Proyecto Final
INTELIGENCIA ARTIFICIAL Y BIG DATA APLICADO AL ÁMBITO BIOSANITARIO

1. Descripción general del Proyecto

El Proyecto tiene como objetivo integrar y demostrar, de forma aplicada, los conocimientos adquiridos a lo largo del curso. El estudiante deberá diseñar, implementar y documentar un flujo de trabajo completo de ciencia de datos, desde la formulación de la pregunta hasta la generación de recomendaciones accionables para un contexto clínico, biosanitario o de negocio.

El proyecto se basa en el análisis de un conjunto de datos tabulares y en la construcción de modelos predictivos y/o explicativos, utilizando Python, R y SQL (al menos uno de los lenguajes principales, preferiblemente combinando varios tal como se ha trabajado en el curso).

2. Elección del conjunto de datos

El estudiante deberá trabajar con un conjunto de datos principal, que servirá como base para todo el proyecto. Existen tres vías posibles para escoger los datos:

- Proponer un CSV propio (de una fuente pública o institucional) sujeto a validación previa del profesor.
- Elegir un proyecto/dataset asociado a una competición de Kaggle.
- Elegir uno de los tres datasets propuestos.

En todos los casos, el dataset debe permitir plantear una pregunta clara, disponer de una variable objetivo (target) bien definida y ofrecer un volumen de datos y variables suficiente como para aplicar las técnicas vistas en el curso (exploración, modelización, evaluación, calibración, umbral de decisión, análisis por subgrupos, etc.).

Fecha límite de elección del dataset: al finalizar la clase del día 18 de noviembre, cada estudiante deberá tener elegido su conjunto de datos y haberlo comunicado y consensuado con el profesor. En caso contrario, el profesor podrá asignar uno de los datasets propuestos por el curso.

2.1 Dataset propio (CSV)

Si el estudiante dispone de un dataset propio, procedente de una fuente pública (por ejemplo, repositorios abiertos, open data institucional) o de un entorno profesional del estudiante, siempre que:

- Se hayan anonimizado correctamente todos los datos personales.
- Exista una variable objetivo (target) razonable y bien definida.
- El tamaño del dataset sea suficiente para aplicar las técnicas del curso (recomendación orientativa: ≥ 1.000 filas y ≥ 8–10 variables relevantes).
- No se infrinjan normas legales (RGPD/LOPDGDD) ni compromisos de confidencialidad.

En este caso, será obligatorio enviar al profesor, dentro del plazo fijado, una breve propuesta con: enlace o fuente de los datos, descripción de la variable objetivo y pregunta principal del proyecto. El profesor deberá aprobar explícitamente la propuesta antes de que el alumno pueda avanzar.

2.2 Proyecto basado en una competición de Kaggle

El estudiante puede optar por utilizar los datos de una competición de Kaggle (https://www.kaggle.com/competitions). En este caso:

- La competición debe estar basada en datos tabulares (no imagen, texto libre o audio como tarea principal), a menos que el profesor autorice explícitamente lo contrario.
- Debe existir una variable objetivo clara (clasificación o regresión).
- El foco del Proyecto Final no es maximizar la posición en el leaderboard, sino diseñar un pipeline riguroso, bien documentado y alineado con los criterios del curso.
- El alumno deberá descargar y conservar una copia local de los datos utilizados, que formará parte del ZIP a entregar, respetando en todo momento las normas de uso de Kaggle.

2.3 Datasets propuestos

Si el estudiante no dispone de un dataset propio o prefiere trabajar con un caso ya curado y conocido, puede elegir uno de los siguientes tres conjuntos de datos. Estos CSV se entregarán junto con el enunciado del Proyecto Final.

- Stroke Prediction Dataset (healthcare-dataset-stroke-data.csv)
  - Dominio biosanitario puro: predicción de ictus.
  - Fuente original: Kaggle – "Stroke Prediction Dataset".
  - Aproximadamente 5.000 pacientes y unas 12 columnas.
  - Variables típicas: edad, sexo, hipertensión, enfermedad cardíaca, tipo de trabajo, índice de masa corporal (BMI), hábito de fumar, etc.
  - Variable objetivo (target): stroke (0 = no, 1 = sí).
  - Este dataset es ideal para un proyecto centrado en la predicción de un evento clínico binario con desbalance de clases, e interpretación de factores de riesgo.
- Obesity Prediction Dataset (ObesityDataSet_raw_and_data_sinthetic.csv)
  - Dominio biosanitario + estilo de vida, con enfoque en hábitos saludables.
  - Basado en el dataset de obesidad de UCI (México, Perú, Colombia).
  - Aproximadamente 2.100 registros y unos 17 atributos (edad, sexo, altura, peso, hábitos alimentarios, actividad física, etc.).
  - Variable objetivo (target): categoría de peso (insuficiente, normopeso, sobrepeso I/II, obesidad I/II/III).
  - Este dataset es especialmente didáctico para estudiar la relación entre estilo de vida y niveles de peso, así como para trabajar con un problema de clasificación multiclase y reflexionar sobre las limitaciones de los datos sintéticos.
- Bank Marketing Dataset (bank.csv)
  - Dominio: marketing bancario (campañas de telemarketing de un banco portugués).
  - Fuente original: UCI / Kaggle – "Bank Marketing Dataset".
  - Aproximadamente 40.000 filas y unas 20 variables (edad, profesión, estado civil, tipo de contrato, número de contactos previos, etc.).
  - Variable objetivo (target): y (yes/no), si el cliente acaba contratando un depósito a plazo.
  - Este dataset ofrece un ejemplo potente fuera del ámbito salud, ideal para trabajar conceptos de coste, selección de umbral de decisión, retorno de la inversión y fairness en un entorno de negocio realista.

