Examen de Certificación en Ciencia de Datos y ML (Versión Alumno V2)

Sin respuestas. Orden aleatorio de preguntas y opciones.

Módulo 1: Fundamentos de Programación en Python (15--20 min)

1) Completa el código para importar la librería NumPy con su alias estándar:

```
import ____ as np
```

A) sklearn

B) numpy

C) pandas

D) np

2) ¿Qué imprime el siguiente código?

```
x = [10, 20, 30] print(x[1])
```

A) 30

B) Error

C) 20

D) 10

3) Completa el código para crear una lista por comprensión con los cuadrados de los números del 1 al 5:

```
squares = [____ for i in range(1,6)]
```

A) i*i

B) i**2

C) pow(i,2)

D) Todas las anteriores son correctas

4) Ordena los pasos para recorrer una lista nums e imprimir sus elementos:

A) 1-2-3

B) 3-1-2

C) 2-1-3

D) 2-3-1

5) ¿Qué librería es más utilizada en Python para manipular datos tabulares (tipo Excel/CSV)?

A) Matplotlib

B) Seaborn

C) Pandas

D) NumPy

6) ¿Cuál de las siguientes opciones es la manera correcta de definir una función en Python?

A) def my_func():

B) func my_func():

C) function my_func():

D) def my_func:

7) En Python, ¿cuál de las siguientes estructuras de datos es mutable?

A) Cadena

B) Tupla

C) Lista

D) Entero

8) ¿Cuál es el error en este código?

```
import numpy as np arr = np.array[1,2,3,4]
```

A) arr no puede contener enteros

B) np.array no existe

C) np debe importarse como nump

D) Falta paréntesis, debería ser np.array([1,2,3,4])

9) ¿Cuál es la forma más eficiente en Python de calcular la suma de una lista x?

A) np.suma(x)

B) suma(x)

C) x.sum()

D) sum(x)

10) Detecta el error en este código:

```
for i in range(5) print(i)
```

A) i no está definido previamente

B) print no acepta enteros

C) range no existe

D) Falta el : al final del for

Módulo 2: Obtención y Preparación de Datos (15--20 min)

11) Completa el código para leer un archivo CSV en un DataFrame llamado df:

```
import pandas as pd df = pd.____("archivo.csv")
```

A) read_csv

B) open_csv

C) csv_reader

D) load_csv

12) ¿Qué método de Pandas devuelve las primeras 5 filas de un DataFrame?

A) df.head()

B) df.begin()

C) df.top()

D) df.first()

13) Completa el código para rellenar valores faltantes con la media en la columna "edad":

```
df["edad"].fillna(____, inplace=True)
```

A) np.mean()

B) df.mean()

C) mean(df)

D) df["edad"].mean()

14) Ordena los pasos correctos para aplicar OneHotEncoder de scikit-learn:

A) 2-1-3

B) 2-3-1

C) 3-2-1

D) 1-2-3

15) ¿Qué opción es correcta para crear variables dummy de la columna "sexo" en df con Pandas?

A) df.to_dummy("sexo")

B) pd.get_dummies(df, columns=["sexo"])

C) pd.dummies(df["sexo"])

D) pd.get_dummies(df, "sexo")

16) Completa el código para aplicar StandardScaler:

```
from sklearn.preprocessing import ____ scaler = StandardScaler() X_scaled =
scaler.fit_transform(X)
```

A) StandardScaler

B) preprocessing

C) pandas

D) sklearn

17) Detecta el error en este código:

```
from sklearn.preprocessing import MinMaxScaler scaler = MinMaxScaler X_scaled = scaler.fit_transform(X)
```

A) MinMaxScaler no existe en sklearn

B) Falta paréntesis: MinMaxScaler()

C) X debe ser un array NumPy

D) fit_transform no acepta X

18) ¿Qué sucede si no escalamos los datos antes de usar KNN?

A) Nada, siempre funciona igual

B) Los datos se vuelven categóricos

C) Una variable con valores grandes dominará la distancia

D) KNN dejará de entrenar

19) ¿Cuál de las siguientes líneas elimina valores nulos en el DataFrame df?

A) df.clean()

B) df.dropna()

C) df.remove_nulls()

D) df.delete_na()

20) Completa el código para normalizar datos en el rango [0,1] usando MinMaxScaler:

```
scaler = MinMaxScaler(feature_range=(0, ____))
```

A) None

B) 100

C) 10

D) 1

Módulo 3: Análisis Exploratorio de Datos (EDA) (15--20 min)

21) Completa el código para graficar un histograma de la columna "edad" con Pandas:

```
df["edad"].____()
```

A) hist()

B) plot_hist()

C) bar()

D) plot.hist()

22) ¿Qué gráfico se utiliza para analizar la correlación entre variables numéricas?

