## Week 2a – Exploring and Visualising Data

1. What is important to determine when initially exploring your data?
   1. Record level.
   2. Shape.
2. Summary statistics tell us everything we need to know about our feature variable values.
   1. False.
3. What enables data visualisation on Spark data sources within Databricks?
   1. Databricks built-in visualisation.

## Week 2b – K-Means Clustering

1. K-Means is an iterative clustering algorithm.
   1. True.
2. What’s true about the k-means clustering algorithm?
   1. It requires all features to be on the same scale.
   2. It recomputes its centroids with each iteration.
3. What attribute can be used to obtain the centroids for a trained k-means object?
   1. KMeans.cluster\_centers\_

## Week 2c – Feature Correlation

1. What are examples of Pearson correlation coefficients representing a strong, direct linear relationship between two features?
   1. 0.87
   2. 0.92
2. Feature interactions can occur between feature variables of different types.
   1. True.
3. What enables a quick view of correlations between multiple features?
   1. Correlation matrix.

## Week 2d – Principal Components Analysis

1. PCA is a technique that can be used to reduce a feature space without losing too much information.
   1. True.
2. What is computed and used during the PCA training process?
   1. Eigenvectors.
   2. Covariance matrix.
3. What metric is helpful in understanding the effectiveness of PCA components?
   1. Percent of explained variance.

## Week 3a – Dimensionality and Feature Importance

1. What is multicollinearity?
   1. The way that features are highly correlated.
2. While fitting a model may be the end of a data science project, it might also be the kickoff to a new set of assumptions and understanding about the data.
   1. True.
3. The curse of dimensionality is caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space.
   1. True.

## Week 3b – Feature Selection in Linear Regression

1. What are linear regression coefficients?
   1. They can be used for feature selection.
   2. They are interpreted as: “for every one unit increase in my data, our target increases by coefficient units.”
2. The p-value in a regression model tests the null hypothesis that a feature isn’t correlated to our target.
   1. True.
3. Stepwise regression finds the worst feature in our training set and then steps down to the best features.
   1. False.

## Week 4a – Algorithm Selection and Decision Trees

1. What are some of the considerations involved in selecting which ML algorithm to apply to a problem?
   1. Training data size.
   2. Need for accuracy vs. interpretability.
   3. Linear assumptions.
   4. Data types (categorical, numeric, continuous, discrete).
   5. Goal (prediction, classification, clustering).
2. If you wanted higher accuracy, potentially at the expense of interpretability, would you choose a linear regression or tree-based model?
   1. Tree-based model.
3. If you wanted higher interpretability, potentially at the expense of accuracy, would you choose a linear regression or tree-based model?
   1. Linear regression.
4. Tree-based models handle non-linearity better than linear regression models.
   1. True.
5. Tree-based models require you to scale the data before training.
   1. False.

## Week 4b – Categorical Features

1. What are two Spark ML techniques that covert categorical features to numeric?
   1. StringIndexer.
   2. VectorIndexer.
2. With scikit-learn, categorical variables don’t have to be encoded for ML algorithms.
   1. False.
3. What problem results from one-hot encoding categorical variables for decision-tree models?
   1. The tree becomes inefficient.
4. String indexing non-ordinal categorical values causes a ML algorithm to interpret these values as ranked or ordered.
   1. True.

## Week 4c – Tree-Based Ensemble Modelling

1. If a decision tree is built with no hyperparameter tuning, what could be the expected accuracy on the training data set?
   1. 100%.
2. Decision trees always have high bias.
   1. False.
3. What is another term for “bootstrap aggregating”?
   1. Bagging.
4. In general, bagging tends to result in lower variance, whilst boosting tends to result in lower bias compared to single decision trees.
   1. True.

## Week 4d - Bootstrap Aggregation

1. In general, bagging tends to result in lower bias, while boosting primarily results in lower variance compared to the performance of single/individual models.
   1. False.
2. Ensemble methods combine multiple “weak learner” models.
   1. True.
3. Bagging combines models in parallel, whilst boosting combines models sequentially.
   1. True.

## Week 4e – Classification Evaluation

1. What is true?
   1. A Type I error means a false positive.
   2. A Type II error means a false negative.
2. Classification accuracy is calculated as the number of correct predictions divided by the total number of predictions.
   1. True.
3. Of the cases that the model predicted to be positive, what percentage were actually positive?
   1. Precision.
4. Of the cases that were actually positive, what percentage did the model correctly identify as positive?
   1. Recall.

## Week 4f – Label Imbalance and Sampling

1. Precision and recall are better metrics than accuracy to evaluate model performance on a dataset with imbalanced classes.
   1. True.
2. What’s true about re-sampling?
   1. Over-sampling increases instances of the minority class.
   2. Under-sampling decreases instances of the majority class.
3. What is a technique used to generate synthetic data samples.
   1. Randomly sampling the attributes from instances in the minority class.
4. SMOTE works by oversampling, where it creates synthetic samples from the minority class instead of creating copies.
   1. True.

## Week 5a – Hyperparameters in Tree-Based Models

1. What is controlled by a hyperparameter?
   1. A model’s training process.
   2. A model’s architecture.
2. Hyperparameters must be specified prior to the beginning of the training process.
   1. True.
3. What hyperparameter allows tree-based models to capture more complex relationships?
   1. max\_depth.

## Week 5b – Grid Search

1. What’s involved in the hyperparameter training process?
   1. A set of input hyperparameter values.
   2. An evaluation metric.
2. The hyperparameter optimisation process can be automated.
   1. True.
3. What’s a benefit of grid search for hyperparameter optimisation?
   1. Automated and organised testing of hyperparameter values.

## Week 5c – Model Generalisation and Validation Set

1. What are benefits of using a validation set?
   1. Avoid data leakage.
   2. Robust to overfitting.
2. Models with high variance fail to capture relationships between their inputs and outputs.
   1. False. These are models with high bias.
3. What’s a limitation of using a validation set?
   1. It reduces data size for each step of training, tuning and evaluation.

## Week 5d – Cross-Validation

1. What are benefits of using cross-validation over a train-validation split?
   1. More data is used during each step.
   2. Each point in cross-validation data is used for training and validation.
2. Cross-validation is more robust to bias in hyperparameter optimisation than the train-validation-test split.
   1. False.
3. If we have 10 unique hyperparameter combinations and are performing 4-fold cross-validation, how many unique models will be trained?
   1. 41.