Prediction for the Car Rental Business with Supervised Techniques

Sandra Zapata-Quentasi, Br[1], Alba Yauri-Ituccayasi,Br[1], Rodrigo Huamani-Avendaño,Br[1] and Jose Sulla-Torres, Dr[1]
[1]Universidad Nacional de San Agustín de Arequipa, Perú, szapataq@unsa.edu.pe, ayaurii@unsa.edu.pe, rhuamaniav@unsa.edu.pe, jsulla@unsa.edu.pe

Abstract -- Car rental is a new trend and is already a reality in many countries, as it is a cheaper option than maintaining your own. The objective of this article is to identify the ideal car for a person, according to the characteristics that you want. In the present work, a study was made of the previous steps involved in the prediction of a car according to the desired characteristics and a comparison of the classification algorithms was carried out to determine which classification is appropriate in terms of the accuracy of the prediction. The steps followed were: Data collection, preprocessing, data preparation and comparison of classification algorithms. The results obtained show that the Random Forest algorithm presents a 95.12% correct classification of the instances and a mean square error of 0.12, which are acceptable results for the tests performed.

Keywords – Prediction, KDD, car rental, supervised techniques.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.371
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

17th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Industry, Innovation, And Infrastructure for Sustainable Cities and Communities", 24-26 July 2019, Jamaica.

Predicción para el Negocio de Alquiler de Automóviles con Técnicas Supervisadas

Sandra Zapata-Quentasi, Br[1], Alba Yauri-Ituccayasi,Br[1], Rodrigo Huamani-Avendaño,Br[1] and Jose Sulla-Torres, Dr[1]

[1]Universidad Nacional de San Agustín de Arequipa, Perú, szapataq@unsa.edu.pe, ayaurii@unsa.edu.pe, rhuamaniav@unsa.edu.pe, jsulla@unsa.edu.pe

***Resumen*—El alquiler de autos es una nueva tendencia y ya es una realidad en muchos países, ya que es una opción más económica que mantener uno propio. El objetivo de este artículo es identificar el auto ideal para una persona, según las características que desee. En el presente trabajo se hizo un estudio de los pasos previos que involucra la predicción de un auto según las características deseadas y se realizó una comparación de los algoritmos de clasificación para determinar que clasificar es el adecuado en cuanto a la precisión de la predicción. Los pasos que se siguieron fueron: La colección de datos, el preprocesamiento, la preparación de datos y la comparación de los algoritmos de clasificación. Los resultados obtenidos muestran que el algoritmo Random Forest presenta un 95.12% de clasificación correcta de las instancias y un error medio cuadrático de 0.12, lo que son unos resultados aceptables para las pruebas realizadas.**

***Palabras Clave*—Predicción, KDD, alquiler de vehículos, técnicas supervisadas.**

***Abstract*— Car rental is a new trend and is already a reality in many countries, as it is a cheaper option than maintaining your own. The objective of this article is to identify the ideal car for a person, according to the characteristics that you want. In the present work, a study was made of the previous steps involved in the prediction of a car according to the desired characteristics and a comparison of the classification algorithms was carried out to determine which classification is appropriate in terms of the accuracy of the prediction. The steps followed were: Data collection, preprocessing, data preparation and comparison of classification algorithms. The results obtained show that the Random Forest algorithm presents a 95.12% correct classification of the instances and a mean square error of 0.12, which are acceptable results for the tests performed.**

***Keywords*— Prediction, KDD, car rental, supervised techniques.**

I. Introducción

Uno de las mayores tendencias en el negocio de alquiler son lo de los vehículos; cada vez más los clientes desean alquilar vehículos de acuerdo a sus necesidades y la las preferencias que necesitan tener una programación de renta [1]. El negocio de alquiler de vehículos asume un proceso de programación simple: el cliente especifica el vehículo que desea, la hora y el lugar donde le gustaría llevarlo, y también un lugar y hora donde le gustaría dejarlo después del alquiler. Tiempo de alquiler. Sin embargo, para resolver este problema se requiere tomar una serie de decisiones, lugar de retiro del vehículo, buscar las preferencias del cliente, las distancias y los costos.

Según un estudio del buscador de viajes Kayak [2], Lima es una de las ciudades en Latinoamerica donde se tienen las mejores ofertas en cuanto al alquiler de un auto. Las empresas encargadas a brindar servicios de alquiler a menudo tienden a poner nuevas sucursales a gestionar el número de trabajadores, a alcanzar un nuevo mercado y tener nuevas estrategias de negocio, la información para este tipo de datos deber ser presentados mediante un cuadro de control (*dashboard)* o un cubo Olap los cuales nos brindan información significativa de manera automatizada y relevante en la era actual de la digitalización y la formación de la economía digital [3].

