Uso de Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama de Ciudad de Cali.

Por: Andres Racines Bolaños

Agenda

1. Planteamiento del Problema
2. Objetivos Generales y Específicos
3. Estado del Arte
4. Marco Teórico
5. Metodología
6. Experimentos
7. Evaluación
8. Conclusiones
9. Trabajo futuro
10. Referencias

1. Planteamiento del Problema

- Riesgo académico en Colombia
- Encuesta Nacional de Deserción Escolar (ENDE), 2014.
  - 21 factores asociados a la deserción
  - Dos Grupos:, (1) gestión y administración de la educación con 24% de impacto sobre la deserción estudiantil, (2) las dificultades académicas con 43% de impacto sobre la deserción estudiantil.
  - factores socioeconómicos y socioculturales son también causantes de las dificultades académicas.
- La Institución Educativa Libardo Madrid Valderrama requiere de un mecanismo para prevenir el riesgo académico

2. Objetivos Generales y Específicos

❑ Objetivo General

- Desarrollar un prototipo de una aplicación Web que incluya un modelo entrenado de minería de datos que pueda ser usado por la Institución Educativa Libardo Madrid Valderrama, para hacer la predicción del rendimiento académico de los estudiantes para el siguiente año lectivo.

❑ Objetivos Específicos

- Aplicar el descubrimiento de conocimiento con el uso de la metodología CRISP-DM.
- Evaluar varios modelos de clasificación de minería de datos para encontrar el que mejor prediga el rendimiento de los estudiantes.
- Desarrollar un prototipo Web de un modelo entrenado de minería de datos que, dadas las características socioculturales, socioeconómicas y las notas finales de periodos anteriores cursados por los estudiantes, prediga el rendimiento académico para el siguiente año lectivo.

3. Estado del Arte

- **Se revisaron al menos quince documentos académicos de los cuales se destacan:**

  - Universidad de Bosque en Bogotá Colombia, Merchán and Duarte, (2016) [11], se demostró que el algoritmo (J48) logró clasificar correctamente alrededor de 78% de las nuevas instancias

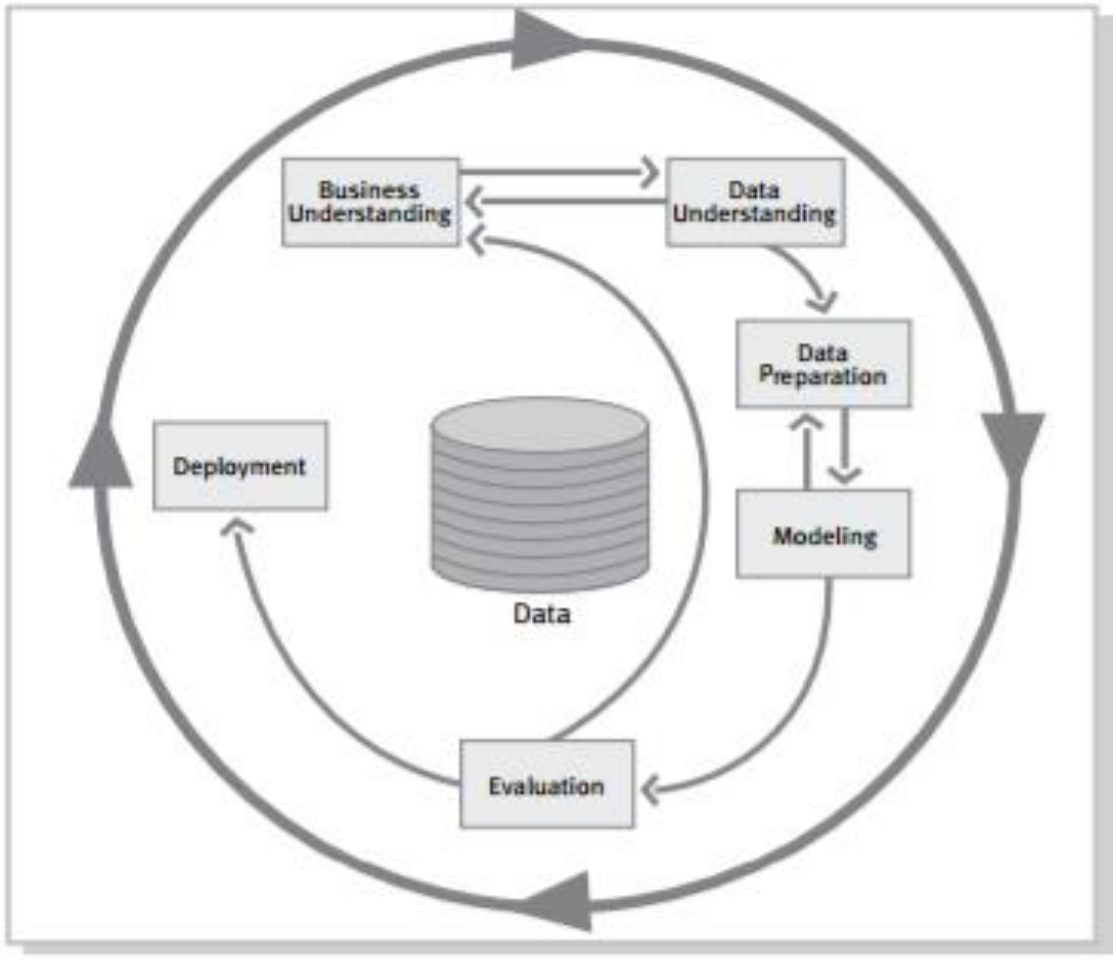
  - Universidad de Minho en Portugal, Cortez and Silva, (2007) [9]. Cuatro técnicas de minería de datos Decision Trees, Random Forest, Neural Networks y Support Vector Machines, logran predecir el 86% de las nuevas instancias para estudiantes matemáticas y leguaje.

