Tecnura
ISSN: 0123-921X
tecnura@udistrital.edu.co
Universidad Distrital Francisco José de Caldas
Colombia

Rosado Gómez, Alveiro Alonso; Verjel Ibáñez, Alejandra
Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander
Tecnura, vol. 19, núm. 45, julio-septiembre, 2015, pp. 101-113
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Tecnura

http://revistas.udistrital.edu.co/ojs/index.php/Tecnura/issue/view/650

DOI: http://dx.doi.org/10.14483/udistrital.jour.tecnura.2015.3.a08

INVESTIGACIÓN

# Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander

## Data mining applied to demand of air transport in Ocaña, North Santander

Alveiro Alonso Rosado Gómez*, Alejandra Verjel Ibáñez**

**Fecha de recepción:** 2 de octubre de 2014 **Fecha de aceptación:** 6 de abril de 2015

**Como citar:** ARosado Gómez, A. A., & Verjel Ibáñez, A. (2015). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Revista Tecnura*, *19*(45), 101-113. doi: 10.14483/udistrital.jour.tecnura.2015.3.a08

### RESUMEN

Este artículo muestra la aplicación de la minería de datos para predecir la demanda del servicio aéreo en Ocaña, Norte de Santander, respecto a los pares origen-destino Ocaña-Bogotá, Ocaña-Bucaramanga, Ocaña-Medellín, Ocaña-Cúcuta, Ocaña-Barranquilla; se utilizan datos de estudios realizados para estimar la demanda de un nuevo medio de transporte en la ciudad. Esta investigación sigue las fases del proceso de extracción del conocimiento en bases de datos. En la etapa de minería de datos se seleccionó como técnica los árboles de decisión y como algoritmo el J48. Con la aplicación de este algoritmo se encontró que las personas que viajan mensual y semanalmente eligen el avión, igual que si viajan por motivos de salud; si trabajan y se dirigen a Barranquilla, Bogotá y Medellín, eligen la buseta.

***Palabras clave*:** árboles de decisión, minería de datos, modos de transporte, WEKA.

### ABSTRACT

This paper shows data mining techniques to predict air tickets demand in Ocaña, North Santander, regarding to origin-destination: Ocaña-Bogota, Ocaña–Bucaramanga, Ocaña–Medellín, Ocaña – Cúcuta and Ocaña-Barranquilla. Some data studies were used to estimate new means of transportation demand in Ocaña. This research follows the process of knowledge extraction in databases. In the data mining process stage, it was selected decision trees technique and as algorithm J48. With this algorithm application, we found that people who travel monthly and weekly choose the plane, alike if they travel by health issues; otherwise, but if they travel to Barranquilla, Bogotá and/or Medellin for work, they choose bus.

***Keywords*:** data mining, decision trees, modes of transport, WEKA.

* Ingeniero de sistemas, magister en Gestión, Aplicación y Desarrollo de Software. Director del Departamento de Sistemas e Informática, Universidad Francisco de Paula Santander Ocaña, Ocaña, Colombia. Contacto: aarosadog@ufpso.edu.co

** Ingeniera de sistemas. Profesional de apoyo, Universidad Francisco de Paula Santander, Ocaña, Colombia. Contacto: averjeli@ufpso.edu.co

## INTRODUCCIÓN

La minería de datos ha sido de gran utilidad para la extracción de conocimiento en muchos ámbitos gracias a su función, ya que es una de las formas más sofisticadas de extraer información importante y relevante a partir de una base de datos (Roland, Uhrmacher, & Saha, 2009), utilizando técnicas para encontrar patrones y crear modelos con dicha información.

Es primordial para el éxito de los negocios conocer el comportamiento de las tendencias y de las personas, por esto la minería de datos se convierte en una herramienta muy útil para identificar las preferencias sobre un determinado bien o servicio, prediciendo y extrayendo información importante, de modo que permite descubrir conocimientos que con otros métodos no es posible (Ferri, Flach, & Hernández-Orallo, 2002).

Este artículo muestra una aplicación de las técnicas de minería de datos, en la cual es posible determinar el comportamiento de la elección modal de transporte interurbano en el municipio de Ocaña, ante escenarios hipotéticos de elección dada la implementación de rutas en transporte aéreo que conducen a los cinco destinos más frecuentes: Bogotá, Medellín, Barranquilla, Bucaramanga y Cúcuta. Las técnicas de minería de datos utilizadas emplean las etapas del proceso de extracción del conocimiento en base de datos —Knowledge Discovery in Databases (KDD)— correspondientes a: selección de datos, preprocesamiento, transformación, minería de datos e interpretación y evaluación (Fayyad, Piatetsky, Smyth, & Uthurusamy, 1996).

Para el municipio de Ocaña, Norte de Santander, es muy importante que exista una nueva alternativa de viaje como lo es, en este caso, el modo aéreo, dado que solo cuenta con transporte vía terrestre (Alcaldía de Ocaña, 2010); por esta razón, muchas personas que deseen tener más comodidad a la hora de viajar o quieran una duración corta de viaje no tienen la opción de escoger entre varios modos al existir solo uno.

Para procesar toda la información se realizó un análisis de las técnicas de predicción disponibles en el entorno para análisis del conocimiento de la Universidad de Waikato, Waikato Environment for Knowledge Analysis (WEKA), se eligió el mejor algoritmo, se generó el modelo y se validó para evaluar su capacidad de clasificar correctamente las instancias.

### Técnicas de minería de datos y su aplicación en modelos de transporte

Dentro de las técnicas de minería de datos, se encuentran las predictivas y las descriptivas (Quinlan, 1986). En esta investigación se emplearon para el análisis las técnicas predictivas, las cuales tienen tareas de clasificación y regresión (Kohavi & Quinlan, 2002), por lo tanto, para la selección de la técnica y el algoritmo más adecuado se estudió el conjunto de datos por medio de métodos bayesianos, árboles de decisión, redes neuronales y regresión logística (Caruana & Niculescu, 2006); el análisis discriminante no se tuvo en cuenta, ya que no está disponible en la versión utilizada de WEKA (Hall, Eibe, Holmes, Pfahringer, Reutemann, & Witten, 2009), siendo este el software utilizado en el proceso de investigación.

