Evaluación de Técnicas de Minería de Datos para la Predicción del Rendimiento Académico

Leticia Laura Ochoa, Mg[1], Karina Rosas Paredes, Mg[2], César Baluarte Araya, Dr[1]
[1]Universidad Nacional de San Agustín, Perú, letymarisol@hotmail.com, cbaluarte@unsa.edu.pe
[2]Universidad Católica de Santa María, Perú, kparedes@ucsm.edu.pe

***Resumen*– En este trabajo se realiza la evaluación comparativa de técnicas de aprendizaje supervisado de minería de datos que permiten la clasificación de más de dos clases o categorías, como son el árbol de decisión, random forest, redes bayesianas, K vecinos más cercanos (KNN) y máquinas de soporte vectorial (SVM) para la predicción del rendimiento académico de los alumnos dentro de las categorías bajo, medio y alto, utilizando el lenguaje R. Para seleccionar el mejor algoritmo de clasificación, se utilizó técnicas de comparación de precisión a partir de las matrices de confusión obtenidas por los diferentes modelos de minería de datos aplicados así como técnicas de validación cruzada, también se comparó los tiempos de ejecución que demoraron en la construcción de los modelos, resultando la técnica de SVM una de las más eficientes en tiempo de ejecución y la más precisa en las pruebas de predicción del rendimiento académico.**

***Palabras claves*– Predicción del rendimiento académico, minería de datos educacionales, técnicas de aprendizaje supervisado, clasificación, deserción estudiantil.**

I. Introducción

La tendencia de la educación superior es garantizar e incrementar la calidad, aumentar la tasa de graduación, reducir el abandono y la deserción [1]. La deserción estudiantil es uno de los problemas que aborda la mayoría de las instituciones de educación superior de Latinoamérica [2][3].

En estudios efectuados acerca de la deserción universitaria, se considera que, en promedio, al menos la mitad de los alumnos que ingresan a la educación postsecundaria abandonan sus estudios antes de lograr el título profesional o grado académico. La mayor proporción de esta cifra corresponde a la deserción que se produce durante el primer año [4]. Esto requiere gestionar estrategias y tomar medidas frente a estos acontecimientos; para ello es posible recurrir al proceso denominado Minería de Datos Educacional (MDE), es decir, la aplicación del proceso de Descubrimiento o Extracción de Conocimiento en Bases de Datos (KDD) en ámbito educativo [5]. En el trabajo de [6] se obtuvo un patrón general de deserción estudiantil determinado por un promedio de calificaciones bajo y el tener materias perdidas en los primeros semestres de la carrera.

Por otra parte, el rendimiento académico ha sido representado de diferentes maneras según los estudios que han abordado el tema. En algunos de ellos, el rendimiento académico es representado por el número de materias aprobadas por un alumno en una carrera, en otros por el resultado de tests específicamente diseñados, así como también, por el promedio de notas de las materias cursadas. Esta variedad de manifestaciones del rendimiento académico está ligada a las particularidades de la investigación en cuestión, referidas al nivel de estudios en el cual se analiza el desempeño de los alumnos, el tiempo de la investigación o el enfoque del investigador [7]. Según [8] las asignaturas son la fuente primaria del rendimiento del estudiante; de las calificaciones reportadas se obtienen los promedios generales que constituyen la fuente esencial de evaluación.

Knowledge Discovery in Databases (KDD), Descubrimiento de conocimiento en base de datos, es el proceso de identificar patrones válidos, nuevos, útiles y comprensibles de grandes volúmenes de datos. La minería de datos es el núcleo matemático del proceso KDD, que comprende los algoritmos que exploran los datos, desarrollan modelos matemáticos y descubren patrones significativos (implícita o explícita), los cuales son la esencia del conocimiento útil [9].

La minería de datos educacional (MDE) es la aplicación de la minería de datos en el ámbito de la educación, para comprender mejor el proceso de aprendizaje de los estudiantes y de su participación global en el proceso, con el objetivo de mejorar la calidad y rentabilidad del sistema educativo [5][10]. Es una disciplina emergente, preocupada por el desarrollo de métodos para explorar los tipos únicos de datos que vienen de los entornos educativos, y con el uso de esos métodos poder comprender mejor a los estudiantes y las características en las que aprenden [11]. La MDE es un área multidisciplinaria en la cual convergen distintos paradigmas de computación como son el desarrollo o construcción de algoritmos de predicción, programación lógica, algoritmos estadísticos, entre otros, con el objetivo de generar principales tareas como; la clasificación, agrupamiento (clustering), estimación, modelado de dependencias, visualización y descubrimiento de reglas con el fin de construir un modelo ajustado a un conjunto de datos sobre un contexto educativo, teniendo como fin último el proporcionar un conocimiento certero del sistema y predecir comportamientos futuros [12].

Según el objetivo del análisis de los datos, los algoritmos utilizados por las técnicas de minería de datos se clasifican en supervisados y no supervisados. Las técnicas de aprendizaje supervisado permiten predecir un dato desconocido a priori a partir de otros datos conocidos y las no supervisadas descubren patrones o tendencias de los datos [13]. En el

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2017.1.1.368
ISBN: 978-0-9993443-0-9
ISSN: 2414-6390

aprendizaje supervisado los datos de entrenamiento tienen etiquetas que indican la clase de las observaciones y los nuevos datos se clasifican según el conjunto de datos de entrenamiento [14].