  - Universidad Complutense de Madrid, Daniel Gonzales, (2008), detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales, logra predecir 69.3% de las nuevas instancias para estudiantes de ingeniería.

4. Marco Teórico

- La minería de datos (DM).
- Descubrimiento de conocimiento (KDD).
- Metodología CRISP-DM.

Fases de la metodología CRISP-DM.



5. Metodología

En el proceso de minera de datos se aplicó el modelo de referencia CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual consta de seis fases. Los resultados de cada fase se describen a continuación:

❑ Comprensión del negocio

- Crear un modelo de minería de datos, para predecir el rendimiento académico de los estudiantes de Institución Educativa Libardo Madrid Valderrama, que permita a docentes y directivos de la Institución establecer estrategias para mejorar la condición académica de los estudiantes a los cuales la predicción del modelo los clasifique con rendimiento bajo y básico, de esta forma contribuir en la prevención de la deserción académica.

❑ Comprensión de los datos

Los datos para realización del trabajo fueron recopilados en tres etapas:

- Se realizaron tres reuniones informativas con la comunidad educativa Docentes, Acudientes y Estudiantes, con el propósito de presentar el proyecto, los objetivos y acuerdos sobre política de protección de datos de la Institución. (consentimiento informado y asentimiento informado).

- La información socioeconómica y sociocultural de los estudiantes fue recogida por medio de la aplicación de una encuesta web, implementada en un formulario de Google.com. En total se recopilaron 290 registros de estudiantes 9°, 10° y 11° del año lectivo 2017.

- Las notas académicas de los estudiantes fueron suministradas por el rector del colegio, estas notas incluyen los resultados por período académico de la materia de matemáticas de los grados 8°,9° y 10° del año lectivo 2016 y la nota final de matemáticas de los mismos estudiantes en año lectivo 2017

Archivo consolidado formato CSV F4P17.csv , a continuación los atributos del conjunto de datos:

