Conference Paper

Árbol De Decisión, Aplicación Con Datos Meteorológicos

Decision Tree, Application With Meteorological Data

Silvia Mariana Haro Rivera

Grupo de Energías Alternativas y Ambiente, Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo. Panamericana Sur Km 1,5 Riobamba, Ecuador, Código postal EC060105

Resumen

La minería de datos es una técnica que hoy en día se aplica en muchas áreas de las ciencias, es por ello que con el objetivo de identificar variables meteorológicas predominantes a ocho intervalos de tiempo se aplicó la técnica supervisada árbol de clasificación en data mining. La información se obtuvo de la estación Alao, misma que se encuentra ubicada a 3064 m.s.m en la provincia de Chimborazo, Ecuador. El estudio se realizó mediante código desarrollado en el software estadístico R; los datos corresponden a información por hora del año 2016, las variables analizadas fueron; temperatura del aire, humedad relativa, presión barométrica, radiación solar difusa, radiación solar global, temperatura del suelo a $-20cm$ y velocidad de viento. El árbol mostró que la principal variable en esta zona es la radiación solar global, a horas comprendidas de 06h00 a 08h00, si ésta es mayor o igual a $120w/m^2$, entonces se puede determinar la presión barométrica de 09h00 a 11h00 de la mañana; y si ésta es mayor o igual que $709w/m^2$, entonces se predice la temperatura del aire. El árbol de decisión es una técnica que permitió identificar variables meteorológicas relevantes, en determinadas horas donde se encuentra ubicada la estación Alao.

Abstract: Data mining is a technique that today is applied in many areas of science, which is why in order to identify predominant meteorological variables at eight time intervals the supervised tree classification technique was applied in data mining. The information was obtained from the Alao station, which is located at 3064 m.s.m in the province of Chimborazo, Ecuador. The study was carried out using a code developed in statistical software R, the data correspond to information by hour of the year 2016, the variables analyzes were air temperature, relative humidity, barometric pressure, diffuse solar radiation, global solar radiation, soil temperature at $-20cm$ and wind speed. The showed that the main variable in this area is the global solar radiation, at hours between 06h00 and 08h00, if it is greater than or equal to $120w/m^2$, then the barometric pressure can be determined from 09h00 to 11h00 of the morning, if, and it is great than or equal to $709w/m^2$, then the air temperature is predicted. The decision tree is a technique that allowed us to identify relevant meteorological variables in certain hours where the Alao station is located.

Palabras clave: árboles de decisión, datos meteorológicos.
Keywords: decision tree, meteorological data.

Corresponding Author:
Silvia Mariana Haro Rivera
s_haro@espoch.edu.ec

Received: 10 January 2020
Accepted: 17 January 2020
Published: 26 January 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the VI Congreso Internacional Sectei 2019 Conference Committee.

1. Introducción

La información que proporcionan las diversas variables meteorológicas tienen gran importancia en temas relacionados con el cambio climático, la determinación de recursos hídricos, eólicos y solares; así como en el planteamiento de políticas relacionadas con el medio ambiente. Es por ello que la detección de variables meteorológicas predominantes en la zona de Alao, provincia de Chimborazo; permitirá establecer condiciones en las que se producen cambios en los parámetros considerados en el estudio.