La clasificación es una de las técnicas más utilizadas en minería de datos, descubrimiento de conocimiento, inteligencia artificial, reconocimiento de patrones y aprendizaje de máquina [15]. Es el proceso de encontrar un conjunto de modelos (o funciones), los cuales describen y distinguen las clases definidas de los datos, con la finalidad de predecir clases de objetos cuyas clasificaciones no se han definido [16]. Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas [17]. Los algoritmos de clasificación corresponden a las técnicas de aprendizaje supervisado o predictivas de minería de datos.

En este trabajo se realiza una evaluación de técnicas de aprendizaje supervisado de minería de datos para la predicción del rendimiento académico utilizando indicadores como la precisión y eficiencia para medir la performance del algoritmo de aprendizaje y utilizando técnicas de evaluación de modelos predictivos como la matriz de confusión y técnicas de validación cruzada, que permita seleccionar el algoritmo de clasificación con la que se obtiene mejor precisión, que pueda utilizarse en las instituciones educativas para clasificar a los estudiantes a partir de datos académicos y establecer estrategias para la mejora de aprendizaje.

II. Metodología

El desarrollo del presente trabajo consta de 5 pasos:

1) Recolección de datos académicos
2) Selección de registros y atributos
3) Transformación y limpieza de datos
4) Selección y aplicación de técnicas y algoritmos de minería de datos
5) Evaluación de los algoritmos de minería de datos

III. Resultados y Discusión

En este trabajo se realizó la evaluación de técnicas de aprendizaje supervisado de minería de datos para la predicción del rendimiento académico utilizando datos de los alumnos del primer año de la Escuela profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María.

A los datos académicos iniciales se les agregó un campo adicional llamado RENDIMIENTO luego de aplicar la técnica supervisada de minería de datos K-means para la agrupación de alumnos, este nuevo campo se utilizará como variable dependiente o variable objetivo para las pruebas de predicción del rendimiento académico.

Luego de la recolección, selección, transformación y limpieza de los datos académicos, se utilizaron las variables que se muestran en la Tabla 1 para las pruebas de predicción.

TABLA 1
Variables Usadas para la Predicción

VARIABLE	DESCRIPCIÓN
CODIGO	Código del Alumno
CALC_DIFE	Cálculo Diferencial
DESA_HUMA	Desarrollo Humano
COMU_ORAL_ESCR	Comunicación Oral y Escrita
FUND_PROG	Fundamentos de Programación
PROG_I	Programación I
ESTR_DISC_I	Estructuras Discretas I
RENDIMIENTO	bajo, medio, alto

A. Selección y Aplicación de Técnicas y Algoritmos de Minería de Datos

La variable RENDIMIENTO a predecir consta de tres valores posibles que son: bajo, medio y alto. La clasificación típica utiliza dos clases como SI/NO, 1/2; por lo cual para poder realizar la predicción del rendimiento académico se requiere utilizar técnicas predictivas que permitan clasificar en más de dos categorías o clases.

Para la construcción de los modelos se dividieron los registros académicos en dos muestras al azar, 80% para la tabla de entrenamiento y 20% para la tabla de prueba

Los modelos de predicción se han construido sobre los registros de entrenamiento y se evaluaron en los registros de prueba para hallar la precisión de los modelos clasificando a los alumnos que no se consideraron en la construcción del modelo.

La construcción del modelo de las técnicas de minería de datos utilizadas para la predicción del rendimiento académico se realizó utilizando el lenguaje R, el cual es el software más utilizado para minería de datos según encuestas realizadas en los últimos años por KDnuggets [18][19].

Las técnicas de clasificación utilizadas en la evaluación comparativa para la predicción del rendimiento son:

1) Árbol de decisión: Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos [17]. Los árboles de decisión recorren las ramas desde la raíz hasta las hojas para lograr una predicción utilizando reglas de decisión. Los algoritmos comunes son: ID3, C4.5 y CART.

Se generó el modelo a partir de la tabla de entrenamiento utilizando el algoritmo CART, luego se realizó la predicción del rendimiento académico con los datos de prueba. En la Fig. 1 se muestra una comparación de los datos reales versus la predicción realizada por el modelo del árbol de decisión,

donde se puede observar que hubo tres errores en la predicción para los códigos de alumnos 17, 52 y 64.

```
> data.frame(prueba$RENDIMIENTO,prediccionArbol)
   prueba.RENDIMIENTO prediccionArbol
4                alto            alto
16               bajo            bajo
17               alto           medio
19               alto            alto
22              medio           medio
30               alto            alto
52               bajo           medio
62               bajo            bajo
63              medio           medio
64               alto           medio
68              medio           medio
69              medio           medio
```

Fig. 1 Rendimiento actual versus predicción del árbol de decisión.

En la Tabla 2 se muestra la matriz de confusión obtenido por el modelo del árbol de decisión, donde se puede apreciar que en los registros de prueba (Valor Actual) existen cinco alumnos de rendimiento alto, tres alumnos de rendimiento bajo y cuatro alumnos de rendimiento medio, mientras que con el modelo del árbol de decisión se obtuvo una predicción de tres alumnos de rendimiento alto, dos alumnos de rendimiento bajo y siete alumnos de rendimiento medio.

