# PROYECTO INTEGRADOR



Imagen by: DALL-E

Sistema de prevención de accidentes

Por: Erick Alexei Cambray Servin

GitHub : <https://github.com/AlexeiEACS/Diplo_CienciaDatos_UNAM>

# Desarrollo

## Justificación de proyecto

Dados los multiples accidentes causados por falta de habilidad humana al manejar, condiciones climáticas y también como es que ha sido influencia de la zona en donde ocurren.

Teniendo esto en mente y con el uso de una base de datos sobre choces en Estados Unidos entre los años 2016 y 2021

Moosavi, S. (2018). US Accidents. Kaggle. <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

Con ayuda de las nuevas tecnologías que se encuentran en desarrollo en los automoviles podemos hacer que con una conexión de red pueda tener ayudas que tengan en alerta al conductor.

## Procesamiento y Limpieza de datos

Las columnas que tenemos son las siguientes:

* *ID: identificador único para cada accidente*
* *Grado Severidad: gravedad del accidente, 0 y 1 donde 1 es muy grave.*
* *Start\_Time y End\_Time: fecha y hora de inicio y fin del accidente.*
* *Start\_Lat y Start\_Lng: Latitud y Longitud de la ubicación donde comenzó el accidente.*
* *End\_Lat y End\_Lng: Latitud y Longitud de la ubicación donde terminó el accidente.*
* *Distance(mi): Distancia en millas entre el inicio y el final del accidente.*
* *Description: descripción textual del accidente.*
* *Number, Street, Side, City, County, State, Zipcode, Country, Timezone, Airport\_Code: detalles sobre la ubicación del accidente.*
* *Weather\_Timestamp: fecha y hora de la información meteorológica relacionada con el accidente.*
* *Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind\_Direction, Wind\_Speed(mph), Precipitation(in), Weather\_Condition: información meteorológica.*
* *Amenity, Bump, Crossing, Give\_Way, Junction, No\_Exit, Railway, Roundabout, Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop: indicadores para identificar características de la ubicación del accidente.*
* *Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, Astronomical\_Twilight: información sobre la posición del sol en relación con la hora del accidente.*

Ahora hacemos una clasificación con la que podemos operar las variables

1. **um** (user-defined): una lista con la variable **ID** que será utilizada como identificador único de cada accidente.
2. **trgt** (target): una lista con la variable **Grado\_Severidad** que será utilizada como la variable objetivo a predecir.
3. **cont\_var** (variables continuas): una lista con variables que representan valores continuos, como la distancia, la temperatura, la velocidad del viento.
4. **cat\_var** (variables categóricas): una lista con variables que representan valores categóricos, como la ciudad, el estado, el código postal, la condición meteorológica.
5. **bin\_var** (variables binarias): una lista con variables que representan valores binarios, como indicadores para identificar características de la ubicación del accidente, como la presencia de una colina, un cruce, etc.

## Limpieza

Los modelos necesitan valores numéricos por lo que las variables faltantes que no sean de identificación las podemos imputar, las variables numéricas serán imputadas con la ‘media’ y las categóricas con ‘el más frecuente’.

Las variables faltantes se tendrían que eliminar en los que no se puede identificar como por ejemplo la variable target, los datos del clima y del día en el momento del choque. Los cuales tendría que ser eliminados ya que al ser variables que no se pueden imputar es importante no suponer valores para ellas.

## Procesamiento

Para poder aplicar la modelación se seguirá la ponderación WoE que involucra:

Normalización:

Para las variables discretas se verá su frecuencia y si es que su frecuencia es menor de 5% de los datos de la columna se agruparán en una categoría llamada “Otros” para poder tener solo las que aporten más información.

Discretización

Los valores continuos se cortarán para poder hacer cortes en intervalos que tengan mayor poder predictivo. Esto se logra gracias al IV (information value).

### Descripción

El IV es una medida de la calidad predictiva de una variable independiente, que refleja la cantidad de información que aporta una variable sobre la variable dependiente. El IV se calcula como la suma de la diferencia entre la probabilidad de la variable dependiente para cada nivel de la variable independiente y la probabilidad de la variable dependiente en general.

El WoE es una transformación de las variables independientes que representa la relación entre la variable dependiente y la variable independiente. El WoE se calcula como el logaritmo de la relación entre la probabilidad de la variable dependiente para cada nivel de la variable independiente y la probabilidad de la variable dependiente en general.

*WoEi = ln((P(y = 1 | x = i)) / P(y = 1))*

## TAD

Una vez que ya tenemos los valores convertidos como se muestra a continuación, podemos entonces prepar el proceso de usar regresión logistica.