Item	Pregunta	Variable	Valor
1	Tipo de vivienda en la que vive el estudiante	TVE	Arrendada; Familiar;Propia; Otro
2	El estudiante convive con la madre	COVM	Si ; No
3	El estudiante convive con el padre	COVP	Si ; No
4	Educacion de la madre del estudiante	EDUM	primaria; secundaria;profesional;posgrado; no sabe
5	Educacion del padre del estudiante	EDUP	primaria;secundaria;profesional;posgrado; no sabe
6	La madre del estudiante trabaja	MTRAB	Si ; No
7	El padre del estudiante trabaja	PTRAB	Si ; No
8	Personas que conviven con el estudiante	#PCONV	numerico 0 a 99
9	Posicion entre los hermanos que ocupa	PHERMANO	numerico 0 a 99
10	Poblacion Tipo	TPOBLACION	Afrodescendiente; Desplazado; Desplazado por la violenci; indigena; Jovenes vulnerables; Ninguna; Raizal
11	Tiempo que demora en llegar al Colegio	TDC	Menos de 15 Minutos; Entre 15 a 30 Minutos; Entre 30 minutos a 1 hora.
12	Horas de estudio extra clases semanal	#HEEC	Entre 1 a 2 horas; Entre 2 a 3 horas; Entre 3 a 5 horas; Entre 5 a 10 horas; Mas de 10 horas
13	El estudiante Trabaja	ET	Si ; No
14	Tiene computador en su casa	LAPTOP	Si ; No
15	Tiene internet en su casa	INTERNET	Si ; No
16	Practica algun deporte	DEPOR	Si ; No
17	Tiene una relacion sentimental	RSENTI	Si ; No
18	Nota de primer periodo 2016	P1-16	numerico 0 a 5
19	Nota de segundo periodo 2016	P2-16	numerico 0 a 5
20	Nota de tercer periodo 2016	P3-16	numerico 0 a 5
21	Nota de cuarto periodo 2016	P4-16	numerico 0 a 5
22	Nota promedio Final 2017	NF	B; BS; A

❑ Preparación de datos

- Carga y codificación de atributos. Se convirtieron los valores categóricos a numéricos. La clase objetivo se categorizo en Bajo" B"(>=0 & <=2,94) ; Básico "BS" (>2,94 & <=3,44) y Alto "A" (>3,44 &<=5)

COVM	COVP	EDUM	EDUP	MTRAB	PTRAB	#PCONV	PHERMANO	TPOBLACION	...	ET	LAPTOP	INTERNET	DEPOR	RSENTI	P1-16	P2-16	P3-16	P4-16	C
0	0	0	0	0	0	9	0	0	...	0	0	0	0	0	2.9	3.4	3.4	3.6	A
1	0	1	1	1	0	3	3	0	...	0	0	1	1	0	2.6	2.4	2.8	3.9	A
1	0	2	0	0	0	5	3	0	...	0	1	1	0	1	3.1	3.4	3.5	3.9	A
1	0	2	2	0	0	4	1	0	...	0	1	1	0	1	3.0	3.1	3.5	3.7	A
1	1	1	1	1	0	2	4	0	...	0	0	0	0	1	2.7	2.9	2.1	3.1	BS

- Se dividió el conjunto de datos en dos subconjuntos usando la técnica de muestreo por clases. Uno para entrenamiento con 70% de los datos y otro para test con el 30% de los datos

```
Total del conjunto de datos 221  Counter({'BS': 89, 'B': 73, 'A': 59})
-------------------------------------------------------------------------
Conteo de clases conjunto train Counter({'BS': 60, 'B': 54, 'A': 40})
Conteo de Clases conjunto test  Counter({'BS': 29, 'A': 19, 'B': 19})
```