TABLA 2
MATRIZ DE CONFUSIÓN DEL ÁRBOL DE DECISIÓN

		Predicción Árbol		
		alto	**bajo**	**medio**
Valor Actual	**alto**	3	0	2
	bajo	0	2	1
	medio	0	0	4

En la Fig. 2 se muestra el cálculo de la precisión y error de la predicción obtenida por el modelo del árbol de decisión, con el que se obtuvo 75% de precisión y 25% de error.

```
> precision <- ((sum(diag(matriz))/length(prueba$RENDIMIENTO))*100)
> precision
[1] 75
> error <- ((1-(sum(diag(matriz))/length(prueba$RENDIMIENTO)))*100)
> error
[1] 25
```

Fig. 2 Precisión y error de la predicción del árbol de decisión.

2) Random Forest: Random Forest, introducido por L. Breiman en 1999, utiliza un conjunto (o bosque) formado por muchos árboles de clasificación (decision trees). Para clasificar un nuevo objeto, cada árbol en el ensamble lo toma como entrada y produce una salida, su clasificación. La decisión del ensamble se toma como la clase con mayoría de votos en el ensamble [20].

Random Forest llamado también "Bosques Aleatorios" o "Selvas Aleatorias" es una técnica predictiva en la cual todos los clasificadores del método de consenso (Bagging) son árboles de decisión. Cada modelo genera una predicción y se selecciona por la mayor cantidad de votos.

Se generó el modelo a partir de la tabla de entrenamiento, luego se realizó la predicción del rendimiento académico con los datos de prueba. En la Fig. 3 se muestra una comparación de los datos reales versus la predicción realizada por el modelo de Random Forest, donde se puede observar que hubo un error en la predicción para el código de alumno 17.

```
> data.frame(prueba$RENDIMIENTO,prediccionRF)
   prueba.RENDIMIENTO prediccionRF
2                bajo         bajo
10               alto         alto
12               bajo         bajo
17               alto        medio
18              medio        medio
22              medio        medio
25              medio        medio
35               bajo         bajo
37              medio        medio
39               alto         alto
49              medio        medio
50               bajo         bajo
60              medio        medio
61              medio        medio
67              medio        medio
```

Fig. 3 Rendimiento actual versus predicción de random forest.

En la Tabla 3 se muestra la matriz de confusión obtenido por el modelo de random forest, donde se puede apreciar que en los registros de prueba (Valor Actual) existen tres alumnos de rendimiento alto, cuatro alumnos de rendimiento bajo y ocho alumnos de rendimiento medio, mientras que con el modelo de random forest se obtuvo una predicción de dos alumnos de rendimiento alto, cuatro alumnos de rendimiento bajo y nueve alumnos de rendimiento medio.

TABLA 3
MATRIZ DE CONFUSIÓN DE RANDOM FOREST

		Predicción Random Forest		
		alto	**bajo**	**medio**
Valor Actual	**alto**	2	0	1
	bajo	0	4	0
	medio	0	0	8

En la Fig. 4 se muestra el cálculo de la precisión y error de la predicción obtenida por el modelo de random forest, con el que se obtuvo 93% de precisión y 7% de error.

```
> precision <- ((sum(diag(matriz))/length(prueba$RENDIMIENTO))*100)
> precision
[1] 93.33333
> error <- ((1-(sum(diag(matriz))/length(prueba$RENDIMIENTO)))*100)
> error
[1] 6.666667
```

Fig. 4 Precisión y error de la predicción de random forest.

3) Redes Bayesianas: Las redes bayesianas modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Estos modelos pueden tener diversas aplicaciones, para clasificación, predicción, diagnóstico [21].

Las redes bayesianas son una técnica de aprendizaje supervisado de minería de datos que permite la predicción de variables desconocidas cuyos resultados se pueden expresar en términos de probabilidad. Utiliza el algoritmo Naive Bayes disponible en muchas herramientas de minería de datos.

Se generó el modelo a partir de la tabla de entrenamiento utilizando el algoritmo Naive Bayes, luego se realizó la predicción del rendimiento académico con los datos de prueba. En la Fig. 5 se muestra una comparación de los datos reales versus la predicción realizada por el modelo de redes bayesianas, donde se puede observar que hubo un error en la predicción para el código de alumno 13.

```
> data.frame(prueba$RENDIMIENTO,prediccionBayes)
   prueba.RENDIMIENTO prediccionBayes
1                alto            alto
2                bajo            bajo
3                alto            alto
4                bajo            bajo
5               medio           medio
6               medio           medio
7                bajo            bajo
8                bajo            bajo
9                bajo            bajo
10               alto            alto
11               bajo            bajo
12              medio           medio
13               alto           medio
14              medio           medio
15              medio           medio
```

Fig. 5 Rendimiento actual versus predicción de redes bayesianas.

En la Tabla 4 se muestra la matriz de confusión obtenido por el modelo de redes bayesianas, donde se puede apreciar que en los registros de prueba (Valor Actual) existen cuatro alumnos de rendimiento alto, seis alumnos de rendimiento bajo y cinco alumnos de rendimiento medio, mientras que con el modelo de redes bayesianas se obtuvo una predicción de tres alumnos de rendimiento alto, seis alumnos de rendimiento bajo y seis alumnos de rendimiento medio.

