Proyectos Finales del Curso de Ciencia de Datos

Lineamientos generales

Cada estudiante desarrollará un proyecto final de análisis aplicado, integrando los conceptos y herramientas vistos a lo largo del curso. El objetivo es que los estudiantes:

- Apliquen el ciclo completo de un proyecto de Ciencia de Datos: carga, limpieza, análisis, modelado y comunicación.
- Expliquen claramente las **técnicas empleadas**, por qué fueron elegidas y qué papel cumplen en el análisis.
- Argumenten de manera crítica las **conclusiones obtenidas**, sustentándolas en evidencia cuantitativa y **inferencia estadística**.

Entregables

- Un **GitHub** bien documentado, que incluya código y explicación de cada paso.
- Un **informe corto** (máx. 10 páginas) con:
  a) Introducción y objetivo del análisis.
  b) Descripción del dataset y limpieza aplicada.
  c) Técnicas analíticas o modelos empleados, con **justificación estadística**.
  d) Visualizaciones y resultados clave, incluyendo **métricas de incertidumbre**.
  e) Conclusiones y discusión crítica, con **limitaciones y supuestos**.

Evaluación sugerida (100 puntos)

- Claridad del objetivo y justificación del enfoque: 20 pts
- Calidad del análisis y limpieza de datos: 25 pts
- Adecuación y **validación estadística** de las técnicas: 25 pts
- Claridad de visualizaciones y comunicación de resultados: 20 pts
- Conclusiones argumentadas y reproducibilidad: 10 pts

1. Proyecto 1: Desigualdad educativa y económica en México

Objetivo

Explorar y analizar la relación entre educación, ingreso y condiciones sociales a nivel municipal o estatal. El estudiante deberá realizar un análisis exploratorio (EDA), emplear

técnicas de visualización y construir una narrativa apoyada en evidencia.

Tareas sugeridas

1. Cargar y combinar datasets socioeconómicos (ver compendio abajo).
2. Limpiar y normalizar variables numéricas y categóricas.
3. Analizar correlaciones, generar mapas de calor y visualizaciones.
4. Si se desea, ajustar un modelo predictivo sencillo (p.ej. regresión lineal sobre ingreso medio).
5. Redactar conclusiones claras justificando qué técnicas se usaron y cómo contribuyeron al análisis.

Compendio de fuentes de datos reales (descargables)

- **INEGI – Indicadores socioeconómicos municipales:** `https://www.inegi.org.mx/app/areasgeograficas/`
- **CONEVAL – Pobreza y carencias sociales:** `https://www.coneval.org.mx/Medicion/Paginas/PobrezaInicio.aspx`
- **SEP – Estadística educativa:** `https://www.planeacion.sep.gob.mx/principalescifras/`
- **ENIGH – Encuesta de Ingresos y Gastos de los Hogares:** `https://www.inegi.org.mx/programas/enigh/`
- **Datos abiertos de México (portal general):** `https://datos.gob.mx/`

Sugerencia

Los alumnos pueden, por ejemplo, analizar si los municipios con mayor ingreso promedio presentan menor rezago educativo o mejor acceso a internet. Se evaluará especialmente la **capacidad para justificar cada técnica usada y conectar los resultados con la pregunta original**.

2. Proyecto 2: Clasificación de riesgo cardiovascular

Objetivo

Aplicar modelos de aprendizaje supervisado para predecir si un paciente presenta riesgo de enfermedad cardíaca a partir de variables biomédicas.

Dataset sugerido

- **Heart Disease UCI Dataset:** `https://archive.ics.uci.edu/ml/datasets/heart+disease`

Énfasis

El estudiante debe explicar por qué eligió determinado modelo y cómo interpreta sus métricas. No se evaluará solo la exactitud, sino la **claridad en la justificación metodológica**.

3. Proyecto 3: Segmentación de países según indicadores de desarrollo

Objetivo

Aplicar técnicas de reducción de dimensión (PCA o NMF) y de agrupamiento (k-means, jerárquico o espectral) para identificar grupos de países con características similares.

Bases de datos recomendadas (descargables)

- **World Bank – World Development Indicators:** `https://data.worldbank.org/`
- **UNESCO Institute for Statistics:** `http://uis.unesco.org/`
- **OECD Data Portal:** `https://data.oecd.org/`
- **Our World in Data:** `https://ourworldindata.org/`

Sugerencia

Alumnos pueden explorar, por ejemplo, si los países se agrupan por nivel de ingreso, esperanza de vida o acceso a internet. Se espera una descripción razonada del proceso: por qué se redujeron las dimensiones, cómo se seleccionó el número de clusters, y qué interpretación se da a cada grupo resultante.

4. Proyecto 4: Procesamiento de imágenes y representación de características

Objetivo

Explorar el uso de técnicas de reducción de dimensión y factorización matricial (como PCA o NMF) aplicadas a imágenes, con el propósito de extraer patrones visuales, comparar representaciones y clasificar imágenes de manera no supervisada o semi-supervisada.