Hoy en día la revolución digital permite que la información sea fácil de almacenar, lo cual genera un gran volumen de datos; es así que, la minería de datos (data mining) busca darle sentido a la masiva información, pues ésta será útil para la toma de decisiones (1). En minería de datos se han realizado varios estudios, es así que (2) emplea las técnicas de clasificación: árbol de decisión, naive bayes aumentado a árbol y regla OneR, para identificar factores que influyen en la deserción de estudiantes universitarios; (3) en su trabajo titulado Árboles de clasificación de Potimirim mexicana (Decapoda: Caridea), organismo hermafrodita protándrico secuencial, determinar la importancia de los árboles en la clasificación de organismos considerando variables morfológicas; (4), Bouza & Santiago realizan aplicaciones de la minería datos, mediante árboles de decisión en aspectos del manejo de hospitales y epidemias; (5), Ruiz y Romero aplican árboles de decisión para mejorar la calidad de la información que pertenece a una base de datos que contiene información bibliográfica del Sistema de Gestión Bibliotecario del Instituto de Información Científico y Tecnológico en Cuba; en el ámbito educativo (6); Uvidea, con el objetivo de identificar áreas de conocimiento en los cursos de nivelación que deben ser fortalecidas aplican algoritmos de clasificación mediante minería de datos; en el área de la meteorología se ha realizado un estudio comparativo de técnico de minería de datos para la predicción de rutas de huracanes aplicando técnicas de predicción: regresión lineal, k vecinos más cercanos y perceptrón multicapa (7).

La minería de datos es un proceso mediante el cual se puede extraer información de grandes volúmenes de datos (8), uno de sus resultados es la clasificación; la cual obtiene un modelo que permite asignar un caso de línea desconocida a una línea concreta (9). Sea $D = \{t_1, t_2, \dots, t_n\}$ la base de datos con registros por hora de las variables meteorológicas: temperatura del aire, humedad relativa, presión barométrica, radiación solar difusa, radiación solar global, temperatura del suelo a $-20cm$ y velocidad de viento; y $C = \{C_1, C_2, \dots, C_8\}$ el conjunto formado por las ocho escalas generadas

con intervalos de tiempo de dos horas, el problema de la clasificación es hallar una función $f : D \rightarrow C$ tal que cada t_i sea asignada en una clase C_j. Como función f se seleccionó el árbol de clasificación; técnica supervisada que determina la decisión que se debe tomar siguiendo condiciones que se cumplen desde la raíz hasta alguna de sus hojas (10), el árbol elige el atributo que mejor clasifica al conjunto de datos (11).

En el estudio se aplicó el algoritmo CART, desarrollado por Breiman (1984) donde el árbol se construye fragmentando sucesivamente el conjunto de datos. La construcción del árbol se realiza siguiendo un enfoque de división binaria recursiva, sea N_j el número de casos en la clase j y $\pi(j) = \frac{N_j}{N}$ las probabilidades de que un dato en la clase esté presente en el árbol, donde N es el número de datos. El estimador de probabilidad de que un caso esté en la clase j dado que se ubicó en el nodo t; esta dado por:

$$p(j \mid t) = \frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N_j} \tag{1}$$

y cumple:

$$\sum_j p(j \mid t) = 1 \tag{2}$$

así, las $p(j \mid t)$ son las proporciones relativas de los casos en la clase j en el nodo t (3).

El objetivo de la investigación es identificar las variables meteorológicas predominantes en las clases (intervalos de horas), mediante un gráfico, (árbol de clasificación), mismo que permite interpretar resultados importantes.

2. Metodología

2.1. Los datos

Para el estudio se empleó una matriz de dimensión 8785x8, es decir 70280 observaciones. Los datos corresponden a información por hora del año 2016, de la estación meteorológica de Alao; ubicada 3064 ms.m en la provincia de Chimborazo. Las variables analizadas fueron: temperatura del aire (℃), humedad relativa (%), presión barométrica (hPa), radiación solar difusa (w/m^2), radiación solar global (w/m^2), temperatura del suelo a $-20cm$ (℃) y velocidad de viento (km/h).

Con el objetivo de identificar variables predominantes cada dos horas se generó una variable categórica denominada ESCALA, misma que se describe en la Tabla. (1).

La base de datos se almacenó en formato CSV y el estudio se realizó mediante código desarrollado en el software estadístico R; para lo cual se procedió a cargar el

TABLE 1: Descripción de la variable ESCALA.

