

Examen de Certificación en Ciencia de Datos y ML (Versión Alumno V3)

Formato: opción múltiple (teoría + código). Secciones de 15–20 min.

Instrucciones: No se muestran respuestas al final. Marque una sola opción por pregunta.

Módulo 9: Casos Prácticos

1. Detecta el error en este código de PCA:

```
pca = PCA(n_components=2)
X_pca = pca.fit(X,y)
```

- A) fit no recibe y
- B) Debe ser $\text{fit}_{transformn_componentsinvlido}$
- C) X debe ser array

2. Completa el código para KMeans:

```
from sklearn.cluster import KMeans
model = KMeans(n_clusters=3)
model.____(X)
```

- A) fit
- B) predict
- C) transform
- D) score

3. Detecta el error en este código de regresión lineal:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression
model.fit(X,y)
```

- A) Faltan paréntesis en LinearRegression()
- B) fit no existe
- C) X debe ser DataFrame
- D) y debe ser categórica

4. Completa el código para calcular accuracy:

```
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, ____))
```

- A) $y_{predy_{train}}$
- B) $X_{testy_{true}}$

5) ¿Qué error hay en este código de RandomForest?

```
forest = RandomForestClassifier()  
forest.train(X_train,y_train)
```

- A) El método correcto es fit()
- B) train no existe
- C) *no es obligatorio que X_train debe escalarse*

6) Completa el código para dividir dataset:

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.25,____=42)
```

- A) random
- B) seed
- C) *random_state shuffle*

7) ¿Qué imprime este código?

```
import numpy as np  
arr = np.arange(4).reshape(2,2)  
print(arr.shape)
```

- A) (4,)
- B) (2,2)
- C) (2,4)
- D) Error

8. Completa el código para gráfico de dispersión:

```
plt.scatter(df["x"], df["y"])  
plt.____()
```

- A) show
- B) plot
- C) scatter
- D) display

9. ¿Qué métrica usarías para un modelo de regresión con valores extremos?

- A) MAE
- B) MSE
- C) Accuracy
- D) Recall

10. ¿Qué significa overfitting?

- A) Modelo se ajusta demasiado al entrenamiento
- B) Modelo generaliza bien
- C) Modelo tiene baja varianza
- D) Modelo con datos faltantes

Módulo 10: Examen Integrador

1. ¿Qué técnica usarías para detectar outliers?

- A) Boxplot
- B) Heatmap
- C) Regresión
- D) Bagging

2. Completa el código para SVM:

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

- A) fit
- B) predict
- C) transform
- D) train

3. ¿Qué algoritmo usarías para segmentar clientes sin etiquetas?

- A) KMeans
- B) Regresión lineal
- C) LogisticRegression
- D) SVM

4. ¿Qué métrica usarías para un dataset de clasificación muy desbalanceado?

- A) Recall
- B) Accuracy
- C) RMSE
- D) MSE

5. ¿Qué significa bagging?

- A) Bootstrap aggregating
- B) Batch training
- C) Boost average
- D) Base gradient

6. Completa el código para matriz de confusión:

```
cm = confusion_matrix(y_true,_____)
```

A) $y_{pred}y_{train}$

B) $X_{test}y_{test}$

7) ¿Qué librería se usa principalmente para Deep Learning?

A) tensorflow

B) sklearn

C) pandas

D) matplotlib

8. ¿Qué significa big data?

A) Volumen, variedad, velocidad

B) Modelo supervisado

C) Base de datos relacional

D) Datos limpios

9. ¿Qué técnica reduce dimensionalidad no lineal?

A) t-SNE

B) PCA

C) Regresión

D) Bagging

10. ¿Qué ventaja tiene ensemble?

A) Mejora generalización

B) Siempre peor

C) Elimina outliers

D) Reduce features