Educación en la Era de los Datos: Cómo la Ciencia de Datos Está Moldeando el Futuro

Juan José Restrepo Rosero

I. Introducción

La Ciencia de Datos ha transformado radicalmente la resolución de problemas complejos y la toma de decisiones en varios sectores. En este informe, se exploran tres proyectos que ejemplifican su impacto en el ámbito académico. El primero optimiza la experiencia del usuario en compras en línea. El segundo predice el rendimiento académico de estudiantes de secundaria y finalmente, el tercer proyecto analiza las tendencias de investigación en inteligencia artificial y su futuro en la educación.

II. Modelo de Caso 1: Optimización de la Experiencia del Usuario en Compras en Línea. [1] [2]

El proyecto surgió por la creciente importancia del comercio electrónico en los consumidores. Fue liderado por el ingeniero Julio Hallo, como parte de su tesis de maestría en donde abordó el problema de la "Experiencia Multiproducto y Multioferta".

El propósito fue utilizar la ciencia de datos y modelos de Machine Learning para detectar y consolidar la presentación de productos similares como una oferta única y mejorar la experiencia al simplificar la toma de decisiones en E-Commerce. Las etapas clave del proyecto incluyeron:

1. **Identificación y Sistematización del Problema** Se definió el problema para establecer objetivos concretos y preguntas de investigación enfocadas en la frecuencia de las compras en línea de los usuarios, su experiencia de compra y si realizaban comparaciones entre productos.

2. **Adquisición y Preprocesamiento de Datos:** Obtener un conjunto de datos lo suficientemente robusto en cantidad de registros, calidad y estandarización fue un desafío inicial. Esto implicó realizar análisis de calidad de los datos y crear conjuntos de entrenamiento, validación y pruebas para el modelo que se utilizaría en la toma de decisiones.

3. **Desarrollo y Evaluación de Modelos:** Se propusieron dos basados en redes neuronales de Perceptrón Multicapa (MLP), con capas densas interconectadas que variaban en arquitectura, hiperparámetros, funciones de activación, tasas de aprendizaje y ajustes nodales específicos, y otros tres fundamentados en LSTM (Long Short-Term Memory) centrándose en el procesamiento del lenguaje natural y la identificación de productos duplicados empleando vectores más potentes para abordar la falta de datos en algunos casos de negocio.

4. **Modelación y Análisis:** Los modelos se evaluaron utilizando métricas de rendimiento, enfocándose en F1-Score. Se compararon los resultados y analizaron las ventajas y desventajas de cada enfoque, como la presencia de falsos positivos o negativos.

5. **Presentación de Resultados:** Los resultados incluyeron la implementación exitosa de cinco modelos para identificar productos duplicados y presentarlos como una oferta única, mejorando la experiencia del usuario y la tasa de conversión en E-Commerce. Los puntajes fueron:

   - **MLP/SM** (F1 Score: 0.818)
   - **LSTM/SM** (F1 Score: 0.785)
   - **LSTM/W2V_501** (F1 Score: 0.816)
   - **MLP/W2V_501** (F1 Score: 0.808)
   - **Bi LSTM/W2V_501** (F1 Score: 0.808)

El proyecto impactos positivos y negativos para cada stakeholder identificado. Estos se presentan a continuación:

TABLA I
Impactos del Proyecto caso 1

Grupos de interés:	Tipos de impacto	Impacto positivo	Impacto negativo
Clientes de Comercio Electrónico	Social	Mayor satisfacción del cliente en la toma de decisiones de compra informadas.	La eliminación de productos duplicados afectaría negativamente la percepción de la plataforma y la disposición a realizar futuras compras.
	Económico y ambiental	1. Mayor frecuencia de compras por la mejora en la experiencia. 2. Disminución en el consumo de recursos en la fabricación de productos por menor duplicación.	Mayor competencia entre vendedores por la eliminación de duplicados.

Plataformas de Comercio Electrónico	Técnico	Mayor eficiencia operativa y visibilidad al reducir los duplicados.	Desafíos técnicos al adaptarse a la nueva estructura de la plataforma.
	Económico y ambiental	1. Aumento de ingresos por un mayor volumen de transacciones. 2. Reducción de demanda de recursos por la disminución de duplicados.	Reducción de ingresos de los no beneficiados por la eliminación de duplicados

Fuente: Elaboración propia

III. Modelo de Caso 2: Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques.[3]

El proyecto se llevó a cabo en Monterrey, México enfocándose en un estudio comparativo de técnicas de aprendizaje automático supervisado para predecir el rendimiento académico de estudiantes de secundaria. Se evaluó y compararon diferentes enfoques de clasificación supervisada para detectar patrones en datos académicos y sociodemográficos.