TABLA 4
MATRIZ DE CONFUSIÓN DE REDES BAYESIANAS

		Predicción Redes Bayesianas		
		alto	**bajo**	**medio**
Valor Actual	**alto**	3	0	1
	bajo	0	6	0
	medio	0	0	5

En la Fig. 6 se muestra el cálculo de la precisión y error de la predicción obtenida por el modelo de redes bayesianas, con el que se obtuvo 93% de precisión y 7% de error.

```
> precision <- ((sum(diag(matriz))/length(prueba$RENDIMIENTO))*100)
> precision
[1] 93.33333
> error <- ((1-(sum(diag(matriz))/length(prueba$RENDIMIENTO)))*100)
> error
[1] 6.666667
```

Fig. 6 Precisión y error de la predicción de redes bayesianas.

4) K Nearest Neighbors (KNN): El algoritmo KNN es uno de los algoritmos más simples y consiste en asignar un objeto a la clase más común entre sus K vecinos más cercanos, siendo K un número entero positivo. Los vecinos se obtienen de un conjunto de objetos (denominados datos de entrenamiento), para los cuales el modelo de clasificación correcto es conocido. Para identificar a los vecinos, los objetos son representados por vectores de posición en un espacio de características multidimensional [22].

K Nearest Neighbors (KNN), también conocido en español como K vecinos más cercanos, es una técnica de aprendizaje supervisado que permite clasificar un registro de prueba en base a su semejanza a los K vecinos más cercanos.

Se generó el modelo a partir de la tabla de entrenamiento utilizando el algoritmo KNN, luego se realizó la predicción del rendimiento académico con los datos de prueba. En la Fig. 7 se muestra una comparación de los datos reales versus la predicción realizada por el modelo de KNN, donde se puede observar que hubo un error en la predicción para el código de alumno 6.

```
> data.frame(prueba$RENDIMIENTO,prediccionKknn)
   prueba.RENDIMIENTO prediccionKknn
1               medio          medio
2               medio          medio
3               medio          medio
4               medio          medio
5                bajo           bajo
6                alto          medio
7                alto           alto
8               medio          medio
9                alto           alto
10               alto           alto
11               alto           alto
12               alto           alto
13               alto           alto
14               bajo           bajo
15              medio          medio
16              medio          medio
17               bajo           bajo
18              medio          medio
19              medio          medio
```

Fig. 7 Rendimiento actual versus predicción de KNN.

En la Tabla 5 se muestra la matriz de confusión obtenido por el modelo de KNN, donde se puede apreciar que en los registros de prueba (Valor Actual) existen siete alumnos de

rendimiento alto, tres alumnos de rendimiento bajo y nueve alumnos de rendimiento medio, mientras que con el modelo de KNN se obtuvo una predicción de seis alumnos de rendimiento alto, tres alumnos de rendimiento bajo y diez alumnos de rendimiento medio.

TABLA 5
MATRIZ DE CONFUSIÓN DE KNN

		Predicción KNN		
		alto	**bajo**	**medio**
Valor Actual	**alto**	6	0	1
	bajo	0	3	0
	medio	0	0	9

En la Fig. 8 se muestra el cálculo de la precisión y error de la predicción obtenida por el modelo de KNN, con el que se obtuvo 95% de precisión y 5% de error.

```
> precision <- ((sum(diag(matriz))/length(prueba$RENDIMIENTO))*100)
> precision
[1] 94.73684
> error <- ((1-(sum(diag(matriz))/length(prueba$RENDIMIENTO)))*100)
> error
[1] 5.263158
```

Fig. 8 Precisión y error de la predicción de KNN.

5) Support Vector Machine (SVM): Conocido también como máquinas de soporte vectorial, es una técnica de clasificación supervisada que busca hallar hiperplanos que separen las clases para la predicción, buscando maximizar el margen se separación entre las clases.

Principalmente las SVM se utilizan como clasificador binario, pero también se pueden utilizar como clasificador multiclases, utilizando los métodos "uno contra todos" ó "uno contra uno", donde el primero consiste en comparar cada clase con todas las demás, mientras que en el segundo cada clase se compara con las restantes individualmente. Es importante destacar que el objetivo fundamental de las SVM para clasificación es encontrar un hiperplano óptimo que separe las clases [23].

SVM es un sistema para entrenar máquinas de aprendizaje lineal de manera eficiente. Tanto para clasificación como para regresión se han encontrado muchas aplicaciones de SVM, como por ejemplo en clasificación de imágenes, en reconocimiento de caracteres, en detección de proteínas, en clasificación de patrones, en identificación de funciones, etc. [22].

Se generó el modelo a partir de la tabla de entrenamiento utilizando el algoritmo SVM, luego se realizó la predicción del rendimiento académico con los datos de prueba. En la Fig. 9 se muestra una comparación de los datos reales versus la predicción realizada por el modelo de SVM, donde se puede observar que no hubo errores en la predicción.

```
> data.frame(prueba$RENDIMIENTO,prediccionSVM)
   prueba.RENDIMIENTO prediccionSVM
2                bajo          bajo
3               medio         medio
6               medio         medio
14               alto          alto
20               alto          alto
22              medio         medio
36              medio         medio
37              medio         medio
47              medio         medio
49              medio         medio
50               bajo          bajo
51               alto          alto
54               bajo          bajo
56              medio         medio
57               bajo          bajo
```

Fig. 9 Rendimiento actual versus predicción de SVM.

En la Tabla 6 se muestra la matriz de confusión obtenido por el modelo de SVM, donde se puede apreciar que en los registros de prueba (Valor Actual) existen tres alumnos de rendimiento alto, cuatro alumnos de rendimiento bajo y ocho alumnos de rendimiento medio, de igual forma con el modelo SVM se obtuvieron los mismos valores en la predicción.

