**Reporte Maestría en Ciencia de Datos**

**Asignatura Retos en Ciencia de Datos**

**La Ciencia de Datos como Aporte al Sector educativo**

Sonia Yurany Gallego Paz

**Introducción**

En educación, la ciencia de datos se relaciona con el desarrollo de métodos para extraer información útil a partir de los datos que se generan en los entornos académicos y como en otros sectores, se usa para la toma de decisiones.

En el sector académico, permite resolver diferentes problemas como la deserción, el riesgo de abandono, el desgranamiento escolar, el rendimiento académico, los factores predictivos de éxito y calidad de las estrategias educativas entre otros [1].

A continuación, se presentan tres proyectos que aplican ciencia de datos, uno relacionado con sector productivo y dos con el sector académico:

1. Identificación de productos duplicados en e-comerce que mejora la toma de decisiones [2, 3].
2. Identificación de variables asociadas al éxito en estudiantes de modalidad e-learning[4].
3. Impacto del uso de una aplicación web en el proceso de enseñanza-aprendizaje sobre la Probabilidad Binomial[5].

**Proyecto 1**[2, 3]**:**

No describe cuando y donde se hizo, probablemente en la ciudad de Cali.

No relacionado con sector académico, se relaciona más con el sector productivo.

**Objetivo:**

Identificar productos duplicados ofertados en e-comerce o marketplace mediante procesos de machine learning para fusionar y presentar como única oferta, mejorando la toma de decisión y aumentar la comisión.

