1. A machine learning engineer has created a Feature Table new\_table using Feature Store Client fs. When creating the table, they specified a metadata description with key information about the Feature Table. They now want to retrieve that metadata programmatically.  
Which of the following lines of code will return the metadata description?

A. There is no way to return the metadata description programmatically.

B. fs.create\_training\_set("new\_table")

**C. fs.get\_table("new\_table").description**

D. fs.get\_table("new\_table").load\_df()

E. fs.get\_table("new\_table")

2. A data scientist has a Spark DataFrame spark\_df. They want to create a new Spark DataFrame that contains only the rows from spark\_df where the value in column price is greater than 0.  
Which of the following code blocks will accomplish this task?

A. spark\_df[spark\_df["price"] > 0]

**B. spark\_df.filter(col("price") > 0)**

C. SELECT \* FROM spark\_df WHERE price > 0

D. spark\_df.loc[spark\_df["price"] > 0,:]

E. spark\_df.loc[:,spark\_df["price"] > 0]

3. A health organization is developing a classification model to determine whether or not a patient currently has a specific type of infection. The organization's leaders want to maximize the number of positive cases identified by the model.  
Which of the following classification metrics should be used to evaluate the model?

A. RMSE

B. Precision

C. Area under the residual operating curve

D. Accuracy

**E. Recall**

4. In which of the following situations is it preferable to impute missing feature values with their median value over the mean value?

A. When the features are of the categorical type

B. When the features are of the boolean type

**C. When the features contain a lot of extreme outliers**

D. When the features contain no outliers

E. When the features contain no missing values

5. A data scientist has replaced missing values in their feature set with each respective feature variable’s median value. A colleague suggests that the data scientist is throwing away valuable information by doing this.  
Which of the following approaches can they take to include as much information as possible in the feature set?

A. Impute the missing values using each respective feature variable’s mean value instead of the median value

B. Refrain from imputing the missing values in favor of letting the machine learning algorithm determine how to handle them

C. Remove all feature variables that originally contained missing values from the feature set

D**. Create a binary feature variable for each feature that contained missing values indicating whether each row’s value has been imputed**

E. Create a constant feature variable for each feature that contained missing values indicating the percentage of rows from the feature that was originally missing

6. A data scientist is wanting to explore summary statistics for Spark DataFrame spark\_df. The data scientist wants to see the count, mean, standard deviation, minimum, maximum, and interquartile range (IQR) for each numerical feature.  
Which of the following lines of code can the data scientist run to accomplish the task?

**A. spark\_df.summary ()**

B. spark\_df.stats()

C. spark\_df.describe().head()

D. spark\_df.printSchema()

E. spark\_df.toPandas()

7. An organization is developing a feature repository and is electing to one-hot encode all categorical feature variables. A data scientist suggests that the categorical feature variables should not be one-hot encoded within the feature repository.  
Which of the following explanations justifies this suggestion?

A. One-hot encoding is not supported by most machine learning libraries.

B. One-hot encoding is dependent on the target variable’s values which differ for each application.

C. One-hot encoding is computationally intensive and should only be performed on small samples of training sets for individual machine learning problems.

D. One-hot encoding is not a common strategy for representing categorical feature variables numerically.

**E. One-hot encoding is a potentially problematic categorical variable strategy for some machine learning algorithms.**

[Reveal Solution](https://www.examtopics.com/exams/databricks/certified-machine-learning-associate/view/2/#)

8. A data scientist has created two linear regression models. The first model uses price as a label variable and the second model uses log(price) as a label variable. When evaluating the RMSE of each model by comparing the label predictions to the actual price values, the data scientist notices that the RMSE for the second model is much larger than the RMSE of the first model.  
Which of the following possible explanations for this difference is invalid?

A. The second model is much more accurate than the first model

**B. The data scientist failed to exponentiate the predictions in the second model prior to computing the RMSE**

C. The data scientist failed to take the log of the predictions in the first model prior to computing the RMSE

D. The first model is much more accurate than the second model

E. The RMSE is an invalid evaluation metric for regression problems

9. A data scientist uses 3-fold cross-validation when optimizing model hyperparameters for a regression problem. The following root-mean-squared-error values are calculated on each of the validation folds:  
• 10.0  
• 12.0  
• 17.0  
Which of the following values represents the overall cross-validation root-mean-squared error?

