EXAMEN FINAL – SAMSUNG INNOVATION CAMP (OPCIONAL)
Curso: Inteligencia Artificial **Instrucciones:** Responde cada pregunta de manera clara y concisa. Justifica tus respuestas y muestra los cálculos cuando sea necesario.
Alumno:

Sección 1: Matemáticas para IA

1. Define el concepto de gradiente y explica su papel en el entrenamiento de redes neuronales.

R= El gradiente es un vector que indica la dirección y la tasa de cambio más notoria o de mayor magnitud de una función en un punto determinado. En el entrenamiento de redes neuronales, se usa porque sirve para ajustar los pesos de la red para minimizar la función de costo y mejorar la precisión del modelo.

2. Menciona 3 algoritmos de optimización que se utilicen en la IA.

- Descenso de gradiente estocástico (SGD)
- Adam
- RMSprop (Root Mean Square Propagation)

3. Calcula la derivada de la función de costo con respecto a w:

$$J(w) = (w^2 - 3w + 4)$$
$$2w - 3$$

Sección 2: Probabilidad y Estadística

4. Explica la diferencia entre probabilidad condicional y probabilidad conjunta.

Probabilidad condicional: La probabilidad de que ocurra un evento A dado que otro evento B ya ha ocurrido.

Probabilidad conjunta: la probabilidad de que ocurran ambos eventos A y B simultáneamente.

5. ¿Qué significa que dos variables sean independientes en términos de probabilidad?

Es cuando el conocimiento de una no afecta la probabilidad de la otra

6. Explica cómo se utiliza la distribución normal en machine learning.

Para modelar datos, detectar valores atípicos y en algoritmos y en normalización de datos.

Sección 3: Limpieza y transformación de datos

A. Manejo de datos nulos

7. Menciona tres técnicas para manejar datos faltantes y explica en qué casos se usaría cada una.
    - Eliminación de datos faltantes: Útil si los valores nulos representan un pequeño porcentaje del conjunto de datos.
    - Imputación con la media o mediana: Aplicable si los datos siguen una distribución normal.
    - Modelos predictivos: Se usa cuando hay suficientes datos para predecir los valores faltantes con algoritmos de machine learning.

8. ¿Por qué es importante verificar que la eliminación de datos nulos no genere sesgo en el modelo?

Porque puede alterar la distribución original de los datos y afectar la capacidad del modelo para generalizar correctamente.

9. Explica qué son los valores atípicos y cómo afectan un modelo de machine learning.

Son datos que difieren significativamente del resto. Pueden afectar el rendimiento de un modelo al distorsionar métricas como la media y la desviación estándar.

B. Normalización y escalado de datos

10. Explica la diferencia entre normalización y estandarización de datos.

La normalización escala los datos entre 0 y 1. Se usa cuando no se conoce la distribución de los datos y la estandarización convierte los datos a una media de 0 y desviación estándar de 1. Se usa cuando los datos siguen una distribución normal.

11. ¿Por qué es importante escalar los datos antes de entrenar un modelo de machine learning?

Para evitar que características con valores grandes dominen el modelo y mejoren la convergencia del algoritmo.

12. Aplica normalización z-score al conjunto de datos:

$$x = [2,5,10,20]$$

```
import numpy as np
import pandas as pd
x = np.array([2, 5, 10, 20])
# Cálculo de la normalización Z-score
mean_x = np.mean(x)
```

```
std_x = np.std(x, ddof=0)
z_score = (x - mean_x) / std_x
df_zscore = pd.DataFrame({"Valores originales": x, "Z-score": z_score})
```

13. A que tipo de variables se le aplica el One-Hot Encoding.

A variables categóricas nominales.

14. Codifica la siguiente columna en formato One-Hot Encoding. (Inserta una tabla si lo consideras necesario)

df[“color”] = [“rojo”, “azul”, “verde”, “azul”, “rojo”]

```
import pandas as pd

df = pd.DataFrame({"color": ["rojo", "azul", "verde", "azul", "rojo"]})

df_encoded = pd.get_dummies(df, columns=["color"])

print(df_encoded)
```

Sección 4: Selección de características

15. ¿Cómo afecta la selección de características al tiempo de entrenamiento de un modelo?

Reduce el tiempo de entrenamiento al disminuir la cantidad de datos procesados.

16. Explica la diferencia entre características relevantes y redundantes.

Las características relevantes contribuyen a la precisión del modelo, mientras que las redundantes no aportan nueva información y pueden eliminarse sin afectar el rendimiento.

17. Explica la diferencia entre selección hacia adelante y selección hacia atrás.

En la de hacia adelante se agregan características una por una hasta encontrar la mejor combinación y en la de hacia atrás se eliminan características de un conjunto completo hasta encontrar la combinación óptima.