Modelado

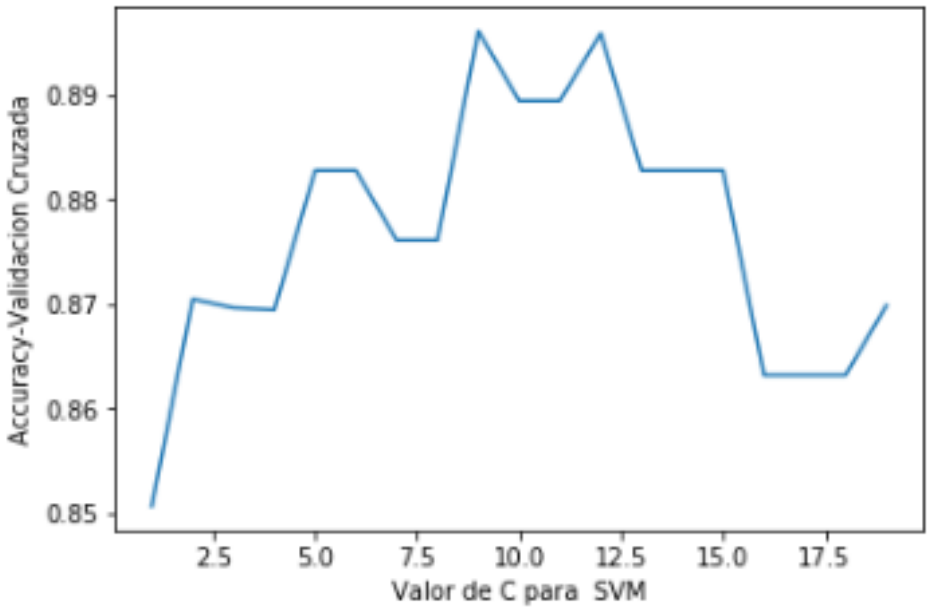
- k- vecinos más cercanos (KNN).
- Máquinas de vectores de soporte (SVM).
- Perceptrón multicapa (MLP)
- Otras consideraciones a la etapa de modelado
  - Selección de atributos
  - Balanceo de clases

- Las técnicas fueron seleccionadas por la capacidad que poseen para resolver problemas que no son linealmente separable, además de lo reducido del número de ejemplos con que se cuentan para entrenar los modelos.

- Cada técnica fue probada con diferentes configuraciones con el objetivo de establecer con cuál configuración se obtendría el mejor rendimiento del modelo. En el ajuste de parámetros se implementó la validación cruzada (K-fold=10).

- Máquinas de vectores de soporte (SVM)

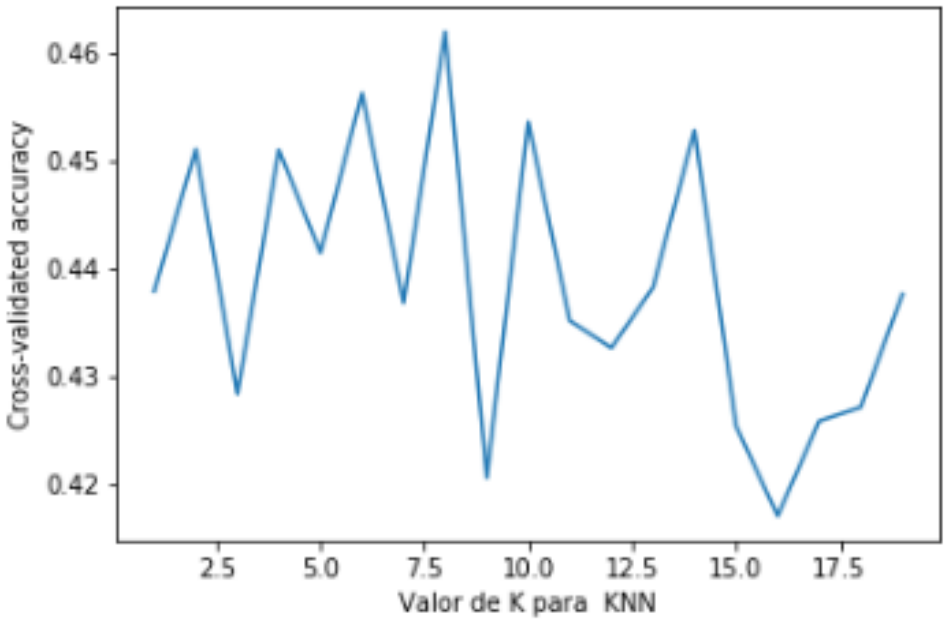
Se realizaron pruebas con diferentes variaciones de los parámetros “C” y gamma. El parámetro gamma fue variado entre 1,10,100, y 1000 obteniendo mejor resultado con gamma =1; c= 8 y kernel lineal



Estimación parámetro C SVM, kernel lineal, gamma = 1

- k- vecinos más cercanos (KNN).