Por toda esta complejidad, es necesario una construcción de un *Datawarehouse* (DW) que ayude a mejorar los procesos de análisis de las empresas de alquiler. Así mismo, se realizará la clasificación de datos con distintos algoritmos como J48, *Random Forest*, comparando los resultados de su eficiencia obtenidos con la herramienta WEKA, obteniendo así el algoritmo que mejor prediga con el uso de los datos relacionados al alquiler del vehículo.

Para desarrollar la propuesta se utilizará la metodología KDD. Dentro de las actividades a desarrollar será la construcción del *Datawarehouse*, la migración de datos de bases de datos operacionales a través del método ETL, defiendo el esquema de DW se hará el análisis para la predicción de puntos de venta y otros factores que ayuden a negocio de alquiler de vehículos.

Se realizará el análisis de resultados del proyecto haciendo uso de la herramienta de Power BI. Finalmente, se mostrará las conclusiones y recomendaciones que se obtienen como resultado del trabajo realizado.

II. Trabajos Relacionados

En esta sección se agregan trabajos que influyeron en la realización del presente artículo.

En el trabajo desarrollado por Mark Ng Monica (2018) se propone un modelo que integra factores de actitud, factores normativos y autocontrol, explica la compra de vehículos eléctricos (VE) por parte de los consumidores. Específicamente, utiliza modelos de ecuaciones estructurales para desarrollar un modelo para identificar las relaciones entre los valores percibidos, las actitudes verdes, los factores

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.371
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

normativos y los beneficios autoexpresivos y la intención de compra de los EV. Se realizó un estudio empírico para probar el marco conceptual y se desarrollaron 11 hipótesis basadas en la literatura. El modelo fue probado con datos de encuestas de 205 autos [4].

Hongfei Zhan Junhe Yu (2017) toma como ejemplo la industria de fabricación de automóviles, basada en el análisis de grandes datos de automóviles de venta, utilizando la tecnología de minería de datos, a través del programa Java para preparar el programa de rastreador web para la recopilación de datos. Para dar algunas sugerencias para la industria de fabricación de automóviles en la producción de automóviles, se reduce el inventario de empresas de automóviles y el desperdicio de recursos [5].

En el trabajo de Al-Noukari (2008), revisa trabajos relacionados a la minería de datos relacionados con modelos comerciales para proporcionar una solución de minería de datos propuesta que se pueda utilizar para el mercado automotriz, así como para muchas otras áreas. Esta solución puede proporcionar a los administradores de inventario un análisis importante utilizando las técnicas de extracción de datos en el campo de la fabricación de automóviles. Tales técnicas ayudarán a los fabricantes a tomar decisiones adecuadas [6].

Otro de los trabajos relacionados se encuentra el desarrollado por Stefan Lessmann (2017), el cual hace mención a un análisis comparativo de la alternativa los métodos de predicción evidencian que la regresión es particularmente efectiva en la reventa previsión de precios También se muestra que el uso de la regresión lineal, el método predominante en trabajo previo, debe ser evitado [7].

Ning Sun (2017) en su trabajo tiene como objetivo establecer un modelo de evaluación de precios de automóviles de segunda mano para obtener el precio óptimo según el tipo del auto, utilizando una red neuronal que selecciona el número óptimo de neuronas ocultas y mejora la precisión del modelo de predicción [8].

Se ha utilizado otras técnicas de minería de datos relacionados al alquiler de vehículos como el uso de máquinas de vectores de soporte [9], construcción de sistemas expertos para predecir los precios en el negocio de los vehículos, utilizando inferencia neuro-difusa adaptativa [10], entre otros para el soporte de decisiones en el alquiler de vehículos [11].

III. Materiales y Métodos

A continuación, se presentan las técnicas y tecnologías utilizadas para la implementación.

La Base de datos de los vehículos cuentan con los atributos: Marca, Modelo, Año, Precio (Valor), Tipo de Seguridad, Aceptabilidad (No apto, Apto, Muy apto)

A. Herramientas

1) Weka

Weka, acrónimo de *Wakaito Environment for Knowledge Analysis*, es un software de código abierto bajo términos de GNU GPL y posee muchas herramientas para minería de datos y aprendizaje automático [12].

2). Azure Machine Learning Studio

Machine Learning Studio es un entorno de creación de arrastrar y soltar visual potente y simple basado en el navegador, donde no se necesita codificación. Pase de la idea al despliegue en cuestión de clics. El principal objetivo es que permite crear, implementar y compartir fácilmente soluciones de análisis predictivo [13].