18. ¿Cómo se puede usar Random Forest para seleccionar características?

Se asigna una importancia a cada característica basada en su impacto en la precisión del modelo.

19. ¿Cuándo es recomendable usar PCA en lugar de eliminar características?

Cuando hay muchas variables correlacionadas y se quiere reducir la dimensionalidad sin perder información.

20. ¿Qué problemas pueden surgir si eliminamos muchas características en un modelo de machine learning?

Puede perderse información valiosa y reducir la capacidad del modelo para generalizar.

Sección 5: Machine Learning

21. ¿Qué entiendes el aprendizaje supervisado y el no supervisado?

El aprendizaje supervizado es el que se entrena con datos etiquetados, mientras que el no supervisado no usa etiquetas y busca patrones en los datos.

22. Explica la diferencia entre clasificación y regresión.

La clasificación predice categorías y la regresión predice valores continuos.

23. Con que métricas se evalúa un modelo de clasificación.

Precisión, recall, F1-score, matriz de confusión, curva ROC-AUC.

24. Con que métricas se evalúa un modelo de regresión.

RMSE, MAE, R^2

25. Menciona 3 algoritmos de clasificación.

Árboles de decisión, SVM, redes neuronales

26. Explica la diferencia entre clustering y reducción de dimensionalidad.

El clustering agrupa datos en categorías sin etiquetas y la reducción de dimensionalidad, reduce el número de variables manteniendo la información.

27. ¿Cuándo es recomendable usar K-means sobre DBSCAN?

Cuando los grupos son esféricos y de tamaño similar.

28. ¿Qué es la validación cruzada y por qué es importante?

Técnica para evaluar modelos dividiendo los datos en múltiples subconjuntos. Mejora la capacidad de generalización.

29. ¿Cómo interpretas la curva ROC y el AUC?

Muestran la relación entre la tasa de verdaderos positivos y falsos positivos. Un AUC cercano a 1 indica un buen modelo.

Sección 6: Minería de texto

30. ¿Qué es la minería de texto y en qué se diferencia del procesamiento de lenguaje natural (NLP)?

La minería de texto extrae patrones y conocimiento de textos y el NLP procesa el lenguaje natural para comprenderlo y manipularlo.

31. ¿Qué técnicas se utilizan para limpiar y preprocesar texto?

Tokenización, lematización, eliminación de stopwords.

32. Explica el concepto de tokenización.

Divide el texto en palabras o frases

33. Explica el concepto de corpus.

Conjunto de textos usados para entrenar modelos de NLP.

Sección 7: Deep Learning

34. ¿Cuál es la diferencia entre una red neuronal densa y una convolucional?

Que la red neuronal densa está conectada con todas las neuronas de la capa anterior y la siguiente y la red convolucional (CNN) usa filtros para detectar patrones locales en los datos, especialmente útil en imágenes.

35. ¿Qué significa backpropagation y cómo funciona en redes neuronales?

Es un algoritmo que ajusta los pesos de la red neuronal minimizando el error de predicción. Se calcula el gradiente del error y se propaga hacia atrás ajustando los pesos mediante el descenso de gradiente

36. ¿Para que sirve el Learning rate?

Controla la magnitud del ajuste de los pesos en cada iteración del entrenamiento.

37. Menciona 2 capas principales de una CNN.

Capa convolucional, capa de pooling

38. ¿Cómo funcionan las capas convolucionales en CNN?

Aplican filtros a la imagen de entrada para extraer características como bordes, texturas y formas, permitiendo que la red aprenda patrones espaciales.

39. ¿Qué es el sobre-ajuste y sub-ajuste (overfitting y underfitting)?

Overfitting: El modelo aprende demasiado los datos que es difícil reconocer los nuevos.

Underfitting: El modelo no aprende lo suficiente de los datos de entrenamiento y tiene un mal rendimiento.

40. Menciona 2 estrategias para solucionar el sobre-ajuste en redes neuronales.

Dropout, aumentación de datos

41. ¿En qué casos se utilizan las redes neuronales recurrentes (RNN)?

En el procesamiento de texto, audio o series de tiempo.

42. Menciona 2 capas de las RNN.

Capa recurrente y capa densa

43. Menciona tres funciones de activación

ReLU (Rectified Linear Unit), sigmoide, softmax

44. ¿Para qué sirve una GAN?

Se usa en generación de imágenes, mejora de calidad de imágenes y creación de datos sintéticos.

45. Describe como es el entrenamiento de una GAN

El generador produce datos falsos, el discriminador evalúa los datos y proporciona retroalimentación, el generador ajusta sus pesos para engañar mejor al discriminador y se repite hasta que el generador produce datos indistinguibles de los reales.