1. AI Winters happened mostly due to the lack of understanding behind the theory of neural networks
2. **True** b. False
3. Most modern applications that use computer vision, use models that were trained using this discipline:
4. Machine learning **b. Deep learning** c. Artificial Intelligience
5. In the Machine Learning Workflow, the main goal of the Data Exploration and Preprocessing step is to:
6. Identify what data that is best suited to find a solution to your business problem
7. **Determine how to clean your data such that you can use it to train a model**
8. What is the goal of supervised learning?
9. Find an underlying structure of the dataset without any labels.
10. Predict the labels c. Find the target **d. Predict the features**
11. What is deep learning?
12. **Deep learning is machine learning that involves deep neural networks.**
13. Deep learning is another name for artificial intelligence.
14. Deep learning includes artificial intelligence and machine learning.
15. None of the above are correct
16. When is a standard machine learning algorithm usually a better choice than using deep learning to get the job done?
17. **When working with small data sets**  c. When working with large data sets
18. When the data is steady over time d. None of the above correct
19. What is a Turing test?
20. It tests images c. It tests and cleans the dataset
21. **It tests a machine’s ability to exhibit intelligent behavior** d. It tests the dataset
22. What are some of the different milestones in deep learning history?
23. Deep Blue defeats a world champion chess player, and Keras is released.
24. **Geoffrey Hinton’s work, AlexNet, and TensorFlow**
25. Deep Blue defeats a world champion chess player, and AlexNet is created.
26. Deep Blue defeats a world champion chess player and TensorFlow is released
27. What is artificial intelligence?
28. A subset of deep learning **c. Any program that can sense, reason, act and adapt**
29. A subset of machine learning d. None of the above
30. What are two spaces within AI that are going through drastic growth and innovation?
31. **Computer vision and natural language processing.**
32. Language processing and deep learning
33. Deep learning and machine learning
34. Computer vision and deep learning
35. Why did AI flourish so much in the last years?
36. Access to hardware for cleaning data
37. Data storage in the cloud is much more expensive
38. Stylish designed computers
39. **Faster and inexpensive computers and data storage**
40. How does Alexa use artificial intelligence?
41. Recognizes faces and pictures **c. Recognizes our voice and answer questions**
42. Suggests who a person on a photo is d. None of these
43. What are the first two steps of a typical machine learning workflow?
44. Problem statement and data clean c. Data collection and data transform
45. **Problem statement and data collection**  d. None of these
46. Which statement about the Pandas read\_csv function is TRUE?
47. It allows only one argument: the name of the file
48. It reads data into a 2-dimensional Numpy array
49. It can only read comma-delimited data
50. **It can read both tab-delimited and space-delimited data**
51. Which of the following is a reason to use JavaScript Object Notation (JSON) files for storing data?
52. Because the data is stored in a matrix format.
53. Because they can store NA values.
54. Because they can store NULL values.
55. **Because they are cross-platform compatible.**
56. The data below appears in 'data.txt', and Pandas has been imported. Which Python command will read it correctly into a Pandas DataFrame?

63.03 22.55 39.61 40.48 98.67 -0.25 AB

39.06 10.06 25.02 29 114.41 4.56 AB

* 1. 2.22 50.09 46.61 105.99 -3.53 AB

a. pandas.read\_csv(‘data.txt’)

b. pandas.read\_csv(‘data.txt’, delim\_whitespace = True)

**c. pandas.read\_csv(‘data.txt’, header = None, sep= ‘’)**

d. pandas.read\_csv(‘data.txt’, header = 0, delim\_whitespace = True)

17. Outliers must be very extreme to noticeably impact the fit of a statistical model.

a. True **b. False**

18. Outliers should always be replaced, since they never contain useful information about the data.

a. True **b. False**

19. Which residual-based approach to identifying outliers compares running a model with all data to running the same model, but dropping a single observation?

a. Standardized residuals **c. Externally-studentized residuals**

b. Unstandardized residuals d. Abnormally-studentized residuals

20. What is a CSV file?

a. CSV files are a standard way to store data across platforms.

**b. CSV files are rows of data or values separated by commas**

c. CSV is a method of JavaScript Object Notation.

d. CSV makes data readily available for analytics, dashboards, and reports

21. What are residuals?

a. Residuals are data removed from the data frame.

**b. Residuals are the difference between the actual values and the values predicted by a given model.**

c. Residuals are a method to standardize data.

d. Residuals are a method for handling identified outliers.

22. If the removal of rows or columns of data is not an option, why must we ensure that information is assigned for missing data?

a. Assigning information for missing data improves the accuracy of the dataset.

b. Information must be assigned to prevent outliers.

c. Missing data may bias the dataset.

**d. Most models will not accept blank values in our data.**

23. What are the two main data problems companies face when getting started with artificial intelligence/machine learning?

a. Lack of training and expertise c. Data sampling and categorization

b. Outliers and duplicated data **d. Lack of relevant data and bad data**

24. What does SQL stand for and what does it represent?

**a. SQL stands for Structured Query Language, and it represents a set of relational databases with fixed schemas.**

b. SQL stands for Structured Query Language, and it represents databases that are not relational, they vary in structure.

c. SQL stands for Sequential Query Language, and it represents a set of sequential databases with fixed schemas.

d. SQL stands for Sequential Query Language, and it represents a set of relational databases with fixed schemas.

25. What does NoSQL stand for and what does it represent?

**a. NoSQL stands for Not-only SQL, and it represents a set of databases that are not relational, therefore, they vary in structure.**

b. NoSQL stands for Non-Structured Query Language, and it represents a set of non-relational databases with varied schemas.

c. NoSQL stands for Non-Structured Query Language, and it represents a set of relational databases with fixed schemas.

d. NoSQL stands for Not-only SQL, and it represents a set of databases that are relational, therefore, they have fixed structure.

26. What is a JSON file?

a. JSON stands for JavaString Object Notation, and it is a standard way to store the data across platforms.

b. JSON stands for JavaScript Object Notation, and it is a non-standard way to store the data across platforms.

c. JSON stands for JavaString Object Notation, and they have very similar structure to Python Dictionaries.

**d. JSON stands for JavaScript Object Notation, and it is a standard way to store the data across platforms.**

27. What is meant by the Messy Data?

a. Duplicated or unnecessary data c. Inconsistent text and typos

b. Missing data **d. All of the above**

28. What is an outlier?

a. Outlier is a data point that has the highest or lowest value in the dataset.

b. Outlier is a data point that is very close to the mean value of all observations.

**c. Outlier is an observation in dataset that is distant from most other observations.**

d. Outlier is a data point that does not belong in our dataset.

29. How do we identify outliers in our dataset?

a. We can identify outliers only by calculating the minimum and maximum values in the dataset.

**b. We can identify outliers both visually and with statistical calculations.**

c. We can only identify outliers visually through building plots.

d. We can only identify outliers by using some statistical calculations.

30. From the options listed below, select the option that is NOT a valid exploratory data approach to visually confirm whether your data is ready for modeling or if it needs further cleaning or data processing:

a. Create a panel plot that shows distributions for the dependent variable and scatter plots for all independent variables

b. Train a model and identify the observations with the largest residuals

c. Create visualizations for scatter plots, histograms, box plots, and hexbin plots

**d. Create a correlation heat map to confirm the sign and magnitude of correlation across your features.**

31. These are two of the most common variables for data visualization:

a. **matplotlib and seaborn**  c. numpy and matplotlib

b. scipy and seaborn d. scipy and numpy

32. You can use the pandas library to use plots.

a. **True** b. False

33. Classification models require that input features be scaled.

a. True **b. False**

34. Feature scaling allows better interpretation of distance-based approaches.

a. **True** b. False

35. Feature scaling reduces distortions caused by variables with different scales.

a. **True** b. False

36. Which scaling approach converts features to standard normal variables?

a. Robust scaling **b. Standard scaling** c. MinMax scaling d. Nearest neighbor scaling

37. Which variable transformation should you use for ordinal data?

a. Standard scaling **b. Ordinal encoding** c. One-hot encoding d. Min-max scaling

38. What are polynomial features?

a. They are logistic regression coefficients. **c. They are higher order relationships in the data**.

b. They are lower order relationships in the data. d. They are represented by linear relationships in the data.