# Modelación

Dada la preparación utlizada se usará el modelo de Regresión Logística

### Justificación

El WoE es una medida de la correlación entre una variable categórica y la variable objetivo. Se usa para transformar las variables categóricas en valores numéricos que pueden ser utilizados en un modelo de regresión logística. El WoE reemplaza los valores de la variable categórica con una medida de la fuerza de la relación entre esa categoría y la variable objetivo. De esta manera, se puede evaluar el impacto de cada categoría en la probabilidad de la variable objetivo y hacer una selección más informada de las variables para incluir en el modelo.

Fuente: IBM Corp. (s.f.). Regresión logística. En SPSS Statistics (Versión en línea). Recuperado de <https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-logistic>

## Pasos para la aplicación del modelo

1. Partición de los datos en entrenamiento y validación.
2. Uso del algoritmo TomekLinks para poder hacer un submuestreo de los datos de clase mayoritaria ( En este caso la gravedad más alta 1)
3. Declaración del modelo
4. Métricas de rendimiento del modelo

### Métricas obtenidas

Accuracy: 89%

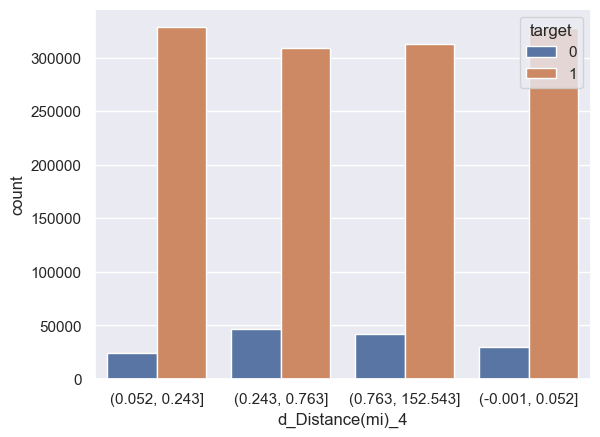
Recall: 99%

F1\_score: 94%

## Gráficas de validación

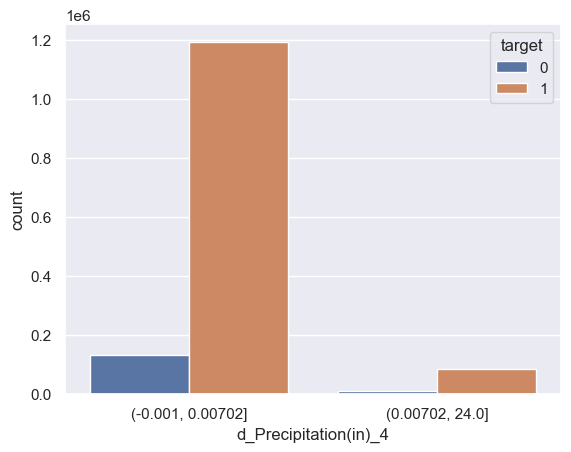
Vamos a ver algunos ejemplos de como es que se ven las gráficas algunas de las columnas usadas para la predicción:

Teniendo la división de la distancia



Como vemos es claro que los accidentes aunque son relevantes en mucho más grande la ocurrencia de los accidentes más graves.

Vemos también que la lluvia cuando es fuerte es cuando tenemos valores más altos de gravedad de los accidentes



# Propuesta para la solución

Tener el dato si es que el accidente será grave o no puede que no sea suficiente para poder dar opciones de uso de estos datos. Pero si la probailidad por los que se presenta la siguiente solucioón:

|  |  |
| --- | --- |
| Probabilidad de accidente grave | Acciones |
| De 0% a 30% | Mostrar en la pantalla del carro una alerta de cuidado por la zona y el clima |
| De 30% a 60% | Reducir la velocidad máxima del carro y seguir con las alertas |
| De 60% a 80% | No permitir conducción autónoma, mantener las restricciones de velocidad y tener |
| Mayor de 80% | Además de las acciones anteriores se tiene que entrar en modo alerta, dejando las luces prendidas y al menor indicio de choque mandar alertas al sistema de seguridad (911) |

# Concluciones

Lo logrado hasta este módulo no solo ha ayudado el uso final de los modelos teniendo usos reales a la programación para poder ayudar a salvar la vida de perosnas.

Lo que se tendrá que hacer para los demás módulos será. Poder armar com más datos nuestra base para tener claro lo que causa este tipo de accidentes y si es que hay alguna forma de poder hacer que dentro del carro como usos de visión computacionar que tengan refuerzos con los datos generados hasta ahora.

Gracias☺