Question 1

L2QM-EP0028-1908

LOS: LOS-6037

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

Which of the following statements regarding false positives and negatives is *correct*?

● Minimizing false positives (FPs) increases the chances of false negatives (FNs).

○ Maximizing false positives (FPs) increases the chances of false negatives (FNs).

○ Minimizing false positives (FPs) decreases the chances of false negatives (FNs).

Rationale

✓ **Minimizing false positives (FPs) increases the chances of false negatives (FNs).**

Correct. Minimizing false positives (FPs) increases the chances of false negatives (FNs).

Rationale

✗ **Maximizing false positives (FPs) increases the chances of false negatives (FNs).**

Incorrect. Minimizing false positives (FPs) increases the chances of false negatives (FNs).

Rationale

✗ **Minimizing false positives (FPs) decreases the chances of false negatives (FNs).**

Incorrect. Minimizing false positives (FPs) increases the chances of false negatives (FNs).

Question 2
L2QM-EP0008-1908
LOS: LOS-6032
Lesson Reference: Lesson 2: Data Preparation and Wrangling
Difficulty: hard
The term "Winsorization" would *most likely* apply to:

○ truncating data with outliers from the dataset.

● replacing outliers with a minimum or maximum value.

○ standardizing outliers to the mean and normalizing them to create a more homogeneous data set.

Rationale

✗ **truncating data with outliers from the dataset.**

Incorrect. Winsorization replaces extremely high outliers with a maximum value and extremely low outliers with a minimum value.

Rationale

✓ **replacing outliers with a minimum or maximum value.**

Correct. Winsorization replaces extremely high outliers with a maximum value and extremely low outliers with a minimum value.

Rationale

✗ **standardizing outliers to the mean and normalizing them to create a more homogeneous data set.**

Incorrect. Winsorization replaces extremely high outliers with a maximum value and extremely low outliers with a minimum value.

Question 3

L2QM-EP0017-1908

LOS: LOS-6034

Lesson Reference: Lesson 4: Model Training

Difficulty: medium

Which of the following types of error is *most likely* to result from an ML model underfitting the training set of data?

● Bias error

○ Variance error

○ Root mean square error (RMSE)

Rationale

✓ **Bias error**

Correct. Bias errors result from underfitting the data. Variance errors result from overfitting the data. RMSE is simply a measure that identifies the degree of error (computed by finding the square root of the mean of the squared differences between actual and predicted values).

Rationale

✗ **Variance error**

Incorrect. Bias errors result from underfitting the data. Variance errors result from overfitting the data. RMSE is simply a measure that identifies the degree of error (computed by finding the square root of the mean of the squared differences between actual and predicted values).

Rationale

✗ **Root mean square error (RMSE)**

Incorrect. Bias errors result from underfitting the data. Variance errors result from overfitting the data. RMSE is simply a measure that identifies the degree of error (computed by finding the square root of the mean of the squared differences between actual and predicted values).

Question 4

L2QM-EP0021-1908
LOS: LOS-6035
Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks
Difficulty: medium
Which of the following statements is *incorrect* with respect to text cleansing?

○ Percentage and dollar symbols are substituted with word annotations.

● Numerical values are kept because they can be used for sentiment prediction.

○ Semi-colons, commas, and special characters such as “+” and “©” are removed.

Rationale

✗ **Percentage and dollar symbols are substituted with word annotations.**
Incorrect. Percentage and dollar symbols are substituted with word annotations.

Rationale

✓ **Numerical values are kept because they can be used for sentiment prediction.**
Correct. Numerical values are removed because they have no use for sentiment prediction. Percentage and dollar symbols are substituted with word annotations before stripping out punctuations. Semi-colons, commas, and special characters such as “+” and “©” are removed.

Rationale

✗ **Semi-colons, commas, and special characters such as “+” and “©” are removed.**
Incorrect. Semi-colons, commas, and special characters such as “+” and “©” are removed.

Question 5

L2QM-EP0016-1908

LOS: LOS-6034

Lesson Reference: Lesson 4: Model Training

Difficulty: medium

With respect to a confusion matrix for data analysis, precision

○ equals the true positive rate.