TABLA 6
MATRIZ DE CONFUSIÓN DE SVM

		Predicción SVM		
		alto	**bajo**	**medio**
Valor Actual	**alto**	3	0	0
	bajo	0	4	0
	medio	0	0	8

En la Fig. 10 se muestra el cálculo de la precisión y error de la predicción obtenida por el modelo de SVM, con el que se obtuvo 100% de precisión y 0% de error.

```
> precision <- ((sum(diag(matriz))/length(prueba$RENDIMIENTO))*100)
> precision
[1] 100
> error <- ((1-(sum(diag(matriz))/length(prueba$RENDIMIENTO)))*100)
> error
[1] 0
```

Fig. 10 Precisión y error de la predicción de SVM.

B. Evaluación de los Algoritmos de Minería de Datos

La evaluación de las técnicas de aprendizaje supervisado de minería de datos aplicadas para la predicción del rendimiento académico se realizó teniendo en cuenta:

- La precisión obtenida por los modelos utilizados.
- La eficiencia de los modelos.

Para evaluar la precisión de los modelos predictivos se utilizó las matrices de confusión y la técnica de validación cruzada dejando uno afuera, llamado también Leave-one-out cross-validation (LOOCV), que ayudó a elegir el modelo más preciso para la predicción del rendimiento académico.

1) Comparación de tiempo de ejecución de la construcción de los modelos: En la Tabla 7 se muestra el tiempo de ejecución tomado por las diferentes técnicas de aprendizaje supervisado de minería de datos que se utilizaron en la construcción de los modelos para la predicción del rendimiento académico. Para hallar el tiempo de ejecución se utilizó la función *proc.time* utilizando el lenguaje R.

TABLA 7
COMPARACIÓN DE TIEMPO DE EJECUCIÓN

Técnica	User time	System time	Elapsed time
Árbol de Decisión	0.02	0.00	0.02
Random Forest	0.03	0.02	0.05
Redes Bayesianas	0.02	0.00	0.02
KNN	0.03	0.00	0.03
SVM	0.01	0.00	0.01

Donde:

- User time, es el tiempo que demora la CPU en la ejecución de las instrucciones para la generación del modelo.
- System time, es el tiempo tomado por el sistema operativo.
- Elapsed time, es el tiempo total desde que se inició el proceso.

Se puede observar que la técnica predictiva que demoró más en la construcción del modelo fue Random Forest con 0.05 segundos, seguido de la técnica de KNN con 0.03 segundos. Así como también se encontró que las técnicas más eficientes en tiempo de ejecución de la construcción del modelo para la predicción fue la máquina de soporte vectorial (SVM) con 0.01 segundos y las técnicas de redes bayesianas y árbol de decisión con 0.02 segundos. Estos tiempos pueden variar dependiendo de las muestras generadas al azar para la tabla de entrenamiento (80%) y tabla de pruebas (20%). En varias pruebas realizadas se obtuvieron resultados similares, resultando las técnicas de redes bayesianas y SVM las más eficientes.

2) Comparación de la precisión de los algoritmos: La precisión global que se obtiene a partir de la matriz de confusión es la métrica de evaluación de rendimiento más ampliamente utilizada [24], la fórmula que se utiliza para hallar la precisión global del modelo de clasificación es:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Donde:

- a = TP (True Positive), que representa la cantidad de registros positivos que fueron clasificados de forma correcta.
- b = FN (False Negative), que representa la cantidad de registros que se han clasificado como negativos de forma incorrecta.
- c = FP (False Positive), que representa la cantidad de registros negativos que se han clasificado como positivos de forma incorrecta.
- d = TN (True Negative), que representa la cantidad de registros negativos que fueron clasificados de forma correcta.

Accuracy es la proporción del número total de predicciones que son correctas [25].

Para comparar la precisión de las diferentes técnicas de minería de datos utilizadas para la predicción del rendimiento académico, se realizó el cálculo a partir de las matrices de confusión, que fueron generadas utilizando el lenguaje R, las cuales son tablas donde se muestra el número de predicciones para cada clase (BAJO, MEDIO, ALTO), representando su diagonal las predicciones que se realizaron de forma correcta.

En la Tabla 8 se muestra el porcentaje de precisión y error hallados a partir de la matriz de confusión obtenidos en la predicción realizada por los diferentes modelos construidos.

TABLA 8
COMPARACIÓN DE LA PRECISIÓN Y ERROR

Técnica	% Precisión	% Error
Árbol de Decisión	75%	25%
Random Forest	93%	7%
Redes Bayesianas	93%	7%
KNN	95%	5%
SVM	100%	0%

En la Fig. 11 se observa que se obtuvo mayor precisión con la técnica SVM.

