Minería de datos para descubrir tendencias en la clasificación de los trabajos de titulación

Silvia Haro[a], Rubén Pazmiño[a] , Miguel Conde[b] , Francisco Peñalvo[c]
[a] Facultad de Ciencias, CIDED, Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador
[b] Department of Computer Science, University of León, León, Spain
[c] Department of Computer Science, University of Salamanca, Salamanca, Spain
s_haro@espoch.edu.ec, rpazmino@espoch.edu.ec, miguel.conde@unileon.es, fgarcia@usal.es

***Resumen*—La minería de datos tiene como fin encontrar patrones que expliquen la tendencia de los datos, es por ello que con el objetivo de extraer conocimiento de los trabajos de titulación de la Facultad de Ciencias de la Escuela Superior Politécnica de Chimborazo; se aplicaron cinco modelos de clasificación: Máquinas de Soporte Vectorial, Redes Neuronales, Árbol de Decisión, Bosque Aleatorio y Potenciación; considerando las líneas Diseño de Experimentos y Análisis Multivariable. Para identificar el modelo óptimo se aplicó Rattle, se calcularon tres medidas de rendimiento, las precisiones: global, positiva y negativa; siendo la curva ROC y los Árboles de Decisión, gráficas que permitieron visualizar el modelo de predicción con mejor ajuste, así como los programas que caracterizan las líneas de investigación. Los resultados mostraron que para la línea Diseño de Experimentos, el modelo con mayor precisión fue Bosque Aleatorio, con un 71,48% de predicciones que son correctas respecto al total; mientras que al considerar la línea Análisis Multivariable no se evidenció diferencia significativa en la precisión global, fluctuando en el 97%; esto significa que con el 97% de certeza la línea de investigación de Análisis Multivariable y con el 71,48% de precisión la línea de Diseño de Experimentos se enmarcan en los programas de investigación institucionales. En el modelo Árbol de Decisión, el nodo principal fue la Carrera cuando se consideró la línea Diseño de Experimentos, debido a que en Bioquímica y Farmacia se impulsa la utilización de estudios de este tipo; y en el caso de la línea Análisis Multivariable fue el programa Consumo Humano para mejorar las condiciones de nutrición y salud, debido a que los trabajos de titulación tienen una baja utilización de técnicas multivariable.**

***Palabras Claves*—minería de datos educativos, técnicas de minería de datos, indicadores educativos, clasificación**

***Abstract*—Data mining has a main purpose which is to find patterns that explain the trend of data, for this reason the aim is to extract knowledge from the degree works of the Faculty of Sciences of the ESPOCH; therefore, it had been applied Five classification models: Vector Support Machines, Neural Networks, Decision Tree, aleatory Forest and Potentiation; considering the Experimental Design line and its Multivariable Analysis. To identify the optimal model Rattle was applied, three measures of performance had been calculated: global, positive and negative precisions; it being the ROC curve and Decision Trees, graphs that allowed to visualize the model prediction with a better adjustment, as well as programs that characterize the lines of investigation as well the results showed that the Experimental Design line was the most accurate model from aleatory Forest, with 71.48% of predictions that are correct with the total; while considering the Multivariate Analysis line. There is no significant difference with the global precision, fluctuating 97%; this means; that the 97% of certain the Multivariable Analysis research line and with the 71.48% accuracy, the Experimental Design line are framed into the institutional research programs. The model of the Decision Tree; the main node was the Career when the Experiments Design line was considered, due to the fact of Biochemistry and Pharmacy it been encouraged to use this type of studies; and with the case of the Multivariable Analysis line was the Human Consumption program to improve the conditions of nutrition and health, due to the degree works it has a low use of multivariable techniques.**