39. What does Boxcox transformation do?

a. It makes the data more left skewed

**b. It transforms the data distribution into more symmetrical bell curve**

c. It makes the data more right skewed.

d. It transforms categorical variables into numerical variables.

40. Select three important reasons why EDA is useful.

a. To examine correlations, to sample from dataframes, and to train models on random samples of data

b. To utilize summary statistics, to create visualizations, and to identify outliers

c. To analyze data sets, to determine the main characteristics of data sets, and to use sampling to examine data

**d. To determine if the data makes sense, to determine whether further data cleaning is needed, and to help identify patterns and trends in the data**

41. What assumption does the linear regression model make about data?

**a. This model assumes a linear relationship between predictor variables and outcome variables.**

b. This model assumes an addition of each one of the model parameters multiplied by a coefficient.

c. This model assumes a transformation of each parameter to a linear relationship.

d. This model assumes that raw data in data sets is on the same scale.

42. What is skewed data?

a. Raw data that may not have a linear relationship.

b. Data that has a normal distribution.

c. Raw data that has undergone log transformation.

**d. Data that is distorted away from normal distribution; may be positively or negatively skewed.**

43. Select the two primary types of categorical feature encoding.

a. Encoding and scaling c. One-hot encoding and ordinal encoding

b. **Nominal encoding and ordinal encoding**  d. Log and polynomial transformation

44. Which scaling approach puts values between zero and one?

a. Robust scaling b. Standard scaling **c. MinMax scaling** d. Nearest neighbor scaling

45. Which variable transformation should you use for nominal data with multiple different values within the feature?

a. **One-hot scaling** b. Standard scaling c. MinMax scaling d. Ordinal encoding

46. In general, the population parameters are unknown.

a. **True** b. False

47. Parametric models have finite number of parameters.

a. **True** b. False

48. The most common way of estimating parameters in a parametric model is:

**a. Using the maximum likelihood estimation**  c. extrapolating a non-parametric model

b. Using the central limit theorem d. extrapolating Bayesian statistics

49. A p-value is:

a. The probability of the null hypothesis being true

b. The probability of the null hypothesis being false

**c. The smallest significance level at which the null hypothesis would be rejected**

d. The smallest significance level at which the null hypothesis is accepted

50. Type 1 Error 1 is defined as:

**a. Saying the null hypothesis is false, when it is actually true**

b. Saying the null hypothesis is false, when it is actually false

51. You find through a graph that there is a strong correlation between Net Promoter Score and the visual time that customers spend on a website. Select the TRUE assertion:

a. To boost the Net Promoter Score of a business, you need to increase the time that customers spend on a website.

**b. There is an underlying factor that explains this correlation, but manipulating the time that customers spend on a website may not affect the Net Promoter Score they will give to the company**

52. Which one of the following is common to both machine learning and statistical inference?

a. Using sample data to make inferences about a hypothesis.

b. Using population data to model a null hypothesis.

c. Using population data to make inferences about a null sample.

**d. Using sample data to infer qualities of the underlying population distribution.**

53. Which one of the following describes an approach to customer churn prediction stated in terms of probability?

a. Predicting a score for individuals that estimates the probability the customer will stay.

b. Churn prediction is a data-generating process representing the actual joint distribution between our x and the y variable.

**c. Predicting a score for individuals that estimates the probability the customer will leave.**

d. Data related to churn may include the target variable for whether a certain customer has left.

54. What is customer lifetime value?

a. The total churn a customer generates in the population.

**b. The total purchases over the time which the person is a customer.**

c. The total value that the customer receives during their life.

d. The total churn generated by a customer over their lifetime.

55. Which one the following statements about the normalized histogram of a variable is true?

**a. It provides an estimate of the variable’s probability distribution.**

b. It is a parametric representation of the population distribution.

c. It is a non-parametric representation of the population variance.

d. It serves as a bar chart for the null hypothesis.

56. The outcome of rolling a fair die can be modelled as a \_\_\_\_\_\_\_ distribution.

a. Poisson b. Normal c. Log-normal **d. Uniform**

57. Which one of the following features best distinguishes the Bayesian approach to statistics from the Frequentist approach?

a. Frequentist statistics incorporates the probability of the hypothesis being true.

b. Frequentist statistics requires construction of a prior distribution.

c. Bayesian statistics is better than Frequentist.

**d. Bayesian statistics incorporate the probability of the hypothesis being true.**

58. Which of the following best describes what a hypothesis is?

a. A hypothesis is a statement about a prior distribution.

b. A hypothesis is a statement about a sample of the population.

**c. A hypothesis is a statement about a population.**

d. A hypothesis is a statement about a posterior distribution.

59. A Type 2 error in hypothesis testing is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_:

a. Correctly rejecting the null hypothesis. c. Incorrectly accepting the alternative hypothesis.

**b. Incorrectly accepting the null hypothesis.**  d. Correctly rejecting the alternative hypothesis.

60. Which statement best describes a consequence of a type II error in the context of a churn prediction example? Assume that the null hypothesis is that customer churn is due to chance, and that the alternative hypothesis is that customers enrolled for greater than two years will not churn over the next year.

a. You incorrectly conclude that there is no effect

b. You correctly conclude that a customer will eventually churn

**c. You incorrectly conclude that customer churn is by chance**

d. You correctly conclude that customer churn is by chance

61. Which of the following is a statistic used for hypothesis testing?

a. The standard deviation. c. The acceptance region

**b. The likelihood ratio** d. The rejection region

62. Which statement about Logistic Regression is TRUE?

**a. Logistic Regression is a generalized linear model.**

b. Logistic Regression models can only predict variables with 2 classes.

c. Logistic Regression models can be used for classification but not for regression.

d. Logistic Regression models can be used for regression but not for classification.

63. Logistic regression is similar to a linear regression, except that it uses a logistic function to estimate probabilities of an observation belonging to a certain class or category.

a. **True**  b. False

64. Usually the first step to fit a logistic regression model using scikit learn is to:

a. **import logistic regression from the sklearn.linear\_model module**

**e.g. from sklearn.linear\_model import LogisticRegression**

b. import Logistic from the sklearn.regression module

e.g. from sklearn.regression import Logistic

c. import logistic regression from the sklearn.linearclassifer module

e.g. from sklearn.linearclassifer import LogisticRegression

65. The output of a logistic regression model applied to a data sample \_\_\_\_\_\_\_\_\_\_\_\_\_.

a. tells you the odds of the sample belonging to a certain class.

b. is the log odds of the sample, which you can use for interpretive purposes.

c. tells you which class the sample belongs to

**d. is the probability of the sample being in a certain class.**

66. Describe how any binary classification model can be extended from its basic form on two classes, to work on multiple classes.

a. Use the coefficients from a linear regression model to weight the classes.

b. Use a process of elimination to discard any unimportant classes.

c. Fit the binary classifier to all of the classes simultaneously.

**d. Use a one-versus-all technique, where for each class you fit a binary classifier to that class versus all of the other classes.**

67. Which tool is most appropriate for measuring the performance of a classifier on unbalanced classes?

a. The true positive rate. c. The false positive rate

b. The Receiver Operating Characteristic(ROC) curve **d. The precision recall curve**

68. One of the requirements of logistic regression is that you need a variable with two classes.

a. True **b. False**

69. The shape of ROC curves are the leading indicator of an overfitted logistic regression.

a. True **b. False**

70. What is the goal of supervised learning?

a. Find an underlying structure of the dataset without any labels. b. Predict the features.

**c. Predict the labels.**  d. Find the target.

71. A simplified way to interpret K Nearest Neighbors is by thinking of the output of this method as a decision boundary which is then used to classify new points.

