

Árboles de decisión sobre la seguridad de vehículos con Python



Complementos de Bases de Datos

Grupo 63

02/07/2021

Índice

1. Introducción.....	3
2. Minería de datos.....	4
Evolución histórica	4
Definición	5
Tareas de la Minería de Datos	7
Árboles de Decisión	8
3. Objetivo	10
4. Implementación.....	11
Análisis y preprocesamiento.....	11
Modelo de clasificación	13
Métricas del modelo	14
5. Pruebas	16
6. Conclusión	17
Bibliografía	18

1. Introducción

La gran importancia que ha adquirido los datos generados en Internet hoy día hace que se desarrollen técnicas para recopilarlos, analizarlos y predecir resultados en base a ellos. La Minería de Datos se relaciona con dichas técnicas y herramientas utilizadas para extraer información útil de grandes volúmenes de datos, además, permite el mejor entendimiento de un negocio y la ejecución de acciones relevantes para guiar a una organización hacia el éxito. Una de las técnicas más usadas son los Árboles de Decisión, los cuales sirven para predecir un resultado en base a un conjunto de datos.

También se destacará la importancia que supone la seguridad en la industria automovilística relacionando el aprovechamiento de los datos para conseguir modelos que puedan predecir objetivos de las empresas, obteniendo así un beneficio en tiempo y resultados.

En el presente trabajo se habla del desarrollo de un modelo haciendo uso de la Minería de Datos, centrado en la creación de un modelo a partir de Árboles de Decisión para obtener como resultado la predicción del nivel de seguridad de los vehículos en base a sus características estructurales o monetarias.

2. Minería de datos

Evolución histórica

El proceso de analizar los datos para así descubrir las conexiones que pueden existir entre ellos y predecir tendencias futuras, nos conduce al término de Minería de Datos o conocido algunas veces como el “el descubrimiento de conocimientos en bases de datos” (Febles Rodríguez, Juan Pedro, & González Pérez, Abel, 2002). La primera aparición de esta expresión se sitúa en la década de 1990 y su base está formada por tres disciplinas entrelazadas: Estadística, Inteligencia Artificial y *Machine Learning*. Lo que se creía como un proceso anticuado volvía a cobrar sentido y utilidad, la Minería de Datos continúa evolucionando para igualar el ritmo de crecimiento del *Big Data* y la accesibilidad a un mayor poder de cómputo.

En esta última década, el avance de la capacidad de procesamiento ha permitido llegar más allá de las prácticas manuales dando como resultado un análisis de datos rápido, fácil y automatizado. Por otro lado existe una relación entre la complejidad de los conjuntos de datos recopilados y el potencial para descubrir las relaciones mas relevantes entre los mismos.

Otro aspecto importante que destacar sobre la Minería de Datos es que las empresas que manejan grandes cantidades de datos como por ejemplo bancos, fabricantes o aseguradoras, entre otros, utilizan la Minería de Datos para detectar relaciones entre todos los datos utilizan la minería de datos para descubrir relaciones entre todas las cosas, desde precios, promociones y demografía hasta la forma en que la economía, el riesgo, la competencia y los medios sociales afectan sus modelos de negocios, ingresos, operaciones y relaciones con clientes (SAS Insights, 2021).

En la Ilustración 1 se puede observar como ha evolucionado la Minería de Datos desde 2010 hasta 2021 en comparación con otros procesos de datos como son el *Big Data* o el *Machine Learning*.