3). Rapid Miner Studio

Es una herramienta de Minería de Datos ampliamente usada y probada a nivel internacional en aplicaciones empresariales, de gobierno y academia. Implementa más de 500 técnicas de pre-procesamiento de datos, modelación predictiva y descriptiva, métodos de prueba de modelos, visualización de datos, etc [14].

4). SQL Server:

Es un sistema de gestión de bases de datos relacionales (RDBMS) de Microsoft que está diseñado para el entorno empresarial. SQL Server se ejecuta en T-SQL (Transact-SQL), un conjunto de extensiones de programación de Sybase y Microsoft que añaden varias características a SQL estándar, incluyendo control de transacciones, excepción y manejo de errores, procesamiento fila, así como variables declaradas. Esta herramienta permitirá crear nuestro modelo estrella y a la vez cargar los datos en el proceso ETL.

B. Métodos

1) KDD:

Proceso de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información [15].

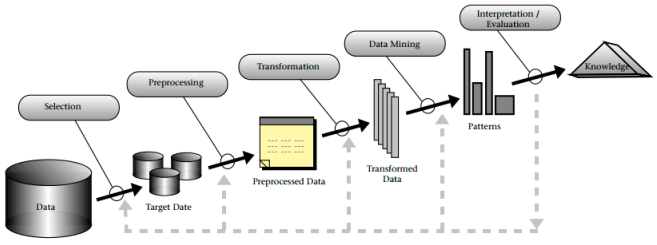


Fig. 1. Proceso KDD [16]

2) ETL:

En la etapa de desarrollo se va a modelar las tablas de dimensiones, primero se procede a extracción, transformación y carga de nuestra base de datos "car.csv" el cual permitirá poblar los datos hacia un gestor MySql mediante la herramienta de *Kettle* esta herramienta lo proporciona Pentaho.

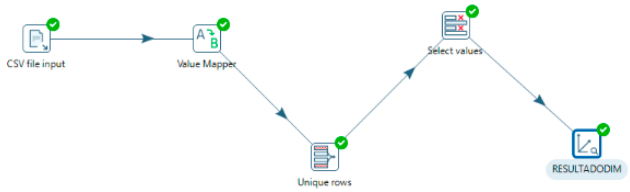


Fig. 2. Proceso ETL

Proceso ETL en Kettle se hace los procesos de transformación a través de la carga de una tabla de hechos para cargar la dimensión con la creación de una nueva tabla en nuestra base de datos "alquiler-auto".

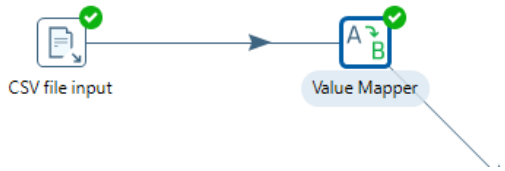


Fig. 3. Proceso de extracción

Las entradas son archivos *csv* por lo que se han transformado algunos valores abreviados cambiándolos con información manejable como vemos en la siguiente Figura.

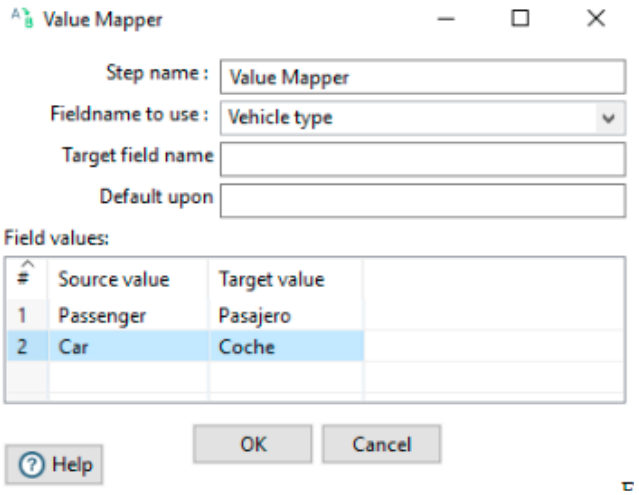


Fig. 4. Creación de conexión a MySQL

Para esta dimensión se hace la carga de estos datos en una tabla llamada "Vehículos" mediante la creación de una conexión a MYSQL.

Finalmente se establece los atributos y la clave primaria se genera el Sql y se crea esta tabla.

