Evaluación Final - Métodos Computacionales

Estimados estudiantes,

La evaluación final del curso de Métodos Computacionales consistirá en la resolución de un caso práctico aplicado a problemas reales del mundo de los datos. El trabajo incluye el análisis, desarrollo del modelo y exposición de los resultados.

El objetivo es que apliquen de manera integrada los conceptos de programación, preprocesamiento, modelos de machine learning y análisis crítico que hemos trabajado a lo largo del curso.

A continuación, se presentan los tres casos disponibles. Deben escoger uno para desarrollar y preparar una exposición con su solución.

Caso 1

Contexto:
Eres analista de datos en una empresa de telecomunicaciones que desea reducir la pérdida de clientes. La empresa cuenta con un conjunto de datos que contiene información relevante sobre el comportamiento, los servicios contratados y los datos demográficos de sus clientes.

Descripción del problema:
A partir del historial de los clientes, debes construir un modelo de clasificación que prediga si un cliente se dará de baja en el próximo mes.

Datos disponibles:

Cada fila corresponde a un cliente y cada columna representa una característica específica del mismo.

El conjunto de datos incluye las siguientes variables:

Churn: indica si el cliente se dio de baja en el último mes (Yes/No).

Servicios contratados: teléfono, múltiples líneas, internet, seguridad en línea, respaldo en línea, protección de dispositivos, soporte técnico, y streaming de TV y películas.

Información de cuenta: duración como cliente (tenure), tipo de contrato (mensual, anual, bienal), método de pago, facturación sin papel, cargos mensuales (MonthlyCharges) y cargos totales (TotalCharges).

Datos demográficos: género, edad (indicador de "SeniorCitizen"), si tiene pareja (Partner) y si tiene dependientes (Dependents).

Su análisis debe incluir:

Preprocesamiento:

Limpia datos faltantes si los hubiera.

Codifica variables categóricas (por ejemplo, mediante one-hot encoding o dummies).

Ajusta formatos numéricos según sea necesario.

Modelado:

Entrena al menos un modelo de clasificación (por ejemplo, regresión logística, árbol de decisión o Random Forest, etc.) para predecir la probabilidad de churn.

Evalúa el rendimiento del modelo mediante métricas como precisión (accuracy), sensibilidad (recall), precisión (precision), F1-score o AUC-ROC.

Interpretación:

Identifica y analiza las variables que más contribuyen a que un cliente abandone.

Sugiere posibles acciones que la empresa podría implementar pensando en los resultados obtenidos (por ejemplo, mejorar contratos flexibles, modificar métodos de pago o servicios complementarios).

Comparación y selección de modelo:

Ajusta el umbral de decisión u otros parametros para optimizar la sensibilidad (capturar más casos de churn), considerando el costo de falsos positivos (campañas innecesarias) y falsos negativos (clientes perdidos sin intervención).

Compara múltiples modelos y comenta cuál es más adecuado según el contexto del negocio.

Caso 2

Contexto:
Eres analista de datos en un centro de salud pública que busca implementar un sistema predictivo para detección temprana de diabetes. El objetivo es identificar los factores que permitan flaggear a los pacientes con mayor riesgo y así implementar intervenciones preventivas.

Descripción del problema:
A partir de variables clínico-demográficas recolectadas en mujeres, construirás un modelo

de clasificación que prediga si una paciente tiene diabetes (Outcome = 1) o no (Outcome = 0).

Datos disponibles:
El conjunto de datos proviene del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales e incluye a 768 mujeres. Cada fila corresponde a una paciente y contiene las siguientes variables

Pregnancies: número de embarazos previos

Glucose: concentración de glucosa en plasma a las 2 horas después de una prueba de tolerancia oral a la glucosa

BloodPressure: presión arterial diastólica (mm Hg)

SkinThickness: grosor del pliegue cutáneo del tríceps (mm)

Insulin: insulina sérica a 2 horas (mu U/ml)

BMI: índice de masa corporal (kg/m^2)

DiabetesPedigreeFunction: función de pedigrí de diabetes (historia familiar)

Age: edad (años)

Outcome: variable objetivo (1 = diabetes, 0 = no diabetes)

Algunos valores como Glucose, BloodPressure, SkinThickness, Insulin o BMI pueden tener ceros que representan datos faltantes

Su análisis debe incluir:

Análisis exploratorio y preprocesamiento

Detecta y maneja los valores faltantes si los hubiera (ej.: ceros), utilizando técnicas adecuadas de imputación.