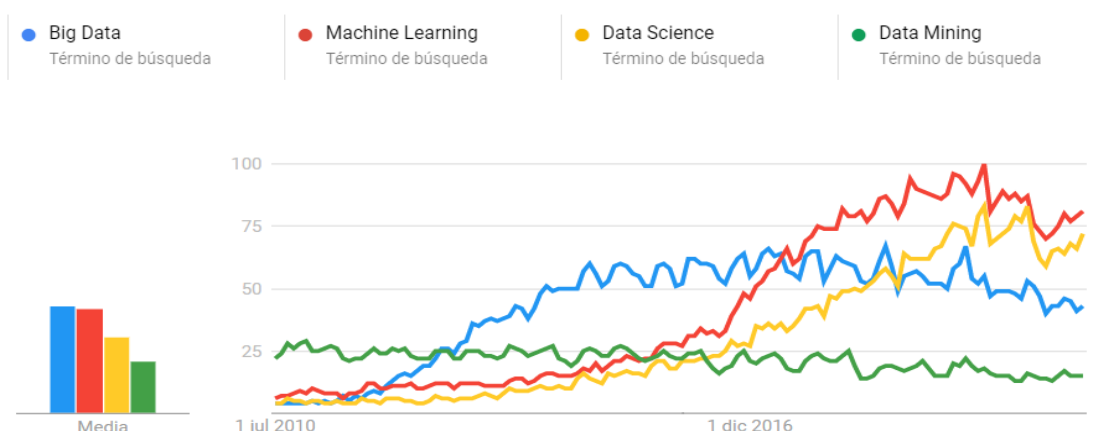


Ilustración 1. Evolución histórica de los diferentes términos sobre el procesamiento de datos

Definición

La aparición de la Minería de Datos surgió por la necesidad de ayudar a comprender grandes cantidades de datos que se puedan utilizar para sacar conclusiones que ayuden a las empresas a mejorar y desarrollarse, especialmente en términos de ventas o fidelización de clientes.

Su objetivo principal es explorar automáticamente enormes bases de datos mediante el uso de diferentes tecnologías y técnicas para encontrar patrones, tendencias o reglas repetitivas que expliquen el comportamiento de los datos recopilados a lo largo del tiempo.

La minería de datos se encuentra dentro de un proceso conocido como Descubrimiento de Conocimiento en Base de Datos o **KDD** (*Knowledge Discovery in Databases*).

Este proceso consta de las siguientes partes:

1. **Abstracción del escenario:** Para poder resolver el problema al cual nos vamos a enfrentar tenemos antes que entender el contexto de este problema y ver distintas soluciones viables. Tendremos que conocer las restricciones y propiedades del problema para después definir los objetivos que se alcanzarán.
2. **Selección de los datos:** Una vez que se recolectan los datos y se han definido los objetivos, se seleccionan los datos que se utilizaran para el estudio del problema en concreto e integrarlos en un solo conjunto. Estos datos pueden venir de varias fuentes y unirlos en un solo conjunto de datos o venir todos de la misma fuente.
3. **Limpieza y preprocesamiento:** Cuando se han recopilado y seleccionado los datos que se van a utilizar, se reescriben los datos o se eliminan variables para descartar información que no sea útil. Así se garantiza que la información a utilizar en este tipo de técnicas es útil y fiable.
4. **Transformación de los datos:** En esta etapa se transforman los datos de manera que se mejore la calidad de estos con el objetivo de que se adapten mejor al problema. Básicamente se trata de realizar cambios como por ejemplo convertir valores numéricos a categóricos.

5. **Selección de la apropiada tarea de Minería de Datos:** A continuación, y una vez que hemos conseguido entender el problema y como vamos a usar los datos, se elige el prototipo apropiado de minería de datos, ya sea la clasificación, regresión o agrupación, depende de si queremos predecir el resultado o para observar el comportamiento del modelo seleccionado.
6. **Elección del algoritmo de Minería de Datos:** Ahora tocará seleccionar el algoritmo o técnica que se va a utilizar para encontrar los patrones en el conjunto de datos y obtener una fuente de conocimiento consistente para utilizar a posteriori. Cada algoritmo funciona mejor o peor según la finalidad que se quiera obtener, pero dentro de la minería de datos podemos encontrar redes neuronales, algoritmo de clústeres o árboles de decisión.
7. **Aplicación del algoritmo:** Ya elegido el algoritmo, lo siguiente será aplicarlo en el conjunto de datos previamente seleccionado y filtrado para que sea aplicable en esta etapa. Una vez que se obtienen los resultados se vuelve a aplicar el algoritmo cambiando algunos de sus parámetros para optimizar los resultados proporcionados.
8. **Evaluación:** Habiendo obtenido los resultados, habrá que realizar una evaluación sobre la calidad del algoritmo, evaluando tanto los patrones obtenidos como su rendimiento. Una de las técnicas más famosas es la Validación Cruzada, la cual divide todo el conjunto de datos en dos partes: Una parte más grande de los datos se destinará a crear el modelo (datos de entrenamiento) y la otra parte, por lo general más pequeña, se encargará de usar el modelo obtenido con los datos de entrenamiento para realizar pruebas para ver que el algoritmo ha funcionado correctamente (datos de prueba).
9. **Aplicación:** Finalmente y si todo ha ido correctamente, el algoritmo se da por finalizado y se utiliza para resolver los problemas en el contexto necesario. Si no se han obtenido los resultados esperados habrá que volver a etapas anteriores y ver si algún cambio en los ajustes o en el modelo hace que el algoritmo satisfaga a las necesidades requeridas.