Este estudio sigue las siguientes etapas:

1. **Recopilación de los Datos:** Se recopilaron datos de 3,726 estudiantes de tercer año de secundaria durante el año académico 2017-2018 por una encuesta que incluyó información sobre antecedentes educativos, situación laboral de los padres y otros factores. El conjunto de datos se agrupó de la siguiente manera: (Ver Tabla II)

TABLA II

DISCRETIZACIÓN DE CARACTERÍSTICAS [3]

Factores			
Académico	Institucional	Sociodemográfico	Económico
1. Promovido al siguiente grado 2. Grado 3. Deseo de abandonar 4. Tasa de aprendizaje 5. Aprendizaje auditivo 6. Aprendizaje Visual 7. Aprendizaje cinestésico 8. Dislexia 9. Escuela secundaria 10. Modalidad 11. Primer grado 12. Tipo de escuela 13. Educación de la madre 14. Educación del padre 15. Hermanos estudiando 16. Horas de sueño 17. Chequeo de tareas	1. Sentimientos sobre la escuela 2. Amistades con compañeros de clase 3. Peleas con compañeros de clase	1. Género 2. Edad 3. Vive con madre 4. Vive con el padre 5. Vive con otros 6. Vivir juntos 7. Vive con hermanos 8. Relación con los padres 9.Relación entre padres 10. Número de hermanos	1. Situación económica 2. Comida 3. Seguridad 4. Condiciones de vivienda 5. Servicio telefónico 6. Servicio de internet 7. Trabajo de la madre 8. Ttrabajo del padre 9. Hermanos trabajando

2. **Preprocesamiento de los datos:** Se eliminaron datos duplicados o inconsistentes, luego se categorizaron en respuestas sobre relaciones familiares como "Excelente", "Buena", "Regular" y "Mala". Se seleccionaron características para identificar aquellas que tenían un mayor impacto en la predicción del rendimiento académico.

3. **Clasificación:** Se aplicaron algoritmos de aprendizaje supervisado para predecir el rendimiento académico de los estudiantes en función de las características seleccionadas, destacando seis: Light Gradient Boosting Machine (lightgbm), Gradient Boosting (Gbc), AdaBoost (Ada), Logistic Regression (Lr), Random Forest (rf) y K-nearest Neighbors (Knn).

4. **Entrenamiento y evaluación de los modelos:** Se entrenaron estos modelos con el 70% de los datos, ajustándolos para aprender patrones y relaciones entre las características y el rendimiento académico. El 30% se usó con métricas de desempeño como Accuracy, Precision, Recall, y F1-score para determinar el mejor desempeño.

5. **Análisis Comparativo de Enfoques de Clasificación:** Se analizaron los resultados de los modelos para identificar el enfoque más efectivo prediciendo el rendimiento académico. Las puntuaciones de todos los algoritmos se aprecian en la siguiente tabla, en donde Gbc fue el mejor y se refuerza con la matriz de confusión y curva ROC.

TABLA III

RESULTADOS DE LOS ALGORITMOS DE CLASIFICACIÓN [3]

Classifiers	**Metrics**			
	Accuracy (%)	***Recall (%)***	***Precision (%)***	***F1(%)***
Gbc	**95.44**	98.33	**96.77**	97.54
Ada	95.28	98	96.51	97.45
Knn	95.21	98.62	96.26	97.42
Lr	95.17	98.25	96.57	97.39
lightgbm	94.98	97.83	96.75	97.28
rf	94.94	98.29	96.30	97.29

FIGURA I
MATRIZ DE CONFUSIÓN DEL ALGORITMO GBC [3]