3. Objetivos de aprendizaje

Al finalizar el Proyecto Final, el estudiante deberá ser capaz de:

- Formular una pregunta clara y relevante a partir de un conjunto de datos tabulares.
- Diseñar y ejecutar un pipeline reproducible de análisis de datos con Python, R y/o SQL.
- Realizar una exploración y limpieza de datos sistemática, documentando las decisiones tomadas.
- Entrenar y comparar varios modelos predictivos (al menos uno lineal y uno basado en árboles/ensembles).
- Evaluar la calidad de los modelos con métricas adecuadas (ROC-AUC, PR-AUC, Brier, etc.).
- Aplicar técnicas de calibración y selección de umbral en problemas de clasificación probabilística.
- Analizar el rendimiento del modelo por subgrupos (edad, sexo, etc.) y discutir aspectos de equidad (fairness).
- Redactar un informe técnico y un resumen ejecutivo comprensibles para un público no experto.

4. Alcance mínimo obligatorio del proyecto

Aunque cada proyecto tendrá sus particularidades, todos los estudiantes deberán cubrir, como mínimo, los siguientes bloques de trabajo.

4.1 Planteamiento del problema y contexto

En el informe (PDF) deberá quedar explícito:

- El contexto del problema (clínico, biosanitario o de negocio).
- La pregunta principal que se desea responder.
- El tipo de problema (clasificación binaria, multiclase, regresión…).
- Las decisiones clave que podrían tomarse a partir del modelo (triage, priorización, oferta de producto, etc.).

4.2 Descripción, inspección y limpieza de datos

Se espera un trabajo sistemático que incluya:

- Descripción del dataset: número de filas y columnas, tipos de variables, prevalencia de la clase positiva (si aplica).
- Tablas y gráficas de exploración (distribuciones, relaciones básicas, correlaciones cuando tenga sentido).
- Análisis de valores perdidos, outliers y codificaciones especiales.
- Decisiones de limpieza (imputaciones, recodificaciones, exclusiones), justificadas brevemente.
- Definición clara de la muestra de análisis y, si corresponde, de train/validation/test o splits temporales.

4.3 Modelado y evaluación

Como mínimo, se deberán entrenar y comparar dos familias de modelos:

- Un modelo lineal o generalizado (por ejemplo, regresión logística).
- Un modelo basado en árboles/ensambles (por ejemplo, random forest, gradient boosting, XGBoost, LightGBM, CatBoost).

Para problemas de clasificación se deberán reportar, como mínimo, las siguientes métricas en una muestra de test no utilizada para el ajuste de hiperparámetros:

- ROC-AUC y, cuando tenga sentido, PR-AUC.
- Brier score (para modelos probabilísticos).
- Matriz de confusión para al menos un umbral de decisión.

En el caso de problemas de regresión, se deberán utilizar métricas apropiadas como RMSE, MAE, R^2, etc., justificando su elección. En este caso, las secciones de calibración, umbral y análisis por subgrupos deberán adaptarse al contexto de regresión (por ejemplo, centrándose en errores relativos por subgrupo, sesgos sistemáticos en la predicción, etc.).

4.4 Calibración y selección de umbral (problemas de clasificación)

En proyectos de clasificación con salidas probabilísticas (p.ej. probabilidad de ictus, obesidad grave o contratación del producto), se espera:

- Evaluar la calibración del modelo (curva de calibración, Brier score, ECE u otras medidas simples).
- Aplicar al menos una técnica de calibración (Platt / regresión logística o calibración isotónica) en un conjunto de validación adecuado.
- Seleccionar uno o varios umbrales operativos basados en un criterio de coste/beneficio y/o restricciones (por ejemplo, sensibilidad mínima y número máximo de alertas/día).
- Discutir las implicaciones prácticas de elegir un umbral más sensible, más específico o más equilibrado.

4.5 Análisis por subgrupos y equidad (fairness)

Se deberá analizar el comportamiento del modelo en al menos dos subgrupos relevantes (por ejemplo, por sexo, franjas de edad, tipo de trabajo, etc.).

