Module 3 Summary and Highlights

Congratulations! You have completed this lesson. At this point in the course, you know:

- Classification is a supervised machine learning method used to predict labels on new data with applications in churn prediction, customer segmentation, loan default prediction, and multiclass drug prescriptions.
- Binary classifiers can be extended to multiclass classification using one-versus-all or one-versus-one strategies.
- A decision tree classifies data by testing features at each node, branching based on test results, and assigning classes at leaf nodes.
- Decision tree training involves selecting features that best split the data and pruning the tree to avoid overfitting.
- Information gain and Gini impurity are used to measure the quality of splits in decision trees.
- Regression trees are similar to decision trees but predict continuous values by recursively splitting data to maximize information gain.
- Mean Squared Error (MSE) is used to measure split quality in regression trees.
- K-Nearest Neighbors (k-NN) is a supervised algorithm used for classification and regression by assigning labels based on the closest labeled data points.
- To optimize k-NN, test various k values and measure accuracy, considering class distribution and feature relevance.
- Support Vector Machines (SVM) build classifiers by finding a hyperplane that maximizes the margin between two classes, effective in high-dimensional spaces but sensitive to noise and large datasets.
- The bias-variance tradeoff affects model accuracy, and methods such as bagging, boosting, and random forests help manage bias and variance to improve model performance.
- Random forests use bagging to train multiple decision trees on bootstrapped data, improving accuracy by reducing variance.