En la literatura se reportan algunas experiencias del uso de la minería de datos en modelos de transporte y aplicaciones afines, como es el caso de Arentze y Timmermans (2007), quienes usaron técnicas basadas en reglas, tales como árboles de decisión, ideales para representar los efectos discontinuos de las variables independientes sobre el comportamiento de elección discreta en transportes o sistemas espaciales; de esta forma se muestra cómo dicho enfoque reemplaza al modelo logit convencional, multinomial logistic regression (MNL) (Akiva & Lerman, 1985) y es utilizado para incorporar la sensibilidad del viaje a atributos como el costo. Los resultados indican que el modelo puede reproducir rangos realistas de elasticidades precio de la demanda de viajes.

Xie, Lu y Parkany (2003) decidieron enfocar su investigación con la premisa de que la elección del modo de transporte puede ser considerada como un problema de reconocimiento de patrones en los que el comportamiento humano refleja múltiples patrones a partir de variables explicativas que determinan las elecciones entre alternativas o clases. Este trabajo investigó la capacidad y el rendimiento en la elección del modo de viaje a partir de dos técnicas, como son los árboles de decisión y patrones de redes neurales. Los resultados de la predicción muestran que los dos modelos de minería de datos son de rendimiento comparable, pero ligeramente mejores que el modelo MNL (el tradicional modelo de elección discreta de raíces econométricas) en términos de los resultados de los modelos; mientras que el modelo de árboles de decisión demuestra eficacia de más alta estimación e interpretación más explícita, el modelo de redes neuronales da una predicción de rendimiento superior en la mayoría de los casos.

Investigaciones más recientes, como las reportadas por De Oña, J., De Oña, R., & Calvo, F. (2012), han aplicado técnicas de minería de datos para estudiar y mejorar los servicios en términos de calidad del servicio para la operación de tránsito en España. La metodología utilizada fue un modelo de árbol de decisión, el cual no necesita ni las hipótesis del modelo ni relaciones predefinidas entre las variables independientes y las variables dependientes. Siguiendo esta dirección y debido a los poderosos resultados obtenidos con el modelo de árbol de decisión, el interés de los autores por otras técnicas de minería de datos se incrementó.

## METODOLOGÍA

### Selección, preprocesamiento y transformación de datos

En las primeras etapas del proceso de extracción del conocimiento en base de datos, se seleccionó el conjunto de datos por evaluar, obtenidos de un estudio previo (Guerrero, Criado, & León, 2013), en el cual se tuvieron en cuenta algunos datos del último viaje interurbano realizado, características socioeconómicas del viajero y del viaje realizado, teniendo como etiqueta de clase el atributo *choice*, el cual incluye los modos de transporte disponibles para realizar dicho estudio, que son el avión (como caso hipotético), el taxi, el bus y la buseta como opciones disponibles en la actualidad.

Para un análisis adecuado en minería de datos es necesario definir los atributos relevantes; es decir, establecer un filtro para seleccionar los atributos que influyen sobre el valor de la clase antes de empezar con el aprendizaje para asegurar la calidad de los datos y mejorar el desempeño predictivo (Castillo, Mendoza, & Poblete, 2011). En este caso, se utilizaron algoritmos de búsqueda y rankeo (Hall & Holmes, 2003), posteriormente se seleccionaron los datos por la frecuencia con la cual aparecen y así tomar los más relevantes (Van Hulse, M. Khoshgoftaar, & Napolitano, 2007) y descartar los atributos que tienen menos de cinco veces de aparición en los resultados de los algoritmos implementados (Guyon & Elisseeff, 2003).

Luego de seleccionar los atributos que intervienen en la clase, se realizó su discretización transformando los datos nominales a discretos formando intervalos (Agrawal & Aggarwal, 2001), debido a que los algoritmos de predicción disponibles en WEKA solo reconocen este tipo de datos (Goebel & Gruenwald, 1999).

### Minería de datos

En esta etapa se realizó un análisis de cada uno de los algoritmos pertenecientes a las técnicas predictivas (Fayyad, Piatetsky, Smyth, & Uthurusamy, 1996), como los métodos bayesianos, los árboles de decisión, las redes neuronales y la regresión logística (Quinlan, 1989). Se aplicó como opción de prueba *percentage split*, con el 80% para la construcción del modelo y el 20% para la evaluación ( Baumgartner & Serpen, 2009), teniendo en cuenta que se emplea como etiqueta de clase el *choice*, es

decir, la elección del modo de transporte, como el avión, el bus, la buseta y el taxi. En las tablas 1, 2, 3 y 4 se presentan los algoritmos de cada técnica con su respectivo porcentaje de acierto y de error.

**Tabla 1.** Árboles de decisión

| Algoritmo | % de acierto | % de error |
|---|---|---|
| BFTree | 70% | 30% |
| Decision Stump | 46,66% | 53,33% |
| Id3 | 70% | 26,66% |
| **J48** | **77,77%** | **22,2222%** |
| LAD Tree | 72,22% | 27,77% |
| LMT | 71,88% | 28,11% |
| NBTree | 67,77% | 32,22% |
| Random Forest | 66,66% | 33,33% |
| Random Tree | 64,44% | 35,55% |
| REP Tree | 64,44% | 35,55% |
| Simple Cart | 73,33% | 26,66% |

**Fuente:** elaboración propia.

**Tabla 2.** Métodos bayesianos

| Algoritmo | % de acierto | % de error |
|---|---|---|
| Bayes Net | 57,77% | 42,22% |
| Naive Bayes | 56,66% | 43,33% |

**Fuente:** elaboración propia.

**Tabla 3.** Redes neuronales

| Algoritmo | % de acierto | % de error |
|---|---|---|
| Multilayer Perceptron | 67,7778% | 32,2222% |

**Fuente:** elaboración propia.

**Tabla 4.** Regresión logística

| Algoritmo | % de acierto | % de error |
|---|---|---|
| Logistic | 75,5556% | 24,4444% |

**Fuente:** Elaboración propia

Todos estos porcentajes demuestran la eficacia del algoritmo J48 para realizar el análisis al conjunto de datos, con un acierto de 77,77% y un 22,22% de error, por lo tanto se seleccionó esta técnica para extraer el conocimiento.