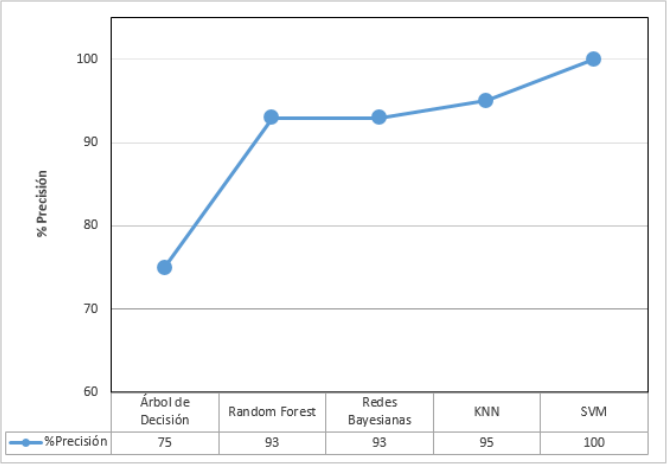

Fig. 11 Comparación de precisión a partir de la matriz de confusión.

En las pruebas realizadas se obtuvo mayor precisión aplicando la técnica de aprendizaje supervisado de minería de datos SVM, obteniendo un 100% de aciertos en las clasificaciones. Sin embargo estos resultados de precisión y error hallados en cada técnica de minería de datos a partir de la matriz de confusión pueden variar dependiendo de las muestras generadas al azar para la tabla de entrenamiento (80%) y tabla de pruebas (20%) por lo que es necesario la utilización de técnicas de validación cruzada para comparar los resultados de los algoritmos de minería de datos y elegir de forma más exacta el modelo que realice la predicción con bastante precisión sobre la clase o categoría a la que pertenece el alumno en base a su rendimiento.

La validación cruzada es un método establecido para evaluar la exactitud de los modelos de minería de datos. En la validación cruzada, divide sucesivamente los datos de la estructura de minería en subconjuntos, genera modelos en los subconjuntos y, a continuación, mide la exactitud del modelo para cada partición. Revisando las estadísticas devueltas, puede determinar el grado de confiabilidad del modelo de minería de datos y comparar más fácilmente los modelos que se basan en la misma estructura [26].

3) Validación cruzada dejando uno afuera

Conocido también como Leave-one-out cross-validation (LOOCV). En cada iteración se toma solamente un registro para la tabla de prueba y el resto de registros para la tabla de entrenamiento.

Se aplicó esta técnica de validación cruzada para comparar los resultados de precisión y error de las técnicas de minería de datos supervisadas que se seleccionó para la clasificación, las cuales son: Árbol de decisión, random forest, redes bayesianas, KNN y SVM.

En las pruebas realizadas para la predicción del rendimiento se utilizó en total 69 registros por lo que se tendrá 69 iteraciones, en donde en cada una de estas se utilizará diferentes datos de entrenamiento y prueba por cada técnica predictiva, como se muestra en la Fig. 12.

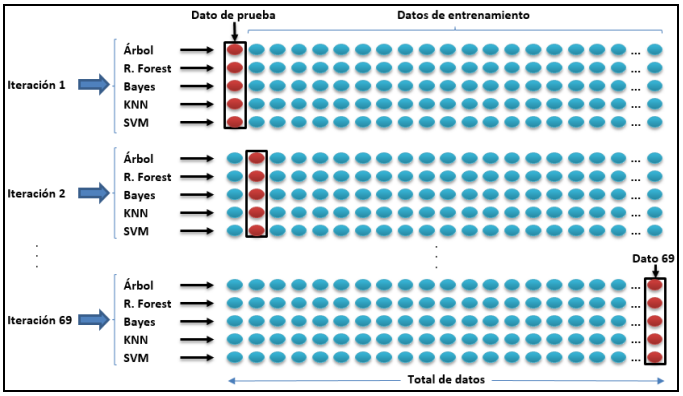

Fig. 12 Validación cruzada dejando uno afuera (LOOCV).

En cada iteración se obtiene cinco resultados de precisión y error correspondientes a las cinco técnicas de minería de datos aplicadas. La instrucción repetitiva que se creó utilizando lenguaje R para la validación cruzada dejando uno afuera se muestra en la Fig. 13.

```
n <- nrow(notas)
for (f in 1:n) {
  prueba <- notas[f, ]
  entrenamiento <- notas[-f, ]

  lstResult <- calcularPrecision('arbol', entrenamiento, prueba)
  precisionArbol <- precisionArbol +  lstResult[1]
  errorArbol <- errorArbol + lstResult[2]

  lstResult <- calcularPrecision('bosque', entrenamiento, prueba)
  precisionForest <- precisionForest +  lstResult[1]
  errorForest <- errorForest + lstResult[2]

  lstResult <- calcularPrecision('bayes', entrenamiento, prueba)
  precisionBayes <- precisionBayes +  lstResult[1]
  errorBayes <- errorBayes + lstResult[2]

  lstResult <- calcularPrecision('knn', entrenamiento, prueba)
  precisionKnn <- precisionKnn +  lstResult[1]
  errorKnn <- errorKnn + lstResult[2]

  lstResult <- calcularPrecision('svm', entrenamiento, prueba)
  precisionSvm <- precisionSvm +  lstResult[1]
  errorSvm <- errorSvm + lstResult[2]
}
```

Fig. 13 Iteraciones para la validación cruzada dejando uno afuera.

La función llamada calcularPrecision retorna la precisión y error de cada técnica de minería de datos utilizada.

Al final se calcula la media aritmética de los resultados por cada técnica predictiva. Se itero cinco veces para verificar que el error es bajo, resultando para casi todos los algoritmos un valor constante.

