Avance Proyecto Final

Integrantes:

Sergio Rojas
Cristhian Amaya
Andrés Beltrán
Gloria Ramos

¿Cuál es la probabilidad de que haya un accidente fatal en un accidente automovilístico en México?

1. **Problema que abordarán y su contexto.**

Contexto:

México es un país latinoamericano con una densa red de carreteras y una alta tasa de movilidad. En 2021, El Sistema Nacional de Seguridad Pública de México registró un total de aproximadamente 296.000 accidentes automovilísticos en el país. De estos accidentes, más de 23.000 resultaron en fatalidades, lo que representó un preocupante aumento en comparación con años anteriores.

Este incremento en las cifras de accidentes fatales ha suscitado una profunda preocupación en las autoridades de tránsito y en la población en general. Se ha identificado que varios factores contribuyen a esta situación, incluyendo el exceso de velocidad, el consumo de alcohol, la falta de cumplimiento de las normas de tránsito y el estado de las carreteras. Además, la falta de implementación efectiva de políticas de seguridad vial ha dejado una huella importante en la seguridad en las carreteras mexicanas.

El impacto social y económico de estos accidentes es significativo. Las fatalidades en accidentes automovilísticos afectan a las familias y comunidades, y también generan costos médicos y rehabilitación considerables, además de pérdidas de productividad en el país.

El presente estudio, que pretende responder a la pregunta de ¿Cuál es la probabilidad de que haya un accidente fatal en un accidente automovilístico en México? Se basa en la necesidad de abordar esta problemática y reducir las consecuencias devastadoras que estos accidentes tienen en México. La información recopilada y analizada proporcionada por El instituto Nacional de Estadística y Geografía (INEGI) de México será fundamental para comprender las tendencias específicas de 2021 y contribuir a la implementación de estrategias efectivas de seguridad vial en el país.

El análisis de datos, respaldado por la información que se capta mediante el SICATUS, permitirá identificar patrones y factores de riesgo asociados a accidentes fatales, lo que, a su vez, proporcionará una base sólida para la toma de decisiones informadas y promoción de prácticas de conducción más seguras en México.

2. **Pregunta de negocio y alcance del proyecto.**

Pregunta de negocio: ¿Cuál es la probabilidad de que algún implicado peatonal, pasajero o conductor resulte en deceso en un accidente automovilístico?

Alcance del proyecto:

El alcance del proyecto se enfocará en los siguientes aspectos:

- Recopilación de datos: Se recopilarán datos de accidentes automovilísticos ocurridos en México durante el año 2021. Estos datos incluirán un total de 46 variables y un total de 3.849 registros.
- Limpieza y preparación de los datos: Preparación de los datos para garantizar la calidad y coherencia de la información recopilada.
- Modelado de los datos: Se aplicarán técnicas estadísticas y modelado de datos para calcular la probabilidad de accidentes fatales en función de las variables recopiladas. Esto incluirá el uso de técnicas de clasificación.
- Identificación de factores de riesgo: Se identificarán los factores de riesgo más influyentes en la probabilidad de accidentes fatales. Esto permitirá comprender qué variables tienen mayor correlación con accidentes mortales.
- Informe final: Construcción de un tablero dinámico que permita evidenciar los resultados.

3. Conjuntos de datos a emplear.

Las fuentes de datos son originadas por INEGI por medio de un cuestionario de accidentes de tránsito la cual recopila dentro de las 46 variables cuatro tipos de datos, registro de localización (código de municipio, zona urbana y zona sub urbana), registro de hora y fecha en que sucede el accidente (Año, hora, minutos, día de la semana, día), datos del conductor y datos del vehículo.

El nombre de la fuente de datos original es “Estadística de Accidentes de Tránsito Terrestre en Zonas Urbanas y SubUrbanas”.