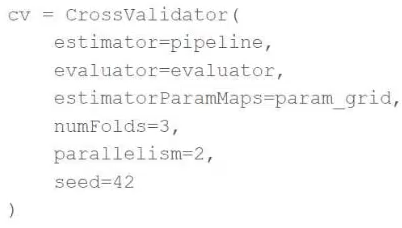
**A. 13.0**

B. 17.0

C. 12.0

D. 39.0

E. 10.0

10. A machine learning engineer is trying to scale a machine learning pipeline pipeline that contains multiple feature engineering stages and a modeling stage. As part of the cross-validation process, they are using the following code block:  
  
A colleague suggests that the code block can be changed to speed up the tuning process by passing the model object to the estimator parameter and then placing the updated cv object as the final stage of the pipeline in place of the original model.  
Which of the following is a negative consequence of the approach suggested by the colleague?

A. The model will take longer to train for each unique combination of hyperparameter values

**B. The feature engineering stages will be computed using validation data**

C. The cross-validation process will no longer be parallelizable

D. The cross-validation process will no longer be reproducible

E. The model will be refit one more per cross-validation fold

11. What is the name of the method that transforms categorical features into a series of binary indicator feature variables?

A. Leave-one-out encoding

B. Target encoding

**C. One-hot encoding**

D. Categorical embeddings

E. String indexing

12. A data scientist wants to parallelize the training of trees in a gradient boosted tree to speed up the training process. A colleague suggests that parallelizing a boosted tree algorithm can be difficult.  
Which of the following describes why?

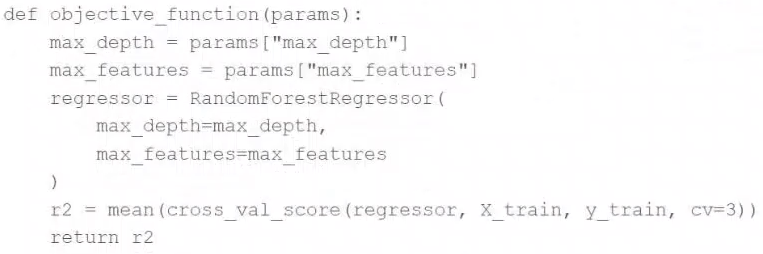
A. Gradient boosting is not a linear algebra-based algorithm which is required for parallelization.

B. Gradient boosting requires access to all data at once which cannot happen during parallelization.

C. Gradient boosting calculates gradients in evaluation metrics using all cores which prevents parallelization.

**D. Gradient boosting is an iterative algorithm that requires information from the previous iteration to perform the next step.**

E. Gradient boosting uses decision trees in each iteration which cannot be parallelized.

13. A data scientist wants to efficiently tune the hyperparameters of a scikit-learn model. They elect to use the Hyperopt library's fmin operation to facilitate this process. Unfortunately, the final model is not very accurate. The data scientist suspects that there is an issue with the objective\_function being passed as an argument to fmin.  
They use the following code block to create the objective\_function:  
  
Which of the following changes does the data scientist need to make to their objective\_function in order to produce a more accurate model?

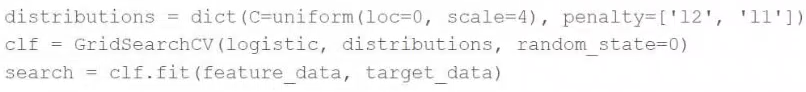
A. Add test set validation process

B. Add a random\_state argument to the RandomForestRegressor operation

C. Remove the mean operation that is wrapping the cross\_val\_score operation

**D. Replace the r2 return value with -r2**

E. Replace the fmin operation with the fmax operation

14. A data scientist is attempting to tune a logistic regression model logistic using scikit-learn. They want to specify a search space for two hyperparameters and let the tuning process randomly select values for each evaluation.  
They attempt to run the following code block, but it does not accomplish the desired task:  
  
Which of the following changes can the data scientist make to accomplish the task?

**A. Replace the GridSearchCV operation with RandomizedSearchCV**

B. Replace the GridSearchCV operation with cross\_validate

C. Replace the GridSearchCV operation with ParameterGrid

D. Replace the random\_state=0 argument with random\_state=1

E. Replace the penalty= ['12', '11'] argument with penalty=uniform ('12', '11')

15. Which of the following tools can be used to parallelize the hyperparameter tuning process for single-node machine learning models using a Spark cluster?

A. MLflow Experiment Tracking

B. Spark ML

C. Autoscaling clusters

**D. Hyperopt**

E. Delta Lake

16. Which of the following describes the relationship between native Spark DataFrames and pandas API on Spark DataFrames?