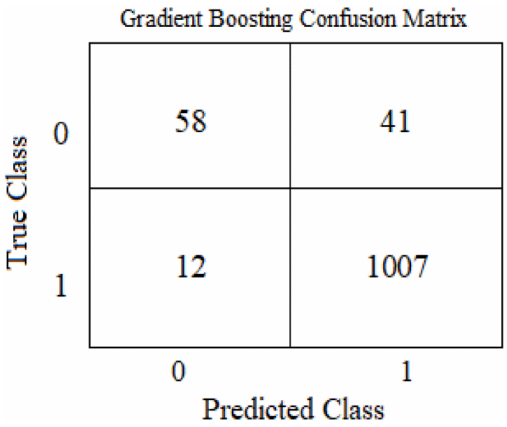
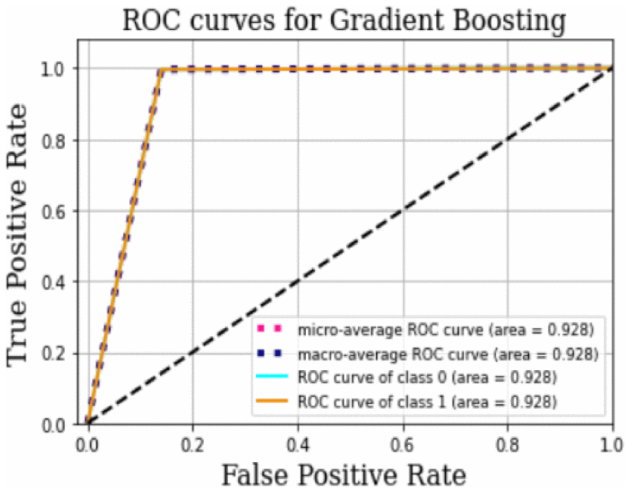


FIGURA II
CURVA ROC DEL RECEPTOR GBC [3]



Los diferentes stakeholders identificados se pueden apreciar en la siguiente tabla:

TABLA IV
IMPACTOS DEL PROYECTO CASO 2

Grupos de interés:	Tipos de impacto	Impacto positivo	Impacto negativo
Estudiantes de secundaria:	Social	Apoyo a los estudiantes con dificultades académicas.	Estigmatización de estudiantes por "etiquetado" de bajo rendimiento.
	Padres de familia	Mayor participación en el rendimiento académico de sus hijos.	Preocupaciones sobre la privacidad de los datos de sus hijos.
Instituciones Educativas	Técnico	1. Mayor capacidad para brindar apoyo a estudiantes con dificultades. 2. Mejora en la reputación de la institución y la efectividad de la enseñanza. 3. Oportunidades de investigación y desarrollo en el campo de la minería de datos educativos.	1. Desafíos técnicos en la implementación y mantenimiento de soluciones de IA. 2. Temores de que la automatización pueda reemplazar la evaluación de los profesores.
	Ético	Promoción de prácticas éticas en el uso de datos educativos.	Preocupaciones éticas relacionadas con la privacidad de los datos y el sesgo algorítmico

Fuente: Elaboración propia

IV. MODELO DE CASO 3: RESEARCH HOTSPOTS AND DEVELOPMENT TRENDS OF AI IN EDUCATION: A STUDY BASED ON KNOWLEDGE MAP AND CO-WORD ANALYSIS [4]

Es un proyecto parte del "Plan de Desarrollo de IA de Nueva Generación" 2017 del Consejo de Estado de Shenzhen, China, que piensa integrar gradualmente la inteligencia artificial en entornos educativos a través de la recopilación y el análisis de Big Data para personalizar la enseñanza de los estudiantes.

Se empleó un software de análisis visual para examinar el estado de la investigación en IA educativa, centrándose en dos aspectos: los puntos críticos de investigación y las fronteras de investigación. Además, se prestó atención a la protección de los datos, siguiendo una metodología que consta de cuatro fases:

1. **Recopilación de los datos**: Se recopilaron 373 documentos y artículos de la base de datos CNKI relacionados con IA y educación de 2012 a 2018.

2. **Análisis comparativo de los puntos críticos de investigación:** Se utilizó la herramienta Citespace para aplicar diversas técnicas para visualizar y comprender la investigación en IA y educación en China. Los pasos incluyeron:

   a. **Generación de Mapas de Co-ocurrencia de Palabras Clave:** Citespace generó mapas de co-ocurrencia para identificar las palabras clave y su relación en la investigación que ganaban relevancia con el tiempo.

   b. **Aplicación de Algoritmo LLR para Agrupación:** Se aplicó el algoritmo LLR (Log-Likelihood Ratio) para detectar palabras emergentes e identificar las que se utilizaban con mayor frecuencia en un período de tiempo específico determinando temas en ascenso en la investigación.

   c. **Generación de Mapas de Secuencia Temporal:** Se crearon mapas de secuencia temporal para identificar tendencias en el tiempo, mostrando cómo evolucionaron las etapas. La de 2012 a 2016, se caracterizó la investigación teórica, la robótica educativa y las influencias de la IA en educación y la

segunda etapa, desde 2017 hasta el presente, experimentó un rápido desarrollo y se centró en aplicaciones prácticas de la IA académica.