ESCALA	Intervalo
I	06:00 - 08:00
II	09:00 - 11:00
III	12:00 - 14:00
IV	15:00 - 17:00
V	18:00 - 20:00
VI	21:00 - 23:00
VII	24:00 - 02:00
VIII	03:00 - 05:00

fichero de datos; se realizó una limpieza para eliminar valores NA's y se ejecutaron los comandos.

Antes de generar el modelo se dividió la base de datos en una tabla de aprendizaje y una tabla de prueba (testing), la primera determina el árbol de decisión y la segunda permite validarlo; es decir verifica si el modelo es bueno o aceptable seleccionando individuos que no fueron parte en la construcción (8). En el estudio se consideró el 80% de la información para la tabla de aprendizaje y el 20% para el testing.

Mediante las funciones *prop.table()* y *table()*, se verificó la aleatoriedad en el conjunto de datos. El modelo se construyó mediante la librería *rpart* y para la visualización del árbol se aplicó *rpart.plot*. Una vez entrenado el modelo se predijeron nuevas instancias, empleando la tabla testing; para ello se utilizó la función *predict()*.

2.2. Matriz de confusión

Para obtener información de las predicciones realizadas por el modelo se obtuvo la matriz de confusión, misma que compara el conjunto de variables meteorológicas de la tabla testing con la predicción obtenida, versus la escala a la que los datos realmente pertenecen (8). Las columnas de la matriz determinan el número de predicciones de cada clase (intervalos de tiempo), y las filas representan las instancias clasificadas en la clase real (12). A continuación, la Tabla. (2) muestra un caso particular de una matriz de confusión con dos clases:

donde:

VN: Verdaderos Negativos

VP: Verdaderos Positivos

FN: Falsos Negativos; y

TABLE 2: Matriz de confusión con dos clases.

		Predicción	
		Negativo	Positivo
Valor Real	Negativo	VN	FP
	Positivo	FN	VP

FP: Falsos Positivos

2.3. Rendimiento del modelo

El rendimiento del modelo se determinó a partir de la matriz de confusión; se calculó la Exactitud (Accuracy), misma que está dada por la ecuación:

$$Accuracy = \frac{VN + VP}{VN + FP + FN + VP} \tag{3}$$

Esta medida determina la proporción de las instancias predichas correctamente VN y VP sobre la suma total de elementos evaluados (13). Estadísticamente la exactitud está relacionada con el sesgo de una estimación, cuando menor es el sesgo más exacta es la estimación.

2.4. Pureza de un nodo

Se dice que un nodo es puro, si todos los individuos con una determinada característica caen dentro de la clase que define esa condición, caso contrario se dice que el nodo es impuro; mientras que un nodo es completamente impuro si el 50% de los individuos caen en la clase y el restante porcentaje no pertenece al grupo. Para identificar la impureza de un nodo, se consideró el error de clasificación y el índice de Gini.

Sea $p(j \mid t)$ la probabilidad de que un caso esté en la clase j dado que se ubicó en el nodo t, se define el error de clasificación mediante la ecuación:

$$Error(t) = 1 - max[p(j \mid t)] \tag{4}$$

y el índice de Gini por:

$$Gini(t) = 1 - \sum_{j} \left[p(j \mid t)\right]^2 \tag{5}$$

Una vez calculado el índice de Gini, se determina en cada nodo el valor total del índice de Gini, conocido como Gini split; mismo que se obtiene mediante la ecuación:

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i) \tag{6}$$

En el estudio se trabajaron con variables numéricas y una división binaria; por lo que, si X_{max} y X_{min} son el máximo y mínimo valor de una de las variables, entonces se prueban con todos los valores comprendidos en el intervalo hasta que uno de ellos minimice el índice de Gini, lo cual permite maximizar la información ganada (14).

Un gráfico que permite identificar si el error decrece o se incrementa es el que muestra el número de interacciones versus el error. Si la figura muestra que a partir de un punto el error sube nuevamente, entonces se procede a la poda del árbol (15), técnica que consiste en cortar ramas o nodos terminales hasta encontrar el árbol adecuado al conjunto de datos. Una técnica empleada es hallar un conjunto de árboles de tamaños decrecientes, mismos que luego son comparados para determinar el adecuado, para lo cual se emplea una función denominada costo de complejidad.