El algoritmo J48 extiende las funcionalidades del algoritmo C4.5, como aceptar la realización del proceso de poda por medio de la reducción del error (Yoav & Schapire, 1996); los árboles de decisión muestran organizados eficientemente los atributos, teniendo en cuenta la entropía, la cual está dada por la ecuación (1).

$$P_n = (1 - P_p) \quad (1)$$

Siendo,

$p_p$ = Es la probabilidad de que las respuestas sean positivas.

$p_n$ = Es la probabilidad de que las respuestas sean negativas.

La entropía se define con base en las probabilidades anteriores (ecuación (2)).

$$H(S) = -p_p - log_2 p_p - p_n log_2 p_n \quad (2)$$

Cuanto menor sea el valor de la entropía, más ordenados se encuentran los datos; para clasificar los datos se utiliza la ganancia de información, la cual es la encargada de reducir la entropía y decidir qué atributo es el más apropiado para usar en cada nodo del árbol (Martínez, Solarte, & Soto, 2011). La fórmula de ganancia de información está dada por la ecuación (3).

$$G(S, A) = H(S) - H(S, A) \quad (3)$$

Siendo,

$H(S)$ = Entropía de S.

$H(S, A)$ = Sumatoria de entropías.

El resultado del algoritmo J48 generó un árbol de 71 hojas y un tamaño de 107 (suma de nodos

internos y nodos hoja), y el tiempo para construir el modelo fue de 0,02 segundos.

El árbol se generó con 450 instancias, de las cuales 360 fueron utilizadas para construir el modelo, es decir, el 80% y 90 para la evaluación con el 20%.

## RESULTADOS

### Interpretación y evaluación

El árbol de decisión generado tuvo como nodo principal la disponibilidad del bus, como lo muestra la figura 1. Cuando las personas valoran el tiempo eligen el avión, mientras que si valoran más el costo que el tiempo eligen la buseta; así mismo, si viajan mensual o semanalmente eligen el avión, y si lo hacen de manera eventual eligen el bus como modo de transporte; cuando el motivo del viaje es por salud eligen el modo hipotético, en este caso, el avión (figura 2).

El resultado de la ejecución del algoritmo J48 en WEKA se puede ver a continuación en la tabla 5. Donde se analizaron diferentes parámetros, como las *instancias correctamente clasificadas*, *las incorrectas*, *kappa statistic*, *mean absolute error*, *root mean squared error*, *relative absolute error* y *root relative squared error*.