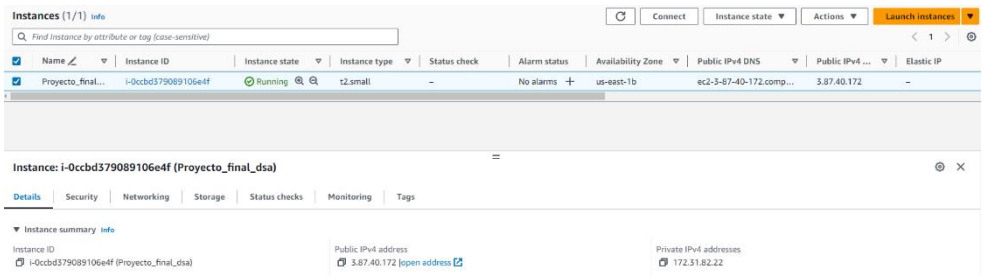
Adicional se cuenta con una tabla de referencia de los Municipios de México, que tiene información del código del municipio y el nombre asociado. Esta base es generada por el Instituto Nacional para el Federalismo y Desarrollo Municipal.

Fuente de datos original y diccionario: https://www.inegi.org.mx/rnm/index.php/catalog/788/data-dictionary/F1?file_name=ATU

4. Repositorio Git en uso para el código y Repositorio DVC en uso para los datos.

URL repositorio:
https://github.com/SergioRojas86/Proyecto_final_Despliegue_de_soluciones_analiticas

Creación Instancia



Inicialización git y dvc

```
(env-dvc) ubuntu@ip-172-31-82-22:~$ git --version
git version 2.34.1
(env-dvc) ubuntu@ip-172-31-82-22:~$ cd ~
(env-dvc) ubuntu@ip-172-31-82-22:~$ mkdir proyecto_final_dsa
(env-dvc) ubuntu@ip-172-31-82-22:~$ ls
env-dvc  proyecto_final_dsa
(env-dvc) ubuntu@ip-172-31-82-22:~$ cd proyecto_final_dsa/
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ mkdir Sergio
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ mkdir Gloria
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ mkdir Andres
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ mkdir Cristhian
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ ls
Andres  Cristhian  Gloria  Sergio
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ cd Sergio
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ git init
hint: Using 'master' as the name for the initial branch. This default branch name
hint: is subject to change. To configure the initial branch name to use in all
hint: of your new repositories, which will suppress this warning, call:
hint:
hint:   git config --global init.defaultBranch <name>
hint:
hint: Names commonly chosen instead of 'master' are 'main', 'trunk' and
hint: 'development'. The just-created branch can be renamed via this command:
hint:
hint:   git branch -m <name>
Initialized empty Git repository in /home/ubuntu/proyecto_final_dsa/Sergio/.git/
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ git branch -m main
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ dvc init
Initialized DVC repository.
```

```
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ git status
On branch main

No commits yet

Changes to be committed:
  (use "git rm --cached <file>..." to unstage)
        new file:   .dvc/.gitignore
        new file:   .dvc/config
        new file:   .dvcignore

(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ git commit -m "Inicialización de DVC"
[main (root-commit) 92f7c48] Inicialización de DVC
 Committer: Ubuntu <ubuntu@ip-172-31-82-22.ec2.internal>
Your name and email address were configured automatically based
on your username and hostname. Please check that they are accurate.
You can suppress this message by setting them explicitly. Run the
following command and follow the instructions in your editor to edit
your configuration file:

    git config --global --edit

After doing this, you may fix the identity used for this commit with:

    git commit --amend --reset-author

 3 files changed, 6 insertions(+)
 create mode 100644 .dvc/.gitignore
 create mode 100644 .dvc/config
 create mode 100644 .dvcignore
```

```
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ dvc add data/atus_anual_2021.csv
100% Adding...|

To track the changes with git, run:

        git add data/.gitignore data/atus_anual_2021.csv.dvc

To enable auto staging, run:

        dvc config core.autostage true
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ dvc add data/inafed_bd_1679023638.xlsx
100% Adding...|

To track the changes with git, run:

        git add data/.gitignore data/inafed_bd_1679023638.xlsx.dvc

To enable auto staging, run:

        dvc config core.autostage true
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ cat data/.gitignore
/atus_anual_2021.csv
/inafed_bd_1679023638.xlsx
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ cd data
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio/data$ ls -la
total 91896
drwxrwxr-x 2 ubuntu ubuntu     4096 Oct 26 02:46 .
drwxrwxr-x 5 ubuntu ubuntu     4096 Oct 26 02:27 ..
-rw-rw-r-- 1 ubuntu ubuntu       48 Oct 26 02:46 .gitignore
-rw-rw-r-- 1 ubuntu ubuntu 93893470 Oct 26 02:46 atus_anual_2021.csv
-rw-rw-r-- 1 ubuntu ubuntu      103 Oct 26 02:46 atus_anual_2021.csv.dvc
-rw-rw-r-- 1 ubuntu ubuntu   183145 Oct 26 02:46 inafed_bd_1679023638.xlsx
-rw-rw-r-- 1 ubuntu ubuntu      107 Oct 26 02:46 inafed_bd_1679023638.xlsx.dvc
```

```
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Sergio$ git push --set-upstream origin main
Username for 'https://github.com': SergioRojas86
Password for 'https://SergioRojas86@github.com':
Enumerating objects: 15, done.
Counting objects: 100% (15/15), done.
Compressing objects: 100% (11/11), done.
Writing objects: 100% (14/14), 1.40 KiB | 717.00 KiB/s, done.
Total 14 (delta 2), reused 0 (delta 0), pack-reused 0
remote: Resolving deltas: 100% (2/2), done.
To https://github.com/SergioRojas86/Proyecto_final_Despliegue_de_soluciones_analiticas.git
   51ecaa4..3445612  main -> main
Branch 'main' set up to track remote branch 'main' from 'origin'.
```