Los procesos anteriormente expuestos han sido recopilados de la siguiente fuente (Landa, 2016). En la Ilustración 2 se puede observar las fases de dichos procesos desde la materia prima, en nuestro caso serán los datos, hasta el producto que será el conocimiento obtenido.



Ilustración 2. Fases del Descubrimiento de Conocimiento en Base de Datos

Tareas de la Minería de Datos

Actualmente, dentro de la minería de datos, nos encontramos campos que combinan la tecnología de base de datos y *Data Warehousing* con técnicas de *Machine Learning* y Estadística. Una posible agrupación de estas tareas podría ser: **tareas descriptivas** y **tareas predictivas**. Las tareas descriptivas son aquellas que tienen como objetivo transformar los datos y obtener información precisa que refleje las propiedades más relevantes y generalidades de los datos. El modelo es construido a partir de ocurrencias del pasado para presentarlos de manera más comprensible. Sin embargo, las tareas predictivas buscan como objetivo un modelo sobre datos ocurridos en el pasado para predecir el comportamiento de nuevos datos del futuro.

Por otro lado, podríamos desglosar la agrupación anterior en la siguiente: **clasificación, estimación, predicción, agrupamiento (clustering), reglas de asociación y correlación**. En la Ilustración 3 se explica de forma visual las agrupaciones definidas de las tareas de la Minería de Datos.



Ilustración 3. Clasificación de tareas en la Minería de Datos

Este trabajo se sitúa dentro del grupo de tareas descriptivas, especificando un poco más, se usará la técnica de clasificación, ya que nuestro objetivo es obtener el nivel de seguridad de un vehículo dada una escala de seguridad. Ampliando un poco más el concepto de clasificación, se puede definir como la acción de identificar características de un objeto con el fin de asignar una clase o una categoría previamente definida. La clasificación puede tener un enfoque descriptivo, como conocer las variables más significativas de cada tipo de clase, o predictivo, usado para asignar una de las clases predefinidas a una nueva instancia.

Árboles de Decisión

Una de las técnicas más utilizadas en la Minería de Datos son los Árboles de Decisión. Un árbol de decisión es un algoritmo de clasificación que da como resultado un modelo esquemático (en forma de árbol) representando diferentes alternativas con posibles resultados elegidos para la misma. En cuanto a su composición, un Árbol de Decisión se compone de nodos, estos podrán ser diferentes según la función que represente. En primer lugar nos encontramos con el nodo principal en el que se produce la primera división en función de la variable más importante dentro del conjunto de datos, también existen los nodos de decisión, que vuelven a dividir el conjunto en función de las variables y finalmente los nodos hojas, cuya función será indicar la clasificación final. La Ilustración 4 muestra la estructura genérica de un Árbol de Decisión según lo explicado previamente.

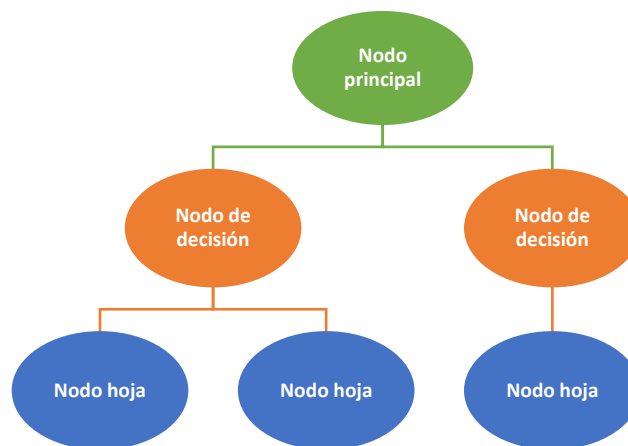


Ilustración 4. Estructura de un Árbol de Decisión

Es importante discutir porqué se ha decidido escoger el uso de esta herramienta para el problema de clasificación en el que nos encontramos, en primer lugar se destaca la facilidad para construir, interpretar y visualizar, además selecciona las variables más importantes dentro de un conjunto de datos, aportando casi la solución de este

proyecto. Por la contrapartida, existen algunas desventajas en el uso de Árboles de Decisión, como por ejemplo el sobreajuste o la creación de sesgos en el modelo si una de las clases es más numerosa que las demás. Para evitar estos problemas se ha llevado a cabo un preprocesamiento en los datos para analizar el número de cada clase dentro del conjunto de datos y una ponderación de pesos a cada clase.

