Técnicas de Minería de Datos en la Industria Automotriz

Data Mining Techniques in the Automotive Industry

Omar Danilo Castrillon, Ph. D[1], Jaime Alberto Giraldo, Ph. D[1], and Jaime Antero Arango, Ph. D[1]
[1] Universidad Nacional de Colombia, Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Industrial, Campus La Nubia Bloque Q piso 2, Manizales, Colombia. odcastrillong@unal.edu.co, jaiagiraldog@unal.edu.co, jaarangom@unal.edu.co

Resumen. Una de las principales preguntas que debe resolver la industria automotriz, es determinar la aceptación que tendrán sus productos en el medio. Así, el objetivo fundamental de esta investigación es dar respuesta a esta pregunta. En esta investigación se parte de 6 variables independientes (costo, mantenimiento, puertas, nro. de personas, baúl y seguridad) con el fin de predecir el comportamiento de una variable dependiente denominada aceptación. En este análisis, se toma una base de datos de 1728 registros y mediante un proceso de selección estadística se establecen las variables más influyentes, con el fin de estructurar un archivo para ser analizado por medio del algoritmo J48 de minería de datos, el cual es ejecutado mediante la plataforma de aprendizaje automático y minería de datos denominada Weka. Como resultado de este proceso se predice (con una efectividad superior al 92%) el comportamiento de la variable dependiente denominada aceptación. Finalmente, los resultados también muestran que las tres variables independientes más influyentes son: costo, nro. de personas, y seguridad.

Palabras claves—Minería de datos, algoritmo J48, Weka, selección estadística, base de datos.

Abstrac. One of the most essential questions that the automotive industry has to solve is determining the acceptance that their products will have in the environment. Thus, the fundamental objective of this research is to answer this question. This research starts with 6 independent variables (cost, maintenance, doors, number of people, trunk, and security) to predict the behavior of a dependent variable called acceptance. In this analysis, a database of 1728 records is taken and the most influential variables are established through a statistical selection process to structure a file that will be analyzed by the J48 data mining algorithm, which is executed by the automatic learning and data mining platform Weka. As a result of this process, the behavior of the dependent variable called acceptance is predicted (with an accuracy of over 92%). Finally, the results also show that the three most influential independent variables are: cost, number of people, and security.

Keywords — Data mining, J48 algorithm, Weka, statistical selection, database.

I. Introducción

En el año 2019 las ventas de automóviles a nivel mundial alcanzaron 90.3 millones de vehículos, una industria que forma parte fundamental de la economía de muchos países como se ilustra en la Tabla 1 (https://datosmacro.expansion.com/negocios/produccion-vehiculos)

TABLA 1.
PRODUCCIÓN MUNDIAL DE VEHÍCULOS (2019)

País	# Vehículos
China	23.362.477
USA	10.533.653
Japón	9.168.651
Alemania	4.661.328
India	4.194.763
Corea del Sur	3.950.617
México	3.772.861
Brasil	2.803.841

La tabla 1, muestra el impacto que esta industria tiene en la economía de todos los países del mundo y en general en la economía mundial. Lo anterior hace que los problemas relacionados con la industria automotriz cobren gran importancia y relevancia en el contexto mundo actual.

No obstante, diseñar un vehículo que logre una mejor aceptación por parte de los usuarios no es una tarea fácil. Hoy en día, existen diferentes alternativas que pueden atraer a los usuarios, como son vehículos eléctricos [1, 2], voladores [3], de diferentes tamaños [4], autónomos [5,6], en especial para el reparto de mercancías [7], y en general vehículos con toda clase de lujos para las personas, como las aplicaciones de navegación móviles entre otros aspectos [8,9]. Sin embargo, determinar si un vehículo tendrá aceptación o no por parte de los usuarios, va más allá de lo anteriormente descrito, el primer problema en resolver es determinar donde se deben hacer las innovaciones esto es: costo, comodidades, seguridad, etc. La respuesta a la anterior pregunta, permitirá priorizar los aspectos más relevantes para el público, con el fin de lograr una mayor aceptación de los vehículos producidos por las casas automotrices.