Clonación repositorio Gloria

```
ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ cd Gloria
ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Gloria$ git clone https://github.com/SergioRojas86/Proyecto_final_Despliegue_de_soluciones_analiticas.git
Cloning into 'Proyecto_final_Despliegue_de_soluciones_analiticas'...
remote: Enumerating objects: 26, done.
remote: Counting objects: 100% (26/26), done.
remote: Compressing objects: 100% (17/17), done.
remote: Total 26 (delta 5), reused 18 (delta 3), pack-reused 0
Receiving objects: 100% (26/26), done.
Resolving deltas: 100% (5/5), done.
```

```
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Gloria/Proyecto_final_Despliegue_de_soluciones_analiticas$ dvc pull
Collecting
Fetching
Building workspace index
Comparing indexes
Applying changes
A       data/atus_anual_2021.csv
A       data/inafed_bd_1679023638.xlsx
2 files added and 2 files fetched
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Gloria/Proyecto_final_Despliegue_de_soluciones_analiticas$ ls -la data
total 91896
drwxrwxr-x 2 ubuntu ubuntu     4096 Oct 26 03:31 .
drwxrwxr-x 5 ubuntu ubuntu     4096 Oct 26 03:13 ..
-rw-rw-r-- 1 ubuntu ubuntu       48 Oct 26 03:13 .gitignore
-rw-rw-r-- 1 ubuntu ubuntu 93893470 Oct 26 03:31 atus_anual_2021.csv
-rw-rw-r-- 1 ubuntu ubuntu      103 Oct 26 03:13 atus_anual_2021.csv.dvc
-rw-rw-r-- 1 ubuntu ubuntu   183145 Oct 26 03:31 inafed_bd_1679023638.xlsx
-rw-rw-r-- 1 ubuntu ubuntu      107 Oct 26 03:13 inafed_bd_1679023638.xlsx.dvc
```