A) Boxplot

B) Heatmap

C) Barplot

D) Pie chart

23) Completa el código para graficar un diagrama de dispersión entre "edad" y "ingresos" con Pandas:

```
df.plot.scatter(x="edad", y="____")
```

A) ingreso

B) salary

C) age

D) ingresos

24) Ordena los pasos para graficar un heatmap de correlaciones con Seaborn:

A) 2-1-3-4

B) 3-2-1-4

C) 2-3-1-4

D) 1-2-3-4

25) ¿Qué función muestra la información de columnas, tipos de datos y valores nulos en un DataFrame?

A) df.describe()

B) df.info()

C) df.structure()

D) df.types()

26) ¿Cuál es el objetivo principal del EDA (Exploratory Data Analysis)?

A) Detectar patrones, relaciones y outliers en los datos

B) Implementar modelos supervisados

C) Normalizar siempre los datos

D) Predecir con alta precisión

27) ¿Cuál es el gráfico más adecuado para ver la distribución de una variable numérica?

A) Gráfico de torta

B) Heatmap

C) Histograma

D) Gráfico de barras

28) Completa el código para obtener estadísticas descriptivas de un DataFrame df:

```
df.____()
```

A) stats

B) info

C) describe

D) summary

29) ¿Qué indica un boxplot con bigotes muy largos y puntos alejados?

A) Datos normalizados

B) Posible presencia de outliers

C) Correlación lineal alta

D) Baja varianza

30) Detecta el error en este código de Matplotlib:

```
import matplotlib.pyplot as plt plt.plot.hist(df["edad"]) plt.show()
```

A) plt.show() no funciona con histogramas

B) plt.plot.hist no existe

C) df no puede graficarse en Matplotlib

D) Falta importar seaborn

Módulo 4: Inferencia Estadística (15--20 min)

31) ¿Qué significa un error tipo I en pruebas de hipótesis?

A) Error en el cálculo del intervalo de confianza

B) Rechazar H0 cuando es verdadera

C) Error en la media muestral

D) Aceptar H0 cuando es falsa

32) Completa el código para obtener la varianza de una columna "ingresos" en Pandas:

```
var = df["ingresos"].____()
```

A) desviacion()

B) variance()

C) var()

D) std()

33) ¿Qué distribución se usa en muestras grandes con varianza conocida?

A) Normal (Z)

B) Exponencial

C) Binomial

D) t de Student

34) Ordena los pasos correctos para calcular una correlación entre dos columnas con Pandas:

A) 2-1-3

B) 1-3-2

C) 1-2-3

D) 3-1-2

35) En una prueba de hipótesis, el p-value indica:

A) Evidencia en contra de la hipótesis nula

B) La probabilidad de que H0 sea verdadera

C) La desviación estándar

D) La media poblacional

36) La Ley de los Grandes Números establece que:

A) La varianza disminuye a medida que aumenta la muestra

B) La media muestral tiende a la media poblacional conforme aumenta el tamaño de muestra

C) Los datos siempre convergen a distribución normal

D) El error estándar aumenta con más datos

37) Completa el código para calcular una prueba t de Student en Python con SciPy:

```
from scipy import stats t_stat, p_value = stats.ttest_ind(grupo1, ____)
```

A) grupo2

B) df

C) p_value

D) group2.mean()

38) Detecta el error en este código de NumPy:

```
import numpy as np datos = [2, 4, 6, 8] desv = np.std{datos}
```

A) datos debe ser un DataFrame

B) Se usan \{ \ en lugar de ( )

C) np.std no existe

D) np no puede calcular desviaciones

39) ¿Qué mide un intervalo de confianza del 95\%?

A) Que la media siempre está dentro del intervalo

B) Que el 95\% de los intervalos calculados contendrán el verdadero parámetro poblacional

C) Que los datos siguen distribución normal

D) Que el 95\% de los datos caen en ese rango

40) Completa el código para calcular la media de una columna "edad" en Pandas:

```
media = df["edad"].____()
```

A) mean()

B) promedio()

C) media()

D) average()

Módulo 5: Aprendizaje Supervisado (15--20 min)

41) ¿Qué métrica es más adecuada para evaluar un modelo de clasificación binaria?

A) RMSE

B) Accuracy

C) SSE

D) R^2

42) ¿Qué función activa usa la regresión logística?