3. Resultados y Discusión

La Fig. (1), muestra el árbol de clasificación de la estación Alao; el 12% de los datos pertenecen a la variable radiación solar global dentro de la clase I, si es mayor o igual a $120w/m^2$, entonces se puede determinar la presión barométrica dentro de la clase II, es decir a horas comprendidas entre las 9 y 11 de la mañana; si ésta es mayor o igual que $709w/m^2$, entonces se puede predecir la temperatura del aire dentro de la misma clase con una probabilidad del 66%, y un 18% de los datos; y si esta es menor que $13℃$ entonces hay un 71% de probabilidad de que de que las variables se encuentren en la clase II con un 11% de la data.

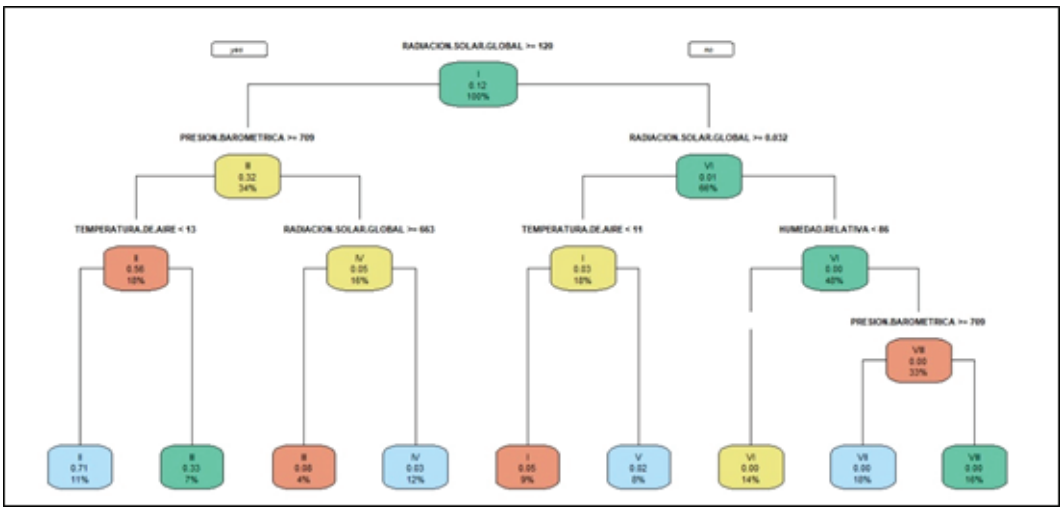

Figure 1: Árbol de clasificación, estación Alao.

La Tabla. (3), muestra la matriz de confusión del modelo. Se puede evidenciar que en la clase I 135 observaciones fueron clasificadas correctamente; mientras que 89 lo

hicieron en otros grupos, en la clase II; 124 datos se clasificaron adecuadamente, en la tercera lo hicieron 115, en la cuarta 119, 107 en la quinta, en la clase VI 132, en la VII 124 y en el horario de 03h00 a 05h00 (clase VIII); 152 datos fueron bien clasificados, 59 están el grupo VII, 14 en el VI y 3 en la clase I.

Table 3: Matriz de confusión.

	I	II	III	IV	V	VI	VII	VIII
I	135	23	0	5	2	0	26	33
II	10	124	44	12	2	0	0	0
III	0	41	115	41	0	0	0	0
VI	0	6	22	119	8	0	0	0
V	10	0	0	10	107	68	5	7
VI	1	0	0	0	0	132	81	20
VII	5	0	0	0	1	40	124	51
VIII	3	0	0	0	0	14	59	152

Table 4: Estadísticas del modelo.