A. pandas API on Spark DataFrames are single-node versions of Spark DataFrames with additional metadata

B. pandas API on Spark DataFrames are more performant than Spark DataFrames

**C. pandas API on Spark DataFrames are made up of Spark DataFrames and additional metadata**

D. pandas API on Spark DataFrames are less mutable versions of Spark DataFrames

E. pandas API on Spark DataFrames are unrelated to Spark DataFrames

17. A data scientist has written a data cleaning notebook that utilizes the pandas library, but their colleague has suggested that they refactor their notebook to scale with big data.  
Which of the following approaches can the data scientist take to spend the least amount of time refactoring their notebook to scale with big data?

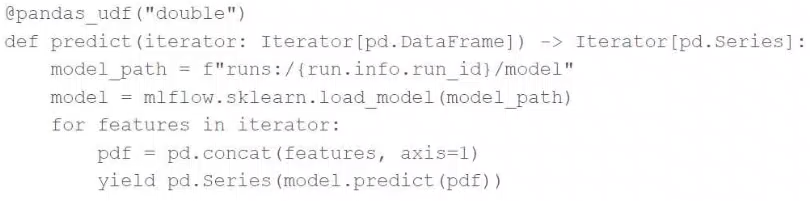
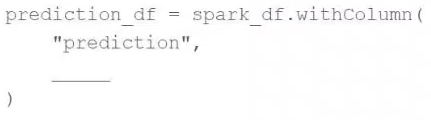
**A.** They can refactor their notebook to process the data in parallel.

**B.** They can refactor their notebook to use the PySpark DataFrame API.

**C.** They can refactor their notebook to use the Scala Dataset API.

**D.** They can refactor their notebook to use Spark SQL.

**E. They can refactor their notebook to utilize the pandas API on Spark.**

18. A data scientist has defined a Pandas UDF function predict to parallelize the inference process for a single-node model:  
  
They have written the following incomplete code block to use predict to score each record of Spark DataFrame spark\_df:  
  
19. Which of the following lines of code can be used to complete the code block to successfully complete the task?

**A.** predict(\*spark\_df.columns)

**B. mapInPandas(predict)**

**C.** predict(Iterator(spark\_df))

**D.** mapInPandas(predict(spark\_df.columns))

**E.** predict(spark\_df.columns)

20. Which of the Spark operations can be used to randomly split a Spark DataFrame into a training DataFrame and a test DataFrame for downstream use?

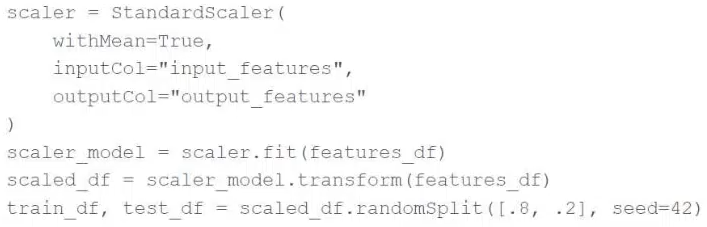
**A.** TrainValidationSplit

**B.** DataFrame.where

**C.** CrossValidator

**D.** TrainValidationSplitModel

**E. DataFrame.randomSplit**

21. A data scientist is using Spark ML to engineer features for an exploratory machine learning project.  
They decide they want to standardize their features using the following code block:  
  
22. Upon code review, a colleague expressed concern with the features being standardized prior to splitting the data into a training set and a test set.  
Which of the following changes can the data scientist make to address the concern?

**A.** Utilize the MinMaxScaler object to standardize the training data according to global minimum and maximum values

**B.** Utilize the MinMaxScaler object to standardize the test data according to global minimum and maximum values

**C.** Utilize a cross-validation process rather than a train-test split process to remove the need for standardizing data

**D.** Utilize the Pipeline API to standardize the training data according to the test data's summary statistics

**E. Utilize the Pipeline API to standardize the test data according to the training data's summary statistics**

23. A machine learning engineer is trying to scale a machine learning pipeline by distributing its feature engineering process.  
Which of the following feature engineering tasks will be the least efficient to distribute?