Los resultados de precisión obtenidos con la técnica de validación cruzada dejando uno afuera (LOOCV) se muestran en la Tabla 9:

TABLA 9
COMPARACIÓN DE PRECISIÓN DE LOS MODELOS

Técnica	Iteración 1	Iteración 2	Iteración 3	Iteración 4	Iteración 5
Árbol de Decisión	86.95652	86.95652	86.95652	86.95652	86.95652
Random Forest	94.20290	92.75362	91.30435	92.75362	92.75362
Redes Bayesianas	97.10145	97.10145	97.10145	97.10145	97.10145
KNN	95.65217	95.65217	95.65217	95.65217	95.65217
SVM	100.0000	100.0000	100.0000	100.0000	100.0000

En la Fig. 14 se puede observar que se obtuvo mayor precisión aplicando la técnica de minería de datos supervisada SVM, obteniendo un 100% de aciertos en las clasificaciones.

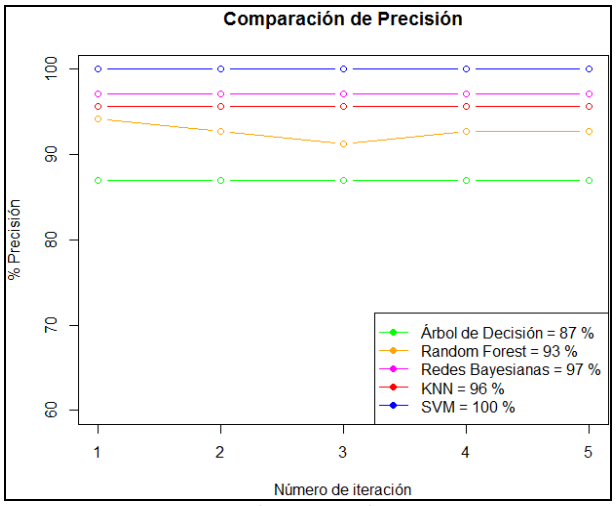

Fig. 14 Comparación de precisión de los modelos.

Los resultados de error obtenidos con la técnica de validación cruzada dejando uno afuera (LOOCV) se muestran en la Tabla 10:

TABLA 10
COMPARACIÓN DE ERROR DE LOS MODELOS

Técnica	Iteración 1	Iteración 2	Iteración 3	Iteración 4	Iteración 5
Árbol de Decisión	13.04348	13.04348	13.04348	13.04348	13.04348
Random Forest	5.797101	7.246377	8.695652	7.246377	7.246377
Redes Bayesianas	2.898551	2.898551	2.898551	2.898551	2.898551
KNN	4.347826	4.347826	4.347826	4.347826	4.347826
SVM	0.000000	0.000000	0.000000	0.000000	0.000000

En la Fig. 15 se puede observar que se obtuvo menor error aplicando la técnica de minería de datos supervisada SVM, obteniendo 0% de error en las clasificaciones.

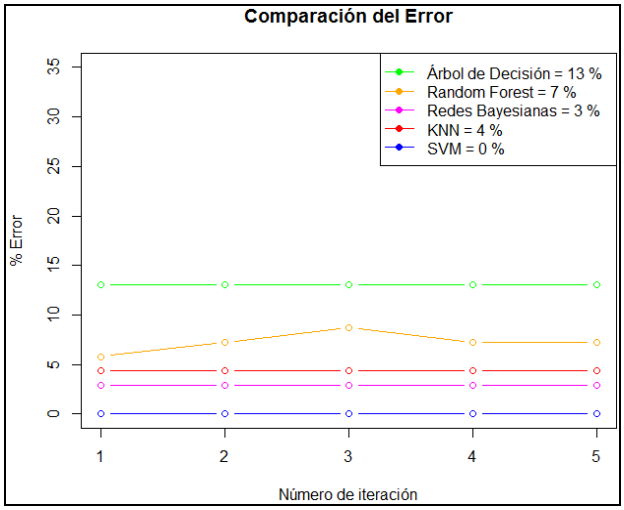

Fig. 15 Comparación de error de los modelos.

En la Tabla 11 se muestra un resumen de los porcentajes promedios de precisión y error obtenidos con la aplicación de las técnicas de aprendizaje supervisado de minería de datos seleccionados.

TABLA 11
PRECISIÓN Y ERROR PROMEDIO DE LOS MODELOS

Técnica	% Precisión	% Error
Árbol de Decisión	87%	13%
Random Forest	93%	7%
Redes Bayesianas	97%	3%
KNN	96%	4%
SVM	100%	0%

Esta técnica de validación cruzada es más confiable ya que el error es muy bajo; pero computacionalmente es más costoso por la cantidad de iteraciones que se pueden requerir dependiendo del número total de registros.

Para la predicción del rendimiento académico se elige el modelo SVM por ser una de las técnicas de aprendizaje supervisado más eficiente en tiempos de ejecución y la más precisa en las pruebas de predicción según los resultados obtenidos a partir de la matriz de confusión y la técnica de validación cruzada dejando uno afuera. Esta técnica de minería de datos podría utilizarse en las instituciones educativas para automatizar sus procesos y clasificar a los estudiantes según sus datos académicos y mejorar el proceso de enseñanza-aprendizaje reforzando a los alumnos de bajo rendimiento.