En este sentido, el objetivo principal de este artículo, es determinar en cuales áreas se deben realizar las innovaciones con el fin de lograr una mayor aceptación del vehículo. En este proceso se analizarán innovaciones referentes al costo, mantenimiento, puertas, nro. de personas, tamaño del baúl y seguridad. Como se expresa en la sección siguiente, este análisis se realizará mediante técnicas de minería de datos, por medio del algoritmo J48 (https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html), el cual, aunque no es un algoritmo bayesiano, presenta

un comportamiento similar a estos, con los cuales es posible obtener muy buenos resultados aun con pocos datos [10].

Se resalta que, en las diferentes revisiones literarias realizadas, existen muy pocos trabajos relacionados [11], los cuales aborden este problema. Este aspecto justifica aún más el estudio de este problema, en especial por medio de técnicas inteligentes donde las soluciones encontradas en la literatura son escasas y la mayoría de ellas centradas en innovaciones referentes en: autos eléctricos, voladores, autónomos, entre otros. Estas innovaciones dejan por fuera un segmento de estudio muy poco explorado, como el analizado en este artículo.

Para su estructuración este documento ha sido organizado de la siguiente forma: En la sección Materiales y métodos se realiza una descripción de la metodología empleada en la solución de este problema, posteriormente se describe la sección de resultados, los cuales son la consecuencia directa de aplicar la metodología, finalmente se encuentran las secciones de discusiones, conclusiones, agradecimientos y referencias.

II. MATERIALES Y METODOS

La metodología propuesta, es desarrollada por medio de la plataforma de aprendizaje automático y minería de datos denominada Weka, la cual es descrita en [12]. Para la predicción de la variable dependiente denominada aceptación, se diseña un archivo. arff el cual es interpretado, desde esta plataforma (Weka), por medio del algoritmo J48, el cual está diseñado con base en las técnicas de árbol de decisión desarrolladas en el algoritmo C4.5. La predicción de esta variable dependiente (aceptación) se realiza con base en seis variables independientes: costo, mantenimiento, puertas, nro. de personas, baúl y seguridad. Finalmente, para lograr una mejor comprensión esta sección es estructurada de la siguiente forma: A) Elaboración de la base de datos. B) diseño del archivo Weka. C) predicción de la variable dependiente. D) identificación de las variables más influyentes. E) árbol de clasificación.

A. Elaboración de la Base de Datos.

En la elaboración de la base de datos, se tomó como referencia el repositorio encontrado en [11]. De este repositorio se tomaron 1728 registros con los datos de las variables independientes y la variable dependiente que serán objeto de análisis en esta investigación.

B. Diseño del archivo Weka.

Tomando como referencia las variables seleccionadas en el paso anterior, se construye el encabezado y el cuerpo del respectivo archivo .arff para ser interpretado desde la plataforma Weka. Esta plataforma es descrita en [1]

C. Predicción de la variable dependiente.

El archivo elaborado en el paso anterior es interpretado por medio del algoritmo J48 desde la plataforma Weka. Esta interpretación permitirá establecer el porcentaje de acierto en la predicción de la variable dependiente. Igualmente, por medio de este algoritmo será factible construir el respectivo árbol de clasificación con el fin de identificar las relaciones, causas y efectos influyentes en la variable dependiente.

D. Identificación de las variables más influyentes.

Por medio de la plataforma Weka se realizará una selección paulatina de las variables independientes con el fin de identificar su influencia sobre la clasificación de la variable dependiente. Esta influencia se establecerá al eliminar la variable independiente deseada y volver a realizar la predicción de la variable dependiente desde la plataforma Weka, sin la variable eliminada. Las 3 variables más influyentes son seleccionadas.

E. Arbol de Clasificación.

Finalmente, con la selección (en el paso anterior) de las 3 variables independientes más influyentes, es construido el respectivo árbol de decisión, el cual permite predecir de una forma visual el comportamiento de la variable dependiente.