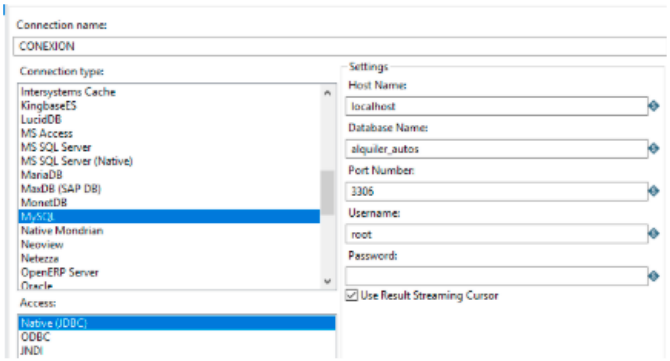


Fig. 5. Proceso de transformación

En la etapa de construcción del *Datawarehouse* se modelan las tablas relacionales en una gran estructura desnormalizada compuesta por tabla de hechos y tablas más pequeñas que definen las dimensiones a esta tabla se le llama tabla de dimensiones.

Las tablas de medidas siempre deben ser numéricas y se almacenan en la tabla de hechos estas tablas contiene los valores de negocio que se deben analizar en la venta de alquiler de autos, mientras que en la tabla de dimensiones se almacenan datos textuales.

1) Diseño del modelo estrella: El esquema estrella separa los datos del proceso de negocios en: hechos y dimensiones. Los hechos contienen datos medibles, cuantitativos, relacionados a la transacción del negocio, y las dimensiones son atributos que describen los datos indicados en los hechos (una especie de meta-datos, o sea datos que describen otros datos).

En nuestro estudio se diseña la base de datos *ventas*, la tabla de hecho y dimensiones está representado de la siguiente forma:

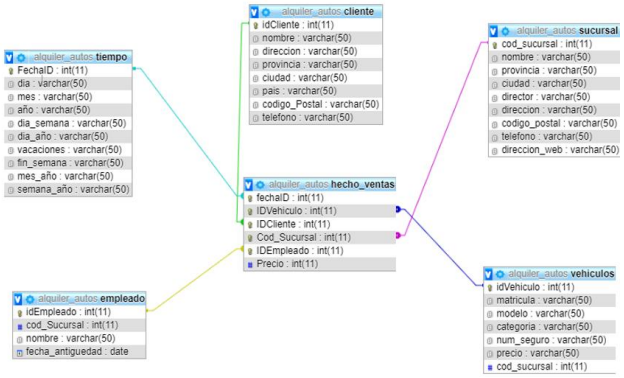


Fig. 6. Diseño Modelo estrella

2) *Carga de las tablas de dimensiones y Hechos*: En esta etapa se lleva a cabo la carga de datos que permite leer las tablas de los sistemas transaccionales para que pueden ser cargadas en las tablas de dimensiones. Se debe evitar la inconsistencia y duplicidad de los datos. En el análisis se dato es recomendable que establecer una actualización de dato mensuales. Se utilizó la herramienta sql server para la carga de datos a las tablas de dimensiones

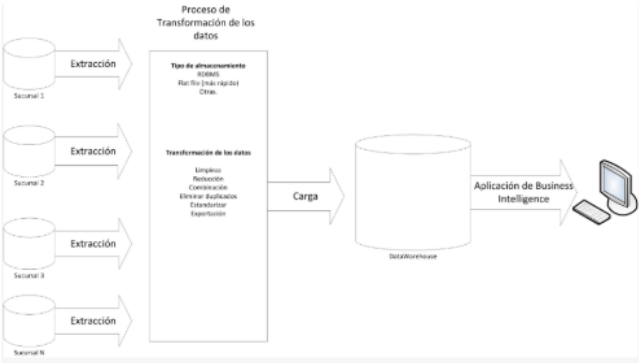


Fig. 7. Esquema ETL

IV. ALGORITMOS DE CLASIFICACION

Clasificación es el proceso de identificar una nueva categoría de observación sobre la base de la formación del conjunto de datos que contiene datos donde las categorías son desconocidas. Primero cargaremos el archivo CSV para realizar el preprocesamiento de datos.

La herramienta WEKA nos muestra todos los datos de los atributos, realizando un histograma para cada uno de ellos.

A. J48

Este sistema es utilizado sobre una base de datos que contiene información sobre la BD de alquiler de autos a través del algoritmo, junto con datos de entrenamiento y validación cruzada, se logra crear un árbol de clasificación que permitirá predecir cómo es que los autos menos costosos tienden a tener una mayor eficiencia en combustible. El algoritmo J48 de clasificación es usado para generar un árbol de decisión [17].