3. **Identificación de Hotspots de Investigación**: Se identificaron hotspots de investigación en el campo de la IA educativa en China y se agruparon en cinco áreas:

    a. **Investigación Teórica sobre el Impacto de la AI en la Educación:** Se analizó cómo la IA afecta a la educación desde una perspectiva teórica, incluyendo desafíos y oportunidades.

    b. **Investigación sobre la Aplicación de Tecnología AI en la Educación:** Se exploró cómo se aplica el aprendizaje automático y la analítica de datos en la enseñanza

    c. **Educación Robótica:** Se investigó el papel de la robótica en la educación, incluyendo la programación de robots y la educación maker.

    d. **Formación de Personal en la Era de la AI:** Se examinó cómo la educación se adapta para formar profesionales en un entorno dominado por la IA.

    e. **Desarrollo de Docentes en un Entorno de AI:** Se analizó cómo los profesores se adaptan a las tecnologías de IA y su evolución en la educación.

4. **Evaluación de resultados:** Se resumieron los resultados, que se habían dividido previamente, destacando tres áreas principales:

    a. **Expansión de la Diversidad de Investigación en AI en Educación:** La investigación se centraba en la educación superior y educación vocacional. Se busca ampliar la aplicación de la IA en todos los niveles educativos y promover el desarrollo integral de los estudiantes.

    b. **Promoción de la Integración de la Teoría y la Práctica de AI en la Educación:** Se señaló que la teoría de IA educativa aún estaba en sus primeras etapas y se instó a profundizar en la investigación teórica combinándola con la práctica.

    c. **Atención a la Protección de Datos de Privacidad:** Se subrayó la importancia de la privacidad de los estudiantes en un entorno donde la IA se usaba para monitorear el aprendizaje y la necesidad de establecer reglas sólidas de privacidad de datos.

Finalmente, se identificaron los impactos positivos y negativos para cada uno de los stakeholders involucrados en esta investigación. Estos se aprecian en la siguiente tabla:

TABLA III
IMPACTOS DEL PROYECTO CASO 3

Grupos de interés:	Tipos de impacto	Impacto positivo	Impacto negativo
Instituciones Educativas	Técnicos y económicos	Mayor eficiencia en los procesos administrativos e implementación de una enseñanza más eficiente.	Costos elevados de implementación y mantenimiento.
	Social y Ético	1. Mejora en la calidad de la educación, personalización del aprendizaje. 2. Promoción del uso ético de la IA en la educación y en otras áreas.	1. Preocupaciones sobre la privacidad y ética del manejo los datos de los estudiantes.
Estudiantes	Técnicos y económicos	1. Posibilidad de aprendizaje personalizado y efectivo. 2. Acceso a recursos de aprendizaje personalizados y actualizados.	1. Posibles dificultades de adaptabilidad del estudiante y presentar problemas técnicos. 2. Posibles costos adicionales asociados a la tecnología.
	Social y ético	1. Enseñanza acerca de la protección de datos y privacidad.	1. Falta de interacción humana en la educación. 2. Posible discriminación algorítmica.

Fuente: Elaboración propia

V. CONCLUSIÓN.

En resumen, estos tres proyectos ejemplifican el potencial transformador de la Ciencia de Datos en el sector académico, desde la optimización de compras en línea, hasta la predicción del rendimiento estudiantil y análisis de tendencias en la investigación educativa. Cada proyecto comparte un compromiso fundamental en la extracción de datos complejos y la aplicación de técnicas avanzadas, proporcionando conocimientos que buscan mejorar continuamente la calidad, la eficiencia y la equidad en la educación, abriendo nuevas puertas para enriquecer el desarrollo de las mentes futuras y el progreso continuo de la academia en la era de la Ciencia de Datos.

IV. BIBLIOGRAFÍA.

[1] Centro Magis [Javeriana Cali], Sector academico: Julio Xavier Hallo Larrea - Parte 1. YouTube, 2022.

[2] Centro Magis [Javeriana Cali], Sector academico: Julio Xavier Hallo Larrea - Parte 2. YouTube, 2022.

[3] N. N. Sánchez-Pozo, J. S. Mejía-Ordóñez, D. C. Chamorro, D. Mayorca-Torres and D. H. Peluffo-Ordóñez, "Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques," 2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop, Monterrey, Mexico, 2021, pp. 1-6, doi: 10.1109/IEEECONF53024.2021.9733756.

[4] X. Lu, "Research Hotspots and Development Trends of AI in Education: A Study Based on Knowledge Map and Co-Word Analysis," 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, 2020, pp. 212-217, doi: 10.1109/AEMCSE50948.2020.00052.