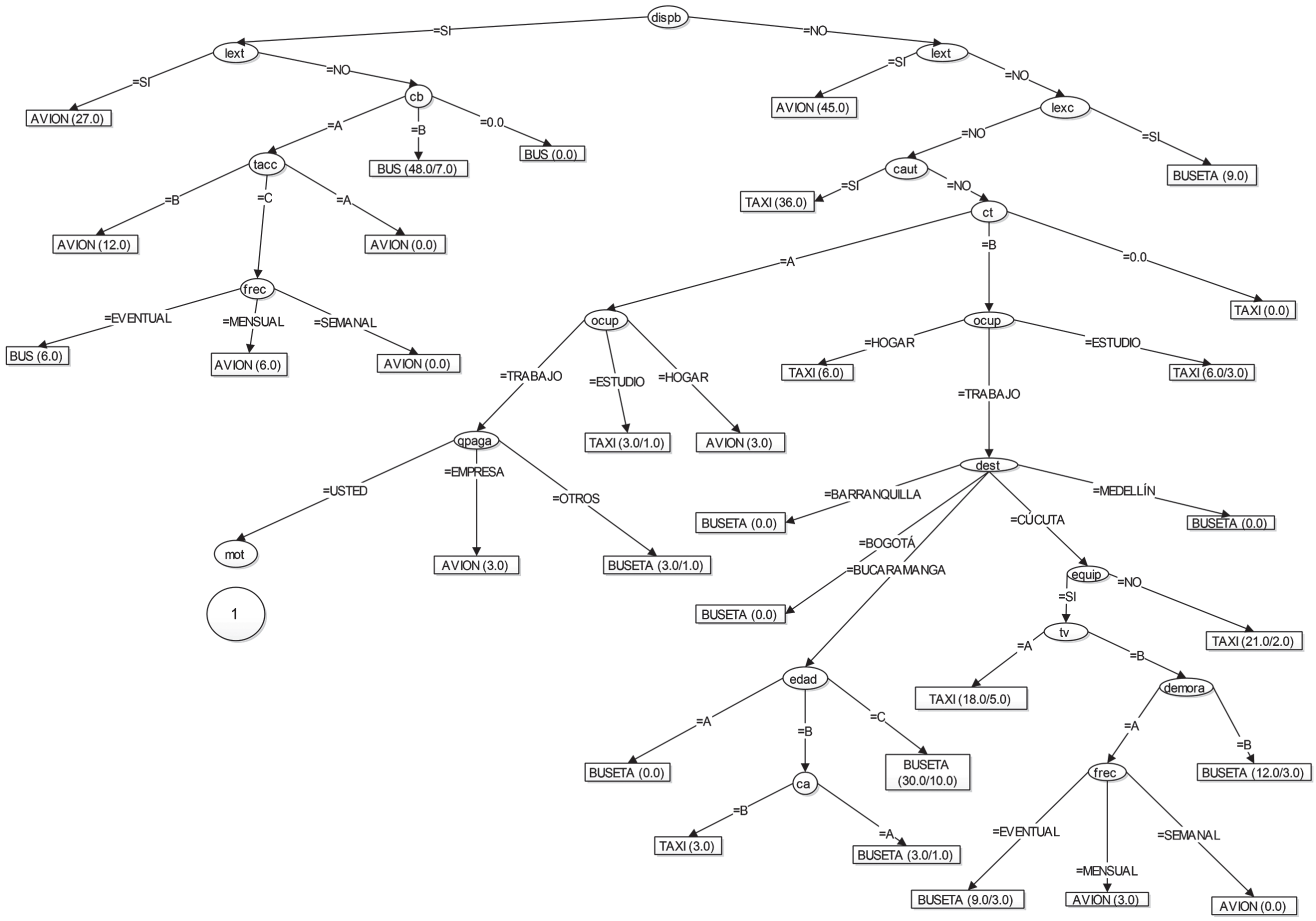


**Figura 1.** Árbol generado vista superior.

**Fuente:** elaboración propia.

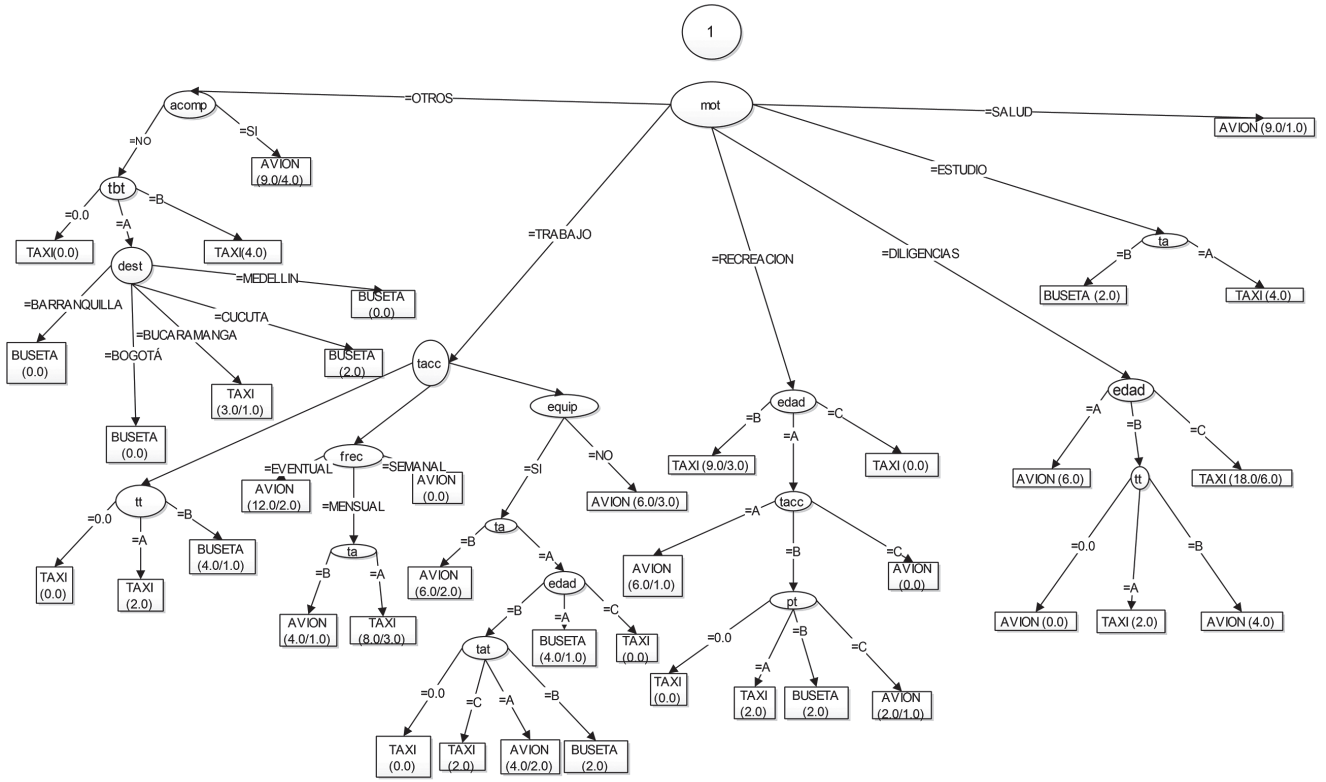


**Figura 2.** Árbol generado vista inferior (continuación).

**Fuente:** elaboración propia.

**Tabla 5.** Evaluación test Split.

| === Evaluation on Test Split === | | |
|---|---|---|
| === Summary === | | |
| Correctly Classified Instances | 70 | 777778% |
| Incorrectly Classified Instances | 20 | 222222% |
| Kappa statistic | 0,6822 | |
| Mean absolute error | 0,1454 | |
| Root mean squared error | 0,3105 | |
| Relative absolute error | 416532% | |
| Root relative squared error | 744492% | |
| Total Number of Instances | 90 | |

**Fuente:** elaboración propia.

## Análisis de parámetros

Las instancias correctamente clasificadas son 70 frente a 20 que fueron clasificadas de forma incorrecta. El total del número de instancias fue de 90.

Los resultados arrojados luego de la ejecución del algoritmo J48 mostraron que el coeficiente de Kappa tuvo un valor de 0,6822, lo cual indica un nivel alto de concordancia entre los datos de prueba y los clasificados por el modelo, ya que cuanto más cercano a 1, mejor es la coincidencia de los datos. El error absoluto medio tuvo un valor de 0,1454, lo cual demuestra que la diferencia entre las instancias de prueba y los valores reales son bajos, es decir, esto revela la eficiencia del modelo (Bresfelean, 2007).

Dentro de los valores numéricos se encuentran el *root mean squared error* con un valor de 0,3105, el *relative absolute error* con un porcentaje de 41,65% y el *root relative squared error* con un porcentaje de 74,44%. Todos estos valores se utilizan para la predicción numérica en lugar de la clasificación; en la predicción numérica, estos errores reflejan una magnitud (Nishimura & Hirose, 2007).

## Detalle de precisión por clase

Los detalles de precisión por clase son otra parte del resultado de la ejecución del algoritmo J48 en WEKA. Estos resultados se pueden visualizar en la tabla 6, la cual muestra los verdaderos positivos (*TP rate*), lo falsos positivos (*FP rate*), la precisión, el *recall*, el *F-Measure* y el área bajo la curva, Receiver Operating Characteristic (ROC).
Como se puede observar en la tabla 6, los verdaderos positivos (*TP rate*) para la clase AVIÓN superan el 74,3%, lo cual quiere decir que el árbol clasifica correctamente las instancias (Witten, Hall, Holmes, Pfahringer, & Reutemann, 2009).

Los falsos positivos (*FP rate*) tienen valores bajos, lo cual demuestra que el modelo pocas veces no clasifica correctamente las instancias (Fawcett, 2004).

Con respecto a la precisión, el porcentaje de acierto del modelo luego de hacer las clasificaciones en cada clase, se puede observar que los valores son elevados, lo cual demuestra que el modelo mide de la mejor manera las instancias correctamente reconocidas respecto al total de instancias predichas.