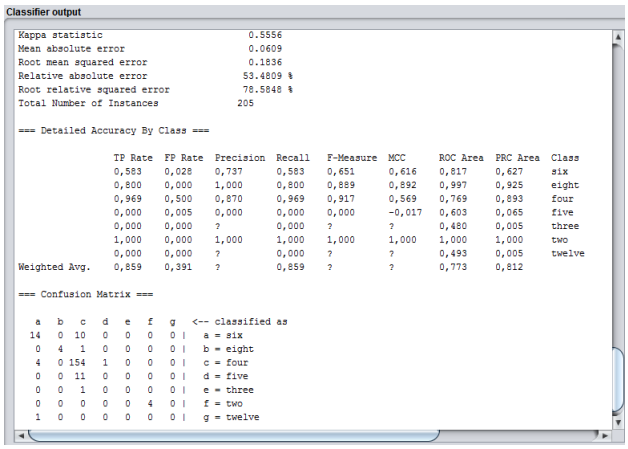


Fig. 8. Resultados del algoritmo con la base de datos alquiler-autos

Obteniendo así el árbol de decisión que se muestra en la Fig. 9

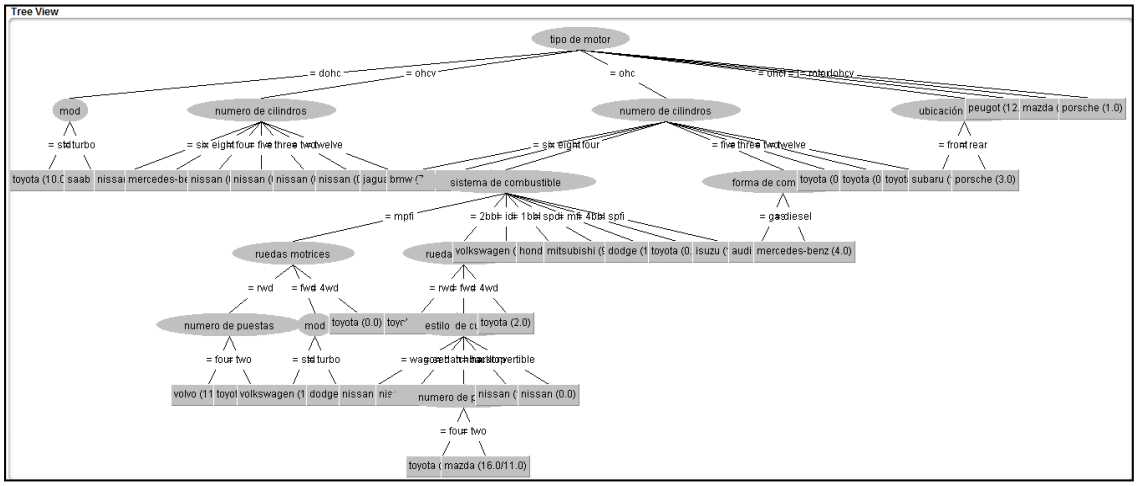
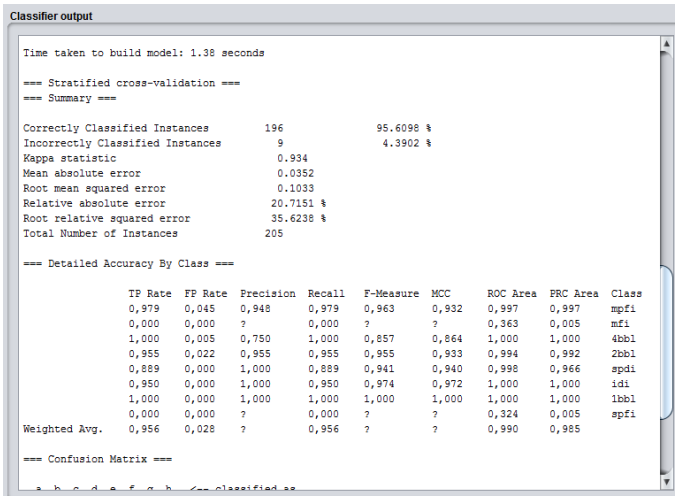


Fig. 9. Árbol de decisión J48

B. Random Forest

Es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de *bagging* que construye una larga colección de árboles no correlacionados y luego los promedia. Usando el algoritmo en WEKA se tienen los siguientes resultados [18].