3. Objetivo

El objetivo del trabajo es realizar un análisis de la seguridad en vehículos según sus características.

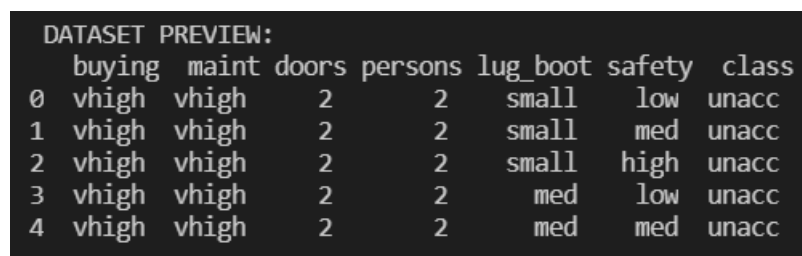
Actualmente la industria del automóvil vive una explosión tecnológica, centrada en mejorar la seguridad que proporcionan los vehículos a todas las personas que son partícipes en un accidente, tanto en el interior del vehículo como en el exterior. El estudio de la seguridad conlleva a una gran inversión por partes de las grandes empresas automovilísticas y hoy día puede ser tratada con Inteligencia Artificial.

Para la realización de este análisis, se recoge un conjunto de datos sobre la evaluación de vehículos basado en características estructurales y monetarias con el fin de poder predecir el nivel de seguridad teniendo en cuenta algunos aspectos como el tamaño del maletero, el precio de compra o el número de ocupantes entre otras cosas. Tras el análisis, se podrá obtener la predicción sobre la seguridad y cuáles son las características que influyen en la seguridad de los vehículos. Además se integrará el uso personalizado de las características con el fin de predecir con nuevos datos por parte del usuario.

4. Implementación

En primer lugar se explicará la estructura del conjunto de datos escogido, se trata de un conjunto de datos derivado de un modelo de decisión jerárquico desarrollado por Marko Bohanec (Bohanec, 1990). La evaluación de los vehículos está construida a partir de las siguientes características:

- **Buying:** Precio de compra del vehículo (low, high, med y vhigh).
- **Maint:** Precio de mantenimiento del vehículo (low, high, med y vhigh).
- **Doors:** Número de puertas del vehículo (2, 3, 4 y 5more).
- **Persons:** Número de personas (2, 4 y more).
- **Lug_boot:** Tamaño del maletero (small, med y big).
- **Safety:** Seguridad estimada (low, med y high).
- **Class:** Variable objetivo (unacc, acc, good, vgood).

A screenshot of a terminal window with a dark background and light-colored text. It displays a dataset preview with 5 rows of data. The first row is the header, and the following four rows are indexed from 0 to 4. Each row contains eight columns of data: buying, maint, doors, persons, lug_boot, safety, and class.

	buying	maint	doors	persons	lug_boot	safety	class
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc

Ilustración 5. Captura de la previsualización de los datos

Para facilitar al usuario una vista principal de los datos que van a ser tratados, la Ilustración 5 muestra una captura obtenida por consola al ejecutar el *script* del proyecto.

Profundizando un poco más en el desarrollo del *script*, se hablará del lenguaje Python en su versión 3.9.4, el cual nos proporciona las librerías necesarias para el análisis, preprocesamiento y aplicación de técnicas de minería de datos e interpretación de dicho conjunto de datos.

