Proyecto 3.
Análisis Comparativo de Modelos de Regresión y Clasificación utilizando el Dataset de E-commerce Brasileño

INSTRUCCIONES:

"La empresa brasileña Olist, líder en comercio electrónico en América Latina, ha contratado a un grupo selecto de estudiantes de ciencia de datos para analizar su extenso conjunto de datos de operaciones comerciales. El objetivo principal es obtener insights valiosos que permitan optimizar sus operaciones y mejorar la experiencia del cliente. Con datos de más de 100,000 órdenes realizadas entre 2016 y 2018, este proyecto representa una oportunidad única para que los estudiantes apliquen sus conocimientos en análisis predictivo, utilizando técnicas tanto de regresión como de clasificación.

Los estudiantes trabajarán en equipos durante cuatro semanas, implementando diversos modelos de machine learning para abordar desafíos específicos como la predicción de precios, la clasificación de la satisfacción del cliente y la optimización de tiempos de entrega. Al final del proyecto, cada equipo presentará sus hallazgos y recomendaciones directamente a los directivos de Olist, quienes evaluarán la viabilidad de implementar las soluciones propuestas en sus operaciones diarias."

DESCRIPCIÓN DEL DATASET

- Disponible en Kaggle: [Brazilian E-commerce Public Dataset by Olist](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)
- Contiene aproximadamente 100,000 órdenes de 2016 a 2018
- Múltiples archivos CSV que cubren órdenes, productos, clientes y reseñas
- Datos comerciales reales, anonimizados para uso público

ACTIVIDADES Y ENTREGABLES

Semana 1: Preparación de Datos y Análisis Inicial

- Entregables:

1. Propuesta inicial del proyecto (2-3 páginas) incluyendo:
    1. Planteamiento del problema
    2. Metodología propuesta
    3. Resultados esperados
2. Reporte de EDA conteniendo:

1. Evaluación de calidad de datos
2. Resúmenes estadísticos
3. Visualización de patrones clave
4. Análisis de correlación de características

3. Dataset preprocesado listo para modelado

Semana 2: Implementación de Modelos de Regresión

- Entregables:

1. Reporte técnico cubriendo:
    1. Implementación de todos los modelos de regresión
    2. Proceso de ajuste de parámetros
    3. Métricas de rendimiento para cada modelo
2. Repositorio de código con:
    1. Implementaciones documentadas de regresión
    2. Pipeline de preprocesamiento de datos
3. Presentación de resultados preliminares (5-10 diapositivas)

Semana 3: Implementación de Modelos de Clasificación

- Entregables:

1. Reporte técnico cubriendo:
    1. Implementación de todos los modelos de clasificación
    2. Proceso de ajuste de parámetros
    3. Métricas de rendimiento para cada modelo
2. Repositorio de código con:
    1. Implementaciones documentadas de clasificación
    2. Procedimientos de validación de modelo
3. Presentación de resultados preliminares (5-10 diapositivas)

Semana 4: Validación Cruzada y Redes Neuronales

- Entregables:

1. Reporte técnico cubriendo:
    1. Implementación de validación cruzada k-fold
    2. Comparación de resultados con/sin CV
    3. Implementación de redes neuronales básicas
    4. Análisis de arquitecturas y resultados
2. Repositorio de código con:
    1. Implementación documentada de CV
    2. Implementación de redes neuronales
    3. Pipeline completo de entrenamiento
3. Presentación de resultados preliminares (5-10 diapositivas)

Semana 5: Análisis Final y Presentación

- Entregables:

1. Reporte final completo (10-15 páginas) incluyendo:
    1. Resumen ejecutivo
    2. Metodología detallada
    3. Análisis comparativo de todos los modelos
    4. Insights de negocio y recomendaciones
2. Presentación final (15-20 minutos)
3. Repositorio de código completo con:
    1. Código bien documentado
    2. Archivo README
    3. Requirements.txt (requerimientos de software, librerías, versiones, etc.)
4. Póster resumiendo hallazgos clave

EVALUACIÓN

Nota Importante: Tiene que poderse comprobar su aporte al trabajo grupal a través de “commits”. Si no existen al menos 10 “commits” con su aporte significativo no va a tener nota en esta entrega. Utilice una herramienta que permita registrar los aportes de cada uno.