Clonación repositorio Cristhian

```
(env-dvc) ubuntu@ip-172-31-82-22:~$ cd proyecto_final_dsa
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ ls
Andres  Cristhian  Gloria  Sergio
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa$ cd Cristhian
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Cristhian$ git clone https://github.com/SergioRojas86/Proyecto_final_Despliegue_de_soluci
ones_analiticas.git
Cloning into 'Proyecto_final_Despliegue_de_soluciones_analiticas'...
remote: Enumerating objects: 30, done.
remote: Counting objects: 100% (30/30), done.
remote: Compressing objects: 100% (21/21), done.
remote: Total 30 (delta 5), reused 22 (delta 3), pack-reused 0
Receiving objects: 100% (30/30), 4.19 KiB | 4.19 MiB/s, done.
Resolving deltas: 100% (5/5), done.
```

```
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Cristhian/Proyecto_final_Despliegue_de_soluciones_analiticas$ dvc pull
Collecting                                                                                                    |2.00 [00:00,  172entry/s]
Fetching
Building workspace index                                                                                      |1.00 [00:00,  103entry/s]
Comparing indexes                                                                                             |4.00 [00:00,  959entry/s]
Applying changes                                                                                              |2.00 [00:00,  15.0file/s]
A       data/inafed_bd_1679023638.xlsx
A       data/atus_anual_2021.csv
2 files added and 2 files fetched
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Cristhian/Proyecto_final_Despliegue_de_soluciones_analiticas$ ls -la data
total 91896
drwxrwxr-x 2 ubuntu ubuntu     4096 Oct 26 04:29 .
drwxrwxr-x 5 ubuntu ubuntu     4096 Oct 26 04:29 ..
-rw-rw-r-- 1 ubuntu ubuntu       48 Oct 26 04:29 .gitignore
-rw-rw-r-- 1 ubuntu ubuntu 93893470 Oct 26 04:29 atus_anual_2021.csv
-rw-rw-r-- 1 ubuntu ubuntu      103 Oct 26 04:29 atus_anual_2021.csv.dvc
-rw-rw-r-- 1 ubuntu ubuntu   183145 Oct 26 04:29 inafed_bd_1679023638.xlsx
-rw-rw-r-- 1 ubuntu ubuntu      107 Oct 26 04:29 inafed_bd_1679023638.xlsx.dvc
```