A) Tangente hiperbólica

B) Softmax

C) Sigmoide

D) ReLU

43) ¿Qué medida de impureza se utiliza en árboles de decisión?

A) R^2

B) Accuracy

C) Índice Gini

D) Varianza explicada

44) ¿Qué tipo de variable predice la regresión lineal?

A) Binaria

B) Categórica

C) Continua

D) Texto

45) Detecta el error en este código de KNN:

```
from sklearn.neighbors import KNeighborsClassifier knn = KNeighborsClassifier(k=5)
knn.fit(X_train, y_train)
```

A) fit no entrena en KNN

B) Falta neighbors=5, no k=5

C) KNeighborsClassifier no existe

D) y_train no puede usarse con KNN

46) Completa el código para entrenar un modelo de regresión logística en scikit-learn:

```
from sklearn.linear_model import LogisticRegression model = LogisticRegression()
model.____(X_train, y_train)
```

A) transform

B) fit

C) predict

D) train

47) Completa el código para importar un Random Forest de clasificación:

```
from sklearn.ensemble import ____
```

A) DecisionForest

B) RandomForestClassifier

C) RFClassifier

D) RandomForest

48) Completa el código para importar la clase LinearRegression desde scikit-learn:

```
from sklearn.linear_model import ____
```

A) LinearRegression

B) LinearRegressor

C) LinearModel

D) RegressionLinear

49) ¿Qué hace un clasificador SVM?

A) Encuentra un hiperplano que maximiza el margen entre clases

B) Calcula centroides para agrupar datos

C) Calcula probabilidades de manera bayesiana

D) Usa reglas if-else para separar datos

50) Completa el código para importar el algoritmo KNN en scikit-learn:

```
from sklearn.neighbors import ____
```

A) NearestNeighbors

B) KNeighborsClassifier

C) KNN

D) KNNClassifier

Módulo 6: Aprendizaje No Supervisado (15--20 min)

51) ¿Cuál es el objetivo principal del algoritmo K-Means?

A) Maximizar la varianza entre variables

B) Clasificar datos supervisados

C) Reducir dimensionalidad

D) Minimizar la distancia entre puntos y sus centroides

52) Completa el código para aplicar PCA a 2 componentes principales:

```
pca = PCA(n_components=2) X_pca = pca.____(X)
```

A) predict

B) transform

C) fit

D) fit_transform

53) Completa el código para importar PCA desde scikit-learn:

```
from sklearn.decomposition import ____
```

A) PCAModel

B) Decomposition

C) PrincipalComponentAnalysis

D) PCA

54) ¿Qué ventaja tiene t-SNE frente a PCA?

A) Captura relaciones no lineales en los datos

B) Solo funciona con 3 dimensiones

C) Es más rápido

D) No requiere parámetros

55) ¿Qué métrica interna se utiliza para evaluar la calidad de un clustering?

A) Inercia (SSE)

B) RMSE

C) Accuracy

D) R^2

56) Completa el código para entrenar un modelo K-Means con 3 clusters:

```
kmeans = KMeans(n_clusters=3) kmeans.____(X)
```

A) fit

B) transform

C) train

D) predict

57) Ordena los pasos correctos para aplicar t-SNE en scikit-learn:

A) 1-3-2

B) 3-2-1

C) 2-1-3

D) 1-2-3

58) En K-Means, ¿qué parámetro define el número de clusters?

A) k_neighbors

B) max_iter

C) n_clusters

D) random_state

59) Completa el código para importar KMeans desde scikit-learn:

```
from sklearn.cluster import ____
```

A) KMeansClustering

B) ClusterK

C) KNN

D) KMeans

60) ¿Qué hace PCA?

A) Calcula distancias entre observaciones

B) Reduce la dimensionalidad explicando la mayor varianza posible

C) Clasifica datos en clusters

D) Normaliza automáticamente los datos

Módulo 7: Ensemble, Bagging y Boosting (15--20 min)

61) Detecta el error en este código de Random Forest:

```
from sklearn.ensemble import RandomForestClassifier model =
RandomForestClassifier(n_estimators=100) model.train(X_train, y_train)
```

A) X_train no puede usarse en Random Forest

B) n_estimators no es un hiperparámetro válido

C) RandomForestClassifier no existe

D) El método correcto es .fit() y no .train()

62) Completa el código para importar un GradientBoostingClassifier:

```
from sklearn.ensemble import ____
```

A) GradientBoosting

B) BoostingClassifier

C) GBMClassifier

D) GradientBoostingClassifier

63) En boosting, ¿cómo se entrenan los modelos?