**a. True**  b. False

72. These are all characteristics of the k nearest neighbors algorithm EXCEPT:

a. It is sensitive to scaling

b. It determines decision boundaries to make predictions

**c. It determines the value for k**

d. It is well suited to predict variables with multiple classes

73. An advantage of k nearest neighbor methods is that they can leverage categorical data without encoding.

a. True **b. False**

74. Usually the first step to fit a k nearest neighbor classifier using scikit learn is to:

a. import KNN from the sklearn.knearestneighbors module

e.g. from sklearn.knearestneighbors import KNN

**b.** **import Classifier from the sklearn.nearestneighbors module**

**e.g. from sklearn.nearestneighbors import Classifier**

c. import KNNClassifier from the sklearn.knearestneighbors module

e.g. from sklearn.knearestneighbors import KNNClassifier

75. Which one of the following statements is true regarding K Nearest Neighbors?

1. **K Nearest Neighbors (KNN) assumes that points which are close together are similar.**
2. The distance between two data points is independent of the scale of their features.
3. For high dimensional data, the best distance measure to use for KNN is the Euclidean distance.
4. The Manhattan distance between two data points is the square root of the sum of the squares of the differences between the individual feature values of the data points.

76. Which one of the following statements is most accurate?

1. Linear regression needs to remember the entire training dataset in order to make a prediction for a new data sample.
2. KNN determines which points are closest to a given data point, so it doesn’t take long to actually perform prediction.
3. **K nearest neighbors (KNN) need to remember the entire training dataset in order to classify a new data sample.**
4. KNN only needs to remember the hyperplane coefficients to classify a new data sample.

77. Which one of the following statements is most accurate about K Nearest Neighbors (KNN)?

1. **KNN can be used for both classification and regression.**
2. KNN is an unsupervised learning method.
3. KNN is a regression model.
4. KNN is a classification model.

78. K Nearest Neighbors with large k tend to be the best classifiers.

a. True **b. False**

79. When building a KNN classifier for a variable with 2 classes, it is advantageous to set the neighbor count k to an odd number.

**a. True** b. False

80. The Euclidean distance between two points will always be shorter than the Manhattan distance:

**a. True** b. False

81. The main purpose of scaling features before fitting a k nearest neighbor model is to:

1. Break ties in case there is the same number of neighbors of different classes next to a given observation
2. Ensure decision boundaries have roughly the same size for all classes
3. **Ensure that features have similar influence on the distance calculation**
4. Help find the appropriate value of k

82. These are all pros of the k nearest neighbor algorithm EXCEPT:

1. It adapts wells to new training data
2. It is simple to implement as it does not require parameter estimation
3. It is easy to interpret
4. **It is sensitive to the curse of dimensionality**

83. (True/False) SVMs calculate predicted probabilities in the range between 0 and 1.

a. True **b. False**

84. All of these are characteristics of SVMs, EXCEPT:

1. Support Vector Machines do not return predicted probabilities.
2. Support Vector Machines use decision boundaries for classification.
3. The algorithm behind Support Vector Machines calculates hyperplanes that minimize misclassification error.
4. **Support Vector Machine models are non-linear.**

85. Any linear model can be turned into a non-linear model by applying a kernel to the model

**a. True**  b. False

86. SVMs with kernels are recommended for large data sets with many features

a. True **b. False**

87. Select the TRUE statement regarding the cost function for SVMs:

1. SVMs use the same loss function as logistic regression
2. **SVMs use the Hinge Loss function as a cost function**
3. SVMs use a loss function that penalizes vectors prone to misclassification
4. SVMs do not use a cost function. They use regularization instead of a cost function.

88. Which statement about Support Vector Machines is TRUE?

1. Support Vector Machine models are non-linear.
2. Support Vector Machine models can be used for regression but not for classification.
3. **Support Vector Machine models rarely overfit on training data.**
4. Support Vector Machine models can be used for classification but not for regression.

89. A large c term will penalize the SVM coefficients more heavily.

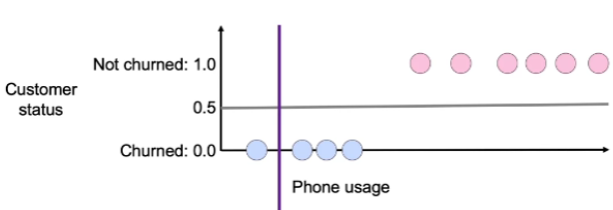
a. True **b. False**

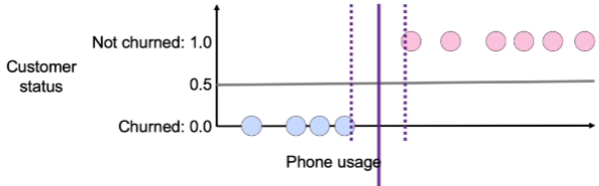
90. Regularization in the context of support vector machine (SVM) learning is meant to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

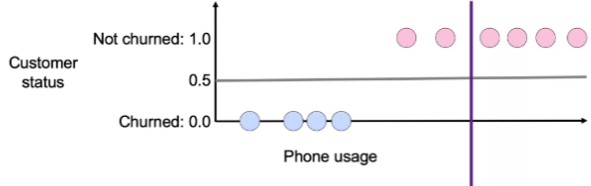
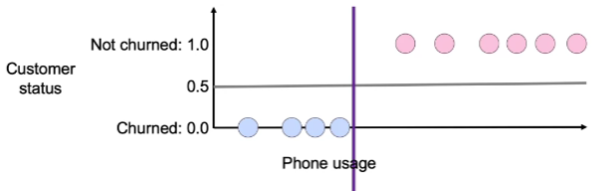
1. **lessen the impact that some minor misclassifications have on the cost function**
2. bring all features to a common scale to ensure they have equal weight
3. smooth the input data to reduce the chance of overfitting
4. encourage the model to ignore outliers during training

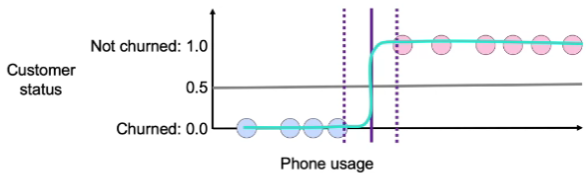
91. Support vector machines can be extended to work with nonlinear classification boundaries by \_\_\_\_\_\_\_.

1. modifying the standard sigmoid function
2. incorporating polynomial regression
3. **using the kernel trick**
4. projecting the feature space onto a lower-dimensional space

92. Select the image that displays the line at the optimal point in the phone usage that the data can be split to create a decision boundary.

a. **c.**

b. d.

93. The below image shows the decision boundary with a clear margin, such decision boundary belongs to what type machine learning model?

**a. Support Vector Machine**

b. Machine Learning

c. Super Vector Machine

94. SVM with kernels can be very slow on large datasets. To speed up SVM training, which methods may you perform to map low dimensional data into high dimensional beforehand?

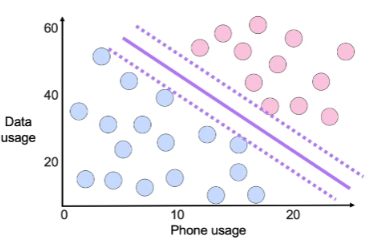
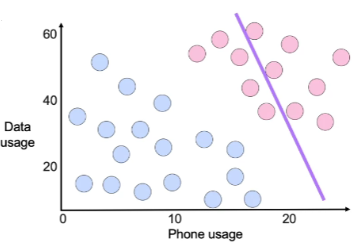
a. Regularization b. Linear SVC **c. RBF Sample d. Nystroem**

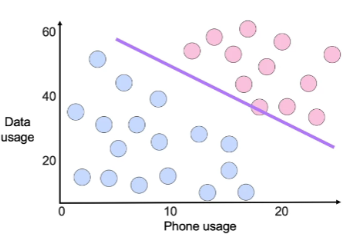
95. Concerning the Machine Learning workflow what model choice would you pick if you have "Few" features and a "Medium" amount of data?

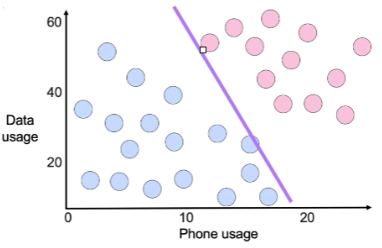
a. LinearSVC, or Kernal Approximation c. Simple, Logistic or LinearSVC

b. Add features or Logistic **d. SVC with RBF**

96. Select the image that best displays the line that separates the classes.

a. b.