  Se realizaron pruebas con diferentes variaciones de los parámetros “K”, el mejor resultado del modelo se obtiene con k =8



KNN estimación parámetro K

- Perceptrón multicapa (MLP)

El mejor rendimiento de la (MLP) ocurre con 10 capas ocultas y 12 unidades ocultas

MLP estimación de parámetros

- Evaluación de modelos
  - Exactitud (Accuracy). $\frac{TP+TN}{TP+TN+FP+FN}$
  - Sensibilidad (Recall). $\frac{TP}{TP+FN}$
  - Precisión $\frac{TP}{TP+FP}$
  - Destacando la importancia para el proyecto de la sensibilidad en las clases Bajo, (B) Básico, (BS).
- Despliegue

  Se creó un prototipo WEB, que incluye un modelo entrenado de minería de datos para predecir el desempeño académico en matemáticas de los estudiantes para el siguiente año lectivo.

6. Experimentos

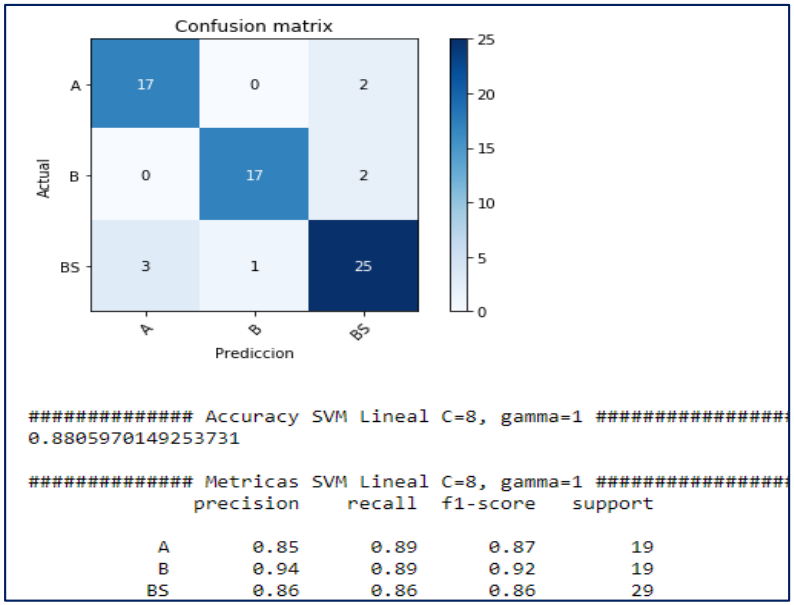
1. Toma todo el conjunto de datos.
2. Selección de atributos en el conjunto de datos, matriz de correlación.
3. Balanceo de clase, con la técnica sobre muestreo (oversampling)
4. Balanceo de clases con submuestreo (undersampling).

Todos los experimentos se hacen con las técnicas SVM, KNN, y MLP, se busca comparar los resultados, para escoger la mejor técnica.

```
Total del conjunto de datos 221  Counter({'BS': 89, 'B': 73, 'A': 59})
-----------------------------------------------------------------------
Conteo de clases conjunto train Counter({'BS': 60, 'B': 54, 'A': 40})
Conteo de Clases conjunto test  Counter({'BS': 29, 'A': 19, 'B': 19})
```

❑ Experimento 1

El entrenamiento de los modelos se realizó con un conjunto de 21 atributos y 154 registros, los modelos se validaron con el conjunto de prueba de 21 atributos y 67 registros.