● is most useful when the cost of a false positive (Type I error) is high.

○ is most useful when the cost of a false negative (Type II error) is high.

Rationale

✗ **equals the true positive rate.**

Incorrect. Precision is the percentage of correctly predicted positive classes (true positive) to all predicted positive classes (true and false positive): Precision (P) = TP/(TP + FP). Precision is most useful when the cost of a false positive is high. Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Recall rather than precision is also known as the *true positive rate* and is most useful when the cost of a false negative is high.

Rationale

✓ **is most useful when the cost of a false positive (Type I error) is high.**

Correct. Precision is the percentage of correctly predicted positive classes (true positive) to all predicted positive classes (true and false positive): Precision (P) = TP/(TP + FP). Precision is most useful when the cost of a false positive is high. Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Recall rather than precision is also known as the *true positive rate* and is most useful when the cost of a false negative is high.

Rationale

✗ **is most useful when the cost of a false negative (Type II error) is high.**

Incorrect. Precision is the percentage of correctly predicted positive classes (true positive) to all predicted positive classes (true and false positive): Precision (P) = TP/(TP + FP). Precision is most useful when the cost of a false positive is high. Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Recall rather than precision is also known as the *true positive rate* and is most useful when the cost of a false negative is high.

Question 6
L2QM-EP0013-1908
LOS: LOS-6033
Lesson Reference: Lesson 3: Data Exploration Objectives and Methods
Difficulty: medium

Which of the following feature selection methods measures how much information a token contributes to a class of texts without concern for the sample size?

○ Chi-square testing

● Mutual information

○ Document frequency

Rationale

✗ **Chi-square testing**

Incorrect. Mutual information (MI) measures how much information a token contributes to a class of texts. The MI approaches 1 as a specific token is identified as occurring only in a specific class of text. Both mutual information and Chi-square testing attempt to determine token membership to a class; Chi-squared testing measures occurrence of the token and occurrence of the class while MI measures the occurrences of the token based on the occurrence of the class. Chi-squared relies on the number of observations to determine the chi-square value.

Rationale

✓ **Mutual information**

Correct. Mutual information (MI) measures how much information a token contributes to a class of texts. The MI approaches 1 as a specific token is identified as occurring only in a specific class of text. Both mutual information and Chi-square testing attempt to determine token membership to a class; Chi-squared testing measures occurrence of the token and occurrence of the class while MI measures the occurrences of the token based on the occurrence of the class. Chi-squared relies on the number of observations to determine the chi-square value.

Rationale

✗ **Document frequency**

Incorrect. Mutual information (MI) measures how much information a token contributes to a class of texts. The MI approaches 1 as a specific token is identified as occurring only in a specific class of text. Both mutual information and Chi-square testing attempt to determine token membership to a class; Chi-squared testing measures occurrence of the token and occurrence of the class while MI measures the occurrences of the token based on the occurrence of the class. Chi-squared relies on the number of observations to determine the chi-square value.

Question 7

L2QM-EP0003-1908

LOS: LOS-6031

Lesson Reference: Lesson 1: Big Data in Investment Management

Difficulty: medium

Machine learning (ML) methods take into account:

○ structured data such as sentiment.

● unstructured data such as news articles.

○ structured data such as what people are talking about.

Rationale

✗ **structured data such as sentiment.**

Incorrect. ML methods take into account unstructured data such as topics (i.e., what people are talking about), sentiment (how people feel), and textual big data (e.g., online news articles, internet financial forums, social networking platforms). Sentiment and what people are talking about are not considered structured data.

Rationale

✓ **unstructured data such as news articles.**

Correct. ML methods take into account unstructured data such as topics (i.e., what people are talking about), sentiment (how people feel), and textual big data (e.g., online news articles, internet financial forums, social networking platforms). Sentiment and what people are talking about are not considered structured data.

Rationale

✗ **structured data such as what people are talking about.**

Incorrect. ML methods take into account unstructured data such as topics (i.e., what people are talking about), sentiment (how people feel), and textual big data (e.g., online news articles, internet financial forums, social networking platforms). Sentiment and what people are talking about are not considered structured data.