1.  d.

97. Which of the following statements about Decision Tree models is TRUE?

1. **Decision Tree models are non-linear.**
2. Decision Tree models rarely overfit training data.
3. Decision Tree models can be used for classification but not for regression.
4. Decision Tree models can be used for regression but not for classification.

98. Decision Trees are considered a greedy algorithm.

**a. True**  b. False

99. These are all characteristics of decision trees, EXCEPT:

1. **They have well-rounded decision boundaries**
2. They can be used for either classification or regression
3. They segment data based on features to predict results
4. They split nodes into leaves

100. Decision trees used as classifiers compute the value assigned to a leaf by calculating the ratio: number of observations of one class divided by the number of observations in that leaf E.g. number of customers that are younger than 50 years old divided by the total number of customers.

How are leaf values calculated for regression decision trees?

1. Median value of the predicted variable c. mode value of the predicted variable
2. Weighted average value of the predicted variable **d. average value of the predicted variable**

101. These are two main advantages of decision trees:

1. **They are very visual and easy to interpret**
2. They output both parameters and significance levels
3. They do not tend to overfit and are not sensitive to changes in data
4. They are resistant to outliers and output scaled features

102. How can you determine the split for each node of a decision tree?

1. Find the split that induces the largest entropy. c. Randomly select the split.
2. **Find the split that minimizes the Gini impurity.** d. Use a nonlinear decision boundary to find the best split.

103. Which of the following describes a way to regularize a decision tree to address overfitting?

1. Reduce the information gain. c. Increase the number of branches.
2. Increase the max depth **d. Decrease the max depth.**

104. What is the disadvantage of decision trees?

1. Scaling is required. **c. They tend to overfit.**
2. They can get too large. d. They are difficult to interpret.

105. What method can you use to minimize the overfitting of a machine-learning model?

1. Decrease the variance of your test data.
2. Choose the hyperparameters that maximize the goodness of fit on your training data.
3. Increase the variance of your training data.
4. **Tune the hyperparameters of your model using cross-validation.**

106. Concerning Classification algorithms, what is a characteristic of K-Nearest Neighbors?

1. **Training data is the model** c. The model is just parameters
2. The fitting can be slow d. Prediction is fast

107. Concerning Classification algorithms, what are the characteristics of Logistic Regression?

1. **The model is just parameters, fitting can be slow, prediction is fast, and the decision boundary is simple and less flexible**
2. The training data is the model, fitting is fast, predicting class for new records can be slow, and the decision boundary is flexible
3. The model is just parameters, fitting is fast, prediction is fast, and the decision boundary is flexible
4. The training data is the model, fitting is fast, prediction is fast, and the decision boundary is flexible

108. When evaluating all possible splits of a decision tree what can be used to find the best split regardless of what happened in prior or future steps?

1. Logistic regression **b. Greedy Search** c. Classification d. Regularization

109. A model that averages the predictions of multiple models reduces the variance of a single model and has high chances to generalize well when scoring new data.

**a. True**  b. False

110. Bagging is a tree ensemble that combines the prediction of several trees that were trained on bootstrap samples of the data.

**a. True**  b. False

111. In general, a random forest can be considered a special case of bagging and it tends to have better out of sample accuracy

**a. True**  b. False

112. Bagging tends to have less overfitting than decision trees

**a. True**  b. False

113. Boosting tend to be well suited for data sets with outliers and rare events.

**a. True**  b. False

114. If you were to combine several logistic regressions using a voting ensemble, you should use a Voting Regressor.

a. True **b. False**

115. All of these are characteristics of boosting algorithms, EXCEPT:

* 1. They use the entire data set, not only bootstrapped samples
  2. They use residuals from previous models
  3. They create trees iteratively
  4. **They create trees independently**

116. The term Bagging stands for bootstrap aggregating.

**a. True**  b. False

117. This is the best way to choose the number of trees to build on a Bagging ensemble.

* 1. Choose a number of trees past the point of diminishing returns
  2. Choose a large number of trees, typically above 100
  3. Prioritize training error metrics over out-of-bag sample
  4. **Tune the number of trees as a hyperparameter that needs to be optimized**

118. Which type of Ensemble modeling approach is NOT a special case of model averaging?

* + - * 1. The Pasting method of Bootstrap aggregation c. Random Forest methods
        2. The Bagging method of Bootstrap aggregation **d. Boosting methods**

119. What is an ensemble model that needs you to look at out of bag error?

1. Logistic Regression b. Stacking **c. Random Forest** d. Out of Bag Regression

120. What is the main condition to use stacking as an ensemble method?

1. **Models need to output predicted probabilities** c. Models need to be parametric
2. Models need to output residual values for each class d. Models need to be nonparametric

121. This tree ensemble method only uses a subset of the features for each tree:

1. Stacking b. Bagging c. Adaboost **d. Random Forest**

122. Order these tree ensembles in order of most randomness to least randomness:

1. **Random Trees, Random Forest, Bagging** c. Bagging, Random Forest, Random Trees
2. Random Forest, Bagging, Random Trees d. Random Forest, Random Trees, Bagging

123. This is an ensemble model that does not use bootstrapped samples to fit the base trees, takes residuals into account, and fits the base trees iteratively:

1. Random Forest b. Bagging c. Random trees **d. Boosting**

124. When comparing the two ensemble methods Bagging and Boosting, what is one characteristic of Boosting?

a. Bootstraped samples **c. Fits entire data set**

b. No weighting used d. Only data points are considered

125. What is the most frequently discussed loss function in boosting algorithms?

a. Gradient Loss Function **c. 0-1 Loss Function**

b. Gradient Boosting Loss Function d. Adaboost Loss Function

126. What type of forest is a classification algorithm that potentially contains hundreds of different decision trees?

a. The Multiple Forest b. Model Forest **c. Random Forest** d. Global Forest

127. When describing models what type of model will feature coefficients help to explain?

a. Global Surrogate model b. KNN **c. Linear Model** d. SVM

128. What type of surrogate model tries to approximate a black-box model globally on every instance in the data set?

a. Strategic Surrogate model c. Local Surrogate model

**b. Global Surrogate model**  d. Complex Surrogate model

129. These are all methods of dealing with unbalanced classes EXCEPT:

a. Down-sampling. **c. A mix of in-sample and out-of-sample**

b. A mix of down-sampling and up-sampling. d. Up-sampling.

130. A best practice to build a model using unbalanced classes is to split the data first, then apply an upsample or undersample technique.

**a. True**  b. False

131. Which of the following statements about Down-sampling is TRUE?

1. Down-sampling results in excessive focus on the more frequently-occurring class.
2. Down-sampling preserves all the original observations.
3. Down-sampling is likely to decrease Recall.
4. Down-sampling is likely to decrease Precision

132. Which of the following statements about Random Upsampling is TRUE?

1. Random Upsampling will generally lead to a higher F1 score.
2. Random Upsampling generates observations that were not part of the original data.
3. **Random Upsampling preserves all original observations.**
4. Random Upsampling results in excessive focus on the more frequently-occurring class.

133. Which of the following statements about Synthetic Upsampling is TRUE?

1. Synthetic Upsampling results in excessive focus on the more frequently-occurring class.
2. **Synthetic Upsampling generates observations that were not part of the original data.**
3. Synthetic Upsampling will generally lead to a higher F1 score.
4. Synthetic Upsampling uses fewer hyperparameters than Random Upsampling

134. What can help humans to interpret the behaviors and methods of Machine Learning models more easily?

**a. Model Explanations** b. Model Debug c. Model Trust d. Explanation Debug

135. What type of explanation method can be used to explain different types of Machine Learning models no matter the model structures and complexity?

a. Model Trust Explanations c. Model Explanations

b. Local Interpretable Model-Agnostic Explanations (LIME) **d. Model-Agnostic Explanations**

136. What reason might a Global Surrogate model fail?

1. Consistency between surrogate models and black-box models
2. Single data instance groups
3. **Large inconsistency between surrogate models and black-box models**
4. Single clusters in the data instance groups

137. When working with unbalanced sets, what should be done to the samples so the class balance remains consistent in both the train and test sets?