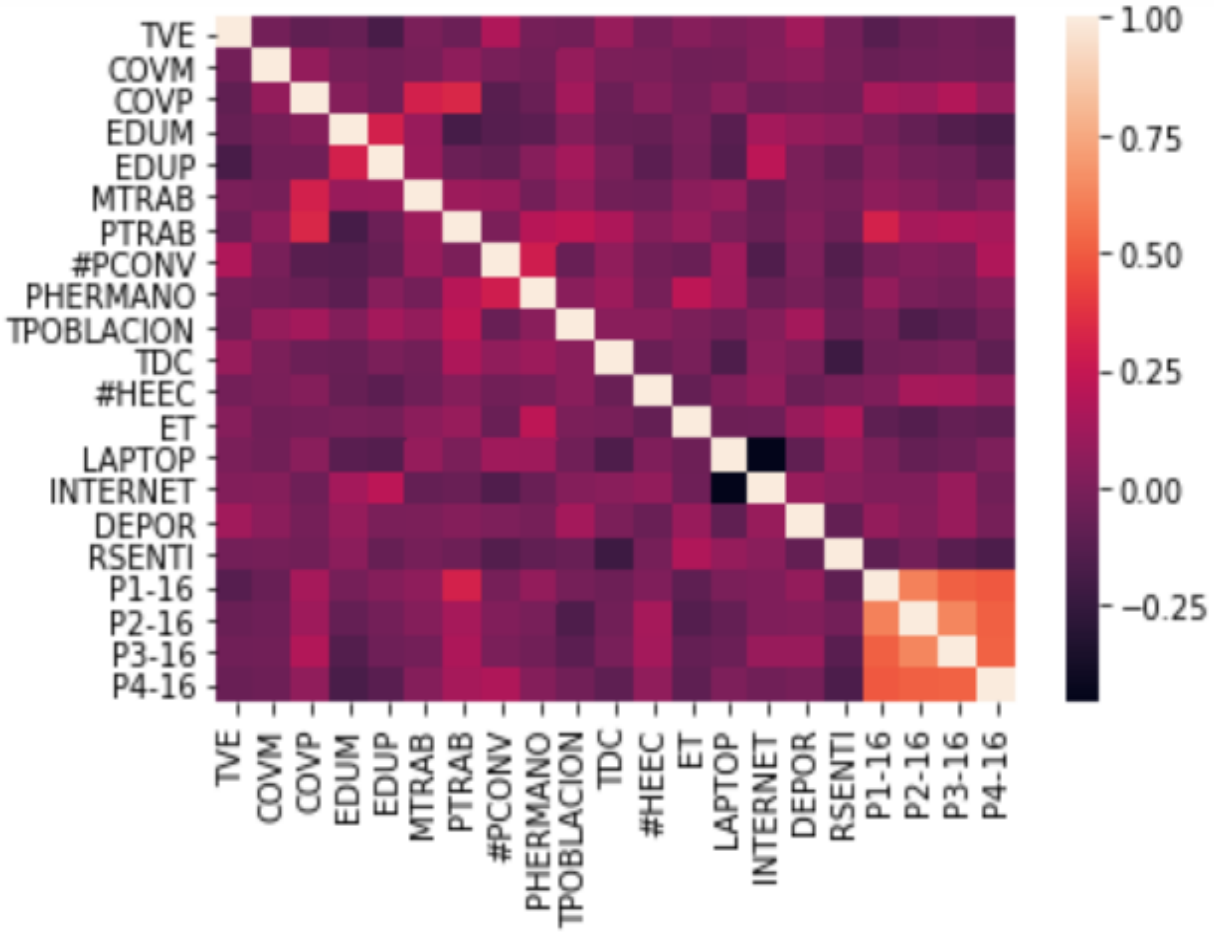



```
############### Accuracy SVM Lineal C=8, gamma=1 ################
0.8805970149253731

############### Metricas SVM Lineal C=8, gamma=1 ################
              precision    recall  f1-score   support

           A       0.85      0.89      0.87        19
           B       0.94      0.89      0.92        19
          BS       0.86      0.86      0.86        29
```

	Sensibilidad (Recall) y Exactitud (accuracy) Experimento 1		
Clase	KNN	SVM	MLP
A	0,47	0,89	0,84
B	0,53	0,89	0,68
BS	0,66	0,86	0,79
Accuracy	0,56	0,88	0,77
	Precision		
A	0,90	0,85	0,84
B	0,48	0,94	0,81
BS	0,53	0,86	0,72

❑ Experimento 2

- Aplicación de técnica de selección de atributos.



a) El estudiantes convive con padre, (COVP) con el atributo: el padre del estudiante trabaja (PTRAB),
b) La educación de la madre (EDUM) con el atributo: la educación del padre (EDUP),
c) Los estudiantes poseen laptop (LAPTOP), con el atributo: los que tienen internet (INTERNET) en la casa.