En los resultados de la cobertura (*recall*) se puede observar que son altos, lo que quiere decir que son favorables porque reconoce las instancias correctamente en cuanto a los términos reales; está dada por:

El *recall* y la precisión están relacionadas entre sí, ya que cuando aumenta el *recall* (la cobertura) disminuye la precisión o al contrario, si disminuye la cobertura aumenta la precisión. Se puede notar que en la clase AVIÓN el *recall* (la cobertura) disminuye, en la clase BUS, se mantiene en 80%, mientras que aumenta en las clases TAXI y BUSETA.

El *F-Measure* muestra la bondad del modelo, en la que cuanto más cercana sea a 1, mayor será la confiabilidad del modelo. Tal como se observa en la tabla 6, el modelo demuestra ser confiable porque todos los valores de la clase se acercan a 1. El *F-Measure* está dado por la ecuación (4).

$$F - Measure = \frac{2 \times recall \times precision}{recall + precision} \quad (4)$$

El ROC area o área bajo la curva entre los verdaderos positivos (eje Y) y los falsos positivos (eje X), cuanto más cercano sea a 1 el test es visto como excelente (Ferri, Flach, & Hernández-Orallo, 2002). En esta investigación los resultados obtenidos son favorables, ya que la mayoría están cercanos a 1. En la tabla 6 se observan también los valores de la curva de todas las clases; notablemente, se puede demostrar la confiabilidad del modelo en cuanto todos los resultados son bastante cercanos a 1. Igualmente, se muestra en las figuras 3, 4, 5 y 6 las gráficas correspondientes a cada clase.

**Tabla 6.** Detalles de precisión por clase

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC area | Class |
|---|---|---|---|---|---|---|
| 0,743 | 0,109 | 0,813 | 0,743 | 0,776 | 0,844 | AVIÓN |
| 0,8 | 0,025 | 0,8 | 0,8 | 0,8 | 0,979 | BUS |
| 0,806 | 0,119 | 0,781 | 0,806 | 0,794 | 0,866 | TAXI |
| 0,786 | 0,066 | 0,688 | 0,786 | 0,733 | 0,92 | BUSETA |