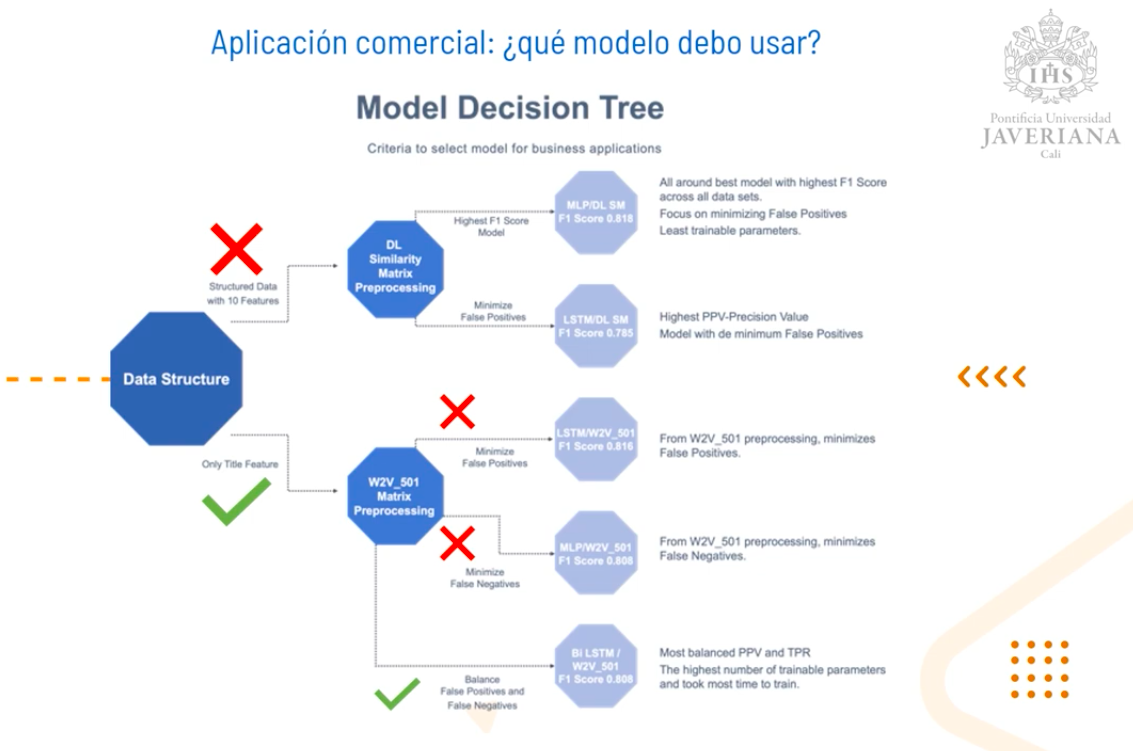
**Etapas:**

1. Identificación y sistematización del problema, definición de objetivos.
2. Obtención y depuración de datos, análisis descriptivo, conocimiento de las variables para definición de conjuntos de entrenamiento, validación y prueba.
3. Estimación de hiperparámetros: Proponer los modelos para identificar los de mejor performance. Primero se define y luego se entrena en parámetros específicos que llevan a un modelo aceptable. Cada grupo debe tener el mismo balance para tener consistencia, es un proceso iterativo.
4. Análisis y comparación de resultados. Utilizaron Deep learning(DL): Red neuronal multicapa que permitía caracterizar las capas, conocer los nodos, identificar el resultado, conocer arquitectura, se fueron ajustando hiperparametros. Con LSD hicieron reconocimiento del lenguaje con vectores más potentes que ayudaban a solucionar la escases de datos. Se identificaron los hiperparametros que más contribuyen según l6a arquitectura y metodología de trabajo. No todos los modelos solucionan todos los problemas. Los resultados deben ser traducibles al negocio, tener en cuenta falsos positivos.

**Resultados:**

* 2 modelos de pre procesamiento de datos,
* Utilizando machine learning, 5 modelos diferentes de identificación de posibles duplicados (Figura 1) con 5 casos de negocio específico que atiende cada modelo y son derivados del tipo de datos disponibles. Uno da prelación al rendimiento total basado en el F1 score sin tener en cuenta las clases, falsos positivos o falsos negativos. Cada modelo tiene un costo en rendimiento. Basados en DL se obtuvieron dos modelos. Los otros tres modelos se basaron en LSD

**Figura 1.** Modelos utilizados



Retos: requerimiento de alto volumen de datos, muchos datos no estaban en español. Fue anticipado desde la planeación. En la construcción de modelos, se requiere investigación bibliográfica exhaustiva. Condensación de datos y capacidad computacional.

**Impacto positivo**:

* Industria: mejorar oferta de un único producto, reducción de productos exhibidos, aumento de conversión en 3X. Oportunidad de nueva implementación con comparación de precios competitivos
* Clientes: mejor experiencia, menos clic para comprar.

**Proyecto 2**[4]**:**

Realizado por el Centro de Educación a Distancia de la Universidad Católica del Norte (CED-UCN) en Chile, con datos del ingreso y resultados finales de estudiantes de CED-UCN entre los años 2000 y 2018.

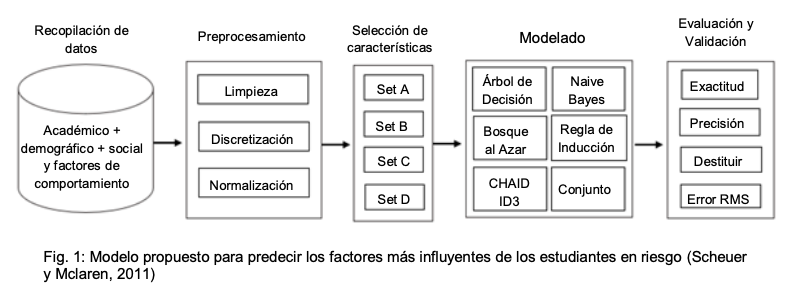
**Objetivo:**

Determinar las variables asociadas al éxito de los estudiantes en programas con modalidad de aprendizaje en línea (e-learning).

**Etapas:**

Metodología de ciencia de datos basada en el modelo CRISP-DM; marco teórico basado en los modelos más influyentes para predecir el éxito de los estudiantes (Figura 1):

