasks as well as other types of problems.	Regression
Precision	A metric that measures the fraction of true positives among all examples predicted to be positive by a classification model.	Evaluating Machine Learning Models
Random Forest	An ensemble learning method that combines multiple decision trees to create a predictive model.	Regression
Recall	Also known as sensitivity or true positive rate, recall measures the fraction of true positives among all actual positive examples.	Evaluating Machine Learning Models
Recommendation Systems	Recommendation systems are applications of clustering that group related items or products based on customer behavior or preferences.	Clustering
Regression	A supervised learning technique that predicts continuous values based on input features, such as predicting the price of a house based on its characteristics.	Introduction to Machine Learning for Everyone
Root Mean Squared Error (RMSE)	The square root of the mean squared error. It has the same unit as the target variable and is easier to interpret than MSE.	Evaluating Machine Learning Models
R-squared	A metric that quantifies the proportion of variance in the dependent variable that can be explained by the independent variable(s) in a regression model. It ranges from 0 to 1, with higher values indicating a better fit.	Evaluating Machine Learning Models
Scatter Plot	A graphical representation of data points on a two-dimensional coordinate system, where each point represents the values of two variables.	Regression
Slope	The slope of the line of best fit represents the rate of change in the dependent variable for a unit change in the independent variable.	Regression
Squared error	A common metric used to evaluate the performance of regression models. It measures the average of the squared differences between the	Evaluating Machine Learning Models

	predicted values and the actual values of the target variable.	
Supervised learning	A category of machine learning where the model is trained using labeled data with known input-output pairs.	Introduction to Machine Learning for Everyone
Support Vector Regression (SVR)	A regression technique that uses support vector machines to create a hyperplane or line that best fits the data points.	Regression
Train/Test Split	The process of dividing a dataset into two separate sets: a training set used to train a machine learning model and a test set used to evaluate the model's performance on new, unseen data.	Evaluating Machine Learning Models
Unsupervised learning	A category of machine learning where the model is trained using unlabeled data, and the algorithms detect patterns and relationships within the data.	Introduction to Machine Learning for Everyone