Análisis y preprocesamiento

Para el análisis del conjunto de datos se hace uso de la librería *Pandas*, obteniendo como resultado el número de filas y columnas del *dataset*. A continuación se eliminarán aquellas entradas que contengan valores nulos en algunos de sus campos, ya que podría producir “ruido” en nuestro conjunto de datos. Finalmente se analiza la frecuencia de los distintos valores en cada una de las columnas, este punto es bastante importante ya que en nuestro caso nos enfrentamos a valores desequilibrados con respecto a la variable objetivo. La Ilustración 7 representa la captura obtenida de la consola para obtener el tamaño del conjunto de datos.

```

VALUES FREQUENCY BY COLUMN:
med      432
low      432
vhigh    432
high     432
Name: buying, dtype: int64

med      432
low      432
vhigh    432
high     432
Name: maint, dtype: int64

2        432
3        432
5more    432
4        432
Name: doors, dtype: int64

2        576
more     576
4        576
Name: persons, dtype: int64

med      576
small    576
big      576
Name: lug_boot, dtype: int64

med      576
low      576
high     576
Name: safety, dtype: int64

unacc    1210
acc       384
good      69
vgood     65
Name: class, dtype: int64

```

Ilustración 7. Frecuencia de los valores del conjunto de datos

Como se muestra en la Ilustración 6 el número de entradas para el valor “unacc” es mucho mayor que para el resto de los valores objetivos pudiéndose producir un sobreajuste en el modelo. Para solucionar este problema se hará uso de técnicas para balancear los pesos de las variables objetivos como se explicará posteriormente.

```

SHAPE OF DATASET (ROWS, COLUMNS):
(1728, 7)

```

Ilustración 6. Tamaño del conjunto de datos

El siguiente paso consistirá en aplicar una codificación de las variables categóricas pertenecientes al conjunto de datos, en este caso todas las columnas están formadas por valores categóricos, para ello se hace uso de la librería *category_encoders*.

Una vez finalizado el proceso de análisis y preprocesamiento de datos se da paso al entrenamiento del árbol de decisión.

Modelo de clasificación

Para la generación de modelos de clasificación como son los Árboles de Decisión, Python nos ofrece la librería *Scikit-learn* que nos proporcionará sencillez y facilidad para crear nuestro modelo.

La Ilustración 9 muestra la creación del Árbol de Decisión mediante el tipo *DecisionTreeClassifier* cuyos parámetros describen el propio comportamiento de la construcción a partir del algoritmo ID3.

```
decision_tree = DecisionTreeClassifier(criterion = 'entropy', random_state = 40, class_weight='balanced')
model = decision_tree.fit(X_train, y_train)
```

Ilustración 8. Creación del Árbol de Decisión

El ID3 construye un árbol de decisión desde arriba hasta abajo, directamente, sin hacer uso de *Backtracking*, y basándose únicamente en los datos iniciales proporcionados. Para ello, usa el concepto de **Ganancia de Información** para seleccionar el atributo más útil en cada paso. En cierta forma, sigue un método voraz para decidir la pregunta que mayor ganancia da en cada paso, es decir, aquella que permite separar mejor los ejemplos respecto a la clasificación final (Caparrini, 2018).

En cuanto a la entropía:

- Una **muestra completamente homogénea** (es decir, en la que todos se clasifican igual) tiene incertidumbre mínima, ya que no hay dudas de cuál es la clasificación de cualquiera de sus elementos. En este caso, la entropía tomará como valor 0.
- Una **muestra igualmente distribuida** (es decir, que tiene el mismo número de ejemplos de cada posible clasificación), tendrá como resultado una incertidumbre máxima, en el sentido de que es la peor situación para poder saber cuál sería la clasificación. Así pues, fijaremos la incertidumbre a 1.

Finalmente el modelo es entrenado con los datos de entrenamiento y las características escogidas con respecto a la entropía se muestran en la Ilustración 9.

FEATURE SCORES:	
safety	0.230466
buying	0.202992
maint	0.175338
persons	0.163519
lug_boot	0.163209
doors	0.064477

Ilustración 9. Ganancia de información por clase

Como se puede observar, la ganancia de información (entropía) por cada clase, en este modelo el atributo con mayor ganancia de información es “*safety*”, bastante igualado con el atributo “*buying*” lo que dará lugar a que el Árbol de Decisión parta desde el nodo raíz “*safety*”.

Métricas del modelo

Una vez se ha construido el modelo clasificador se analizará un conjunto de métricas obtenidas tras hacer uso del Árbol de Decisión para predecir los datos de pruebas.