Rúbrica Detallada por Semana:

Semana 1: Preparación de Datos y Análisis Inicial (100%)

- Propuesta inicial del proyecto (30%)
  - Planteamiento claro del problema (10%)
  - Metodología bien definida (10%)
  - Objetivos y resultados esperados (10%)
- Reporte EDA (40%)
  - Calidad del análisis exploratorio (15%)
  - Visualizaciones efectivas (10%)
  - Análisis estadístico (15%)
- Dataset preprocesado (30%)
  - Limpieza de datos (15%)
  - Transformación de variables (15%)

Semana 2: Modelos de Regresión (100%)

- Implementación de modelos (40%)
  - Código bien documentado (15%)
  - Implementación correcta de todos los modelos (25%)
- Reporte técnico (35%)
  - Descripción de metodología (10%)
  - Análisis de resultados (15%)
  - Justificación de decisiones técnicas (10%)
- Presentación preliminar (25%)
  - Claridad en la presentación (10%)
  - Calidad del material visual (5%)
  - Dominio del tema (10%)

Semana 3: Modelos de Clasificación (100%)

- Implementación de modelos (40%)
  - Código bien documentado (15%)
  - Implementación correcta de todos los modelos (25%)
- Reporte técnico (35%)
  - Descripción de metodología (10%)
  - Análisis de resultados (15%)
  - Justificación de decisiones técnicas (10%)
- Presentación preliminar (25%)
  - Claridad en la presentación (10%)
  - Calidad del material visual (5%)
  - Dominio del tema (10%)

Semana 4: Validación Cruzada y Redes Neuronales (100%)

- Implementación de modelos (40%)
  - Implementación de validación cruzada (20%)
  - Implementación de redes neuronales (20%)
- Reporte técnico (35%)
  - Descripción de metodología (10%)
  - Análisis comparativo de resultados (15%)
  - Justificación técnica de decisiones (10%)
- Presentación preliminar (25%)
  - Claridad en la presentación (10%)
  - Calidad del material visual (5%)
  - Dominio del tema (10%)

Semana 5: Análisis Final y Presentación (100%)

- Reporte final completo (40%)
  - Integración del trabajo (15%)
    - Coherencia general (5%)
    - Flujo lógico (5%)
    - Completitud (5%)
  - Análisis comparativo (15%)
    - Comparación exhaustiva de modelos (5%)
    - Análisis de métricas (5%)
    - Interpretación de resultados (5%)
  - Conclusiones y recomendaciones (10%)
    - Insights relevantes (5%)
    - Recomendaciones prácticas (5%)
- Presentación final (30%)
  - Claridad y profesionalismo (10%)
    - Estructura clara (3%)
    - Comunicación efectiva (4%)
    - Manejo del tiempo (3%)
  - Calidad del material visual (10%)
    - Diseño profesional (3%)
    - Visualizaciones efectivas (4%)
    - Organización de información (3%)
  - Dominio del tema (10%)
    - Respuesta a preguntas (5%)
    - Manejo de conceptos técnicos (5%)
- Repositorio y documentación (30%)
  - Organización del código (15%)
    - Estructura de archivos (5%)
    - Documentación clara (5%)

  - Control de versiones (mínimo 9 commits significativos por estudiante) (5%)
- Calidad del póster (15%)
  - Diseño visual (5%)
  - Contenido relevante (5%)
  - Claridad en la presentación (5%)

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con la información requerida por la empresa.
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó jupyter notebooks o rmd debe añadir el html que se genera)
- Enlace de controlador de versiones utilizado.

FECHAS DE ENTREGA

NOTA: Para poder tener nota completa debe entregar las asignaciones en el tiempo adecuado. No se calificarán las entregas si no fueron subidas en tiempo, aunque estén en el repositorio.

Sugerencia: Los días lunes de clase, presencial, tendrá un tiempo de aclaración de dudas con el profesor. Se sugiere que el resto de la semana avance en la resolución práctica de los pasos del contenido teórico para que aclare todas sus dudas al respecto en dicho espacio.