**A.** One-hot encoding categorical features

**B. Target encoding categorical features**

**C.** Imputing missing feature values with the mean

**D.** Imputing missing feature values with the true median

**E.** Creating binary indicator features for missing values

24. Which of the following is a benefit of using vectorized pandas UDFs instead of standard PySpark UDFs?

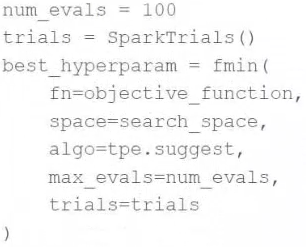
**A.** The vectorized pandas UDFs allow for the use of type hints

**B. The vectorized pandas UDFs process data in batches rather than one row at a time**

**C.** The vectorized pandas UDFs allow for pandas API use inside of the function

**D.** The vectorized pandas UDFs work on distributed DataFrames

**E.** The vectorized pandas UDFs process data in memory rather than spilling to disk

25. A data scientist wants to tune a set of hyperparameters for a machine learning model. They have wrapped a Spark ML model in the objective function objective\_function and they have defined the search space search\_space.  
As a result, they have the following code block:  
  
Which of the following changes do they need to make to the above code block in order to accomplish the task?

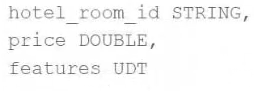
**A. Change SparkTrials() to Trials()**

**B.** Reduce num\_evals to be less than 10

**C.** Change fmin() to fmax()

**D.** Remove the trials=trials argument

**E.** Remove the algo=tpe.suggest argument

26. A machine learning engineer would like to develop a linear regression model with Spark ML to predict the price of a hotel room. They are using the Spark DataFrame train\_df to train the model.  
The Spark DataFrame train\_df has the following schema:  
  
The machine learning engineer shares the following code block:  
  
Which of the following changes does the machine learning engineer need to make to complete the task?

**A.** They need to call the transform method on train\_df

**B. They need to convert the features column to be a vector**

**C.** They do not need to make any changes

**D.** They need to utilize a Pipeline to fit the model

**E.** They need to split the features column out into one column for each feature

27. Which of the following tools can be used to distribute large-scale feature engineering without the use of a UDF or pandas Function API for machine learning pipelines?

**A.** Keras

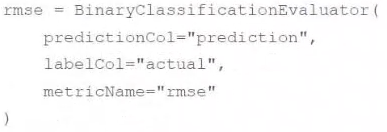
**B.** pandas

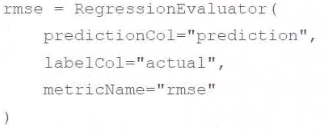
**C.** PyTorch

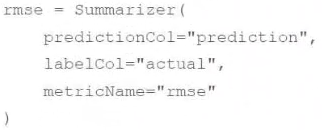
**D. Spark ML**

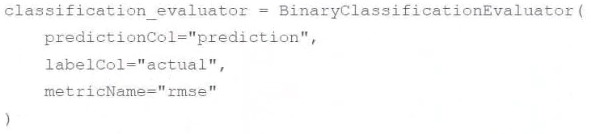
**E.** Scikit-learn

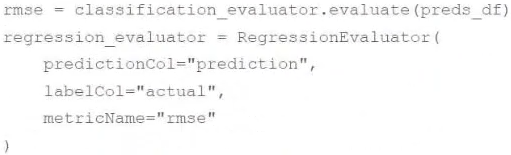
28. A data scientist has developed a linear regression model using Spark ML and computed the predictions in a Spark DataFrame preds\_df with the following schema: prediction DOUBLE actual DOUBLE  
Which of the following code blocks can be used to compute the root mean-squared-error of the model according to the data in preds\_df and assign it to the rmse variable?

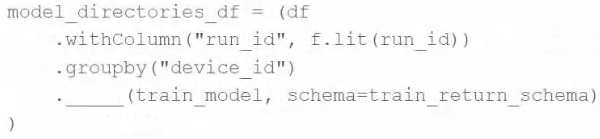
**A.** 

**B.** 

**C.** 

**D.** 

**E.** 

29. A machine learning engineer wants to parallelize the training of group-specific models using the Pandas Function API. They have developed the train\_model function, and they want to apply it to each group of DataFrame df.  
They have written the following incomplete code block:  
  
Which of the following pieces of code can be used to fill in the above blank to complete the task?