**Fuente:** elaboración propia.

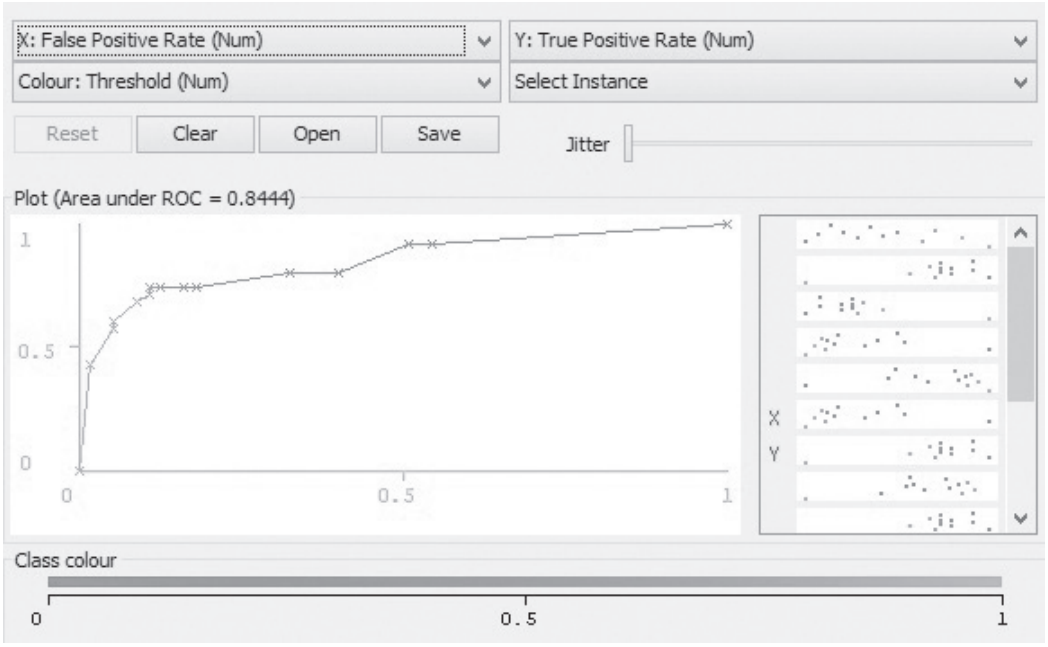


**Figura 3.** ROC area-Avión

**Fuente:** elaboración propia.

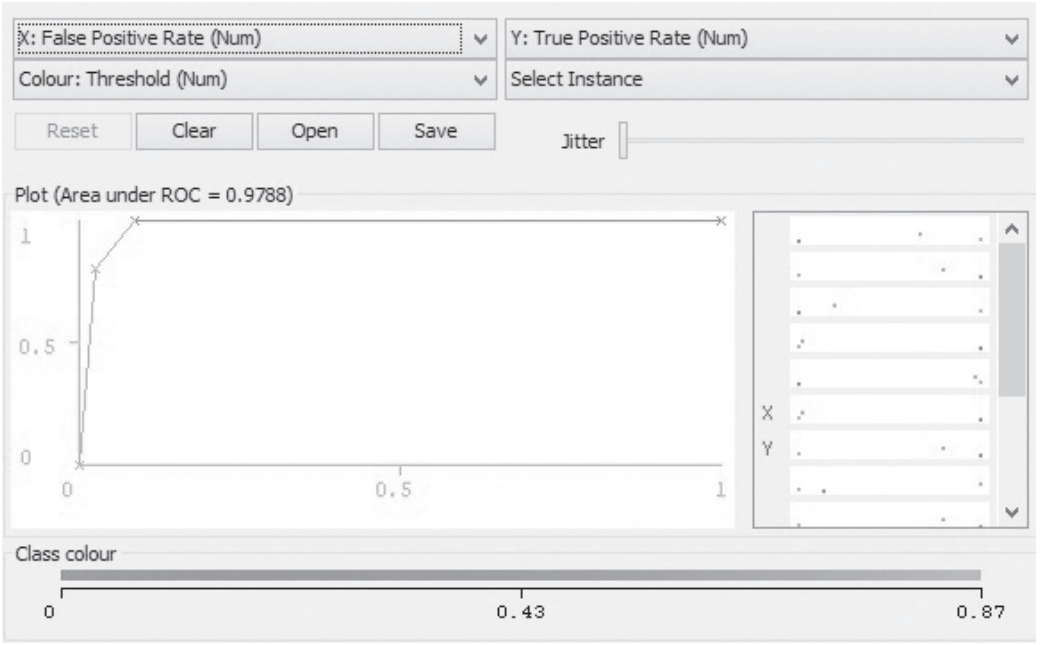


**Figura 4.** ROC area-Bus.

**Fuente:** elaboración propia.

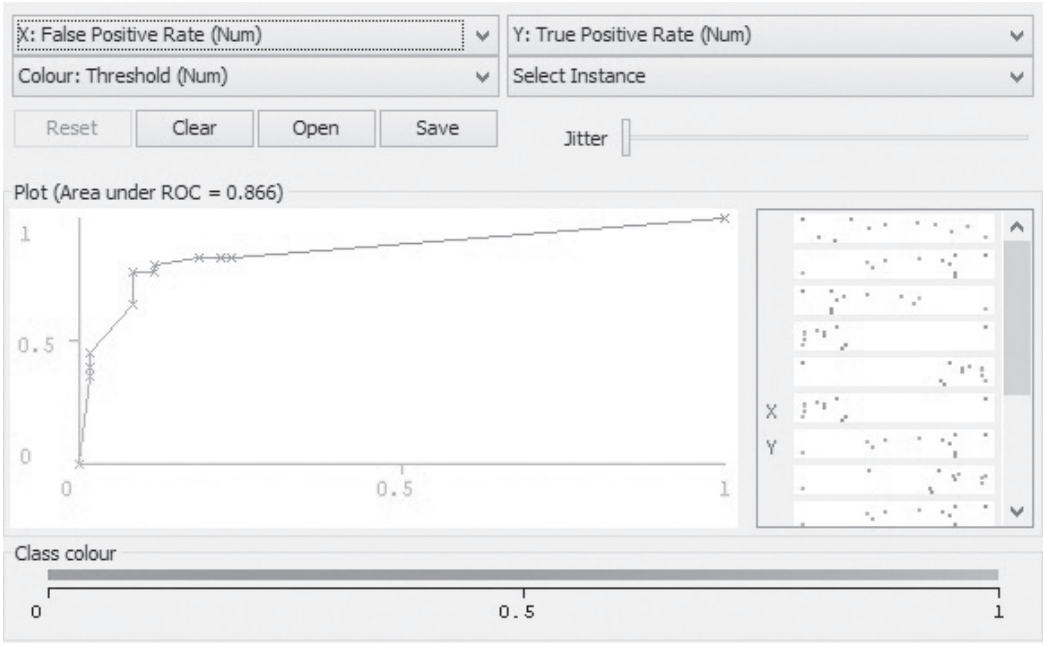


**Figura 5.** ROC area-Taxi.

**Fuente:** elaboración propia.

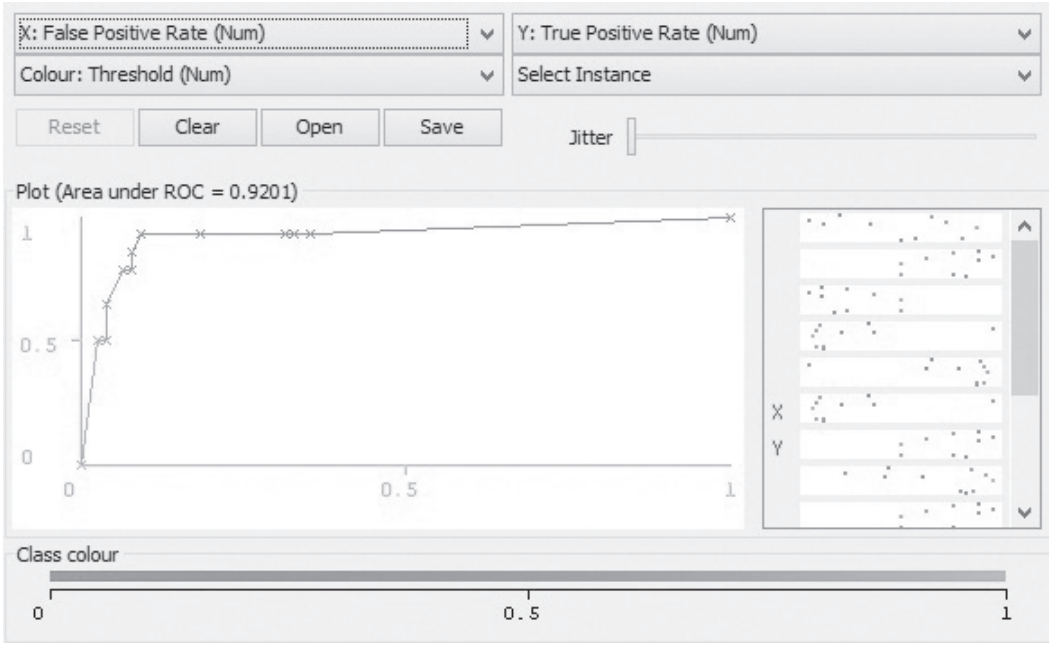


**Figura 6.** ROC area-Buseta.

**Fuente:** elaboración propia

## Matriz de confusión

La matriz de confusión resultante del análisis se puede ver en la tabla 7, en la cual se estableció las instancias clasificadas correctamente, que son las que están en la diagonal; el resto de valores son los errores.

**Tabla 7.** Matriz de confusión

| === Confusion Matrix === | | | |
|---|---|---|---|
| a | b | c | d \| <— classified as |
| 26 | 2 | 5 | 2 \| a = AVIÓN |
| 2 | 8 | 0 | 0 \| b = BUS |
| 3 | 0 | 25 | 3 \| c = TAXI |
| 1 | 0 | 2 | 11 \| d = BUSETA |

**Fuente:** Elaboración propia.

Para la clase AVIÓN se clasificaron correctamente 26 instancias y 9 incorrectamente; para la clase BUS las correctas fueron 8 y las incorrectas 2; con respecto a la clase TAXI se clasificaron correctamente 25 y 6 incorrectas, y por último, en la clase BUSETA 11 instancias fueron clasificadas correctamente y solo 3 incorrectamente.