Classifier output

```
Time taken to build model: 1.38 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         196               95.6098 %
Incorrectly Classified Instances         9                4.3902 %
Kappa statistic                          0.934
Mean absolute error                      0.0352
Root mean squared error                  0.1033
Relative absolute error                 20.7151 %
Root relative squared error             35.6238 %
Total Number of Instances              205

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,979     0,045     0,948      0,979    0,963      0,932    0,997     0,997     mpfi
               0,000     0,000     ?          0,000    ?          ?        0,363     0,005     mfi
               1,000     0,005     0,750      1,000    0,857      0,864    1,000     1,000     4bbl
               0,955     0,022     0,955      0,955    0,955      0,933    0,994     0,992     2bbl
               0,889     0,000     1,000      0,889    0,941      0,940    0,998     0,966     spdi
               0,950     0,000     1,000      0,950    0,974      0,972    1,000     1,000     idi
               1,000     0,000     1,000      1,000    1,000      1,000    1,000     1,000     1bbl
               0,000     0,000     ?          0,000    ?          ?        0,324     0,005     spfi
Weighted Avg.  0,956     0,028     ?          0,956    ?          ?        0,990     0,985

=== Confusion Matrix ===
```

Fig. 10. Resultado del algoritmo Random Forest

En la Fig. 10 se puede apreciar que una vez ejecutado el algoritmo *Random Forest* se obtiene un 95.6% de clasificación correcta de las instancias y un error medio cuadrático de 0.10, lo que son unos resultados aceptables para las pruebas realizadas.

D. Clustering

Es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio, por lo general son de distancia o similitud.

1) KMeans

Tiene como objetivo la partición de un conjunto de *n* observaciones en *k* grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos. Luego, utilizamos la función CLUSTER.

2) Cluster usando Rapid Miner

Utilizaremos la base de datos de automóviles para clasificar datos en un número determinado de clases que tengan características en común.

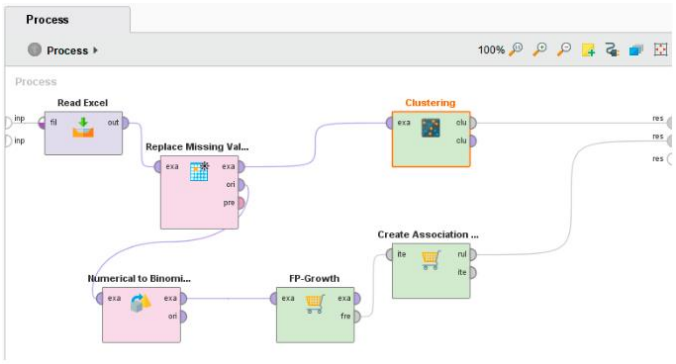


Fig. 11. Diagrama de procesos

Los resultados obtenidos muestran 5 clase (*k*=5) referente al precio de los automóviles.

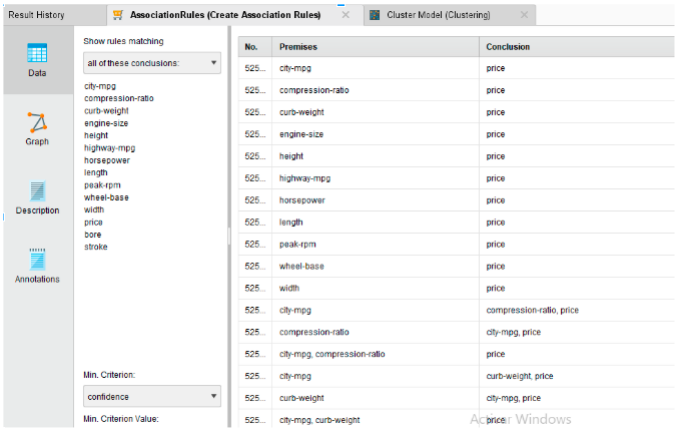


Fig. 12. Resultados de clustering a través de Rapid Miner

3) Lógica difusa

Se basa en lo relativo de lo observado como posición diferencial. Este tipo de lógica toma dos valores aleatorios, pero contextualizados y referidos entre sí. Así, por ejemplo, una persona que mida dos metros es claramente una persona alta, si previamente se ha tomado el valor de persona baja y se ha establecido en un metro. Ambos valores están contextualizados a personas y referidos a una medida métrica lineal

Fuzzy en Matlab Aplicando la lógica al proyecto basado en su índice de seguridad y precio. En la siguiente imagen se designan los parámetros de entrada