***Keywords*—data mining education, data mining techniques, education indicators, classification**

I. Introducción

La minería de datos es un proceso que permite extraer conocimiento de bases de datos [1], su objetivo es descubrir situaciones anómalas y/o interesantes, tendencias, patrones y secuencias en los datos. Uno de sus resultados es la clasificación [3, 4], misma que trata de obtener un modelo que permita asignar un caso de línea desconocida a una línea concreta [2]. Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de registros (trabajos de titulación) y un conjunto de líneas $C = \{C_1, C_2, \dots, C_m\}$,, el problema de la clasificación es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada en una línea C_j. Para el estudio se consideraron dos líneas: Diseño Experimental y Análisis Multivariable; y como función f se tomaron los modelos de clasificación:

- Máquinas de Soporte Vectorial (SVM): Técnica que permite extraer información relevante a partir de conjuntos de datos y construir algoritmos eficientes y rápidos. Se basa en encontrar un hiperplano que separe las líneas en el mayor margen posible, la frontera de decisión se establece mediante los patrones que más resaltan la distribución de líneas [5].
- Redes Neuronales: Son sistemas dinámicos auto adaptativos [6]. Es un conjunto de nodos interconectados y enlaces ponderados. Los nodos de salida son la suma de cada uno de los valores de entrada de acuerdo con los pesos de sus vínculos.
- Árbol de Decisión: Es una técnica de clasificación supervisada [7], que permite determinar la decisión que se debe tomar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas [8]. Esta técnica recursiva considera el criterio de la mayor proporción de ganancia de información, es decir, elige el atributo que mejor clasifica los datos [9].
- Bosque Aleatorio: Es una combinación de árboles predictivos, trabaja con una colección de árboles incorrelados y los promedia. Cada árbol depende de los valores de un vector aleatorio de la muestra de forma independiente y con la misma distribución de todos los árboles en el bosque [7].
- Potenciación: Este modelo toma una muestra aleatoria

de los datos originales y aplica sobre ésta un método clasificatorio, luego aumenta el peso (potenciar) a los individuos mal clasificados para que en la siguiente aplicación del método clasificatorio se enfoque más en estos individuos mal clasificados, mejorando su clasificación; y así sucesivamente.

II. Método

La metodología utilizada fue la propuesta por Williams Graham en su trabajo titulado "Data Mining with Rattle and R, The Art of excavating data for knowledge discovery"; donde muestra el paquete creado en R y sus aplicaciones [10]. Ésta consistió en emplear un porcentaje de datos para probar los modelos: Máquinas de Soporte Vectorial, Redes Neuronales, Árbol de Decisión, Bosque Aleatorio y Potenciación; y el restante para validarlos. Se evaluó su rendimiento mediante las precisiones global, positiva y negativa; y mediante la curva ROC para evidenciar gráficamente si los modelos son o no adecuados.

Para la investigación se empleó una base de datos que corresponde a 24200 observaciones, misma que está conformada por 968 trabajos de titulación [11] de las carreras de: Bioquímica y Farmacia, Biofísica, Ingeniería en Biotecnología Ambiental, Ingeniería en Estadística Informática, Ingeniería en Química y Licenciatura en Educación Ambiental; de la Facultad de Ciencias de la Escuela Superior Politécnica de Chimborazo (ESPOCH). Se seleccionaron los trabajos de titulación presentados durante los años 2012 – 2016. Las variables utilizadas fueron los 12 programas: "Biodiversidad sustentable, Energías alternativas, Gestión y tratamiento del aire, agua y suelo, Biorremediación del ambiente, Desarrollo de aplicaciones de software para procesos de gestión y administración pública y privada. Educación, Análisis Estadístico Implicativo y Computacional, Evaluación del estado de la seguridad alimentaria, Consumo humano para mejorar las condiciones de nutrición y salud, Evaluación del estado de salud y nutrición, Administración de servicios de salud, nutrición y alimentación, Biofísica aplicada a la medicina y Desarrollo de fitofármacos" [12].