**Figura 1.** Modelo propuesto

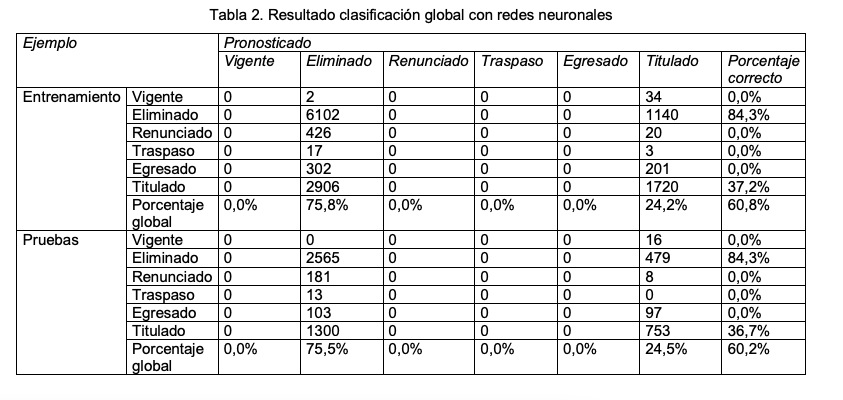


1. Comprensión del Negocio: conocimiento de la institución CED-UCN, ofrece programas 100% on-line.
2. Comprensión de Datos: Obtención e integración de datos de estudiantes matriculados, definieron criterios de exclusión.
3. Preparación de Datos: limpieza de los datos, generación de variables adicionales.
4. Modelamiento: el modelo de clasificación predice “el perfil del estudiante” asociado al éxito en los programas con modalidad de aprendizaje en línea, y cuáles son aquellos atributos que derivan tal perfil. Se utilizaron redes neuronales, arboles de decisión con algoritmo C5.0 y CHAID
5. Evaluación.
6. Implantación.

**Resultados:**

El resultado obtenido con redes neuronales clasifica un 60,8% de predicciones correctas (Figura 2).

**Figura 2.** Resultados de redes neuronales



El modelo de importancia de las variables estableció que el éxito o fracaso depende en gran medida de variables como edad, sexo, profesión, nivel de escolaridad y región.

La comprensión de los factores relevantes en el éxito académico en cada tipo de programa es determinante a la hora de seleccionar a los estudiantes y en los procesos de difusión de los programas.

**Impacto positivo:**

* Organización: permitirán apoyar al know-how respecto al establecimiento de políticas de difusión y mantenimiento de estudiantes en modalidad e-learning. Focalizar los esfuerzos de admisión y retención de estudiantes con mayor riesgo potencial de deserción.
* Estudiantes: enfoque de docentes sobre el riesgo individual, favoreciendo la permanencia en los programas académicos.

**Proyecto 3**[5]

Realizado en México con 61 alumnos de las Licenciaturas en Administración, Comercio, Contaduría, Informática y Mercadotecnia, que cursaron la asignatura Instrumentación estadística para los negocios en una universidad privada de la Ciudad de México durante el ciclo escolar 2018.

GeoGebra es una aplicación web gratuita que permite innovar el proceso educativo relacionado con las matemáticas.

**Objetivo:**

Analizar el impacto de la aplicación GeoGebra en el proceso de enseñanza-aprendizaje sobre la Probabilidad Binomial por medio de la ciencia de datos y el aprendizaje automático (regresión lineal)

**Etapas:**

1. Identificación de participantes,
2. Definición de procedimiento de investigación: Metodologías de aprendizaje, hipótesis de investigación, definición de herramientas de análisis de datos.
3. Recolección de datos: validación de instrumentos de medición.
4. Análisis de datos: análisis cualitativo, con machine learning se realizó el cálculo de regresiones lineales por medio de la segmentación de la muestra. modelos predictivos con regresión lineal (Figura 1), técnica de árbol de decisión para los diversos modelos.

Los modelos predictivos fueron sobre contenidos, estética y simulación de la aplicación GeoGebra y el proceso educativo:

Modelos predictivos sobre los Contenidos de la aplicación GeoGebra y el proceso educativo:

1: Impacto (contenidos) para la asimilación del conocimiento sobre la Probabilidad Binomial.

2: Impacto (contenidos) para el desarrollo de las habilidades matemáticas sobre la Probabilidad Binomial.