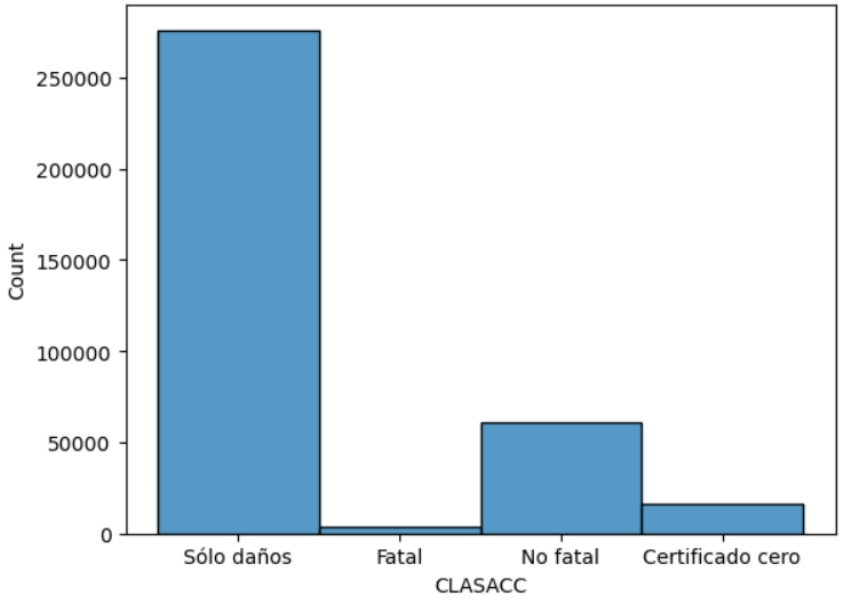
Clonación repositorio Andrés

```
rm: remove write-protected regular file 'Proyecto_final_Despliegue_de_soluciones_analiticas/.git/objects/pack/pack-8983a6218992bfb934096681a677fc79c07e38df.idx'? y
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres$ y
y: command not found
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres$ git clone https://github.com/SergioRojas86/Proyecto_final_Despliegue_de_soluciones_analiticas.git
Cloning into 'Proyecto_final_Despliegue_de_soluciones_analiticas'...
remote: Enumerating objects: 30, done.
remote: Counting objects: 100% (30/30), done.
remote: Compressing objects: 100% (21/21), done.
remote: Total 30 (delta 5), reused 22 (delta 3), pack-reused 0
Receiving objects: 100% (30/30), 4.19 KiB | 4.19 MiB/s, done.
Resolving deltas: 100% (5/5), done.
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres$ ls
Proyecto_final_Despliegue_de_soluciones_analiticas
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres$ cd Proyecto_final_Despliegue_de_soluciones_analiticas
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas$ dvc remote list
fuentelocal     /tmp/dvcstore
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas$ dvc pull
Collecting                                                                                                    |2.00 [00:00,  180entry/s]
Fetching
Building workspace index                                                                                      |1.00 [00:00,  101entry/s]
Comparing indexes                                                                                             |4.00 [00:00,  985entry/s]
Applying changes                                                                                              |2.00 [00:00,  17.5file/s]
A       data/atus_anual_2021.csv
A       data/inafed_bd_1679023638.xlsx
2 files added and 2 files fetched
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas$ ls -la data
total 91900
drwxrwxr-x 2 ubuntu ubuntu     4096 Oct 26 04:03 .
drwxrwxr-x 5 ubuntu ubuntu     4096 Oct 26 04:02 ..
-rw-rw-r-- 1 ubuntu ubuntu       48 Oct 26 04:02 .gitignore
-rw-rw-r-- 1 ubuntu ubuntu 93893470 Oct 26 04:03 atus_anual_2021.csv
-rw-rw-r-- 1 ubuntu ubuntu      103 Oct 26 04:02 atus_anual_2021.csv.dvc
-rw-rw-r-- 1 ubuntu ubuntu   183145 Oct 26 04:03 inafed_bd_1679023638.xlsx
-rw-rw-r-- 1 ubuntu ubuntu      107 Oct 26 04:02 inafed_bd_1679023638.xlsx.dvc
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas$ mkdir maqueta
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas$ cd maqueta
(env-dvc) ubuntu@ip-172-31-82-22:~/proyecto_final_dsa/Andres/Proyecto_final_Despliegue_de_soluciones_analiticas/maqueta$
```

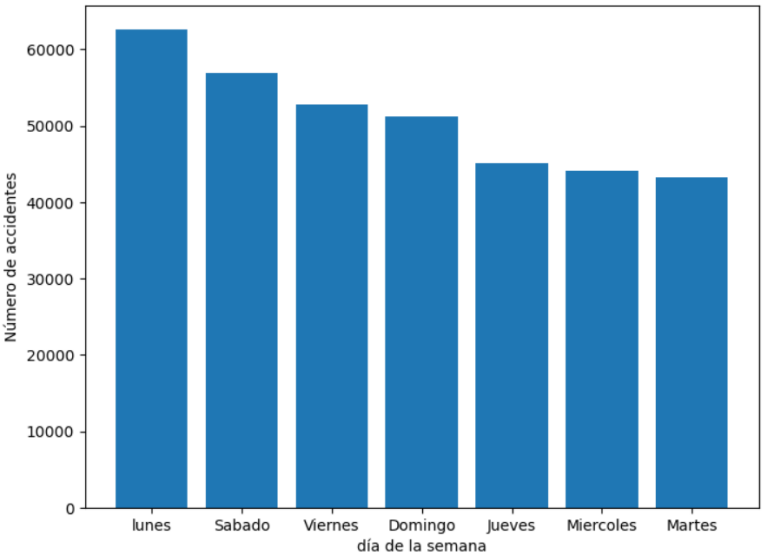
4. **Exploración de los datos.**

   A continuación, se realizara un breve resumen del análisis exploratorio de estos datos. Para ver el análisis exploratorio completo, por favor dirigirse al repositorio.