1. Apply weighted observations **c. Stratify the samples**
2. Use oversampling d. Use a combination of oversampling and undersampling

138. What approach are you using when trying to increase the size of a minority class so that it is similar to the size of the majority class?

1. Synthetic Oversampling b. Random Oversampling c. Undersampling **d. Oversampling**

139. What approach are you using when you create a new sample of a minority class that does not yet exist?

1. Random Oversampling **b. Synthetic Oversampling** c. Oversampling d. Weighting

140. What intuitive technique is used for unbalanced datasets that ensures a continuous downsample for each of the bootstrap samples?

1. SMOTE b. Upsampling c. Downsampling **d. Blagging**

141. Predicting payment default, whether a transaction is fraudulent, and whether a customer will be part of the top 5% spenders on a given year, are examples of:

1. **Classification**  b. Regression

142. It is less concerning to treat a Machine Learning model as a black box for prediction purposes, compared to interpretation purposes:

**a. True**  b. False

143. Predicting total revenue, number of customers, and percentage of returning customers are examples of:

1. Classification **b. Regression**

144. (True/False) The Sum of Squared Errors (SSE) can be used to select the best-fitting regression model.

**a. True**  b. False

145. (True/False) The R-squared value from estimating a linear regression model will almost always increase if more features are added.

**a. True**  b. False

146. (True/False) The Total Sum of Squares (TSS) can be used to select the best-fitting regression model.

a. True **b. False**

147. You can use supervised machine learning for all of the following examples, EXCEPT:

1. **Segment customers by their demographics.**
2. Predict the number of customers that will visit a store on a given week.
3. Predict the probability of a customer returning to a store.
4. Interpret the main drivers that determine if a customer will return to a store.

148. The autocorrect on your phone is an example of:

1. Unsupervised Learning **c. Supervised learning**
2. Semi-supervised learning d. Reinforcement learning\

149. This is the type of Machine Learning that uses both data with labeled outcomes and data without labeled outcomes:

1. Unsupervised Machine Learning c. Supervised Machine learning
2. **Semi-Supervised Machine learning** d. Mix-Supervised Machine learning

150. This option describes a way of turning a regression problem into a classification problem:

1. **Create a new variable that flags 1 for above a certain value and 0 otherwise**
2. Use outlier treatment
3. Use missing value handling
4. Create a new variable that uses autoencoding to transform a continuous outcome into categorical

151. This is the syntax you need to predict new data after you have trained a linear regression model called LR :

1. LR = predict(X\_test) **b. LR.predict(X\_test)** c. LR.predict(LR, X\_test) d. predict(LR, X\_test)

152. All of these options are useful error measures to compare regressions except:

1. SSE b. R squared c. TSS d. ROC index

153. All of the listed below are part of the Machine Learning Framework, except:

1. Observations b. Features c. Parameters **d. None of the above**

154. Select the option that is the most INACCURATE regarding the definition of Machine Learning:

1. Machine Learning allows computers to learn from data
2. Machine Learning allows computers to infer predictions for new data
3. Machine Learning is a subset of Artificial Intelligence
4. **Machine Learning is automated and requires no programming**

155. In Linear Regression, which statement about model evaluation is the most accurate?

1. Model selection involves choosing a model that minimizes the cost function.
2. **Model estimation involves choosing parameters that minimize the cost function.**
3. Model estimation involves choosing a cost function that can be compared across models.
4. Model selection involves choosing modeling parameters that minimize in-sample validation error.

156. When learning about regression we saw the outcome as a continuous number. Given the below options, what is an example of regression?

1. A fraudulent charge
2. Under certain circumstances determine if a person is a Republican or Democrat
3. Customer churn
4. **Housing prices**

157. What is another term for the testing data:

1. Training data **b. Unseen data** c. Corroboration data d. Cross validation data

158. (True/False) The ShuffleSplit will ensure that there is no bias in your outcome variable.

**a. True**  b. False

159. Select the option that has the syntax to obtain the data splits you will need to train a model having a test split that is a third the size of your available data.

1. X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.5)
2. **X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.33)**
3. X\_train, y\_test = train\_test\_split(X, y, test\_size=0.33)
4. X\_train, y\_test = train\_test\_split(X, y, test\_size=0.5)

160. What is the main goal of adding polynomial features to linear regression?

1. Remove the linearity of the regression and turn it into a polynomial model.
2. **Capture the relation of the outcome with features of higher order.**
3. Increase the interpretability of a black box model.
4. Ensure similar results across all folds when using K-fold cross-validation.

161. What are the most common sklearn methods to add polynomial features to your data?

Note: polyFeat =PolynomialFeatures(degree)

1. polyFeat.add and polyFeat.transform c. polyFeat.add and polyFeat.fit
2. **polyFeat.fit and polyFeat.transform**  d. polyFeat.transform

162. How can you adjust the standard linear approach to regression when dealing with fundamental problems such as prediction or interpretation?

1. Create a class instance
2. **Add some non-linear patterns, i.e., polynomial features**
3. Import the transformation method
4. By transforming the data

163. The main purpose of splitting your data into training and test sets is:

1. To improve accuracy **c. To avoid overfitting**
2. To improve regularization d. To improve cross-validation and overfitting

164. What term is used if your test data leaks into the training data?

1. Test leakage b. Training leakage **c. Data leakage** d. Historical data leakage

165. Which one of the below terms uses a linear combination of features?

1. Binomial Regression **b. Linear Regression** c. Multiple Regression d. Polynomial Regression

166. When splitting your data, what is the purpose of the training data?

1. Compare with the actual value **c. Fit the actual model and learn the parameters**
2. Predict the label with the model d. Measure errors

167. Polynomial features capture what effects?

1. **Non-linear effects** b. Linear effects c. Multiple effects d. Regression effects

168. Which fundamental problems are being solved by adding non-linear patterns, such as polynomial features, to a standard linear approach?

1. Prediction b. Interpretation **c. Prediction and Interpretation** d. None of the above

169. A testing data could be also reffered to as:

1. Training data **b. Unseen data** c. Corroboration data d. Cross validation data

170. What is the correct sklearn syntax to add a third degree polynomial to your model?

1. polyFeat = polyFeat.add(degree=3) c. polyFeat = polyFeat.fit(degree=3)
2. **polyFeat = PolynomialFeatures(degree=3)** d. polyFeat = polyFeat.transform(degree=3)

171. Complete the following sentence: The training data is used to fit the model, while the test data is used to:

1. measure the parameters and hyperparameters of the model
2. tweak the model hyperparameters
3. tweak the model parameters
4. **measure error and performance of the model**

172. In model complexity versus error diagram, the model compexity increases as the training error decreases.

1. True **b. False**

173. In model complexity versus error diagram, there is an inflection point after which, as the cross validatio error increases, so does the complexity of the model.

1. **True** b. False

174. In the model complexity versus error diagram, the right side of the curve is where the model is underfitted and the left side of the curve, is where the model is overfitted.

1. True **b. False**

175. In K-fold cross-validation, how will increasing k affect the variance (across subsamples) of estimated model parameters?

1. Increasing k will not affect the variance of estimated parameters.
2. Increasing k will usually reduce the variance of estimated parameters.
3. **Increasing k will usually increase the variance of estimated parameters.**
4. Increasing k will increase the variance of estimated parameters if models are underfitting, but reduce it if models are overfitting.

176. Which statement about K-fold cross-validation below is TRUE?

1. Each subsample in K-fold cross-validation has at least k observations.
2. Each of the k subsamples in K-fold cross-validation is used as a training set.
3. **Each of the k subsamples in K-fold cross-validation is used as a test set.**
4. Each subsample in K-fold cross-validation has at least k-1 observations.

177. Which of the following statements about a high-complexity model in a linear regression setting is TRUE?

1. Cross-validation with a small k will reduce or eliminate overfitting.
2. **A high variance of parameter estimates across cross-validation subsamples indicates likely overfitting.**
3. A low variance of parameter estimates across cross-validation subsamples indicates likely overfitting.
4. Cross-validation with a large k will reduce or eliminate overfitting.