III. RESULTADOS

Como consecuencia de aplicar la metodología descrita en el numeral 2, sobre la base de datos seleccionada, se obtienen los siguientes resultados:

A. Elaboración de la Base de Datos.

En relación directa con el paso *A* de la metodología, la siguiente base de datos es elaborada (ver Tabla 2 - Por razones de espacio solo se muestra parte del archivo):

TABLA 2.
MUESTRA DE LA BASE DE DATOS ANALIZADA [2]

Costo	Mante	Puertas	Personas	Baul	Seguri	aceptacion
Vhigh	Vhigh	2	2	Small	Low	unacc
Vhigh	Vhigh	2	2	Small	Med	unacc
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Vhigh	Vhigh	2	2	Small	High	unacc
Vhigh	Vhigh	2	2	Med	Low	unacc
Vhigh	Vhigh	2	2	Med	Med	unacc

B. Diseño del archivo Weka.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.47
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

La elaboración del archivo Weka (.arff) para ser interpretado por medio del algoritmo J48, desde la plataforma Weka, es estructurado en dos partes, las cuales son ilustradas en las Tabla 3 y 4 respectivamente (Por razones de espacio solo se muestra parte del archivo):

TABLA 3. ENCABEZADO DEL ARCHIVO .ARFF

Valor	Mante	Puertas
@relation	relation	
@attribute	Costo	{"vhigh", "high", "med", "low"}
@attribute	Mant	{"vhigh", "high", "med", "low"}
@attribute	Puertas	NUMERIC
@attribute	Personas	{"2","4","more"}
@attribute	Baul	{"small", "med", "big"}
@attribute	Seguridad	{"low", "med", "high"}

En la siguiente tabla (4), las variables independientes V1, V2, V3, V4, V5 y V6 representan: Costo, Mantenimiento, Puertas, Personas, Baúl, y Seguridad.

TABLA 4. DETALLE DEL ARCHIVO. ARFF [2]

Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7
vhigh	vhigh	2	2	small	low	unacc
vhigh	vhigh	2	2	small	med	unacc
.	.	.	.	.	.	.
vhigh	vhigh	2	2	med	low	unacc
vhigh	vhigh	2	2	med	med	unacc

C. Predicción de la variable dependiente.

Por medio de la plataforma Weka y el algoritmo de clasificación J48 se realiza la predicción de la variable dependiente empleando todas las variables independientes. Esta predicción se realiza por medio de una validación cruzada 80% y 20%. Como resultado de este proceso se obtienen los porcentajes de clasificación y precisión ilustrados en las Tabla 5 y 6 respectivamente

TABLA 5. PORCENTAJE DE CLASIFICACIÓN

Variable	Número	Porcentaje
Correctly Classified Instances	1606	92.9398 %
Incorrectly Classified Instances	122	7.0602 %
Kappa statistic	0.8483	
Mean absolute error	0.0398	
Root mean squared error	0.1663	
Relative absolute error	17.3738 %	
Root relative squared error	49.1832 %	
Total Number of Instances	1728	

TABLA 6. PRECISIÓN DE LA CLASIFICACIÓN

+	-	Preci	Recal	Med F	ROC	Clase
0.96	0.039	0.983	0.96	0.971	0.985	unacc
0.901	0.049	0.84	0.901	0.869	0.965	acc
0.609	0.011	0.689	0.609	0.646	0.925	good
0.877	0.01	0.77	0.877	0.82	0.995	vgood
0.929	0.039	0.931	0.929	0.93	0.978	Prom

D. Identificación de las variables más influyentes.

Con el fin de identificar las variables más influyentes en este proceso, cada una de las variables independientes empleadas (sección anterior) fueron suprimidas una por una, repitiéndose los pasos desde el numeral *A* al *C*, con el fin de determinar la importancia de esta variable. Los resultados obtenidos son ilustrados en la Tabla 7. En esta tabla se ilustra con una x las variables empleadas en cada clasificación.

TABLA 7. RESULTADOS DE CLASIFICACION

V1	V2	V3	V4	V5	V6	%Acierto
	x	x	x	x	x	**80.90%**
x		x	x	x	x	82.46%
x	x		x	x	x	92.93%
x	x	x		x	x	**77.14%**
x	x	x	x		x	85.47%
x	x	x	x	x		**67.18%**

Nota: V1: Costo, V2: Mantenimiento, V3: Puertas, V4: Personas, V5: Baúl, V6: Seguridad.

La Tabla 7 muestra que al suprimir las variables V1, V4 y V6 se obtiene los resultados más bajos de clasificación, esto indica que estas variables presentan una mayor influencia. Cuando se realiza la clasificación con estas tres variables se obtiene una clasificación del 81.94%, como se ilustra en la sección siguiente (*E*).

E. Arbol de Clasificación.

Con las variables seleccionadas en el paso anterior se construye el respectivo árbol de decisión desde el programa Weka, como se ilustra en la figura 1. Aunque el porcentaje de clasificación con estas tres variables se reduce al 82% como se ilustra en la Tabla 8, los resultados son más controlables.