La aplicación se realizó mediante Rattle, para la evaluación de los modelos se requirió de una tabla de aprendizaje, misma que entrena (aprende) el modelo de predicción, es decir, a partir de esta tabla se calcula f; y de una tabla testing, la cual permite validar el modelo y es seleccionada internamente por el software, misma que verifica que los resultados en individuos que no participaron en la construcción del modelo es bueno o aceptable [1]. En el estudio se empleó el 75% de los datos para la tabla de aprendizaje y el 25% para la tabla testing.

Para evaluar el rendimiento de los modelos se determinaron las medidas: Precisión Global (Exactitud), que muestra el número total de predicciones que son correctas al total, la Precisión Positiva (Sensibilidad), que es la proporción de casos positivos que fueron identificados correctamente y la Precisión Negativa (Especificidad); que indica la proporción de casos negativos que fueron identificados correctamente [1]. Estas medidas se calcularon a partir de la matriz de confusión; misma que contiene información acerca de las predicciones realizadas por un modelo o sistema de clasificación, comparando para el conjunto de trabajos de titulación la tabla de aprendizaje o testing, con la predicción dada versus la línea a la que estos realmente pertenecen. En la Tabla I, se muestra la matriz de confusión para un clasificador de dos líneas:

TABLA I
MATRIZ DE CONFUSIÓN

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	VN (Verdadero Negativo)	FP (Falso Positivo)
	Positivo	FN (Falso Negativo)	VP (Verdadero Positivo)

Un gráfico que muestra el rendimiento de los modelos fue la curva ROC [10]. Compara la tasa de falsos positivos con la de verdaderos positivos, ubica en el eje Y= Sensibilidad y en el X = 1 – Especificidad. Sirve para seleccionar los modelos óptimos en problemas de clasificación binarios, este método considera el área bajo la curva; si ésta es 1 el modelo es ideal para el conjunto de datos, si es mayor que 0.5 es óptimo y el modelo es despreciable para su uso si su valor es inferior a 0.5.

III. Análisis de Resultados

Mediante Rattle se determinó la matriz de confusión para cada clasificador. Se muestra un caso particular para el modelo Máquinas de Soporte Vectorial cuando se aplicó la línea Diseño de Experimentos, Tabla II:

TABLA II
MATRIZ DE CONFUSIÓN, SVM; LÍNEA DISEÑO DE EXPERIMENTOS

		Predicción	
		0 No es un diseño experimental	1 Es un diseño experimental
Valor Real	0 No es un diseño experimental	63	53
	1 Es un diseño experimental	33	93

Se calcularon las medidas de rendimiento mediante las ecuaciones Precisión Global P (Exactitud), Precisión Positiva (Sensibilidad) y Precisión Negativa (Especificidad), (1), (2) y (3), respectivamente, como sigue:

$$P = \frac{VN + VP}{VN + FP + FN + VP} \quad (1)$$

$$PP = \frac{VP}{FN + VP} \quad (2)$$

$$PN = \frac{VN}{VN + FP} \quad (3)$$

Los resultados de estos indicadores para los cinco modelos, aplicando las dos líneas se muestran en la Tabla III; debido a la versión de Rattle no se pudo determinar la matriz de confusión para el modelo Potenciación y por ende no se calcularon las precisiones.

TABLA III
RESULTADOS DE MEDIDAS DE RENDIMIENTO

CLASIFICADOR (MODELO)	Precisión Global (P%)	Precisión Positiva. *Sensibilidad* (PP%)	Precisión Negativa. *Especifidad* (PN%)	ROC
LÍNEA: Diseño de Experimental				
Máquinas de Soporte Vectorial	64.46	54.31	73.81	0.72
Redes Neuronales	71.48	67.24	75.39	0.76
Árbol de Decisión	64.05	50.86	76.19	0.69
Bosque Aleatorio	68.59	58.62	77.77	0.77
Potenciación				0.78
LÍNEA: Análisis Multivariable				
Máquinas de Soporte Vectorial	97.93	99.08	86.95	0.96
Redes Neuronales	97.11	97.26	95.65	0.97
Árbol de Decisión	97.93	99.09	86.95	0.93
Bosque Aleatorio	97.93	99.09	86.95	0.99
Potenciación				0.99

