

Proyecto Final de IA

Se debe desarrollar un proyecto de investigación el cual debe cumplir mínimamente lo siguiente:

1. Descripción Detallada del Dataset

- Origen y Fuente: Especifique la procedencia del conjunto de datos utilizado en el estudio (por ejemplo, Kaggle, UCI Machine Learning Repository, entre otros). Incluir detalles sobre la cantidad de instancias y atributos presentes, así como cualquier sesgo o peculiaridad que pueda influir en los resultados.
- Características del Dataset: Detallar las variables presentes en el dataset, clasificándolas según su tipo (numéricas, categóricas, binarias, etc.). Indicar si el conjunto de datos presenta alguna característica adicional que requiera atención particular (como desbalanceo de clases, datos faltantes, etc.).
- Propósito del Dataset: Explicar el objetivo del dataset, señalando si está orientado a problemas de clasificación, regresión, agrupamiento, etc. Es importante que el propósito esté alineado con el objetivo de la investigación.

2. Objetivo General de la Investigación

- Planteamiento del Objetivo: Definir de manera precisa el objetivo que se busca alcanzar a partir del análisis del conjunto de datos. Por ejemplo: "Desarrollar un modelo predictivo para clasificar a los individuos según características socioeconómicas", "Determinar el factor principal que predice el rendimiento de los estudiantes", entre otros.
- Justificación del Estudio: Explicar de manera detallada por qué es relevante abordar este problema con el conjunto de datos seleccionado. Discutir cómo los resultados obtenidos pueden influir en la toma de decisiones en el ámbito académico, empresarial o social.

3. Preprocesamiento de los Datos

- Identificación y Tratamiento de Valores Faltantes: Describir cómo se gestionan los valores faltantes, especificando si se emplea imputación, eliminación de registros o cualquier otro enfoque.
- Normalización y/o Estandarización: Indicar si se realizó alguna normalización o estandarización de las variables, especialmente si estas tienen escalas muy dispares, con el objetivo de mejorar la convergencia de los modelos.
- Transformación de Variables: Explicar las transformaciones realizadas a las variables, como la conversión de variables categóricas a variables

dummy, la creación de nuevas características o la aplicación de técnicas de reducción de dimensionalidad.

- Balanceo de Datos: Si el conjunto de datos presenta clases desbalanceadas, describir las técnicas utilizadas para equilibrar los datos, tales como submuestreo, sobremuestreo o la técnica SMOTE.

4. Selección del Clasificador

- Elección del Clasificador: Definir qué clasificador o modelo de aprendizaje automático se empleará (por ejemplo, árboles de decisión, máquinas de soporte vectorial (SVM), k-NN, regresión logística, etc.), justificando la elección según las características del conjunto de datos.
- Justificación Académica: La elección del clasificador debe basarse en principios teóricos sólidos, los cuales serán sustentados con fuentes académicas (por ejemplo, artículos con DOI o libros con ISBN). Este análisis puede ocupar un máximo de dos páginas, explicando por qué el clasificador seleccionado es adecuado para el tipo de problema y las características del dataset.

5. Primera Ejecución del Modelo

- Confiabilidad del Modelo: Se debe incluir una evaluación inicial del modelo utilizando métricas estándar como la precisión (accuracy), recall, F1-score, entre otras, para determinar la efectividad del modelo en la tarea de clasificación.
- Matriz de Confusión: Incluir la matriz de confusión para obtener una visión clara de los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, lo cual es fundamental para evaluar el rendimiento del modelo de manera más precisa.

6. Evaluación a través de Divisiones de Datos

- Proporción de Datos para Entrenamiento y Prueba: Para la primera ejecución, utilizar una división de 80% para entrenamiento y 20% para prueba, lo cual es adecuado para validar el rendimiento del modelo en un entorno académico.
- Segunda Ejecución para Investigación: Para una validación más robusta, utilizar una división 50%/50% entre entrenamiento y prueba en la segunda ejecución, lo que permite un análisis equilibrado de los datos.
- Evaluación con Múltiples Asignaciones: Realizar al menos 100 asignaciones aleatorias de los datos y calcular la mediana de la confiabilidad obtenida, con el fin de evaluar la estabilidad del modelo bajo diferentes condiciones.

7. Aplicación de Análisis de Componentes Principales (PCA)

- Fundamentos de PCA: Explicar el concepto de Análisis de Componentes Principales (PCA), detallando cómo se lleva a cabo mediante álgebra lineal (por ejemplo, descomposición en valores singulares, eigenvectores y eigenvalores). El propósito de PCA es reducir la dimensionalidad del dataset mientras se conserva la mayor parte de la varianza.
- Optimización de Componentes: Realizar experimentos con diferentes números de componentes principales (por ejemplo, 12, 10, 9, 5, 3) y evaluar el impacto en el rendimiento del modelo. Incluir al menos cinco ejecuciones para comparar los resultados obtenidos.