Realiza análisis descriptivo (medias, medianas, outliers) y visualizaciones relevantes para entender la distribución de las variables clave (especialmente con respecto a Outcome).

Modelado

Entrena al menos dos modelos clasificatorios distintos (por ejemplo, regresión logística, Random Forest, SVM, etc).

Evalúa el rendimiento mediante métricas adecuadas: precisión (accuracy), sensibilidad (recall), precisión (precision), F1-score y AUC-ROC.

Interpretación e impacto clínico

Identifica las variables más influyentes en la predicción de diabetes (utiliza importancia de variables, coeficientes o técnicas como SHAP/LIME si lo consideras).

Discutir cómo estos resultados podrían orientar decisiones clínicas o estrategias de intervención preventiva.

Comparación de modelos y ajuste

Ajusta el umbral de decisión u otros parametros si corresponde, justificando su selección en función de la estrategia clínica (por ejemplo, privilegiar sensibilidad si el objetivo es minimizar falsos negativos).

Compara los modelos entre sí y explica cuál recomendarías para desplegar en el sistema clínico, considerando tanto métricas como costo o consecuencia de errores.

Caso 3

Contexto:
Eres analista de datos en un Ministerio de Educación que busca entender qué factores influyen en el rendimiento de los estudiantes para diseñar políticas educativas más efectivas. Tu tarea es construir un modelo que prediga la nota final de un estudiante en matemáticas, en función de variables socioeconómicas y hábitos de estudio.

Descripción del problema:
Se dispone de un conjunto de datos con información de estudiantes de secundaria de dos colegios de Portugal. La variable objetivo es la nota final en matemáticas (G3), que va de 0 a 20 puntos.

Datos disponibles:
Cada fila representa un estudiante y las columnas incluyen información como:

school: escuela a la que asiste el estudiante

sex: género

age: edad

studytime: horas de estudio semanales

failures: número de materias reprobadas anteriormente

famsup: apoyo educativo de la familia (sí/no)

goout: frecuencia de salidas con amigos

health: estado de salud autodeclarado (1–5)

absences: número de ausencias a clase

G1, G2: calificaciones parciales anteriores

G3: calificación final en matemáticas (variable objetivo)

Su análisis debe incluir:

Preprocesamiento de datos

Maneja valores faltantes o inconsistencias si las hubiera.

Codifica variables categóricas y estandariza variables numéricas cuando sea necesario.

Explora gráficamente la relación entre algunas variables clave (ej.: horas de estudio, ausencias, notas previas) y el rendimiento final.

Modelado

Entrena al menos dos modelos de regresión distintos (por ejemplo: regresión lineal múltiple, Random Forest Regressor, Gradient Boosting).

Evalúa el desempeño con métricas de regresión: RMSE, MAE y R^2.

Interpretación de resultados

Identifica las variables que más influyen en el rendimiento académico.

Discute si los factores identificados coinciden con lo esperado en un contexto educativo (ej.: importancia de las notas previas o de la asistencia).

Comparación y selección de modelo

Compara los modelos construidos y selecciona el más adecuado para este problema, justificando tu elección con base en métricas y utilidad práctica.

Propón recomendaciones de política educativa o acciones pedagógicas basadas en los hallazgos.

Indicaciones de entrega

- El trabajo debe ser desarrollado en un Jupyter Notebook o Google Colab.
- La entrega incluye: código bien documentado, análisis de resultados y conclusiones.
- Cada estudiante debe preparar una exposición oral (25 minutos) explicando su solución.
- Se evaluará el preprocesamiento, construcción y validación del modelo, interpretación de resultados y la claridad de la exposición.