Descripción

Este proyecto busca que el estudiante trabaje con imágenes reales, aplicando transformaciones, reducciones de dimensión y visualizaciones de componentes latentes. El objetivo no es construir una red neuronal profunda, sino comprender cómo la información visual puede representarse mediante métodos estadísticos y lineales.

Tareas sugeridas

1. Cargar un conjunto de imágenes (ver opciones abajo) y convertirlas a matrices numéricas (escala de grises o RGB).
2. Aplicar técnicas de reducción de dimensión (PCA, NMF, o t-SNE) para visualizar el espacio latente de las imágenes.
3. En caso de NMF, visualizar los **componentes base** para interpretar qué estructuras visuales representan (bordes, texturas, rasgos comunes).
4. Realizar una clasificación o agrupamiento simple de imágenes.
5. Discutir los resultados e interpretar visualmente los patrones descubiertos.

Datasets sugeridos

- **MNIST Digits:** conjunto clásico de dígitos escritos a mano (`https://www.openml.org/d/554`).
- **Fashion-MNIST:** imágenes de prendas de vestir (camisas, zapatos, bolsos, etc.) `https://github.com/zalandoresearch/fashion-mnist`
- **Olivetti Faces Dataset (scikit-learn):** rostros humanos en blanco y negro, útil para PCA o NMF. `https://scikit-learn.org/stable/datasets/real_world.html#olivetti-faces-dataset`
- **CIFAR-10:** objetos comunes (animales, vehículos, etc.) `https://www.cs.toronto.edu/~kriz/cifar.html`

Énfasis metodológico

El estudiante deberá explicar con claridad:

- Qué transformación aplicó a las imágenes (normalización, vectorización, etc.).
- Qué técnica de reducción de dimensión utilizó y por qué.
- Cómo interpretó las componentes o clusters resultantes.

Recomendación importante

Nota: Este proyecto implica un mayor costo computacional.

5. Propuesta de proyecto propio

Los estudiantes pueden proponer un proyecto diferente, el cual deberá cumplir con los siguientes criterios:

Requisitos mínimos

- **Enfoque estadístico**: Debe incluir técnicas de inferencia, modelado o pruebas de hipótesis
- **Complejidad adecuada**: Comparable en dificultad a los proyectos propuestos
- **Datos accesibles**: Fuentes públicas o disponibles para evaluación
- **Validación rigurosa**: Métodos para evaluar la robustez de los resultados

Elementos a considerar en la propuesta

- Justificación del problema y su relevancia estadística
- Descripción de las técnicas de inferencia a emplear
- Plan para la validación de supuestos y resultados
- Métricas de evaluación estadísticamente sólidas

Notas finales

Los proyectos buscan fomentar el razonamiento analítico y la capacidad de comunicar resultados con claridad. No se evaluará únicamente el resultado numérico o gráfico, sino la **capacidad para articular un argumento analítico sólido, apoyado en evidencia y justificación técnica**.

Énfasis en la práctica estadística: Se valorará especialmente:

- La comprensión de los supuestos detrás de cada método
- La evaluación de la incertidumbre en las estimaciones
- La interpretación contextual de los resultados estadísticos
- La comunicación honesta de limitaciones metodológicas

Presentación final y evaluación del dominio técnico

Cada alumno dispondrá de **10 minutos** para presentar los hallazgos clave de su proyecto, seguidos de 5 minutos de preguntas del profesor y compañeros.

Estructura sugerida para la presentación

- Introducción y planteamiento del problema
- Metodología y justificación de técnicas estadísticas

- Resultados principales y visualizaciones clave
- Conclusiones y lecciones aprendidas

Criterios de evaluación del dominio técnico

Se evaluará específicamente la capacidad de:

- **Explicar el fundamento estadístico**: Comprender y comunicar los principios teóricos detrás de las técnicas empleadas
- **Justificar decisiones metodológicas**: Argumentar por qué se eligió un modelo específico sobre alternativas
- **Interpretar resultados en contexto**: Relacionar hallazgos estadísticos con el problema de investigación original
- **Responder preguntas técnicas**: Demostrar comprensión profunda durante la sesión de Q&A
- **Reconocer limitaciones**: Identificar y comunicar las restricciones metodológicas y supuestos del análisis

Puntos clave en la evaluación

- ¿El estudiante demuestra comprensión conceptual de las técnicas empleadas?
- ¿Puede explicar los resultados más allá de la implementación computacional?
- ¿Reconoce las implicaciones prácticas y limitaciones de sus hallazgos?
- ¿Responde preguntas técnicas con precisión y confianza?

Nota: La presentación no es solo un resumen del informe, sino una oportunidad para demostrar dominio profundo del enfoque estadístico y capacidad de pensamiento crítico sobre el análisis realizado.