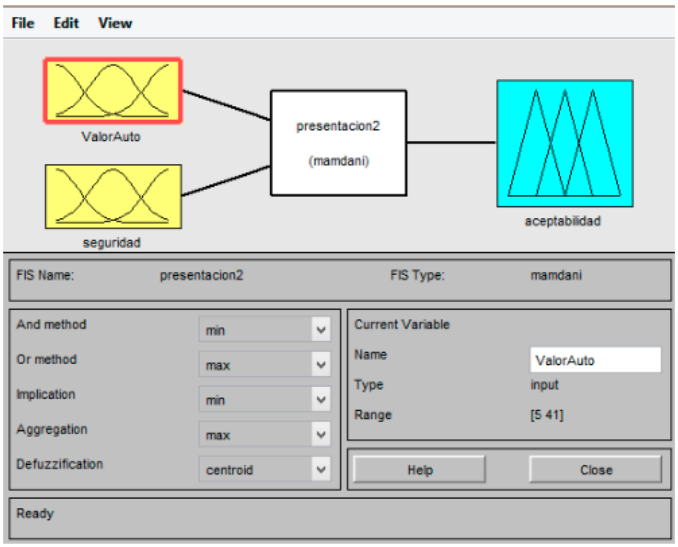


Fig. 13. Entradas del proceso

A continuación, designaremos las reglas condicionales para muestro proyecto.

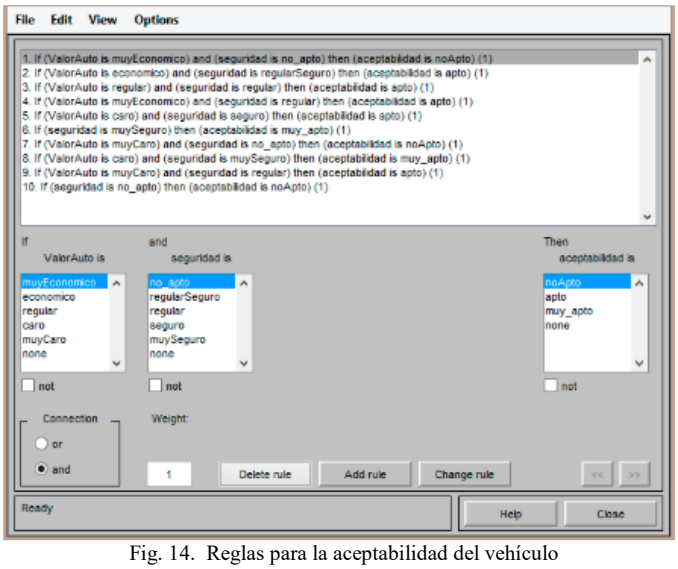


Fig. 14. Reglas para la aceptabilidad del vehículo

Se mostrarán los valores asignados a entradas y salidas para realizar el análisis.

Para finalizar se muestra la vista de superficie del proceso, para la determinación de la aceptabilidad en el alquiler de un vehículo.

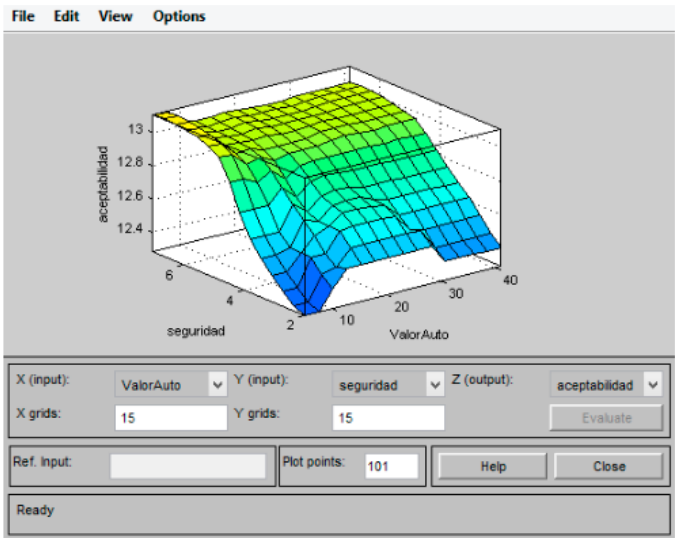


Fig. 15. Superficie del proceso

IV. Resultados

En esta sección se mostrará la utilidad de nuestra implementación, para ello utilizaremos la herramienta *Power BI*. *Power BI* es una solución de análisis empresarial que permite visualizar los datos y compartir información en toda su organización, o incorporarlos en su aplicación o sitio web. Se contacta a cientos de fuentes de datos y haga que sus datos cobren vida con paneles e informes en vivo. A continuación, se visualizarán los resultados del análisis del procesamiento de datos

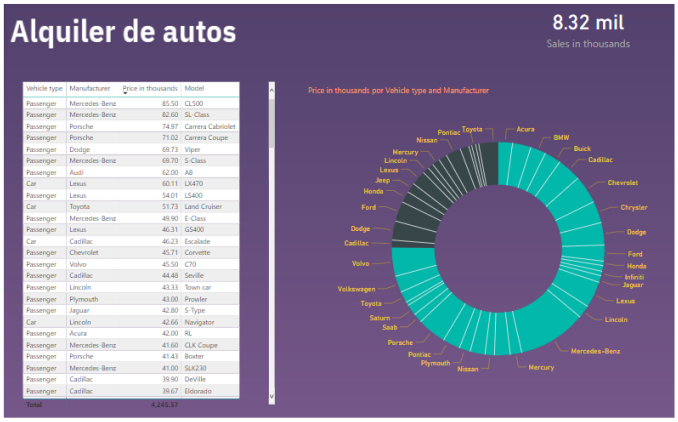


Fig. 16. Análisis de vehículo por precio.