En la tabla 7 se observan los valores de la diagonal, de las opciones AVIÓN, BUS, TAXI y BUSETA, en la cual aparecen seleccionadas correctamente 26, 8, 25 y 11, respectivamente, lo cual hace que el modelo generado sea confiable y de un acierto alto de clasificación.

## Validación del modelo

Para la validación del modelo generado a través del software WEKA, es necesario tomar un archivo de entrenamiento y uno de prueba. Luego de guardar el modelo se pueden realizar las predicciones para un conjunto de pruebas, si ese grupo se compone de valores de clase válidos o no, la salida contendrá tanto la clase real como la predicha. En esta investigación se optó por dejar la clase de prueba con un valor "*?*", por lo tanto es importante aclarar que para la etiqueta de la clase *actual* de cada instancia no contendrá información útil, pero la etiqueta *predicho* (*predicted*) sí lo hará (Siddharthan & Katsos, 2010). Al cambiar la etiqueta de clase por un signo de pregunta se puede validar el modelo: el proceso consiste en que al cambiar el valor de la clase por un signo, se evalúa la capacidad de predicción del modelo, es decir, se comprueba el porcentaje de clasificación correcta de cada clase.

La salida de la validación del modelo fue la siguiente, como se observa en la tabla 8, con una primera columna que es la instancia; la segunda no se tiene en cuenta porque todos los atributos fueron marcados con un "*?*", por ende la columna *actual* puede ser ignorada, se limita a establecer que cada clase pertenece a una clase desconocida; la columna *predicted* muestra la predicción de cada instancia, y la columna *error prediction* refleja la probabilidad de que la instancia en realidad pertenezca a la clase (Marozzo, Talia, & Trunfio, 2013).

Las primeras 11 instancias predicen como clase el AVIÓN y que la probabilidad de que eso sea efectivo es del 100%. Tal como lo muestra la tabla 8, la mayoría de los valores son superiores al 85%, lo cual comprueba que el modelo tiene un nivel de predicción alto.

**Tabla 8.** Validación del modelo

| **=== Predictionson Test Data ===** | | | |
|---|---|---|---|
| 1 | 1:? | 1:AVIÓN | 1 |
| 10 | 1:? | 1:AVIÓN | 1 |
| 11 | 1:? | 1:AVIÓN | 1 |
| 12 | 1:? | 2:BUS | 0,854 |
| 13 | 1:? | 1:AVIÓN | 1 |
| 14 | 1:? | 1:AVIÓN | 1 |
| 15 | 1:? | 2:BUS | 0,854 |
| 16 | 1:? | 1:AVIÓN | 1 |
| 17 | 1:? | 1:AVIÓN | 1 |
| 18 | 1:? | 2:BUS | 0,854 |
| 19 | 1:? | 1:AVIÓN | 1 |
| 2 | 1:? | 1:AVIÓN | 1 |
| 20 | 1:? | 2:BUS | 0.854 |
| 21 | 1:? | 2:BUS | 0.854 |
| 22 | 1:? | 1:AVIÓN | 1 |

| === Predictionson Test Data === | | | |
|---|---|---|---|
| 23 | 1:? | 2:BUS | 0.854 |
| 24 | 1:? | 2:BUS | 0.854 |
| 25 | 1:? | 1:AVIÓN | 1 |
| 26 | 1:? | 2:BUS | 0.854 |
| 27 | 1:? | 2:BUS | 0,854 |
| 28 | 1:? | 1:AVIÓN | 1 |
| 29 | 1:? | 1:AVIÓN | 1 |
| 3 | 1:? | 1:AVIÓN | 1 |
| 30 | 1:? | 1:AVIÓN | 1 |
| 31 | 1:? | 1:AVIÓN | 1 |
| 32 | 1:? | 1:AVIÓN | 1 |
| 33 | 1:? | 1:AVIÓN | 1 |
| 34 | 1:? | 1:AVIÓN | 1 |
| 35 | 1:? | 1:AVIÓN | 1 |
| 36 | 1:? | 1:AVIÓN | 1 |
| 37 | 1:? | 2:BUS | 1 |
| 38 | 1:? | 2:BUS | 0,854 |
| 39 | 1:? | 2:BUS | 0,854 |
| 4 | 1:? | 1:AVIÓN | 1 |
| 40 | 1:? | 2:BUS | 1 |
| 41 | 1:? | 2:BUS | 0,854 |
| 42 | 1:? | 2:BUS | 0,854 |
| 43 | 1:? | 2:BUS | 1 |
| 44 | 1:? | 2:BUS | 0,854 |
| 45 | 1:? | 2:BUS | 0,854 |
| 46 | 1:? | 2:BUS | 1 |
| 47 | 1:? | 2:BUS | 0,854 |
| 48 | 1:? | 2:BUS | 0,854 |
| 49 | 1:? | 2:BUS | 1 |
| 5 | 1:? | 1:AVIÓN | 1 |
| 50 | 1:? | 2:BUS | 0,854 |
| 51 | 1:? | 2:BUS | 0,854 |
| 52 | 1:? | 2:BUS | 1 |
| 53 | 1:? | 2:BUS | 0,854 |
| 54 | 1:? | 2:BUS | 0,854 |
| 55 | 1:? | 1:AVIÓN | 1 |
| 56 | 1:? | 2:BUS | 0,854 |
| 57 | 1:? | 2:BUS | 0,854 |
| 58 | 1:? | 1:AVIÓN | 1 |
| 59 | 1:? | 2:BUS | 0,854 |
| 6 | 1:? | 1:AVIÓN | 1 |
| 60 | 1:? | 2:BUS | 0,854 |
| 61 | 1:? | 1:AVIÓN | 1 |
| 7 | 1:? | 1:AVIÓN | 1 |
| 8 | 1:? | 1:AVIÓN | 1 |
| 9 | 1:? | 1:AVIÓN | 1 |
| Inst# | Actual | Predicted Error Prediction | |

**Fuente:** elaboración propia.

## CONCLUSIONES

Este estudio mostró cómo implementar minería de datos para la estimación de un modelo de elección de transporte, siguiendo las etapas del KDD. Estableció que para este tipo de conjunto de datos, el algoritmo J48 perteneciente a la técnica de árboles de decisión fue el más adecuado por tener un porcentaje de acierto alto con respecto a los demás algoritmos de predicción.