3: Impacto (contenidos) para el rol activo del estudiante

Modelos predictivos sobre la Estética de la aplicación GeoGebra y el proceso educativo sobre la Probabilidad Binomial:

4: Impacto (estética) para la asimilación del conocimiento sobre la Probabilidad Binomial

5: Impacto (estética) para el desarrollo de las habilidades matemáticas sobre la Probabilidad Binomial

6: Impacto (estética) para el rol activo del estudiante,

Modelos predictivos sobre la Simulación de la aplicación GeoGebra y el proceso educativo son:

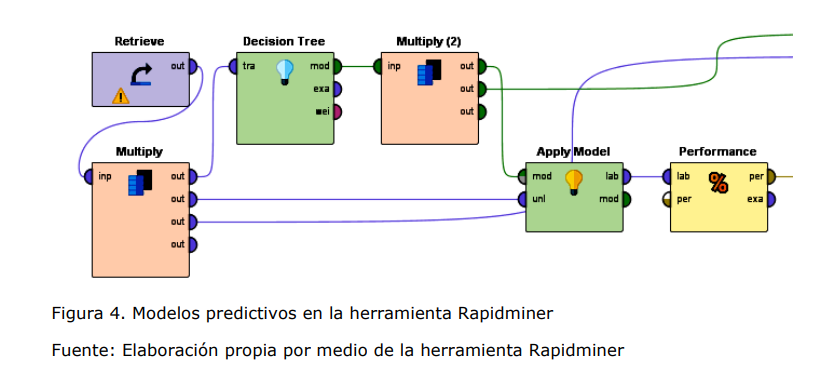
7: Impacto (simulación) para la asimilación del conocimiento sobre la Probabilidad Binomial

8: Impacto (simulación) para el desarrollo de las habilidades matemáticas sobre la Probabilidad Binomial

9: Impacto (simulación) para el rol activo del estudiante.

1. Generación de resultados.

**Figura 1.** Modelos predictivos en el estudio.



**Resultados:**

En el artículo se describen los resultados para cada uno de los 9 modelos predictivos propuestos.

Se encontró que la aplicación de GeoGebra es idónea para el proceso de enseñanza-aprendizaje sobre la Probabilidad Binomial.

Los resultados de la regresión lineal (aprendizaje automático con 60%, 70% y 80% de entrenamiento) indican que los contenidos, la estética y la simulación de esta aplicación influyen positivamente la asimilación del conocimiento, el desarrollo de las habilidades matemáticas y el rol activo de los estudiantes.

**Impacto positivo:**

* Docentes: al identificar las condiciones favorables para el proceso de enseñanza-aprendizaje en temas específicos.
* Estudiantes: Al identificar su satisfacción y motivación con los métodos educativos

**Conclusiones**

La utilización de la minería de datos en el sector educativo es un gran aporte a la sistematización y eficiencia en la identificación de patrones en los datos en la educación, permite identificar variables relacionadas con el éxito de los estudiantes e intervenir en aquellos con mayor riesgo de deserción escolar o identificar el éxito en la aplicación de herramientas educativas en el proceso de enseñanza - aprendizaje, tanto para docentes como para estudiantes.

Como en otros sectores, la ciencia de datos en el sector educativo tiene múltiples aplicaciones y potencialidades de uso.

**Referencias bibliográficas**

1. M. D. Panizzi, "Establecimiento del estado del arte sobre la Minería de Datos Educacional en el Nivel Superior: Un Estudio de Mapeo Sistemático," *Revista de Investigaciones Científicas de la Universidad de Morón*, vol 4, pp. 51-60, 2019
2. J. Hallo, " Sector académico: Julio Xavier Hallo Larrea - Parte 1" Centro Magis [Javeriana Cali]. YouTube, [2022]. Disponible en: <https://www.youtube.com/watch?v=j0omZg-faWI>
3. J. Hallo, " Sector académico: Julio Xavier Hallo Larrea - Parte 2" Centro Magis [Javeriana Cali]. YouTube, [2022]. Disponible en: <https://www.youtube.com/watch?v=43_q36hWGXk>
4. G. Mancilla-Vela, P. Leal-Gatica, A. Sánchez-Ortiz, y C. Vidal-Silva, "Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: un análisis en minería de datos," *Formación Universitaria*, vol. 13, no. 6, pp. 23-36, 2020.
5. Rueda, R. A. S., Rueda, R. D. S. Uso de la ciencia de datos y el aprendizaje automático para analizar la aplicación GeoGebra en el proceso educativo" *Digital Education Review,* 36, pp. 117-151, 2019.