CONFUSION MATRIX:				
[95	3	16	3]
[2	18	0	0]
[8	0	353	0]
[0	0	0	21]]
ACCURACY :				
93.83429672447014				
REPORT :				
	precision	recall	f1-score	support
acc	0.90	0.81	0.86	117
good	0.86	0.90	0.88	20
unacc	0.96	0.98	0.97	361
vgood	0.88	1.00	0.93	21
accuracy			0.94	519
macro avg	0.90	0.92	0.91	519
weighted avg	0.94	0.94	0.94	519

Ilustración 10. Métricas del modelo

En la Ilustración 10 se detalla en primer lugar la matriz de confusión del modelo, aportando la información del desempeño del algoritmo. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real (Wikipedia, 2021). Como se puede comprobar el modelo clasifica el valor “*unacc*” en 353 muestras y solo falla en 24 (16 + 8), dando lugar a una buena evaluación del modelo, pudiéndose contrastar con el valor de precisión de un 93.8% en los datos de pruebas. Finalmente se agregan algunas métricas de calidad como son: *Precision* ($TP/(TP+FP)$), *Recall* ($TP/(TP+FN)$) y *F1-Score* ($((2 \times Precision \times Recall)/(Precision + Recall))$).

En la Ilustración 11 se puede observar una matriz de confusión genérica que servirá para explicar las métricas anteriormente expuestas, dicha matriz esta compuesta por filas que representan el número de predicciones por cada una de las clases, mientras que cada columna representa el número de clases reales. La métrica *Precision* medirá la calidad del modelo en cuanto a clasificación, *Recall* informará sobre la cantidad que el modelo es capaz de identificar y *F1* proporcionará un valor referente a la combinación de las anteriores.

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Ilustración 11. Matriz de confusión genérica

5. Pruebas

En esta sección se usará el modelo de clasificación para predecir el nivel de seguridad de un vehículo a través de los parámetros que le pasará el usuario por consola. En la primera prueba reflejada en la Tabla 1 el usuario proporciona las entradas expuestas en la tabla dando como resultado una predicción como un vehículo con nivel de seguridad “*unacc*” mientras que en la segunda prueba capturada en la Tabla 2, los datos que introduce el usuario dan lugar a un vehículo con el mayor nivel de seguridad.

PRUEBA 1		
buying	low	PRUEBA MANUAL: Precio de compra (low, high, med, vhigh): low Precio de mantenimiento (low, high, med, vhigh): low Número de puertas (2, 3, 4, 5more): 2 Número de personas (2, 4, more): 2 Tamaño del maletero (small, med, big): small Nivel de seguridad estimado (low, high, med): low Predicción del nivel de seguridad: unacc
maint	low	
doors	2	
persons	2	
lug_boot	small	
safety	low	
Predicción		unacc (Inaceptable)

Tabla 1. Prueba 1

PRUEBA 2		
buying	med	PRUEBA MANUAL: Precio de compra (low, high, med, vhigh): med Precio de mantenimiento (low, high, med, vhigh): med Número de puertas (2, 3, 4, 5more): 2 Número de personas (2, 4, more): 4 Tamaño del maletero (small, med, big): big Nivel de seguridad estimado (low, high, med): high Predicción del nivel de seguridad: vgood
maint	med	
doors	2	
persons	4	
lug_boot	big	
safety	high	
Predicción		vgood (Muy buena)

Tabla 2. Prueba 2

6. Conclusión

A modo de concluir todo lo expuesto en este trabajo, se puede decir que es posible predecir el nivel de seguridad de un vehículo basándonos en factores estructurales y monetarios, además con la ayuda del lenguaje de Python y sus librerías el grupo de estudiantes ha podido desarrollar un modelo con métricas bastante buenas, ampliando el conocimiento en el campo de la Minería de Datos.

Desde el punto de vista de la industria automovilística podría ser un gran avance el estudio de los diversos factores en el diseño de vehículos ya que podrían determinar la seguridad de un vehículo basándose en el tipo de estructura o material utilizado. Además hoy día se ha progresado bastante en los coches autónomos, reforzando la importancia de la seguridad de las personas en los mismos.

Bibliografía

Bohanec, M. (1990). Retrieved from <https://archive.ics.uci.edu/ml/datasets/car+evaluation>

Caparrini, F. S. (2018). Retrieved from <http://www.cs.us.es/~fsancho/?e=104>

Febles Rodríguez, Juan Pedro, & González Pérez, Abel. (2002). Retrieved from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000200003

Landa, J. (2016). Retrieved from <http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/>

SAS Insights. (2021). Retrieved from https://www.sas.com/es_es/insights/analytics/data-mining.html

Wikipedia. (2021). Retrieved from https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n