El resultado de todo este análisis demostró que aquellas personas que basan su elección de viaje con motivo de salud eligen el avión, al igual que si tienen una frecuencia de viaje mensual y semanal; por otro lado, las personas que viajan a destinos largos, como Barranquilla, Bogotá y Medellín, eligen viajar en buseta; sin embargo, a los viajeros cuyo tiquete es comprado por la empresa donde trabajan, eligen el modo aéreo, hecho que podría deberse a que en esta situación el individuo no percibe directamente el costo del pasaje, lo cual hace que muy probable base su decisión de elección en el costo de este. Asimismo, se demuestra que las personas cuya ocupación es estudiante eligen la alternativa de transporte taxi; esto se puede deber a que los estudiantes, al no tener ingresos propios ni pertenecer a una empresa que costee sus gastos, optan por elegir modos más acordes a sus posibilidades económicas.

En general, la minería de datos sirve para predecir el comportamiento de las personas cuando se debe elegir entre ciertas alternativas. En este estudio las alternativas por elegir eran avión, bus, buseta y taxi, con lo cual se concluye que la demanda del transporte aéreo tiene una gran acogida en la ciudad de Ocaña, Norte de Santander.

Por lo general la técnica correspondiente a los árboles de decisión se destacó frente a las demás por sus altos porcentajes de acierto, ya que en promedio tienen el 68% de instancias correctamente clasificadas, lo cual demuestra que esta técnica aplicada al conjunto de datos puede ser utilizada para realizar aprendizaje supervisado.

Como futura línea de investigación sobre este tema, se puede aplicar aprendizaje no supervisado para validar si con estas técnicas es posible describir el comportamiento oculto del conjunto de datos.

## FINANCIAMIENTO

Es el resultado del proyecto titulado "Generación de un modelo para predecir la demanda del servicio aéreo en la ciudad de Ocaña aplicando técnicas de minería de datos", patrocinado con recursos de la Universidad Francisco de Paula Santander Ocaña.

## REFERENCIAS

Agrawal, D., & Aggarwal, C. (2001). On the design and quantification of privacy preserving data mining algorithms. *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. (págs. 247-255). ACM.

Akiva, B., & Lerman, S. (1985). *Discrete choice analysis: theory and application to travel demand*. (Vol. 9). MIT press.

Alcaldía de Ocaña. (2010). *http://ocana-nortedesantander.gov.co/*. Recuperado el 29 de Septiembre de 2014, de http://ocana-nortedesantander.gov.co/apc-aa-files/38343339653963383637363461323363/INFORME_GENERAL_DEL_MUNICIPIO.pdf

Arentze, T., & Timmermans, H. (2007). Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Research Part B: Methodological, 41*, 772-783.

Baumgartner, D., & Serpen, G. (2009). Large Experiment and Evaluation Tool for WEKA Classifiers. *DMIN, 16*, 340-346.

Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in Weka environment. *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on. IEEE*, (págs. 51-56).

Caruana, R., & Niculescu, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*. (págs. 161-168). ACM.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *WWW '11 Proceedings of the 20th international conference on World wide web* (págs. 675-684 ). New York: ACM.

De Oña, J., De Oña, R., & Calvo, F. (2012). A classification tree approach to identify key factors of transit service quality. *Expert Systems with Applications 39.12*, 11164-11171.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning, 31*, 1-38.

Fayyad, U. M., Piatetsky, G., Smyth, P., & Uthurusamy, R. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*, 37-54.

Ferri, C., Flach, P., & Hernández-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. *ICML., 2*, 139-146.

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter, 1*(1), 20-33.

Guerrero, T., Criado, E., & León, I. (2013). Análisis de la demanda de viajes interurbanos combinando datos de diferentes fuentes. *Ingeniería y competitividad-Universidad del Valle (En revisión)*.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157-1182.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on, 15*(6), 1437-1447.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*, 10-18.

Kohavi, R., & Quinlan, R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. En W. Klösgen, & J. Zytkow, *Handbook of data mining and knowledge discovery* (págs. 267-276).

Marozzo, F., Talia, D., & Trunfio, P. (2013). Scalable script-based data analysis workflows on clouds. *8th Workshop on Workflows in Support of Large-Scale Science* (págs. 124-133). New York: ACM, New York.

Martínez, G., Solarte, R., & Soto, J. (2011). Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares. Scientia et Technica, 3(49), 104-109.

Nishimura, K., & Hirose, M. (2007). The study of past working history visualization for supporting trial and error approach in data mining. *Proceedings of the 2007 conference on Human interface: Part I* (págs. 327-334 ). ACM.

Quinlan, R. (1986). Induction of decision trees. *Machine learning, 1*, 81-106.

Quinlan, R. (1989). Unknown attribute values in induction. En M. Kaufmann, *Machine Learning Proceedings 1989* (págs. 164-168).

Roland, E., Uhrmacher, A., & Saha, K. (2009). Data mining for simulation algorithm selection. *Proceeding Simutools '09 Proceedings of the 2nd International Conference on Simulation Tools and Techniques* (pág. Article No. 14 ). Brussels: ICST.

Siddharthan, A., & Katsos, N. (2010). Reformulating Discourse Connectives for Non-Expert Readers. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* (págs. 1002–1010). Los Angeles: Association for Computational Linguistics.

Van Hulse, J., M. Khoshgoftaar, T., & Napolitano, A. (2007). Experimental Perspectives on Learning from Imbalanced Data. *International Conference on Machine Learning* (págs. 935-942). New York: ACM New York.

Witten, I., Hall, M., Holmes, G., Pfahringer, B., & Reutemann, P. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*, 10-18.

Xie, C., Lu, J., & Parkany, E. (2003). Work travel mode choice modeling with data mining: decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board, 4*, 50-61.

Yoav, F., & Schapire, R. (1996). Experiments with a new boosting algorithm. *ICML, 96*, 148-156.