A) Todos en el mismo conjunto y se promedian

B) Únicamente con variables categóricas

C) Todos en paralelo en subconjuntos de datos

D) Secuencialmente, cada modelo corrige los errores del anterior

64) ¿Qué significa el método ensemble en aprendizaje automático?

A) Reducir la dimensionalidad de los datos

B) Usar únicamente modelos de regresión

C) Combinar varios modelos para mejorar el desempeño

D) Usar un único modelo optimizado

65) ¿Cuál de las siguientes es una desventaja de los métodos de boosting como XGBoost?

A) Tienden a sobreajustar si no se regulan adecuadamente

B) Siempre generan peor desempeño que un solo árbol

C) No pueden usarse en clasificación

D) No funcionan con datos numéricos

66) ¿Qué ventaja tiene un Random Forest frente a un árbol de decisión único?

A) No necesita datos de entrenamiento

B) Siempre entrena más rápido

C) Reduce el overfitting al promediar múltiples árboles

D) No requiere hiperparámetros

67) Completa el código para crear un modelo AdaBoost en scikit-learn:

```
from sklearn.ensemble import AdaBoostClassifier model =
AdaBoostClassifier(n_estimators=50, ____=1.0)
```

A) learning_rate

B) alpha

C) beta

D) gamma

68) El concepto de bagging consiste en:

A) Entrenar múltiples modelos en subconjuntos de datos con reemplazo y promediar

B) Usar pesos para dar más importancia a errores

C) Usar modelos no supervisados para clasificación

D) Entrenar un solo modelo en todos los datos

69) Completa el código para crear un Random Forest con 200 árboles:

```
forest = RandomForestClassifier(n_estimators=____)
```

A) 200

B) 2

C) 2000

D) 20

70) Completa el código para importar un RandomForestClassifier:

```
from sklearn.ensemble import ____
```

A) ForestClassifier

B) BaggingClassifier

C) RandomForest

D) RandomForestClassifier

Módulo 8: Métricas de Desempeño (Regresión y Clasificación) (15--20 min)

71) Completa el código para calcular la precisión (precision):

```
from sklearn.metrics import precision_score prec = precision_score(y_true, y_pred, average="____")
```

A) weighted

B) multi

C) accuracy

D) binary

72) ¿Qué significa un valor de R^2 cercano a 1?

A) Los datos no están correlacionados

B) El modelo tiene un mal ajuste

C) Los errores son muy grandes

D) El modelo explica casi toda la variabilidad de los datos

73) ¿Qué representa la sensibilidad (recall) en un modelo de clasificación?

A) La proporción de verdaderos positivos sobre los positivos predichos

B) La proporción de verdaderos negativos sobre los negativos predichos

C) La proporción de verdaderos positivos sobre los positivos reales

D) La proporción de verdaderos negativos sobre todos los negativos

74) ¿Qué métrica de regresión calcula el Error Absoluto Medio?

A) MAE

B) R^2

C) MSE

D) RMSE

75) ¿Qué diferencia existe entre el MSE y el RMSE?

A) El MSE solo se usa en clasificación

B) El RMSE no depende del MSE

C) El RMSE es la raíz cuadrada del MSE

D) El MSE es la raíz cuadrada del RMSE

76) Completa el código para importar la matriz de confusión:

```
from sklearn.metrics import ____
```

A) confusion

B) accuracy_score

C) confusion_matrix

D) classification_report

77) ¿Qué representa el AUC (Área Bajo la Curva ROC)?

A) El número de clusters encontrados

B) La capacidad del modelo para distinguir entre clases

C) El área entre 0 y la curva de precisión

D) La varianza explicada

78) Completa el código para calcular el R^2 en scikit-learn:

```
from sklearn.metrics import r2_score r2 = r2_score(y_true, ____)
```

A) y_pred

B) X_train

C) X_test

D) y_true

79) ¿Qué curva se utiliza para evaluar el desempeño de un clasificador binario en distintos umbrales?

A) Gráfico de dispersión

B) Curva ROC

C) Histograma

D) Boxplot

80) Completa el código para calcular el MSE en scikit-learn:

```
from sklearn.metrics import mean_squared_error mse = mean_squared_error(y_true, ____)
```

A) y_pred

B) y_train

C) X_train

D) X_test

Módulo 9: Casos Prácticos de Código (15--20 min)

81) Ordena los pasos para aplicar un pipeline con escalado y regresión logística:

A) 1-2-4-3

B) 1-2-3-4

C) 1-3-2-4

D) 2-1-3-4

82) Completa el código para obtener un reporte de clasificación en sklearn:

```
from sklearn.metrics import classification_report print(classification_report(y_true, ____))
```

A) y_true

B) y_pred

C) X_train

D) X_test

83) ¿Qué librería de Python se recomienda usar para visualizaciones avanzadas de datos junto con Matplotlib?