De la Tabla III, para la Línea: Diseño de Experimentos; la mayor precisión global se presentó en el modelo Redes Neuronales con un 71.48% de predicciones que son correctas respecto al total; seguido por Bosque Aleatorio con un 68.59%. Entre los casos positivos que fueron identificados correctamente, Redes Neuronales fue el modelo que alcanzó el mayor porcentaje con un 67.24% y el más bajo fue el Árbol de Decisión con un 50.86%. El modelo que presentó la mayor proporción de casos negativos identificados correctamente fue Bosque Aleatorio con un 77.77%; mientras que SVM alcanzó el porcentaje más bajo con un 73.81%. Se puede observar también en la Tabla III los resultados de la curva ROC para cada uno de los modelos con las dos líneas.

La Fig. 1 muestra la gráfica de esta curva para el modelo SVM con el clasificador Diseño de Experimentos.

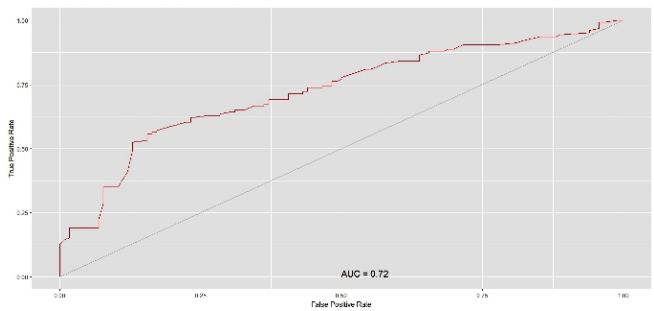

Fig. 1. Curva ROC, modelo SVM; Línea Diseño de Experimentos

La curva ROC mostró que los modelos adecuados para esta línea son Potenciación y Bosque Aleatorio con un área bajo la curva de 0.78 y 0.77 respectivamente; mientras que SVM es el que presenta la menor área (Fig. 1). Para la línea Análisis Multivariable; se evidenció que la Exactitud (97.93%) y la Especificidad (86.95%) son iguales para SVM y la Sensibilidad (99.09%) para Árbol de Decisión y Bosque Aleatorio. La curva ROC mostró que los modelos con mayor área bajo la curva fueron Bosque Aleatorio y Potenciación con un valor cercano a 1; cabe indicar que todos los modelos mostraron áreas superiores a 0.90.

