

## Objetivo de la investigación

El objetivo principal de esta investigación fue doble:

1. **Desarrollar y evaluar modelos predictivos** utilizando técnicas de minería de dato (data mining) para predecir con una precisión razonable la severidad de las lesiones (mortal, grave, moderada o leve) en futuros accidentes de tráfico en Abu Dhabi, Emiratos Árabes Unidos (EAU).
2. **Identificar los factores más influyentes** que contribuyen a la gravedad de los accidentes. El propósito final era establecer un conjunto de reglas claras que pudieran ser utilizadas por las agencias de tráfico de los EAU para diseñar estrategias más efectivas, mejorar la seguridad vial y asignar recursos de manera más eficiente.

## Metodología Aplicada

Para lograr sus objetivos, los investigadores siguieron un proceso de estructurado basado en el análisis de datos históricos:

- **Conjunto de datos:** Se utilizó una base de datos que contenía 5,973 registros de accidentes de tráfico ocurridos en Abu Dhabi durante un período de 6 años, desde 2008 hasta 2013. Cada registro incluía múltiples variables relacionadas con el conductor, el vehículo, la vía y las circunstancias del accidente.
- **Software y Técnicas:** Emplearon el software de minería de datos WEKA (Waikato Environment for Knowledge Analysis) para construir y probar los modelos. Se aplicaron cuatro algoritmos de clasificación bien conocidos:
  - **Árbol de Decisión (J48):** Un modelo basado en reglas que divide los datos según las variables más significativas.
  - **Inducción de Reglas (PART):** Un algoritmo que genera un conjunto de reglas IF-THEN para clasificar los datos.
  - **Naive Bayes:** Un clasificador probabilístico basado en el teorema de Bayes.
  - **Multilayer Perceptron (MLP):** Un tipo de red neuronal artificial, un modelo de aprendizaje automático más complejo.
- **Proceso de Evaluación:** La efectividad de cada modelo se evaluó de tres maneras para asegurar la robustez de los resultados:
  - **Entrenamiento con todo el dataset:** Se usó el conjunto de datos completo para entregar el modelo, como una prueba inicial.
  - **Validación Cruzada (10-fold cross-validation):** El método estándar, donde los datos se dividen en 10 partes, usando 9 para entrenar y 1 para probar, repitiendo el proceso 10 veces para obtener una medida de precisión más fiable.
  - **Re-muestreo y Validación Cruzada:** Dado que los accidentes leves eran mucho más comunes que los mortales (datos desequilibrados), los investigadores re-muestrearon los datos para crear una distribución más uniforme. Luego, aplicaron de nuevo la validación cruzada. Este paso fue crucial para mejorar la capacidad de los modelos para predecir las clases menos frecuentes (accidentes graves y mortales).
- **Extracción de Conocimiento:** Para identificar los factores clave, se analizaron en detalle las reglas generadas por el algoritmo de Árbol de Decisión (J48), ya que son fáciles de interpretar.

## Conclusiones Principales

La investigación llegó a varias conclusiones significativas tanto en el rendimiento de los modelos como en los factores de riesgo identificados:

- **Rendimiento de Modelos**
  - Los modelos de Árboles de Decisión (J48), Inducción de Reglas (PART) y la Red Neuronal (MLP) mostraron una precisión predictiva general similar y robusta (entre 71% y 74% con validaciones cruzadas). El modelo Naive Bayes tuvo un rendimiento notablemente inferior.
  - La precisión de los modelos mejoró significativamente después de re-muestrear los datos para equilibrar las clases. Por ejemplo, la precisión del modelo J48 aumento de un 73.6% a un 88.1%, demostrando ser una técnica muy efectiva para este tipo de problema.
- **Factores Clave de Severidad**
  - Los factores más determinantes para predecir la severidad de un accidente fueron: la edad del implicado, el género, la nacionalidad, el año del accidente, el tipo de víctima (conductor, pasajero o peatón) y el tipo de colisión.
  - El grupo de edad más vulnerable es el de 18 a 30 años, lo que sugiere la necesidad de campañas de concienciación y educación vial dirigidas a jóvenes conductores.
  - Los hombres se ven involucrados en accidentes con más frecuencia que las mujeres.
  - Los conductores son el tipo de víctima más común, seguido de los pasajeros y los peatones.
  - Se observó una tendencia clara a la reducción del número de accidentes a lo largo del período de estudio (2008-2013), lo que podría atribuirse a nuevas regulaciones de tráfico implementadas.
  - Las colisiones en ángulo recto fueron las más numerosas, seguidas de los atropellos a peatones y las colisiones por alcance.