178. Which of the following statements about cross-validation is/are True?

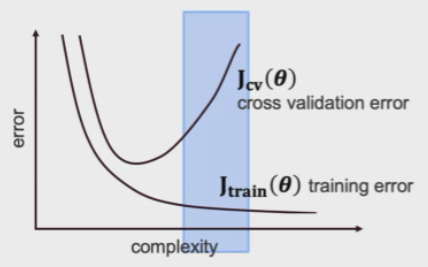
1. Cross-validation is an essential step in hyperparameter tuning.
2. We can manually generate folds by using KFold function.
3. GridSearchCV is commonly used in cross-validation.
4. **All of the above is True.**

179. Which of the following statements about GridSearchCV is/are True?

1. GridSearchCV scans over a dictionary of parameters.
2. GridSearchCV finds the hyperparameter set that has the best out-of-sample score.
3. GridSearchCV retrains on all data with the "best" hyper-parameters.
4. **All of the above is True.**

180. Which of the below functions, randomly selects data to be in the train/test folds?

1. `StratifiedKFold` c. `GroupKFold`
2. **'KFold' and `StratifiedKFold`**  d. 'KFold'

181. Reviewing the below graph, what is the model considered when associated with the right side of the cross validation error?

1. **Overfitting** c. Training error
2. Polynomial Regression d. Underfitting

182. If a low-complexity model is underfitting during estimation, which of the following is MOST LIKELY true (holding the model constant) about K-fold cross-validation?

1. **K-fold cross-validation will still lead to underfitting, for any k.**
2. K-cross-validation with a small k will reduce or eliminate underfitting.
3. K-fold cross-validation with a large k will reduce or eliminate underfitting.
4. None of the above.

183. Which of the following functions perform K-fold cross-validation for us, appropriately fitting and transforming at every step of the way?

1. 'cross\_val' b. 'cross\_validation' c. 'cross\_validation\_predict' **d. 'cross\_val\_predict'**

184. (True/False) The variance of a model is determined by the degree of irreducible error.

1. True **b. False**

185. (True/False) As more variables are added to a model, both its complexity and its variance generally increase.

1. **True** b. False

186. (True/False) Model adjustments that decrease bias also decrease variance, leading to a bias-variance trade off.

1. True **b. False**

187. Regularization zeroes out or gets model’s coefficients closer to zero and, in such a way, it avoids the data being overfitted.

1. **True** b. False

188. Scaling the features is not very important before using regularization techniques.

1. True **b. False**

189. (True/False) A model with high variance is characterized by sensitivity to small changes in input data.

1. **True** b. False

190. Which of the following statements about model complexity is TRUE?

1. Higher model complexity leads to a lower chance of overfitting
2. **Higher model complexity leads to a higher chance of overfitting.**
3. Reducing the number of features while adding feature interactions leads to a lower chance of overfitting.
4. Reducing the number of features while adding feature interactions leads to a higher chance of overfitting

191. Which of the following statements about model errors is TRUE?

1. Underfitting is characterized by lower errors in both training and test samples.
2. **Underfitting is characterized by higher errors in both training and test samples.**
3. Underfitting is characterized by higher errors in training samples and lower errors in test samples.
4. Underfitting is characterized by lower errors in training samples and higher errors in test samples.

192. Which of the following statements about regularization is TRUE?

1. Regularization always reduces the number of selected features.
2. Regularization increases the likelihood of overfitting relative to training data.
3. **Regularization decreases the likelihood of overfitting relative to training data.**
4. Regularization performs feature selection without a negative impact on the likelihood of overfitting relative to the training data.

193. Which of the following statements about scaling features prior to regularization is TRUE?

1. Feature scaling is not recommended prior to regularization.
2. Features should rarely or never be scaled prior to implementing regularization.
3. **The larger a feature’s scale, the more likely its estimated impact will be influenced by regularization.\**
4. The smaller a feature’s scale, the more likely its estimated impact will be influenced by regularization.

194. Which one of the 3 Regularization techniques: Ridge, Lasso, and Elastic Net, performs the fastest under the hood?

1. **Ridge** b. Lasso c. Elastic Net d. None of the above

195. Which of the following statements about Elastic Net regression is TRUE?

1. **Elastic Net combines L1 and L2 regularization**.
2. Elastic Net does not use L1 or L2 regularization.
3. Elastic Net uses L2 regularization, as with Ridge regression.
4. Elastic Net uses L1 regularization, as with Ridge regression.

196. BOTH Ridge regression and Lasso regression

1. Do not adjust the cost function used to estimate a model.
2. **Add a term to the loss function proportional to a regularization parameter.**
3. Add a term to the loss function proportional to the square of parameter coefficients.
4. Add a term to the loss function proportional to the absolute value of parameter coefficients.

197. Compared with Lasso regression (assuming a similar implementation), Ridge regression is:

1. Less likely to overfit training data. c. More likely to overfit to training data.
2. **Less likely to set feature coefficients to zero.** d. More likely to set feature coefficients to zero.

198. Which of the following about Ridge Regularization is TRUE?

1. It enforces the coefficients to be lower, but not 0
2. It minimizes irrelevant features
3. It penalizes the size magnitude of the regression coefficients by adding a squared term
4. **All of the above**

199. Which of the below statements is correct?

1. Neither RidgeCV nor LassoCV uses the L1 regularization function.
2. Both RidgeCV and LassoCV use the L1 regularization function.
3. Only RidgeCV uses the L1 regularization function.
4. **Only LassoCV uses the L1 regularization function.**

200. (True/False) In an Analytic View, increasing L2/L1 penalties force coefficients to be smaller, restricting their plausible range.

1. **True** b. False

201. (True/False) Under the Geometric formulation, the cost function minimum is found at the intersection of the penalty boundtry and a contour of the traditional OLS cost function surface.

1. **True** b. False

202. (True/False) Under the Probabilistic formulation, L2 (Ridge) regularization imposes Gaussian prior on the coefficients, while L1 (Lasso) regularization imposes Laplacian prior.

1. **True** b. False

203. When working with regularization, what is the view that illuminates the actual optimization problem and shows why LASSO generally zeros out coefficients?

1. Analytical view **b. Geometric view** c. Probabilistic view d. Regression view

204. When working with regularization, what is the view that recalibrates our understanding of LASSO and a Ridge, as a base problem, where coefficients have particular prior distributions?

1. Analytical view b. Geometric view **c. Probabilistic view**  d. Regression view

205. When working with regularization, what is the logical view of how to achieve the goal of reducing complexity?

1. **Analytical view** b. Geometric view c. Probabilistic view d. Regression view

206. All of the following statements about Regularization are TRUE except:

1. Optimizing predictive models is about finding the right bias/variance tradeoff.
2. **Features should rarely or never be scaled prior to implementing regularization.**
3. We need models that are sufficiently complex to capture patterns in data, but not so complex that they overfit.
4. Regularization techniques have an analytical, geometric, and probabilistic interpretation.

207. When working with regularization and using the geometric formulation, what is found at the intersection of the penalty boundary and a contour of the traditional OLS cost function surface?

1. **The cost function minimum** c. A smaller range of coefficients
2. The prior distribution of β d. A peaked density

208. Which statement under the Probabilistic View is correct?

1. Regularization imposes certain errors on the regression coefficients. Feedback: Incorrect! Please review the further Details of Regularization lessons.
2. **Regularization imposes certain priors on the regression coefficients.**
3. Regularization uses some regression coefficients to inflate the errors.
4. Regularization coefficients do not take into consideration prior probabilities.

209. Increasing L2/L1 penalties force coefficients to be smaller, restricting their plausible range. This statement is part of what View?

1. **Analytical view** b. Geometric view c. Probabilistic view d. Regression view

210. What does a higher lambda term mean in the Regularization technique?

1. **Higher lambda decreases variance which means smaller coefficients.**
2. Higher lambda increases variance, which means smaller coefficients.
3. Higher lambda decreases variance which means larger coefficients.
4. Higher lambda decreases the prior probability.