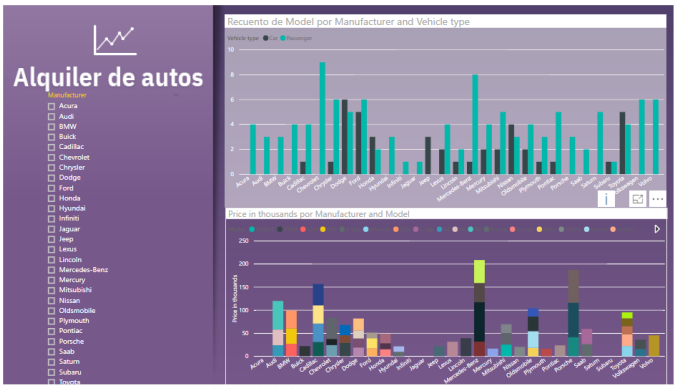


Fig. 17. Análisis del detalle de tipo de vehículo y precio

V. Conclusiones

El desarrollo de este trabajo ha permitido analizar y explorar los datos, proporcionando un análisis de diferentes herramientas de predicción, construcción, etc. En desarrollo de almacén de datos se debe tener muy claro la información que se pretende analizar ya que el proceso ETL permite establecer información en la tabla de hechos y dimensiones definidas en el almacén de datos. Los sistemas de Inteligencia de negocios ayudan a hacer más competitiva la estrategia de las empresas. Con el uso del algoritmo de minería de datos *Random Forest* se obtuvo una precisión de 95.12% de clasificación correcta de las instancias y un error medio cuadrático de 0.12, lo que son unos resultados aceptables para las pruebas realizadas. La lógica difusa permitió mostrar la aceptabilidad de un modo más natural.

Referencias

[1] S. Andreev, G. Rzevski, P. Shviekin, P. Skobelev, and I. Yankov, "A multi-agent scheduler for rent-a-car companies," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009.

[2] C. G. M. Aparicio, "A new way to do business in tourism: the searchers in the web," *Int. J. Sci. Manag. Tour.*, vol. 3, no. 2, pp. 101–120, 2017.

[3] A. V. Babkin, D. D. Burkaltseva, A. V. Betskov, H. S. Kilyaskhanov, A. S. Tyulin, and I. V. Kurianova, "Automation digitalization blockchain: Trends and implementation problems," *Int. J. Eng. Technol.*, 2018.

[4] M. Ng, M. Law, and S. Zhang, "Predicting purchase intention of electric vehicles in Hong Kong," *Australas. Mark. J.*, 2018.

[5] Q. Zhang, H. Zhan, and J. Yu, "Car Sales Analysis Based on the Application of Big Data," in *Procedia Computer Science*, 2017.

[6] M. Al-Noukari and W. Al-Hussan, "Using data mining techniques for predicting future car market demand," in *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA*, 2008.

[7] S. Lessmann and S. Voß, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *Int. J. Forecast.*, 2017.

[8] N. Sun, H. Bai, Y. Geng, and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," in *Proceedings - 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2017*, 2017.

[9] M. Listiani, "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application," 2009.

[10] J. Da Wu, C. C. Hsu, and H. C. Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," *Expert Syst. Appl.*, 2009.

[11] S. Lessmann, M. Listiani, and S. Voß, "Decision support in car leasing: A forecasting model for residual value estimation," *Icis*, 2010.

[12] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[13] R. Barga, V. Fontama, and W. H. Tok, *Predictive Analytics with Microsoft Azure Machine Learning*. 2015.

[14] J. Vijayalakshmi and E. Ramaraj, "Comparative study of big data analytical tools," *Int. J. Pure Appl. Math.*, 2018.

[15] R. Kalavathy, R. M. Suresh, and R. Akhila, "KDD and data mining," *IET-UK Int. Conf. Inf. Commun. Technol. Electr. Sci. (ICTES 2007)*, no. Ictes, pp. 1105–1110, 2007.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[17] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 6, pp. 2277–128, 2013.

[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.