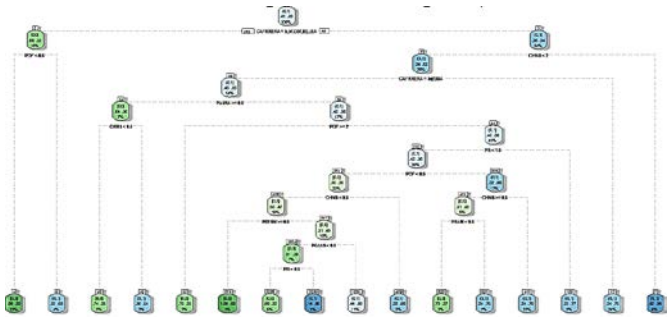

Fig. 2. Árbol de decisión para Diseño de Experimentos

La Fig. 2 muestra la gráfica del Árbol de Decisión para la línea Diseño de Experimentos, el nodo principal es la variable Carrera, existe el 41% de probabilidad de que los datos correspondan a 0; es decir, el trabajo de titulación NO es Diseño Experimental y el restante 59% para 1; es decir, SI lo es. Si la Carrera es Bioquímica y Farmacia nos dirigimos al Programa de desarrollo fitofármaco (PDF) donde cayeron el 16% de los datos; este nodo predice que hay un 68% de probabilidad de que el trabajo de titulación sea un Diseño Experimental, caso contrario se predice que es un programa de Consumo humano para mejorar las condiciones de nutrición y salud (CHNS) donde caen el 84% de los datos y establece que existe un 64% de probabilidad de que los trabajos no sean un Diseño. La Fig. 3 muestra el Árbol de Decisión para la línea Análisis Multivariante, el nodo principal es el indicador Consumo humano para mejorar las condiciones de nutrición y salud (CHNS), existe el 89% de probabilidad de que los datos sean 0; es decir NO es un Análisis Multivariable y el restante a 1. Si la puntuación de CHNS es menor que 2 entonces se predice que el trabajo de titulación es un Programa de desarrollo fitofármaco (PDF), donde existe el 94% de probabilidad de que no sea un Análisis Multivariable y un 0,06% de que si lo sea; si el indicador PDF obtiene una puntuación menor que 2 se predice que el tema de investigación es un programa de Análisis Estadístico Implicativo Computacional (PAEIC) con un 92% de los datos.

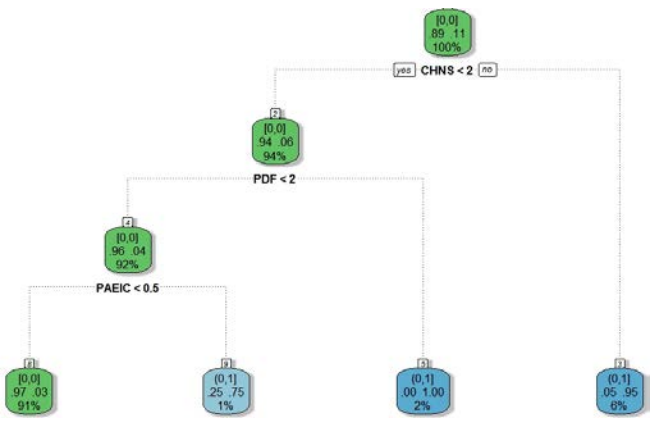

Fig. 3. Árbol de decisión para Análisis Multivariable

Para el modelo de Potenciación se realizaron gráficas que permitieron identificar la importancia de las variables y la curva del error. Las Fig. 4 y Fig. 5 muestran un caso

particular para la línea Diseño de Experimentos.

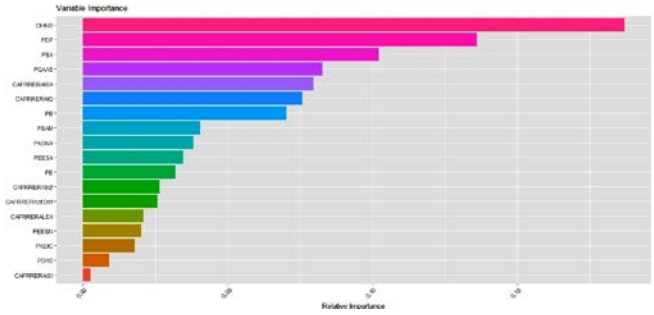

Fig. 4. Importancia de los programas para la línea Diseño de Experimentos

La Fig. 4 muestra que los cuatro programas más importantes dentro del modelo de Potenciación son: Consumo humano para mejorar las condiciones de nutrición y salud, Desarrollo fitofármaco, Biorremediación del ambiente, Gestión y tratamiento del aire, agua y suelo y la Carrera de Ingeniería en Biotecnología Ambiental; y la Fig. 5 muestra que este método se estabiliza aproximadamente con 35 interacciones.