211. What concept/s under Probabilistic View is/are True?

1. We can derive the posterior probability by knowing the probability of the target and the prior distribution.
2. The prior distribution is derived from independent draws of a prior coefficient density function that we choose when regularizing.
3. L2 (ridge) regularization imposes a Gaussian prior on the coefficients, while L1 (lasso) regularization imposes a Laplacian prior.
4. **All of the above**

212. What statement is True?

1. **We reduce the complexity of the model by minimizing the error on our training set.**
2. By penalizing the cost function, we increase the complexity of the model.
3. The goal of Regularization is always going to be to optimize our complexity trade-off, so we can minimize error on the hold-out set.
4. Introducing Regularization will increase bias and variance.

213. Which statement about unsupervised algorithms is TRUE?

1. Unsupervised algorithms are relevant when we have outcomes we are trying to predict.
2. **Unsupervised algorithms are relevant when we don’t have the outcomes we are trying to predict and when we want to break down our data set into smaller groups.**
3. Unsupervised algorithms are typically used to forecast time-related patterns like stock market trends or sales forecasts
4. Unsupervised algorithms are relevant in cases that require explainability, for example comparing parameters from one model to another.

214. What is one of the real-world solutions to fix the problems of curse dimensionality?

1. Increase the size of the data set c. Use more computational power
2. Reduce the dimension of the data set. d. Balance the classes of a data set

215. Which statement is a common use of Dimension Reduction in the real world?

1. **Image tracking**
2. Explaining the relation between the amount of alcohol consumption and diabetes.
3. Deep Learning
4. Predicting whether a customer will return to a store to make a major purchase.

216. Which statement describes better “the smarter initialization of K-mean clusters?

1. “Draw a line between the data points to create 2 big clusters.”
2. “After we find our centroids, we calculate the distance between all our data points.”
3. **“Pick one random point, as the initial point, and for the second point, instead of picking it randomly, we prioritize by assigning the probability of the distance.”**
4. “We start by having two centroids as far as possible between each other.”

217. Which of the following statements best describes the iterative part of the K-means algorithm?

1. The k-means algorithm assigns a number of clusters at random.
2. **The k-means algorithm adjusts the centroids to the new mean of each cluster, and then it keeps repeating this process until no example is assigned to another cluster.**
3. The k-means algorithm iteratively deletes outliers.
4. The k-means algorithm iteratively calculates the distance from each point to the centroid of each cluster.

218. What happens with our second cluster centroid when we use the probability formula?

1. When we use the probability formula, we put less weight on the points that are far away. So, our second cluster centroid is likely going to be closer.
2. **When we use the probability formula, we put more weight on the points that are far away. So, our second cluster centroid is likely going to be more distant.**
3. When we use the probability formula, we put more weight on the lighter centroids, because it will take more computational power to draw our clusters. So, the second cluster centroid is likely going to be less distant.
4. When we use the probability formula, we put less weight on the points that are far away. So, our second cluster centroid is likely going to be more distant.

219. What is the implication of a small standard deviation of the clusters?

1. A small standard deviation of the clusters defines the size of the clusters.
2. The standard deviation of the cluster defines how tightly around each one of the centroids are. With a small standard deviation, the points will be closer to the centroids.
3. **The standard deviation of the cluster defines how tightly around each one of the centroids are. With a small standard deviation, we cannot find any centroids.**
4. A small standard deviation of the clusters means that the centroids are not close enough to each other.

220. After we plot our elbow and we find the inflection point, what does that point indicate to us?

1. **The ideal number of clusters.** c. The data points we need to form a cluster
2. How we can reduce our number of clusters. d. Whether we need to remove outliers.

221. What is one of the most suitable ways to choose K when the number of clusters is unclear?

1. You can start by choosing a random number of clusters.
2. **By evaluating Clustering performance such as Inertia and Distortion.**
3. By increasing the number of clusters calculate the square root.
4. You can start by using a k-nearest neighbor method.

222. Which statement describes correctly the use of distortion and inertia?

1. When the sum of the points equals a prime number, use inertia, and when the sum of the points equals a pair number, use distortion.
2. When we can calculate a number of clusters higher than 10, we use distortion, when we calculate a number of clusters smaller than 10, we use inertia.
3. When outliers are a concern use inertia, otherwise use distortion.
4. When the similarity of the points in the cluster is more important, you should use distortion, and if you are more concerned that clusters have similar numbers of points, then you should use inertia.

223. Which method is commonly used to select the right number of clusters?

1. **The elbow method.** c. The ROC curve.
2. The perfect Square Method d. The Sum of Square Method

224. What is the other name we can give to the L1 distance?

1. Hamming Distance b. Euclidean Distance **c. Manhattan Distance** d. Mahalanobis Distance

225. What is the key feature of the Jaccard Distance?

1. It takes into account the angle between the 2 points.
2. **It looks at the difference and similarities between sets of values.**
3. It describes distance by squaring each term, adding and squaring them.
4. It is obtained by adding up the absolute value of each term.

226. What is the advantage of the L1 distance over the L2?

1. It's better for data where the location of occurrence is less important.
2. It's useful for coordinate-based measurements.
3. It shows the difference between sets of values.
4. **It can better handle high-dimensional data.**

227. What is the other name we can give to the L2 distance?

1. Hamming Distance **b. Euclidean Distance** c. Manhattan Distance d. Mahalanobis Distance

228. Which of the following statements is a business case for the use of the Manhattan distance (L1)?

* 1. **We use it in business cases where there is very high dimensionality.**
  2. We use it in business cases with outliers.
  3. We use it in business cases where the dimensionality is unknown.
  4. We use it in business cases where there is low dimensionality.

229. What is the key feature of Cosine Distance?

1. **The Cosine Distance takes into account the angle between 2 points.**
2. The size of the curve.
3. It is sensitive to the size of the data set.
4. It is not sensitive to the size of the data set.

230. The following statement is an example of a business case where we can use the Cosine Distance

1. **Cosine is better for data such as text where the location of occurrence is less important.**
2. Cosine is useful for coordinate-based measurements.
3. Cosine distance is more sensitive to the curse of dimensionality
4. Cosine distance is less sensitive to the curse of dimensionality

231. Which distance metric is useful when we have text documents and we want to group similar topics together?

1. Euclidean b. Mahalanobis Distance c. Manhattan Distance **d. Jaccard**

232. How is a core point defined in the DBSCAN algorithm?

1. **A point that has more than n\_clu neighbors in their Ɛ-neighborhood.**
2. A point that has no points in its Ɛ-neighborhood.
3. A point that has the same amount of n\_clu neighbors within and outside the Ɛ-neighborhood.
4. An Ɛ-neighbor point has fewer than n\_clu neighbors itself.

233. Why do we need a stopping criterion when we are using the HAC?

1. The algorithm will turn our data into small clusters.
2. **The algorithm will turn our data into just one cluster.**
3. The algorithm will not start working if we don’t assign a number of clusters.
4. The stopping criterion ensures centroids are calculated correctly.

234. According to the DBSCAN required inputs, which statement describes the n\_clu input?

1. It's the function to calculate distance.
2. It's the radius of the local neighborhood.
3. **It determines the density threshold (for fixed Ɛ) (The minimum amount of points for a particular point to be considered a core point of a cluster)**
4. It's the maximum amount of observations for a particular point to be considered a core point of a cluster.

235. Which of the following statements is a characteristic of the DBSCAN algorithm?

1. **Can handle tons of data and weird shapes.**
2. Finds uneven cluster sizes (one is big, some are tiny).
3. It will do a great performance finding many clusters.
4. It will do a great performance finding a few clusters.

236. Which of the following statements is a characteristic of the Hierarchical Clustering (Ward) algorithm?

1. If we use a mini-batch to find our centroids and clusters this will find our clusters fairly quickly.
2. **It offers a lot of distance metrics and linkage options.**
3. Too small epsilon (too many clusters) is not trustworthy.
4. Too large epsilon (too few clusters) is not trustworthy.