Question 8

L2QM-EP0012-1908

LOS: LOS-6033

Lesson Reference: Lesson 3: Data Exploration Objectives and Methods

Difficulty: medium

Which of the following statements regarding feature engineering techniques is *incorrect*?

● N-grams are discriminative three-word patterns.

○ Numbers are converted into a token such as “/number/.”

○ Name entity recognition (NER) algorithm analyzes individual tokens and their surrounding semantics to tag an object class to the token.

Rationale

✓ **N-grams are discriminative three-word patterns.**

Correct. N-grams are discriminative multi-word patterns.

Rationale

✗ **Numbers are converted into a token such as “/number/.”**

Incorrect. Numbers are converted into a token such as “/number/.”

Rationale

✗ **Name entity recognition (NER) algorithm analyzes individual tokens and their surrounding semantics to tag an object class to the token.**

Incorrect. Name entity recognition (NER) algorithm analyzes individual tokens and their surrounding semantics to tag an object class to the token.

Question 9

L2QM-EP0001-1908

LOS: LOS-6031

Lesson Reference: Lesson 1: Big Data in Investment Management

Difficulty: medium

Identify which of the following accurately describes the term velocity in the context of big data management.

○ The internal and external availability of structured, semi-structured, and unstructured data.

○ The credibility and reliability of various data sources and the need to find quality data within a large quantity of data.

● The speed at which data are created, such as how many “tweets” or internet searches occur on a daily basis.

Rationale

✗ **The internal and external availability of structured, semi-structured, and unstructured data.**

Incorrect. **Variety** refers to the internal and external availability of structured, semi-structured, and unstructured data, including traditional transactional data; user-generated text, images, and videos; social media; sensor-based data; web and mobile clickstreams; and spatial-temporal data.

Rationale

✗ **The credibility and reliability of various data sources and the need to find quality data within a large quantity of data.**

Incorrect. **Veracity** refers to the credibility and reliability of various data sources and the need to find quality data within a large quantity of data. For example, as much as 10%–15% of social media is fake, spam accounts for 20% of Internet content, and clickstreams are very susceptible to noise.

Rationale

✓ **The speed at which data are created, such as how many “tweets” or internet searches occur on a daily basis.**

Correct. **Velocity** is the speed at which data are created, such as how many “tweets” or internet searches occur on a daily basis. This information has important implications for real-time predictive analytics. Analyzing “data-in-motion” is challenging because finding relevant patterns and insights is a moving target relative to “data-at-rest.”

Question 10

L2QM-EP0025-1908

LOS: LOS-6036

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

Higher term frequency-inverse document frequency (TF–IDF) values indicate words that appear

○ more frequently within a larger number of documents, signifying relatively important, unique terms.

○ less frequently within a smaller number of documents, signifying relatively important, unique terms.

● more frequently within a smaller number of documents, signifying relatively important, unique terms.

Rationale

✗ **more frequently within a larger number of documents, signifying relatively important, unique terms.**

Incorrect. Higher TF–IDF values indicate words that appear more frequently within a smaller number of documents, signifying relatively important, unique terms.

Rationale

✗ **less frequently within a smaller number of documents, signifying relatively important, unique terms.**

Incorrect. Higher TF–IDF values indicate words that appear more frequently within a smaller number of documents, signifying relatively important, unique terms.

Rationale

✓ **more frequently within a smaller number of documents, signifying relatively important, unique terms.**

Correct. Higher TF–IDF values indicate words that appear more frequently within a smaller number of documents, signifying relatively important, unique terms.

Question 11

L2QM-EP0009-1908

LOS: LOS-6032

Lesson Reference: Lesson 2: Data Preparation and Wrangling

Difficulty: medium

Which of the following statements regarding text wrangling is *incorrect*?

● Stop words are kept to increase the number of tokens involved in ML training.

○ Stemming would convert the words “analyzed” and “analyzing” to “analyz.”

○ Lemmatization would convert the words “analyzed” and “analyzing” to “analyze.”

Rationale

✓ **Stop words are kept to increase the number of tokens involved in ML training.**

Correct. Stop words are usually removed to reduce the number of tokens involved in ML training.