- Comparar métricas básicas (sensibilidad, especificidad, tasa de falsos positivos, AUC…) entre subgrupos.
- Comentar posibles problemas de equidad y sesgo, y proponer medidas mitigadoras (si procede).
- Cuando el contexto y los datos lo permitan, se priorizará el análisis de equidad entre subgrupos relevantes; si no es posible (por ejemplo, ausencia de variables sensibles), el estudiante deberá explicarlo explícitamente.

4.6 Interpretabilidad del modelo

Se espera que el proyecto incluya herramientas de interpretación como:

- Coeficientes e odds ratios en modelos lineales.
- Importancia de variables (VIP) en modelos de árboles/ensambles.
- Gráficas que ayuden a entender el efecto de algunas variables clave (por ejemplo, plots marginales o similares).

5. Formato de entrega y estructura del ZIP

La entrega del Proyecto Final se realizará en un único archivo comprimido con extensión .zip. El nombre del archivo deberá seguir la convención: **PF_Apellidos_Nombre.zip**

Dentro del ZIP deberá incluirse, como mínimo, la siguiente estructura recomendada:

- data/raw/ → Datos originales (CSV descargados, tal como se obtuvieron de la fuente).
- data/processed/ → Datos procesados/listos para modelizar (si aplica).
- notebooks/ → Cuadernos Jupyter, RMarkdown o equivalentes utilizados en el análisis.
- scripts/ → Scripts .py, .R, .sql u otros utilizados para automatizar partes del pipeline.
- models/ → Modelos entrenados (por ejemplo, ficheros .joblib, .pkl, .rds).
- reports/ → Informe en PDF y, opcionalmente, versiones intermedias o HTML.
- figures/ → Gráficos finales incluidos en el informe.
- config/ → Ficheros de configuración (por ejemplo, config.yaml) si se utilizan.

Se recomienda incluir un fichero README (README.txt o README.md) con instrucciones breves sobre cómo reproducir los resultados principales (orden de ejecución de los notebooks/scripts, dependencias clave, etc.).

6. Requisitos del informe en PDF

El informe escrito es una parte fundamental del Proyecto Final. Deberá entregarse en formato PDF e incluir, al menos, las siguientes secciones (los títulos pueden adaptarse ligeramente):

- Resumen ejecutivo (máx. 1 página).
- Introducción y contexto.
- Datos y metodología.
- Resultados.
- Discusión, implicaciones prácticas y recomendaciones.
- Limitaciones del estudio y líneas futuras.
- Conclusiones.
- Referencias y anexos (opcional).

El tono del informe debe ser profesional, claro y riguroso, acorde con un proyecto profesional. Las figuras y tablas deberán estar correctamente referenciadas en el texto.

7. Consideraciones adicionales

- El proyecto es individual, salvo que el profesor autorice explícitamente un trabajo en grupo (solo si se trata de una competición de Kaggle y con la intención de darle continuidad después de finalizado el curso).
- Se permite el uso de librerías habituales de Python (scikit-learn, xgboost, lightgbm, catboost, etc.) y R (tidymodels, ranger, xgboost, etc.), así como SQL/DuckDB u otros motores afines.
- Es obligatorio citar adecuadamente cualquier recurso externo utilizado (artículos, blogs, código de ejemplo, notebooks de Kaggle, etc.).
- El uso de herramientas de IA generativa (incluyendo ChatGPT) debe declararse en el informe, indicando para qué se ha utilizado (por ejemplo, ayuda en la redacción, generación de código base, revisión de texto…).
- No se permitirá entregar proyectos basados únicamente en herramientas AutoML de tipo "caja negra" (por ejemplo, asistentes que prueban muchos modelos automáticamente y solo devuelven un ranking). Es obligatorio que el estudiante diseñe y programe su propio flujo de trabajo (preparación de datos, elección de modelos, validación, métricas, etc.), demostrando comprensión y criterio propio.

Cualquier duda sobre el alcance, la elección del dataset o la interpretación de este enunciado deberá consultarse con el profesor con la máxima antelación posible.

8. Plazos de entrega y presentación oral

Fecha límite de entrega del proyecto
La entrega del archivo comprimido (PF_Apellidos_Nombre.zip) deberá realizarse como máximo el día 25 de noviembre a las 9:00 h, a través del aula virtual del curso (justo antes de comenzar el Examen Final Teórico).

Presentación oral del proyecto
El mismo 25 de noviembre, durante la clase presencial de ese día, cada estudiante realizará una presentación oral de su proyecto ante el resto del grupo.

- El tiempo máximo total por estudiante será de **15 minutos**, incluyendo:
  - Preparación del material en el aula,
  - Exposición,
  - Turno de preguntas/respuestas.
- Se valorará especialmente la capacidad de síntesis, la claridad en la explicación de la pregunta, la metodología seguida y las conclusiones prácticas.