TABLA 8. PORCENTAJE DE CLASIFICACIÓN.

Correctly Classified Instances	1416	81.9444 %
Incorrectly Classified Instances	312	18.0556 %
Kappa statistic	0.5996	
Mean absolute error	0.1222	
Root mean squared error	0.25	
Relative absolute error	53.3742 %	
Root relative squared error	73.9258 %	

IV. DISCUSIONES

Los resultados mostrados en la figura 1, permiten inferir que la seguridad es el factor más importante, para lograr la aceptación de un vehículo. En esta gráfica se observa que los autos con baja seguridad (low) no son aceptados (unacc). En este mismo sentido, un análisis de la figura 1, también permite concluir que los autos diseñados para dos pasajeros presentan un porcentaje de aceptación más bajo que sus homólogos de cuatro o más personas. Igualmente, la tercera variable influyente, en la aceptación de un automóvil es el valor (costo), teniendo un mayor nivel de aceptación los autos de un valor medio o bajo.

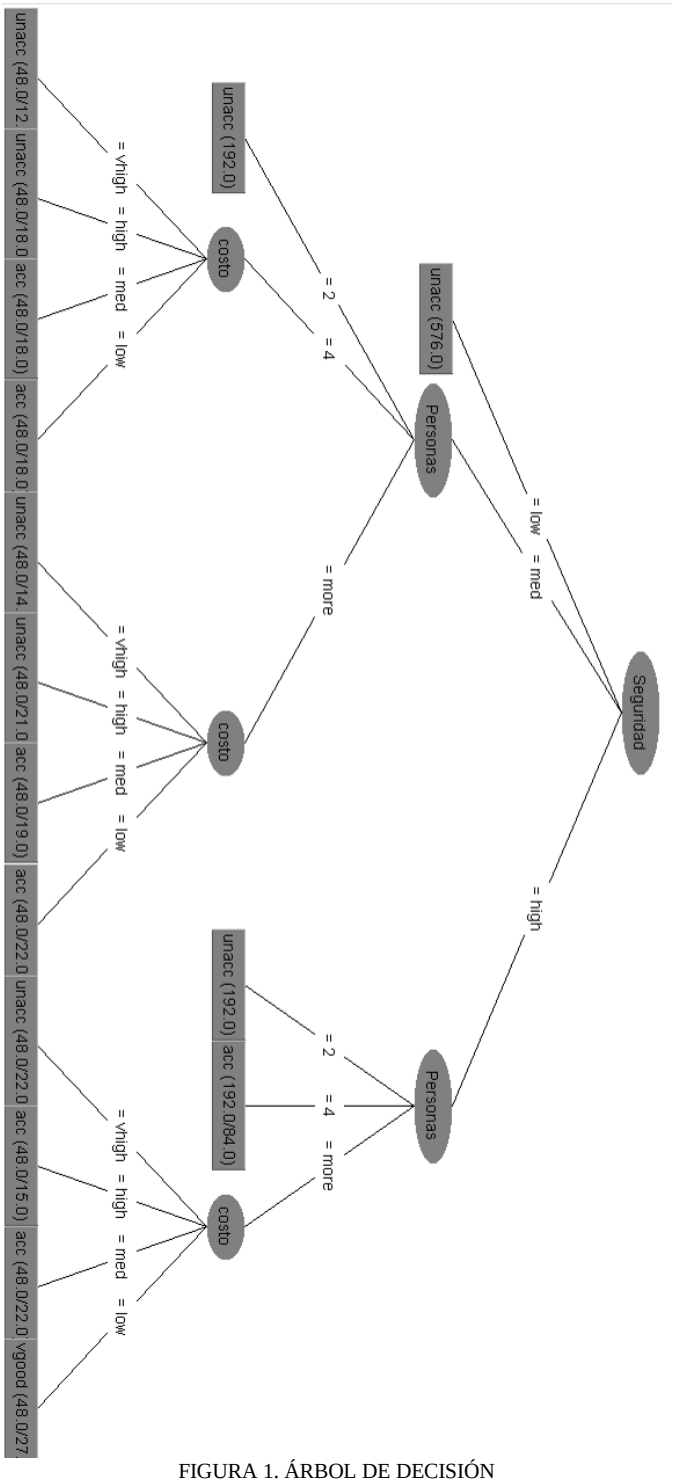

FIGURA 1. ÁRBOL DE DECISIÓN
VARIABLES: VALOR, PERSONAS, SEGURIDAD.