Rationale

✗ **Stemming would convert the words “analyzed” and “analyzing” to “analyz.”**

Incorrect. Stop words are usually removed to reduce the number of tokens involved in ML training.

Rationale

✗ **Lemmatization would convert the words “analyzed” and “analyzing” to “analyze.”**

Incorrect. Stop words are usually removed to reduce the number of tokens involved in ML training.

Question 12

L2QM-EP0004-1908

LOS: LOS-6031

Lesson Reference: Lesson 1: Big Data in Investment Management

Difficulty: easy

Web spidering (scraping or crawling) programs extract:

● raw content.

○ processed content.

○ numerical content only.

Rationale

✓ **raw content.**

Correct. Web spidering (scraping or crawling) programs extract raw content in any form.

Rationale

✗ **processed content.**

Incorrect. Web spidering (scraping or crawling) programs extract raw content in any form. This content may be used for later processing.

Rationale

✗ **numerical content only.**

Incorrect. Web spidering (scraping or crawling) programs extract raw content in any form.

Question 13

L2QM-EP0020-1908

LOS: LOS-6035

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

“Corpus” could *best* be described as any collection of

● raw text data.

○ wrangled text data.

○ preprocessed text data.

Rationale

✓ **raw text data.**

Correct. A corpus is any collection of raw text data.

Rationale

✗ **wrangled text data.**

Incorrect. A corpus is any collection of raw text data.

Rationale

✗ **preprocessed text data.**

Incorrect. A corpus is any collection of raw text data.

Question 14

L2QM-EP0014-1908

LOS: LOS-6033

Lesson Reference: Lesson 3: Data Exploration Objectives and Methods

Difficulty: medium

Identify which of the following statements regarding standardization is correct.

○ Data must be lognormally distributed to be used in standardization.

● Data must be normally distributed to be used in standardization.

○ Any type of data can be used in standardization.

Rationale

✗ **Data must be lognormally distributed to be used in standardization.**

Incorrect. Data must be normally distributed for use in standardization. Standardization divides the distance from the mean by the standard deviation of the dataset. This is only useful if the dataset is normally distributed.

Rationale

✓ **Data must be normally distributed to be used in standardization.**

Correct. Data must be normally distributed for use in standardization. Standardization divides the distance from the mean by the standard deviation of the dataset. This is only useful if the dataset is normally distributed.

Rationale

✗ **Any type of data can be used in standardization.**

Incorrect. Data must be normally distributed for use in standardization. Standardization divides the distance from the mean by the standard deviation of the dataset. This is only useful if the dataset is normally distributed.

Question 15

L2QM-EP0007-1908

LOS: LOS-6032

Lesson Reference: Lesson 2: Data Preparation and Wrangling

Difficulty: hard

Which of the following is *correct* with respect to min–max normalization?

○ It is not a type of scaling.

○ Outliers may remain in the data for processing.

● Min-max normalization may be used for non-normal data distributions.

Rationale

✗ **It is not a type of scaling.**

Incorrect. The purpose of scaling is to wrangle data values into a range that will be more useful to the machine learning process. Outliers must be removed before normalizing the data, and both min-max normalization and standardization (another type of scaling) are affected by outliers. Normalization may be used with non-normal data distributions. The formula to normalize variable X for each observation (X_i) is:

$X_{i\ (normalized)} = (X_i - X_{min}) / (X_{max} - X_{min})$

Rationale

✗ **Outliers may remain in the data for processing.**

Incorrect. The purpose of scaling is to wrangle data values into a range that will be more useful to the machine learning process. Outliers must be removed before normalizing the data, and both min-max normalization and standardization (another type of scaling) are affected by outliers. Normalization may be used with non-normal data distributions. The formula to normalize variable X for each observation (X_i) is:

$X_{i\ (normalized)} = (X_i - X_{min}) / (X_{max} - X_{min})$

Rationale

✓ **Min-max normalization may be used for non-normal data distributions.**

Correct. The purpose of scaling is to wrangle data values into a range that will be more useful to the machine learning process. Outliers must be removed before normalizing the data, and both min−max normalization and standardization (another type of scaling) are affected by outliers. Normalization may be used with non-normal data distributions. The formula to normalize variable X for each observation (X_i) is:

$X_{i\ (normalized)} = (X_i - X_{min}) / (X_{max} - X_{min})$

Question 16

L2QM-EP0023-1908

LOS: LOS-6036

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

Term frequency (TF) values range from:

● 0 to 1.