237. Which of the following statements is a characteristic of the Mean Shift algorithm?

1. **Does not require to set the number of clusters; the number of clusters will be determined.**
2. Bad with non-spherical cluster shapes.
3. You need to decide the number of clusters on your own, choosing the numbers directly or the minimum distance threshold.
4. Good with non-spherical cluster shapes.

238. When using DBSCAN, how does the algorithm determine that a cluster is complete and is time to move to a different point of the data set and potentially start a new cluster?

1. When the algorithm requires you to change the input.
2. When the algorithm forms a new cluster using the outliers.
3. **When no point is left unvisited by the chain reaction.**
4. When the solution converges to a single cluster.

239. Which of the following statements correctly defines the strengths of the DBSCAN algorithm?

1. No need to specify the number of clusters (cf. K-means), allows for noise and can handle arbitrary-shaped clusters.
2. Do well with different densities, works with just one parameter, and the n\_clu defines itself.
3. The algorithm will find the outliers first, draw regular shapes, and works faster than other algorithms.
4. The algorithm is computationally intensive, it is sensitive to outliers, and it requires a few hyperparameters to be tune

240. Which of the following statements correctly defines the weaknesses of the DBSCAN algorithm?

1. The clusters it finds might not be trustworthy, it needs noisy data to work, and it can’t handle subgroups.
2. **It needs two parameters as input, finding appropriate values of Ɛ and n\_clu can be difficult, and it does not do well with clusters of different densities.**
3. The algorithm will find the outliers first, it draws regular shapes, and works faster than other algorithms.
4. The algorithm is computationally intensive, it is sensitive to outliers, and it requires too many hyperparameters to be tuned.

241. (True/false) Does complete linkage refer to the maximum pairwise distance between clusters?

1. **True** b. False

242. Which of the following measure methods computes the inertia and pick the pair that is going to ultimately minimize the inertia value?

1. Single linkage b. Average linkage **c. Ward linkage** d. Complete linkage

243. What is the purpose of dimensionality reduction in enterprise datasets?

1. **To improve model performance by reducing the number of features used.**
2. To create clusters for grouping data points.
3. To predict the target with the best accuracy.
4. To improve model performance by providing a ranking of the features and maximizing the features used

244. (True/False) Principal Component Analysis reduces dimensions by identifying features that can be excluded.

1. True **b. False**

245. Let’s say that PCA found two principal components v1 and v2. v1 accounts for 0.5 of the total amount ò variance in our dataset and v2 accounts for 0.24. Which one is more important and why?

1. **v1 because we will be able to maintain more of the original variance in the dataset**
2. v1 because it reduces 50 of the total variance in the data set.
3. v2 because it accounts for lower variance in the dataset
4. v2 because it reduces the amount of variance in the dataset

246. Select the option that best completes the following sentence: For data with many features, principal components analysis

1. identifies which features can be safely discarded
2. reduces the number of features without losing any information.
3. establishes a minimum number of viable features for use in the analysis.
4. **generates new features that are linear combinations of the original features.**

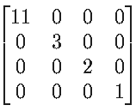
247. Which option correctly lists the steps for implementing PCA in Python?

1. Fit PCA to data 2. Scale the data

3. Determine the desired number of components based on total explained variance

4. Define a PCA object

1. 4, 1, 2, 3 b. 4, 1, 3, 2 **c. 2, 4, 1, 3** d. 2, 1, 3, 4

248. Given the following matrix for lengths of singular vectors, how do we rank the vectors in terms of importance?

1. **v1, v2, v3, v4**
2. v4, v3, v2, v1
3. v1, v4, v3, v2
4. v2, v3, v4, v1

249. (True/False) In PCA, the first principal component represents the most important feature in the dataset.

1. True **b. False**

250. Given two principal components v1, v2, let’s say that

Feature f1 contributed 0.15 to v1 and 0.25 to v2.

Feature f2 contributed -0.11 to v1 and 0.4 to v2

Which feature is more important according to their total contribution to the components?

1. v1 because 0.15 + 0.25 > -0.11 + 0.4 c. v2 because -0.11 + 0.4 < 0.15 + 0.25
2. **v2 because | -0.11| + | 0.4 | > | 0.15 | + | 0.25 |** d. Neither

251. (True/False) If the number of components is equal to the dimension of the original features, kernel PCA will reconstruct the data, returning the original.

1. True **b. False**

252. How does the goal of MDS (Multidimensional Scaling) compare to PCA?

1. PCA tries to maintain geometric distances between data points, whereas MDS tries to preserve variance within data.
2. Both MDS and PCA try to preserve variance within data.
3. **MDS tries to maintain geometric distances between data points, whereas PCA tries to preserve variance within data.**
4. Both MDS and PCA try to maintain geometric distances between data points.

253. Given the data visualized below with the classes represented by different colors, should PCA or kernel PCA be used, and why?

1. **Kernel PCA because the data is not linearly separable.**
2. Neither because the data cannot be projected onto a lower dimension.
3. PCA because the data is clearly separable when projected onto a lower dimension.
4. Either is fine because the two classes are clearly separable.

254. What is the main difference between kernel PCA and linear PCA?

1. The objective of linear PCA is to decrease the dimensionality of the space whereas the objective of Kernel PCA is to increase the dimensionality of the space.
2. Kernel PCA tends to uncover non-linearity structures within the dataset by increasing the dimensionality of the space thanks to the kernel trick.
3. Kernel PCA and Linear PCA are both Linear dimensionality reduction algorithms but they use a different optimization methods.
4. **Kernel PCA tends to preserve the geometric distances between the points while reducing the dimensionality of the space.**

255. (True/False) Multi-Dimensional Scaling (MDS) focuses on maintaining the geometric distances between points.

1. True **b. False**

256. Which of the following data types is more suitable for Kernel PCA than PCA?

1. **Data where the classes are not linearly separable.**
2. Data with linearly separable classes.
3. Data that do not need to be mapped to a higher dimension to distinguish categories.
4. None; they can be used interchangeably.

257. By applying for MDS, you are able to:

1. Attain higher dimensions for the features.
2. Maximize the distance between data points in a lower dimension.
3. Preserve variance within the original data.
4. **Find embeddings for points so that their distance is the most similar to the original distance.**

258. Which one of the following hyperparameters is NOT considered when using GridSearchCV for Kernel PCA?

1. n\_clusters **b. n\_components** c. gamma d. kernels

259. In which case would you prefer using PCA over NMF?

1. When canceling out with negative values is not desired.
2. **When you have a linear combination of features.**
3. When the original decomposition strictly contains positive values.
4. When you want to decompose videos, music, or images.

260. Which of the following is the most suitable for NMF?

1. **Reconstruct a text document with learned topics (features).**
2. Analyze potential movements and relationships of multiple stocks.
3. Predict the price of a rental space based on location, facility, and average rent in the surrounding area.
4. Learn features for a dataset in which negative values are highly insightful and valuable.

261. What is the key difference between NMF and PCA?

1. NMF decomposes the original matrix, whereas PCA does not.
2. PCA finds a representation of the data in a lower dimension, whereas NMF does not.
3. **The input matrix for NMF consists of only positive values.**
4. NMF requires orthogonal vectors created, whereas such constraint doesn't apply for PCA.

262. (True/False) In some applications, NMF can make for more human interpretable latent features.

1. **True** b. False

263. Which of the following set of features is the least adapted to NMF?

1. Word Count of the different words present in a text. c. Pixel color values of an Image.
2. Spectral decomposition of an audio file. **d. Monthly returns of a set of stock portfolios**.

264. (True/False) The NMF can produce different outputs depending on its initialization.

1. True **b. False**

265. In Practice lab: Non-Negative Matrix Factorization, why did we use "pairwise\_distances" from sci-kit-learn?

1. To calculate the pairwise distance between points of the NMF-encoded version of the original dataset.
2. **To calculate the pairwise distance between the NMF encoded version of the original dataset and the encoded query dataset.**
3. To calculate the pairwise distance between data points for eliminating outliers.
4. To calculate the maximum pairwise distance between points in the dataset.

266. (True/False) Machine Learning consists in programming computers to learn from real-time human interactions

1. True **b. False**

267.