Projecto.Rmd

Source

Visual

24 # 1.1

25 library(tidyverse)

26 library(dplyr)

27 library(ggplot2)

28 ~

29 ~

30 ~

31 # 1.2

32 cities <- read\_csv("../Dataset/cities.csv", show\_col\_types = FALSE)

33 meal\_types <- read\_csv("../Dataset/meal\_types.csv", show\_col\_types = FALSE)

34 meals <- read\_csv("../Dataset/meals.csv", show\_col\_types = FALSE)

35 members <- read\_csv("../Dataset/members.csv", show\_col\_types = FALSE)

36 monthly\_member\_totals <- read\_csv("../Dataset/monthly\_member\_totals.csv", show\_col\_types = FALSE)

37 order\_details <- read\_csv("../Dataset/order\_details.csv", show\_col\_types = FALSE)

38 orders <- read\_csv("../Dataset/orders.csv", show\_col\_types = FALSE)

39 restaurant\_types <- read\_csv("../Dataset/restaurant\_types.csv", show\_col\_types = FALSE)

40 restaurants <- read\_csv("../Dataset/restaurants.csv", show\_col\_types = FALSE)

41 serve\_types <- read\_csv("../Dataset/serve\_types.csv", show\_col\_types = FALSE)

42 ~

43 ~

44 ~

45 # 1.3

46 # 1.4

47 cities

48 head(cities)

49 str(cities)

50 ~

spec\_tbl\_df

tbl\_df

R Console

spec\_tbl\_ [5 x 2] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

\$ id : num [1:5] 1 2 3 4 5

\$ city: chr [1:5] "Tel Aviv" "Ramat Gan" "Ramat Hasharon" "Herzel" ...

- attr(\*, "spec")=

.. cols(

.. id = col\_double(),

.. city = col\_character()

.. )

- attr(\*, "problems")=externalptr

329/32

Chunk 24

R Markdown

Console

Terminal

Background Jobs

R 4.4.2

> print("Guillermo Rene Medinilla Hernández", "999017289")

Error in print.default("Guillermo Rene Medinilla Hernández", "999017289") :

invalid printing digits 999017289

> print("Guillermo Rene Medinilla Hernández - 999017289")

[1] "Guillermo Rene Medinilla Hernández - 999017289"

> |

Environment

History

Connections

Tutorial

R

Global Environment

Data

airlines 14 obs. of 2 variables

airports 1251 obs. of 8 variables

all\_wars\_ma\_ num [1:3, 1:3] 461 290 309 314\_

answer\_coun\_ 243930 obs. of 2 variables

answers 380643 obs. of 4 variables

by\_continent 5 obs. of 3 variables

by\_continen\_ 5 obs. of 3 variables

by\_year 12 obs. of 3 variables

by\_year\_con\_ 60 obs. of 3 variables

chickwts 71 obs. of 2 variables

cities 5 obs. of 2 variables

city\_expense 5 obs. of 2 variables

colors 179 obs. of 3 variables

Files

Plots

Packages

Help

Viewer

Presentatio

Home

Name

Size

Modifier

.RData 78.9 MB Apr 8,

.Rhistory 16.8 KB Apr 9,

Applications

Applications (Parallels)

Desktop

Documents

Downloads

GNS3

Library

MacBook-Pro-de-71826.local.x... 101.8 KB Dec 15

Movies

Music

Parallels

Pictures

Postman

Public

VirtualBox VMs

Advanced settings