Nodo	CP	nsplit	Rel error	xerror	Xstad
1	0.134028	0	1.00000	1.00556	0.0047
2	0.100868	1	0.86597	0.87205	0.0060
3	0.089757	2	0.76510	0.77205	0.0066
4	0.080556	3	0.67535	0.69948	0.0069
5	0.073264	4	0.59479	0.61441	0.0070
6	0.048264	5	0.52153	0.52587	0.0070
7	0.019271	6	0.47326	0.47778	0.0069
8	0.018750	7	0.45399	0.47049	0.0069
9	0.010000	8	0.43524	0.44896	0.0068

Mediante la matriz de estadísticas, Tabla. (4); se puedo observar que el error desciende y se mantiene; resultado que se puede corroborar mediante el gráfico del error (Fig. 3), esta consecuencia indica que el modelo se estabiliza, por lo que el árbol no necesita de una poda. La Fig. (2), muestra el error por nodo, mismo que es del 86.83%.

En la Fig. (4) se muestra el resultado de la exactitud, se observa que el rendimiento del modelo es del 60,79% de datos predichos correctamente en relación a la suma total de observaciones evaluadas.

```
Variables actually used in tree construction:
[1] HUMEDAD.RELATIVA       PRESION.BAROMETRICA    RADIACION.SOLAR.GLOBAL
[4] TEMPERATURA.DE.AIRE

Root node error: 5760/6633 = 0.86839
```

Figure 2: Estadísticas, estación Alao.

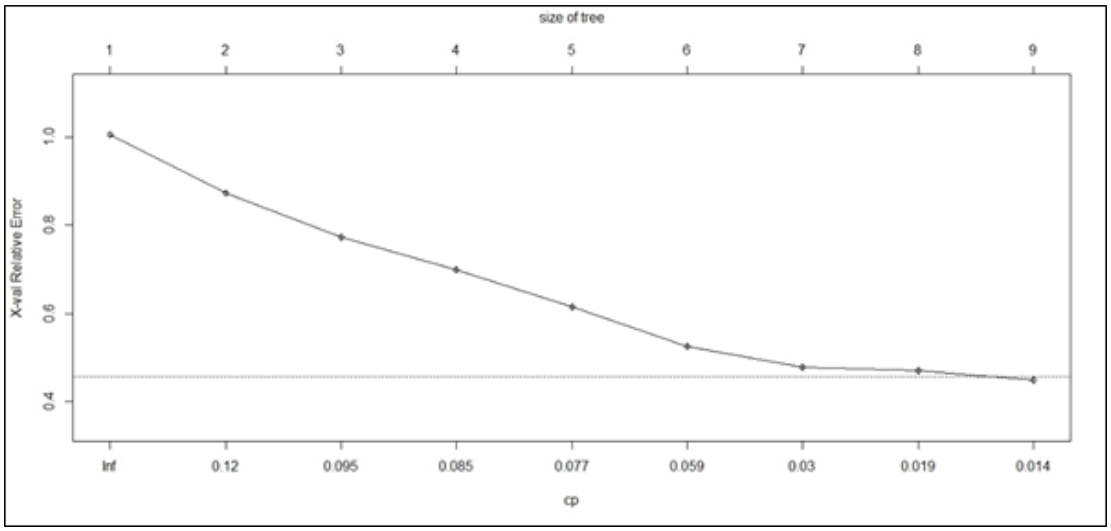

Figure 3: Gráfica del error, estación Alao.

```
> #Rendimiento del modelo
> accuracy_Test <- sum(diag(table_mat)) / sum(table_mat)
> print(paste('Accuracy for test', accuracy_Test))
[1] "Accuracy for test 0.607961399276236"
```

Figure 4: Exactitud, estación Alao.