Se eliminaron los atributos (COVP, EDUM y LAPTOP) .

❑ Experimento 2

- Aplicación de técnica de selección de atributos.

	Sensibilidad (Recall) y Exactitud (accuracy) Experimento 2		
Clase	KNN	SVM	MLP
A	0,68	1,00	0,79
B	0,63	0,95	0,84
BS	0,62	0,83	0,86
Accuracy	0,64	0,91	0,83
	Precision		
A	0,76	0,90	0,94
B	0,57	0,86	0,80
BS	0,62	0,96	0,81

❑ Experimento 3

- Técnica de sobre muestreo (oversampling).

```
Total del conjunto de datos 221  Counter({'BS': 89, 'B': 73, 'A': 59})
-----------------------------------------------------------------------
Conteo de clases conjunto train Counter({'BS': 60, 'B': 54, 'A': 40})
Conteo de Clases conjunto test  Counter({'BS': 29, 'A': 19, 'B': 19})

Total del conjunto de datos 267  Counter({'BS': 89, 'B': 89, 'A': 89})
-----------------------------------------------------------------------
Conteo de clases conjunto train Counter({'A': 63, 'B': 62, 'BS': 61})
Conteo de Clases conjunto test  Counter({'BS': 28, 'B': 27, 'A': 26})
```

	Sensibilidad (Recall) y Exactitud (accuracy) Experimento 3		
Clase	KNN	SVM	MLP
A	0,88	1,00	0,96
B	0,74	0,93	0,85
BS	0,39	0,89	0,89
Accuracy	0,66	0,93	0,90
	Precision		
A	0,82	0,90	1,00
B	0,62	1,00	0,88
BS	0,52	0,93	0,83

❑ Experimento 4

- Técnica de submuestreo (undersampling)

```
Total del conjunto de datos 221  Counter({'BS': 89, 'B': 73, 'A': 59})
-----------------------------------------------------------------------
Conteo de clases conjunto train Counter({'BS': 60, 'B': 54, 'A': 40})
Conteo de Clases conjunto test  Counter({'BS': 29, 'A': 19, 'B': 19})

Total del conjunto de datos 177  Counter({'BS': 59, 'B': 59, 'A': 59})
-----------------------------------------------------------------------
Conteo de clases conjunto train Counter({'A': 48, 'B': 39, 'BS': 36})
Conteo de Clases conjunto test  Counter({'BS': 23, 'B': 20, 'A': 11})
```

	Sensibilidad (Recall) y Exactitud (accuracy) Experimento 4		
Clase	KNN	SVM	MLP
A	0,82	0,91	0,64
B	0,35	0,75	0,70
BS	0,09	0,74	0,65
Accuracy	0,33	0,77	0,66
	Precision		
A	0,30	0,71	0,58
B	0,50	0,88	0,82
BS	0,20	0,74	0,60

7. Evaluación

	Sensibilidad (Recall) y Exactitud (accuracy)			
	KNN			
Clase	Experimento 1	Experimento 2	Experimento 3	Experimento 4
A	0,47	0,68	0,88	0,82
B	0,53	0,63	0,74	0,35
BS	0,66	0,62	0,39	0,09
Accuracy	0,56	0,64	0,66	0,33
	SVM			
A	0,89	1,00	1,00	0,91
B	0,89	0,95	0,93	0,75
BS	0,86	0,83	0,89	0,74
Accuracy	0,88	0,91	0,93	0,77
	MLP			
A	0,84	0,79	0,96	0,64
B	0,68	0,84	0,85	0,70
BS	0,79	0,86	0,89	0,65
Accuracy	0,77	0,83	0,90	0,66

8. Conclusiones

- los modelos basados en distancia como KNN, estuvieron aproximadamente 32%, por debajo de rendimiento de los modelos generados con las SVM y MLP, en promedio en los cuatro experimentos.
- Los modelos generados con las SVM estuvieron aproximadamente 10%, por encima del rendimiento de los modelos generados con MLP, en promedio en los cuatro experimentos
- En el experimento 3, se obtuvieron los mejores resultados incluso la diferencia entre el modelo generado con SVM solo estuvo un 4% por encima de los resultados obtenidos por el MLP.
- Se logro desarrollar el prototipo de una aplicación Web que incluye un modelo entrenado de SVM (experimento 3), con una sensibilidad del 94% aproximadamente. Este puede ser usado por la Institución Educativa Libardo Madrid Valderrama, para hacer la predicción del rendimiento académico en matemáticas de los estudiantes para el siguiente año lectivo
- Los colegios públicos en Colombia no tienen los sistemas de información que faciliten automatizar el proceso, para continuar haciendo aprendizaje para otras áreas.