Se resalta que, en las diferentes revisiones literarias realizadas, los estudios similares encontrados son muy escasos [11], muchas de estos estudios se centran en determinar la aceptación de diferentes clases de vehículos como voladores [3], autónomos [5], eléctricos [1], entre otros.

Sin embargo, como se muestra en este estudio para lograr una adecuada aceptación de un vehículo, las compañías automotrices se deben centrar en el control de las tres variables priorizadas en este estudio: Seguridad, personas y costo, en el mismo orden de importancia. Adicionalmente se resalta que, como futuras líneas de investigación se debe analizar el impacto de otros aspectos en la aceptación de los vehículos como: carros voladores, eléctricos, autónomos y todo tipo de lujos y accesorios electrónicos.

IV. CONCLUSIONES

En este artículo por medio de una metodología basada en minería de datos, se logra determinar con gran efectividad las variables más influyentes en la aceptación de un vehículo, las cuales en su orden de importancia son: Seguridad, personas y costo. Aspectos de gran importancia los cuales permitirán un mayor control y aceptación de los vehículos producidos por la industria automotriz. Finalmente, se expresa que esta investigación puede ser replicada en otros contextos de la industria automotriz, para lo cual solo es necesario repetir cada uno de los pasos estructurados en la metodología propuesta.

IV. AGRADECIMIENTOS

Se agradece la colaboración a la Universidad Nacional de Colombia y así como: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

REFERENCIAS

[1] Wolff, S., Madlener, R.: Driven by change: Commercial drivers' acceptance and efficiency perceptions of light-duty electric vehicle usage in Germany. Transportation Research Part C 262–282 (2019)

[2] Du, M., Cheng, L., Li, X., Xiong, J.: Analyzing the acceptance of electric ridesharing by drivers with and without local registered permanent residence, Journal of Cleaner Production 265, 121868 (2020)

[3] Eker, U., Fountas, G., Anastasopoulos, CH.: An exploratory empirical analysis of willingness to pay for and use flying cars, Aerospace Science and Technology. 104, 105993 (2020)

[4] Setiyo, M., Widodo, E., Rosyidi M., Waluyo, B., Pambuko Z., Tamaldin, N.: Feasibility study on small cars as an alternative to conventional fleets due to low occupancy: case study in Indonesia. Heliyon, 6, e03318 (2020)

[5] Zhang,T., Tao,D., Qu, X., Zhang,X., Zeng,J., Zhu,H., Zhu, H.: Automated vehicle acceptance in China: Social

influence and initialtrust are key determinants, Transportation Research Part C, 112, 220–233 (2020)

[6] Rezaei, A., Caulfield, B.: Examining public acceptance of autonomous mobility, Travel Behaviour and Society. 21, 235–246 (2020)

[7] Kapser, S., Abdelrahman, M.: Acceptance of autonomous delivery vehicles for last-mile delivery in Germany – Extending UTAUT2 with risk perceptions. Transportation Research Part C 210-225 (2020)

[8] Yang,L., Bian, Y., Zhao,X., Liu,X., Yao, X.: Drivers' acceptance of mobile navigation applications: An extended technology acceptance model considering drivers' sense of direction, navigation application affinity and distraction perception, International Journal of Human-Computer Studies. 145, 102507 (2021)

[9] Urbinati A, Franzo S, Chiaroni D,: An empiricalanalysis in the Italian automotive industry, Sustainable Production and Consumption. (2021) Article in press. doi: https://doi.org/10.1016/j.spc.2021.01.022

[10] Valencia, M., Correa, J., Díaz, F.: Métodos Estadísticos Clásicos y Bayesianos para el Pronóstico de Demanda. Un Análisis Comparativo, Revista Facultad de Ciencias Universidad Nacional de Colombia, 4(1), 52 -67 (2015)

[11] Dua, D., Graff, C.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. (2019)

[12] Witten, I., Frank, E. y otros dos autores, Data Mining Practical Machine Learning Tools and Techniques, Morgan and Kaufman publication (Elsevier), ISBN-13: 978-0128042915, Cambrige, USA (2017)