4. Conclusiones

De acuerdo al árbol generado por el modelo se pudo determinar que las variables meteorológicas predominantes en la estación Alao fueron: radiación solar global, presión barométrica y temperatura del aire. La clase IV, con horas comprendidas entre las 15h00 y 17h00; fue la que registró mayor frecuencia en los nodos del árbol. Alao tiene un clima templado y cálido, es por ello que el estudio fue importante, pues se identificaron horarios donde la radiación y temperatura son altas, lo cual puede tener repercusiones en la salud de sus habitantes.

El modelo tiene un rendimiento de aproximadamente el 61%, lo cual podría ser de utilidad para pronóstico de sucesos futuros, pues un aporte que genera esta técnica de clasificación en minería de datos, es la predicción; por lo que con información recaba en estos últimos años se puede identificar si han existido variantes en la zona de estudio.

El árbol de clasificación es una técnica que permitió visualizar de manera sencilla, variables meteorológicas relevantes a distintos intervalos de tiempo, esta información permite predecir situaciones climáticas en la zona de estudio. A futuro se pretenden realizar estudios similares que involucren las restantes zonas donde se encuentran ubicadas las estaciones meteorológicas monitoreadas por el Centro de Energías Alternativas y Ambiente.

Agradecimientos

Al Grupo de Energías Alternativas y Ambiente de la Facultad de Ciencias (CEAA), y a su director Dr. Celso Recalde.

Conflicto intereses

No existen intereses particulares por parte del autor que puedan afectar directa o indirectamente a los resultados obtenidos en la investigación.

References

[1] Riqueime J, Ruiz R, Gilbert K. Minería de Datos: Conceptos y Tendencias. Iberoamericana de Inteligencia Artificial. 2006; p. 11-18.

[2] Eckert K, Suénaga R. Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. Formación Universitaria. 2015; p. 3-12.

[3] Bortolini J, Alonso P, Álvarez F. Árboles de clasificación de Potimirim mexicana (Decapoda: Caridea), organismo Hermafrodita protándrico secuencial. Latin American Journal of Aquatic Research. 2013; p. 739-745.

[4] Bouza C, Santiago A. La minería de datos: Árboles de decisión y su aplicación en estudios médicos. CHAP. 2012; p. 64-78.

[5] Ruiz E, Romero C. Resultados obtenidos en un proceso de minería de datos aplicado a una base de datos que contiene información bibliográfica referida a cuatro segmentos de la ciencia. Journal of Information Systems and Technology Management. 2018; p. 1-11.

[6] Uvidia M, Cisneros A, Viñán J. Minería de datos de la evaluación integral del desempeño académico de la unidad de nivelación. Descubre. 2017; p. 44-54.

[7] Coronado M, Bianchi V, Vivas J, Perera M. Estudio comparativo de técnicas de minería de datos para la predicción de rutas de huracanes. CONAIC. 2017; p. 43-52.

[8] Haro S, Pazmiño R, Conde M, Peñalvo F. Data mining to discover the classification trend in titling works. En: XIII Congreso de Ciencia & Tecnología, ESPEQuito: Universidad de las Fuerzas Armadas, ESPE; 2018; p. 125-128.

[9] Witten I, Frank E, Hall M, Pal C. Data mining: Practical Machine Learning Tools and Techniques: Elsevier Inc.; 2016.

[10] Robles Y, Sotolongo A. Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL. Gestión de Tecnología y Sistemas de Información. 2013; p. 389-406.

[11] Valero S, Vargas A, García M. 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. 2010; p. 30-33.

[12] Graham W. Data Mining with Rattle and D New York: Springer; 2011.

[13] Haro S, Zúñiga L, Meneses A, Vera L, Escudero A. Métodos de clasificación en minería de datos meteorológicos. Perfiles. 2018; p. 107-113.

[14] Raileanu L, Stoffel K. Theoretical Comparison between the Gini Index and Information Gain Crieteria. The Swiss National Science Foundation.

[15] Gámez M, Cortés E, Alfaro JL, García N. Árboles de clasificación para el análisis de gráficos de control multivariante. Matemática: Teoría y aplicaciones. 2008; p. 30-42.