9. Trabajo futuro

- Crear modelos para otras áreas del conocimiento, y poder, predecir la clasificación de los estudiantes en diferentes materias simultáneamente.

10. Referencias

- [1] S. M. Merchán and J. A. Duarte, “Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance,” IEEE Lat. Am. Trans., vol. 14, no. 6, pp. 2783–2788, 2016.
- [2] P. Sectorial, “Encuentro Regional 2011.” [Online]. Available: http://www.mineducacion.gov.co/cvn/1665/articles-279754_archivo_pdf_ministra.pdf. [Accessed: 22-May-2017].
- [3] Tan, M. & Shao, P. (2015). Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. International Journal of Emerging Technologies in Learning (iJET), 10(1), 11-17. Kassel, Germany: International Association of Online Engineering
- [4] “Zeti.” [Online]. Available: https://zeti.net.co/. [Accessed: 16-May-2017].
- [5] J. S. Henao Parra, “Las redes neuronales y su desempeño bajo la estrategia de Neuroevolución,” web, 2013. [Online]. Available: https://repository.javeriana.edu.co/handle/10554/12100?show=full. [Accessed: 22-May-2017].
- [6] D. Santín González, “Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales / Daniel Santín González,” Madrid : Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, 1999, 2008. [Online].Available: http://cisne.sim.ucm.es/search~S6*spi/?searchtype=h&searcharg=W28+%289902%29&searchscope=6&sortdropdown=-

- [8] N. Pukkhem, “A semantic-based approach for representing successful graduate predictive rules,” 16th Int. Conf. Advanced Communications Technology., pp. 222–227, 2014.
- [9] Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In Evolution of teaching and learning paradigms in intelligent environment (pp. 183-221). Springer Berlin Heidelberg.
- [10] “Manual CRISP-DM de IBM SPSS Modeler.” [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf. [Accessed: 17-May-2017].
- [11] R. Jindal and M. D. Borah, “A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS" Rajni vol. 5, no. 3, pp. 53–73, 2013.
- [12] Baker, R.S.J.D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A review and future visions. Journal of educational Data Mining, 1, 3- 17.
- [13] P. G. Espejo, S. Ventura and F. Herrera, "A Survey on the Application of Genetic Programming to Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 2, pp. 121-144, March 2010.
- [14] “Análisis comparativo de algoritmos de aprendizaje para predecir la evolución de pacientes con Daño Cerebral Adquirido.” [Online]. Available: http://oa.
- [15] “KDD Proceso de Extracción de conocimiento WebMining.” [Online]. Available:http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/. [Accessed: 31-May-2017].
- [16] K. Tan, Steinbach, “Data Mining Classification: Basic Concepts, Decision Trees, and Model Evaluation Lecture Notes for Chapter 4 Introduction to DataMining.”[Online]Avaliable:http://www.users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf . [Accessed: 31-May-2017].

Fin.

En general los modelos obtenidos con las SVM obtienen el mejor desempeño en todos los experimentos, se incluye una tabla para comparar la diferencia entre el recall, de los modelos SVM en los diferentes experimentos vs el recall de KNN y MLP.

	Diferencia SVM vs KNN			
Clase	Experimento 1	Experimento 2	Experimento 3	Experimento 4
A	0,42	0,32	0,12	0,09
B	0,36	0,32	0,19	0,40
BS	0,20	0,21	0,50	0,65
	Diferencia SVM vs MLP			
A	0,05	0,21	0,04	0,27
B	0,21	0,11	0,08	0,05
BS	0,07	-0,03	0,00	0,09