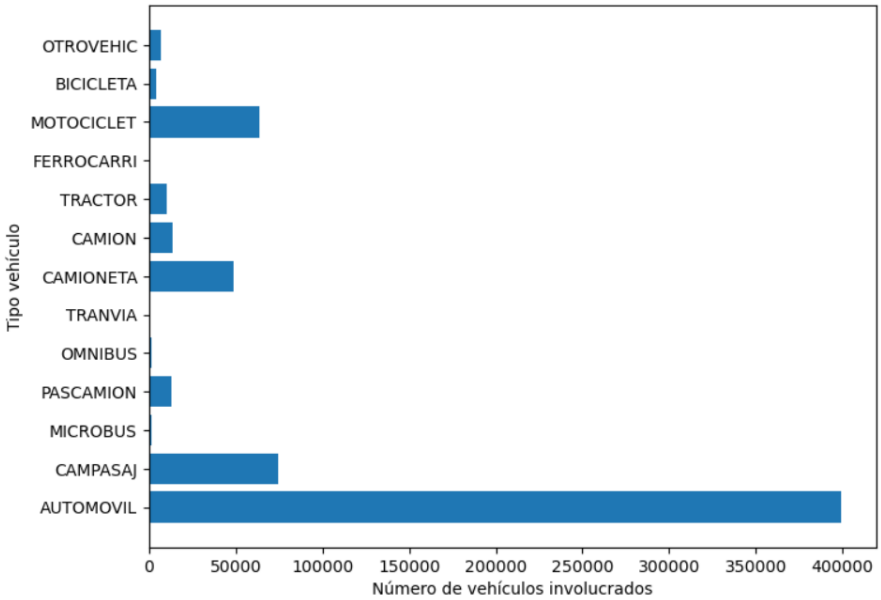
   Realizamos un análisis de distribución de la variable de respuesta.



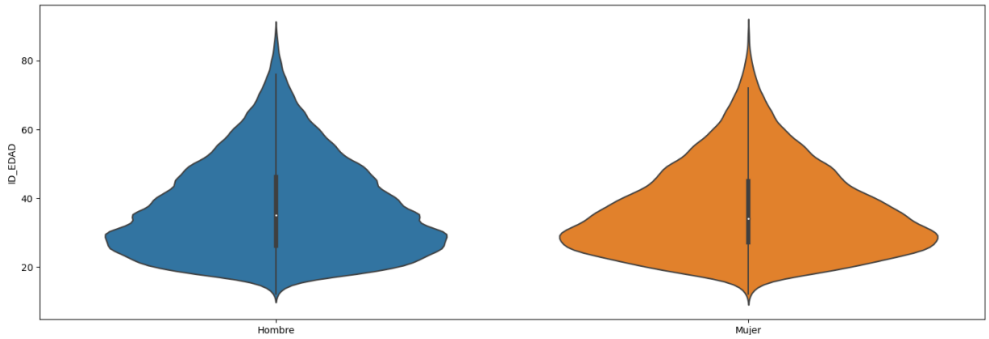
   La variable respuesta, representa el 1% del total de los datos. Ahora, miramos la distribución de las variables predictoras.