A) scikit-learn

B) Pandas

C) NumPy

D) Seaborn

84) Detecta el error en este código de KNN:

```
knn = KNeighborsClassifier(n_neighbors=5) knn.fit(X_train)
```

A) X_train debe estar normalizado

B) Falta el argumento y_train en .fit()

C) n_neighbors=5 no es válido

D) KNeighborsClassifier no existe

85) Completa el código para dividir un dataset en entrenamiento y prueba:

```
from sklearn.model_selection import train_test_split X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, ____=42)
```

A) random

B) seed

C) random_state

D) shuffle

86) Detecta el error en este código de Random Forest:

```
forest = RandomForestClassifier() forest.fit(X_train, y_train) y_pred = forest.predict(X_test, y_test)
```

A) RandomForestClassifier no existe

B) fit no funciona en Random Forest

C) predict no recibe y_test como argumento

D) X_test debe estar escalado

87) Ordena los pasos para entrenar y evaluar un modelo en scikit-learn:

A) 1-2-4-3-5

B) 1-2-3-4-5

C) 2-1-3-4-5

D) 1-3-2-4-5

88) Detecta el error en este código para entrenar un modelo de regresión lineal:

```
from sklearn.linear_model import LinearRegression model = LinearRegression
model.fit(X_train, y_train)
```

A) Falta instanciar: LinearRegression()

B) X_train debe ser un DataFrame

C) fit no funciona en regresión

D) LinearRegression no existe

89) ¿Cuál es la mejor práctica antes de aplicar PCA?

A) Entrenar un árbol de decisión

B) Aumentar el número de dimensiones

C) Convertir todas las variables en categóricas

D) Normalizar los datos con StandardScaler

90) ¿Qué error hay en este código de PCA?

```
from sklearn.decomposition import PCA pca = PCA(n_components=2) X_pca = pca.fit(X_train,
y_train)
```

A) PCA no existe en sklearn

B) PCA no acepta y_train en .fit()

C) n_components=2 no es válido

D) X_train debe estar escalado con RobustScaler

Módulo 10: Examen Integrador (15--20 min)

91) Completa el código para calcular la matriz de confusión:

```
from sklearn.metrics import confusion_matrix cm = confusion_matrix(y_true, ____)
```

A) y_test

B) y_pred

C) X_train

D) X_test

92) Ordena los pasos correctos para aplicar PCA seguido de K-Means:

A) 1-3-2-4

B) 1-2-3-4

C) 4-1-2-3

D) 2-1-3-4

93) ¿Qué método ensemble entrena modelos en paralelo para luego promediar resultados?

A) PCA

B) Boosting

C) Stacking

D) Bagging

94) Completa el código para inicializar un KMeans con 4 clusters y semilla fija:

```
kmeans = KMeans(n_clusters=4, ____=42)
```

A) seed

B) init

C) random_state

D) shuffle

95) Completa el código para entrenar un modelo de SVM lineal:

```
from sklearn.svm import SVC svm = SVC(kernel="____") svm.fit(X_train, y_train)
```

A) poly

B) sigmoid

C) rbf

D) linear

96) ¿Qué técnica de reducción de dimensionalidad es no lineal?

A) Bagging

B) Normalización Min-Max

C) PCA

D) t-SNE

97) Detecta el error en este código de AdaBoost:

```
from sklearn.ensemble import AdaBoostClassifier model =
AdaBoostClassifier(n_estimators=50) model.predict(X_train, y_train)
```

A) predict no recibe y_train como argumento

B) n_estimators no es válido

C) fit debe llamarse antes de predict

D) AdaBoostClassifier no existe

98) ¿Qué algoritmo es más adecuado para predecir el precio de una casa?

A) t-SNE

B) K-Means

C) PCA

D) Regresión lineal

99) ¿Qué ventaja tiene Gradient Boosting frente a Random Forest?

A) Se entrena secuencialmente corrigiendo errores previos

B) Siempre entrena más rápido

C) No necesita parámetros de ajuste

D) Siempre da menor error que cualquier otro modelo

100) ¿Qué métrica es más adecuada para un problema de clasificación desbalanceada?

A) RMSE

B) Accuracy

C) R^2

D) Recall