IV. CONCLUSIONES

Se realizó una evaluación comparativa de técnicas de aprendizaje supervisado de minería de datos que permiten clasificaciones de más de dos clases, como son el árbol de decisión, random forest, redes bayesianas, K vecinos más cercanos (KNN) y máquinas de soporte vectorial (SVM) para realizar la predicción del rendimiento académico utilizando técnicas de comparación de precisión a partir de las matrices de confusión obtenidas por los diferentes modelos de minería de datos aplicados, también se comparó los tiempos de ejecución que demoraron en la construcción de los modelos, resultando la técnica de máquina de soporte vectorial (SVM) una de las más eficientes en tiempo de ejecución y la más precisa en las pruebas de predicción. Finalmente se utilizó la técnica de validación cruzada dejando uno afuera (Leave-one-out cross-validation) para obtener resultados más confiables de precisión y error de las técnicas de minería de datos aplicadas, con la que se logró obtener valores constantes de precisión y error que permitieron seleccionar el modelo SVM por ser la técnica de aprendizaje supervisado con mayor precisión para la predicción del rendimiento académico.

REFERENCIAS

[1] Universia Perú, Tendencias de la educación superior en el siglo XXI, http://noticias.universia.edu.pe/vida-universitaria/noticia/2007/06/19/747349/tendencias-educacion-superior-siglo-xxi.html, Revisado en Mayo del 2016.

[2] UNESCO: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, Repetition at high cost in Latin America and the Caribbean, IESALC - UNESCO, 2004

[3] R. Zárate and E. Mantilla, "La deserción estudiantil UIS, una mirada desde la responsabilidad social universitaria," *Zona Próxima - Revista del Instituto de Estudios en Educación Universidad del Norte*, no. 21, 2014.

[4] E. Himmel, "Modelos de análisis de la deserción estudiantil en la educación superior," *Revista Calidad en la Educación*. Consejo Superior de Educación. Ministerio de Educación, Chile, no. 17, pp. 91-108, 2002.

[5] K. Eckert and R. Suénaga, "Aplicación de técnicas de minería de datos al análisis de situación y comportamiento académico de alumnos de la UGD," in *Proc. XV Workshop de Investigadores en Ciencias de la Computación*, RedUNCI, pp. 92-96, 2013.

[6] R. Timarán and J. Jiménez, "Extracción de perfiles de deserción estudiantil en la institución universitaria CESMAG," *Investigium IRE*, vol. 6, no 1, pp. 30-44, 2015.

[7] E. A. Porcel, G. N. Dapozo and M. V. López, "Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa," *Revista electrónica de investigación educativa - REDIE*, vol. 12, no. 2, 2010.

[8] M. Rojas and D. C. González, "Rendimiento y calificación, dos aspectos problemáticos de la evaluación en la universidad," *Revista Virtual Universidad Católica del Norte*, vol. 27, pp. 1-21, 2009.

[9] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook, Second Edition*, New York, EEUU: Springer, 2010.

[10] Investigación u-Learning, EDM http://ingenieriaeducacion.blogspot.pe/2010/06/edm.html, Revisado en Diciembre del 2016.

[11] International Educational Data Mining Society, Educational Data Mining, http://www.educationaldatamining.org/, Revisado en Diciembre del 2016.

[12] A. Ballesteros, D. Sánchez-Guzmán and R. García, "Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un context educativo," *Latin-American Journal of Physics Education - LAPJE*, vol. 7, no. 4, Diciembre 2014.

[13] S. Weiss and N. Indurkhya, Predictive Data Mining: A Practical Guide, San Francisco, CA, EEUU: Morgan Kaufmann, 1998.

[14] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, *3rd ed.*, San Francisco, CA, EEUU: Morgan Kaufmann, 2011.

[15] J. Thongkam, "Towards Breast Cancer Survivability Prediction Models in Thai Hospital Information Systems," Tesis para obtener el grado de Doctor de Filosofia, School of Engineering and Science, Faculty of Health, Engineering and Science, Victoria University, Australia, 2009.

[16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, San Francisco, CA, EEUU: Morgan Kaufmann, 2000.

[17] M. Maneiro, "Minería de Datos," Monografía de Adscripción en Licenciatura de Sistemas de Información, Universidad Nacional del Nordeste, Facultad de Ciencias Exactas, Naturales y Agrimensura, Argentina, 2008.

[18] KDnuggets, "Analytics, Data Mining, Data Science software/tools used in the past 12 months," http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html, Revisado en Enero del 2016.

[19] KDnuggets, "R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results," http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html, Revisado en Noviembre del 2016.

[20] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[21] B. Sierra, *Aprendizaje Automático*, Madrid, España: Pearson, 2006.

[22] M. Gallardo, "Aplicación de técnicas de clustering para la mejora del aprendizaje," Proyecto de fin de carrera en Ingeniería de Telecomunicación, Universidad Carlos III de Madrid, España, 2009.

[23] E. Brito, J. Sánchez, P. Guillén and C. Torres, "Predicción de patrones de flujo bifásico mediante máquinas de soporte vectorial," *Conference: X Congreso Internacional de Métodos Numéricos en Ingeniería y Ciencias Aplicadas*, CIMENICS 2010, Venezuela, 2010.

[24] P. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, New York, EEUU: Pearson Education, 2006.

[25] M. Sánchez, Matriz de Confusión, http://myslide.es/documents/matriz-de-confusion-listo.html, Revisado en Marzo del 2016.

[26] Microsoft TechNet, Novedades (Analysis Services - Minería de datos), https://technet.microsoft.com/es-es/library/bb510513%28v=sql.105%29.aspx, Revisado en Marzo del 2016.