○ −1 to 0.

○ −1 to 1.

Rationale

✓ **0 to 1.**

Correct. TF values range from 0 to 1.

Rationale

✗ **−1 to 0.**

Incorrect. TF values range from 0 to 1.

Rationale

✗ **−1 to 1.**

Incorrect. TF values range from 0 to 1.

Question 17

L2QM-EP0027-1908

LOS: LOS-6037

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: easy

With respect to method selection for evaluating a machine learning model, logistic regression coefficients:

○ indicate the probability of positive sentiment.

● are used to find the probability of positive sentiment.

○ cannot be used to find the probability of positive sentiment.

Rationale

✗ **indicate the probability of positive sentiment.**

Incorrect: Logistic regression coefficients are used to find the probability of positive sentiment. A mathematical function is used to convert the regression coefficient to a probability.

Rationale

✓ **are used to find the probability of positive sentiment.**

Correct: Logistic regression coefficients are used to find the probability of positive sentiment. A mathematical function is used to convert the regression coefficient to a probability.

Rationale

✗ **cannot be used to find the probability of positive sentiment.**

Incorrect. Logistic regression coefficients are used to find the probability of positive sentiment. A mathematical function is used to convert the regression coefficient to a probability.

Question 18

L2QM-EP0024-1908

LOS: LOS-6036

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: easy

The process that differentiates necessary tokens (features) for a dataset from noise could best be described as:

○ Propagation

○ Lemmatization

● Frequency analysis

Rationale

☒ **Propagation**

Incorrect. Frequency analysis filters unnecessary tokens (features); that is, those which are too frequent to provide meaning or too infrequent to indicate belonging. Lemmatization converts inflected forms of a word into its root.

Rationale

☒ **Lemmatization**

Incorrect. Frequency analysis filters unnecessary tokens (features); that is, those which are too frequent to provide meaning or too infrequent to indicate belonging. Lemmatization converts inflected forms of a word into its root.

Rationale

✓ **Frequency analysis**

Correct. Frequency analysis filters unnecessary tokens (features); that is, those which are too frequent to provide meaning or too infrequent to indicate belonging. Lemmatization converts inflected forms of a word into its root.

Question 19

L2QM-EP0006-1908

LOS: LOS-6032

Lesson Reference: Lesson 2: Data Preparation and Wrangling

Difficulty: easy

The date of birth shown in a dataset was outside a normal human lifespan. This would *most likely* be considered an example of data:

● invalidity.

○ inaccuracy.

○ inconsistency.

Rationale

✓ **invalidity.**

Correct. Invalidity errors occurs when the data are outside of a meaningful range. Inaccuracy occurs when data are not a measure of true value such as with a numeric value in a true/false response area. Inconsistency occurs when data conflicts with other data points or reality such as when a female solicitation (e.g., Mrs.) has been used with a male name.

Rationale

✗ **inaccuracy.**

Incorrect. Invalidity errors occurs when the data are outside of a meaningful range. Inaccuracy occurs when data are not a measure of true value such as with a numeric value in a true/false response area. Inconsistency occurs when data conflicts with other data points or reality such as when a female solicitation (e.g., Mrs.) has been used with a male name.

Rationale

✗ **inconsistency.**

Incorrect. Invalidity errors occurs when the data are outside of a meaningful range. Inaccuracy occurs when data are not a measure of true value such as with a numeric value in a true/false response area. Inconsistency occurs when data conflicts with other data points or reality such as when a female solicitation (e.g., Mrs.) has been used with a male name.

Question 20

L2QM-EP0026-1908

LOS: LOS-6037

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

Which of the following statements about partitioning cleansed and preprocessed data is *correct?*

○ Test set comprises 60% of data.

○ Training set comprises 20% of data.