**A. applyInPandas**

**B.** mapInPandas

**C.** predict

**D.** train\_model

**E.** groupedApplyIn

30. Which of the following statements describes a Spark ML estimator?

**A.** An estimator is a hyperparameter grid that can be used to train a model

**B.** An estimator chains multiple algorithms together to specify an ML workflow

**C.** An estimator is a trained ML model which turns a DataFrame with features into a DataFrame with predictions

**D. An estimator is an algorithm which can be fit on a DataFrame to produce a Transformer**

**E.** An estimator is an evaluation tool to assess to the quality of a mode

31. A data scientist has been given an incomplete notebook from the data engineering team. The notebook uses a Spark DataFrame spark\_df on which the data scientist needs to perform further feature engineering. Unfortunately, the data scientist has not yet learned the PySpark DataFrame API.  
Which of the following blocks of code can the data scientist run to be able to use the pandas API on Spark?

**A. import pyspark.pandas as ps**  
**df = ps.DataFrame(spark\_df)**

**B.** import pyspark.pandas as ps  
df = ps.to\_pandas(spark\_df)

**C.** spark\_df.to\_sql()

**D.** import pandas as pd  
df = pd.DataFrame(spark\_df)

**E.** spark\_df.to\_pandas()

32. A data scientist has produced two models for a single machine learning problem. One of the models performs well when one of the features has a value of less than 5, and the other model performs well when the value of that feature is greater than or equal to 5. The data scientist decides to combine the two models into a single machine learning solution.  
Which of the following terms is used to describe this combination of models?

**A.** Bootstrap aggregation

**B.** Support vector machines

**C.** Bucketing

**D. Ensemble learning**

**E.** Stacking

33. Which statement describes a Spark ML transformer?

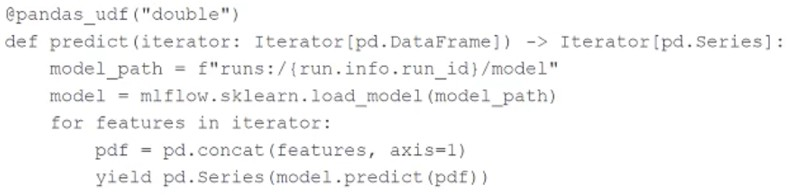
**AA transformer is an algorithm which can transform one DataFrame into another DataFrame**

**B**A transformer is a hyperparameter grid that can be used to train a model

**C**A transformer chains multiple algorithms together to transform an ML workflow

**D**A transformer is a learning algorithm that can use a DataFrame to train a model

34. A machine learning engineer is using the following code block to scale the inference of a single-node model on a Spark DataFrame with one million records:



35. Assuming the default Spark configuration is in place, which of the following is a benefit of using an Iterator?

**A**The data will be limited to a single executor preventing the model from being loaded multiple times

**B**The model will be limited to a single executor preventing the data from being distributed

**CThe model only needs to be loaded once per executor rather than once per batch during the inference process**

**D**The data will be distributed across multiple executors during the inference process

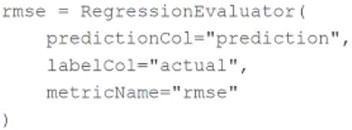
36. A data scientist has developed a linear regression model using Spark ML and computed the predictions in a Spark DataFrame preds\_df with the following schema:

prediction DOUBLE

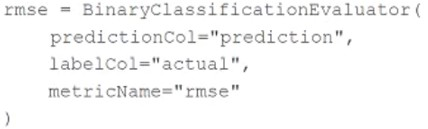
actual DOUBLE

Which of the following code blocks can be used to compute the root mean-squared-error of the model according to the data in preds\_df and assign it to the rmse variable?

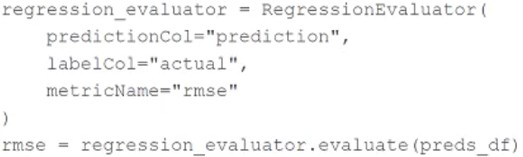
A)



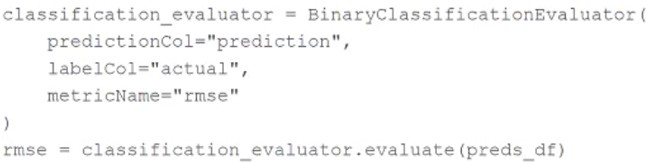
B)



C)



D)



37. A data scientist has written a feature engineering notebook that utilizes the pandas library. As the size of the data processed by the notebook increases, the notebook's runtime is drastically increasing, but it is processing slowly as the size of the data included in the process increases.

Which of the following tools can the data scientist use to spend the least amount of time refactoring their notebook to scale with big data?

**A**PySpark DataFrame API

**Bpandas API on Spark**

**C**Spark SQL

**D**Feature Store