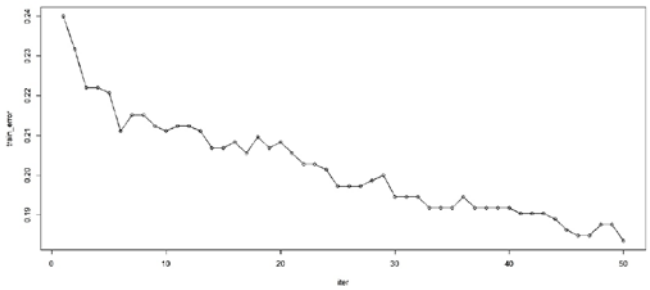

Fig. 5. Gráfico de errores para indicar estabilidad del método

IV. Conclusiones

Los métodos de clasificación permitieron identificar información relevante sobre los indicadores educativos cuando se trabajaron las líneas Diseño de Experimentos y Análisis Multivariable. Mediante Rattle se identificaron los modelos adecuados para el conjunto de datos. Con los Árboles de Decisión se visualizaron variables representativas que pueden ayudar a predecir eventos futuros. Rattle es un paquete con una interfaz de fácil manejo y no presenta inconveniente alguno. Es importante considerar que la base de datos debe ser guardada con extensión CSV y al momento de cargar el archivo se deben identificar los signos empleados como separador de datos y decimales. Para el estudio se procedió a cargar el archivo de datos, luego se seleccionaron los porcentajes para la tabla testing y aprendizaje, posteriormente se aplicaron los modelos de clasificación y mediante Excel se calcularon las medidas de rendimiento. Con este trabajo se ha determinado la tendencia de futuros trabajos de titulación, ya que dependiendo de la Carrera se podrá identificar si éstos van a aplicar Diseño de Experimentos o un Análisis Multivariante. El estudio realizado es un preámbulo para luego realizar minería de texto con todos los documentos electrónicos de los 1090 trabajos de titulación de la Facultad de Ciencias en el período 2012-2017 y se luego se puede ampliar al considerar los indicadores vigentes desde finales del 2018 hasta el 2022. A futuro se pretende realizar minería de texto con los trabajos de titulación de la Facultad de Ciencias y determinar la impureza de la clasificación mediante el error de clasificación Split, Índice Gini y Entropía.

Referencias

[1] Y. Zheng, “Trajectory data mining: An overview,” *ACM Trans. Intell. Syst. Technolog.*, vol. 6, no. 3, 2015.

[2] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques: Elsevier Inc., 2016.

[3] B. Frénay and M. Verleysen, “Classification in the presence of label noise: A survey,” *IEEE Trans. Neural Networks Learn. Sys.*, vol. 25, no. 5, pp. 845–869, 2014.

[4] A. M. Cheriyadat, “Unsupervised feature learning for aerial scene classification,” *IEEE Trans Geosci Remote Sens*, vol. 52, no. 1, pp. 439–451, 2014.

[5] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” *Expert Sys Appl*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

[6] G. B. Ermentrout, S. E. Folias, and Z. P. Kilpatrick, “Spatiotemporal pattern formation in neural fields with linear adaptation,” in Neural Fields: Theory and Applications: Springer-Verlag Berlin Heidelberg, 2014, pp. 119–151.

[7] R. Medina and C. Ñique, “Bosques Aleatorios como extensión de los árboles de clasificación con los programas R y Python,” *Interfases*, pp. 165–189, 2017.

[8] Y. Robles and A. Sotolongo, “Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL,” *Gestión de Tecnología y Sistemas de Información*, pp. 389–406, 2013.

[9] S. Valero, A. Vargas, and M. García, “Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos,” *Recursos Digitales para la Educación y la Cultura*, 33-30, 2010.

[10] W. Graham, Data Mining with Rattle and R. New York, USA: Springer, 2011.

[11] IDI, Reglamento del Instituto de Investigaciones de la Escuela Superior Politécnica de Chimborazo 2014. Ecuador.

[12] IDI, Plan de investigación ESPOCH 2014-2018. Ecuador, vol. 2018-10-04.