● Cross-validation set comprises 20% of data.

Rationale

✗ **Test set comprises 60% of data.**

Incorrect. Training sets, cross-validation sets, and test sets are partitioned using a common ratio of 60:20:20, respectively.

Rationale

✗ **Training set comprises 20% of data.**

Incorrect. Training sets, cross-validation sets, and test sets are partitioned using a common ratio of 60:20:20, respectively.

Rationale

✓ **Cross-validation set comprises 20% of data.**

Correct. Training sets, cross-validation sets, and test sets are partitioned using a common ratio of 60:20:20, respectively.

Question 21

L2QM-EP0018-1908

LOS: LOS-6034

Lesson Reference: Lesson 4: Model Training

Difficulty: medium

Which of the following is the *most appropriate* measure of a good model when a class dataset is unequally distributed and it is necessary to balance bias and variance errors?

○ Recall

● F1 score

○ Accuracy

Rationale

✗ **Recall**

Incorrect. F1 score would be most appropriate in this circumstance. F1 score is the harmonic mean of precision P and recall R; therefore, F1 score = (2 × P × R)/(P + R). Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Accuracy is the percentage of correctly predicted positive classes out of all predictions: Accuracy = (TP +TN)/(TP + FP + TN + FN).

Rationale

✓ **F1 score**

Correct. F1 score would be most appropriate in this circumstance. F1 score is the harmonic mean of precision P and recall R; therefore, F1 score = (2 × P × R)/(P + R). Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Accuracy is the percentage of correctly predicted positive classes out of all predictions: Accuracy = (TP +TN)/(TP + FP + TN + FN).

Rationale

✗ **Accuracy**

Incorrect. F1 score would be most appropriate in this circumstance. F1 score is the harmonic mean of precision *P* and recall *R*; therefore, F1 score = (2 × P × R)/(P + R). Recall is the percentage of correctly predicted positive classes to all positive classes (true positives and false negatives): Recall (P) = TP/(TP + FN). Accuracy is the percentage of correctly predicted positive classes out of all predictions: Accuracy = (TP +TN)/(TP + FP + TN + FN).

Question 22

L2QM-EP0019-1908

LOS: LOS-6035

Lesson Reference: Lesson 5: Classifying and Predicting Sentiment for Stocks

Difficulty: medium

Stop words are *most likely* to be included or excluded during the

● data exploration stage.

○ machine learning (ML) stage.

○ text wrangling (preprocessing) stage.

Rationale

✓ **data exploration stage.**

Correct. No words are removed during the text wrangling (preprocessing) stage, although they may be marked. Words like “a,” “an,” and “the” may be removed during the data exploration stage if they are found to be of no value in the process. Custom stop words will also be identified during the data exploration stage.

Rationale

✗ **machine learning (ML) stage.**

Incorrect. No words are removed during the text wrangling (preprocessing) stage, although they may be marked. Words like “a,” “an,” and “the” may be removed during the data exploration stage if they are found to be of no value in the process. Custom stop words will also be identified during the data exploration stage.

Rationale

✗ **text wrangling (preprocessing) stage.**

Incorrect. No words are removed during the text wrangling (preprocessing) stage, although they may be marked. Words like “a,” “an,” and “the” may be removed during the data exploration stage if they are found to be of no value in the process. Custom stop words will also be identified during the data exploration stage.

Question 23

L2QM-EP0011-1908
LOS: LOS-6033
Lesson Reference: Lesson 3: Data Exploration Objectives and Methods
Difficulty: easy

The term frequency of a token that occurs 10 times in a dataset consisting of 100 tokens is *closest to*:

○ 10

● 10%

○ 1000

Rationale

✗ **10**

Incorrect. Term frequency (TF) is the ratio of how often a particular token occurs to the total number of tokens in the dataset: 10/100 = 10%.

Rationale

✓ **10%**

Correct. Term frequency (TF) is the ratio of how often a particular token occurs to the total number of tokens in the dataset: 10/100 = 10%.

Rationale

✗ **1000**

Incorrect. Term frequency (TF) is the ratio of how often a particular token occurs to the total number of tokens in the dataset: 10/100 = 10%.