8. Aprendizaje No Supervisado

- Análisis No Supervisado: Ejecutar un análisis no supervisado, utilizando técnicas como el agrupamiento (por ejemplo, K-means, DBSCAN) sin tener en cuenta la variable objetivo ("y"). Este tipo de análisis permite identificar patrones inherentes en los datos sin requerir etiquetas previas.
- Análisis y Resultados: Evaluar los clusters obtenidos y discutir su utilidad para explorar la estructura subyacente de los datos.

9. Redacción del Artículo de Investigación

- Documento Académico: Elaborar un artículo académico de al menos cuatro páginas en el que se detalle el proceso llevado a cabo, los resultados obtenidos y su interpretación. El artículo debe abordar tanto la metodología como las conclusiones del análisis realizado.
- Código en Repositorios Abiertos: Subir el código desarrollado a plataformas como GitHub, Kaggle o CodeLab para garantizar la transparencia y la reproducibilidad de los resultados.

10. Resolución del Problema de las n-caballos

- Enfoque Combinatorio: Resolver el problema de las N-caballos utilizando métodos de análisis combinatorio, analizando las distintas formas de posicionar los caballos en un tablero de ajedrez de tamaño $n \times n$ sin que se ataquen entre sí.
- Simulado Recocido: Implementar el algoritmo de Simulated Annealing para encontrar una solución aproximada al problema de las N-Reinas. Explicar cómo este enfoque busca optimizar la colocación de las reinas mediante la simulación de un proceso físico de enfriamiento.

Formato de entrega

Se debe entregar todo en un solo archivo comprimido

- Documento PDF: informe académico estructurado con todos los apartados del proyecto.
- Notebook (.ipynb o .py): código ejecutable con comentarios y resultados.

- Apoyo de artículos científicos: referencias con DOI o ISBN que sustenten teoría, modelos y metodología.

Evaluación (30% del 100%)

| Nº | Criterio | Subcriterios principales | Puntaje Máximo |
|----|---|--|----------------|
| 1 | Descripción Detallada del Dataset | - Origen y fuente correctamente citados. - Variables descritas y clasificadas. - Identificación de sesgos, desbalance o valores faltantes. | 3 pts |
| 2 | Objetivo General de la Investigación | - Objetivo claro, medible y coherente con el dataset. - Justificación sólida y contextualizada. | 3 pts |
| 3 | Preprocesamiento de los Datos | - Manejo de valores faltantes. - Normalización/estandarización. - Transformaciones adecuadas (dummies, reducción, etc.). - Balanceo de clases si corresponde. | 4 pts |
| 4 | Selección del Clasificador | - Modelo elegido según el tipo de problema. - Sustento teórico con artículos o libros académicos (DOI/ISBN). | 3 pts |
| 5 | Primera Ejecución del Modelo | - Uso de métricas estándar (accuracy, recall, F1-score). - Matriz de confusión interpretada correctamente. | 2 pts |
| 6 | Evaluación de Divisiones de Datos | - Uso correcto de divisiones 80/20 y 50/50. - Evaluación de estabilidad (≥ 100 asignaciones). | 3 pts |
| 7 | Aplicación de PCA | - Fundamento teórico y aplicación matemática. - Pruebas con componentes distintos. - Análisis comparativo. | 3 pts |
| 8 | Aprendizaje No Supervisado | - Implementación (K-means, DBSCAN, etc.). - Análisis e interpretación de los clusters. | 2 pts |
| 9 | Redacción del Artículo de Investigación | - Documento ≥ 4 páginas. - Presenta resultados, discusión y conclusiones. - Código en repositorio abierto (GitHub/Kaggle). | 4 pts |
| 10 | Resolución del Problema de las N-Caballos / N-Reinas | - Correcta aplicación del método combinatorio. - Implementación funcional de Simulated Annealing. | 3 pts |
| 11 | Apoyo con Artículos Científicos | - Uso de fuentes académicas relevantes (con DOI o ISBN). - Aplicación de ideas o métodos tomados de esos estudios. - Citación correcta (APA/IEEE). | 4 pts |

Disertación Oral (70% del 100%)

| Nº | Criterio | Descripción de Evaluación | Pts Máx |
|----|--|--|---------|
| 1 | Claridad y dominio conceptual | Explica con precisión los fundamentos teóricos del dataset, PCA, clasificadores y métodos utilizados. | 4 pts |
| 2 | Presentación estructurada | Orden lógico en la exposición: introducción, metodología, resultados y conclusiones bien enlazadas. | 3 pts |
| 3 | Argumentación y justificación académica | Justifica correctamente decisiones metodológicas (clasificador, métricas, división de datos, PCA, etc.) usando respaldo científico (DOI/ISBN). | 3 pts |
| 4 | Interpretación de resultados | Analiza críticamente los resultados obtenidos (métricas, matriz de confusión, clustering, etc.) y su impacto. | 3 pts |
| 5 | Capacidad de respuesta | Responde preguntas con seguridad, precisión y relación con la teoría y práctica. | 3 pts |
| 6 | Expresión y lenguaje técnico | Usa vocabulario adecuado, lenguaje académico y recursos visuales (diapositivas o notebook). | 2 pts |
| 7 | Innovación y profundidad analítica | Propone ideas originales, mejoras o reflexiones más allá del contenido base. | 3 pts |