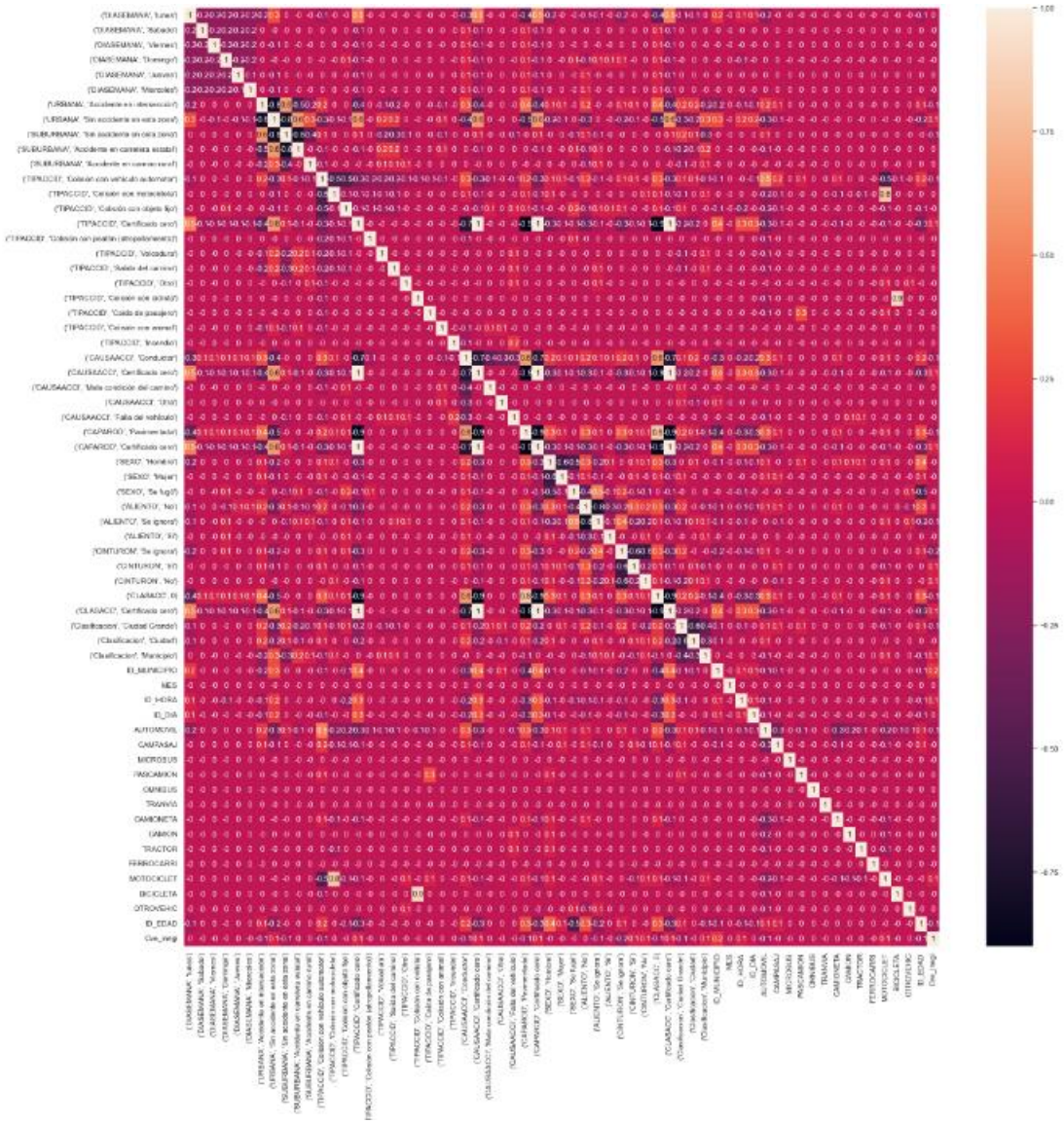
El gráfico de barras permite identificar que los lunes y los sábados son los días con mayor número de accidentes.



Asimismo, que el automóvil es el medio con mayores accidentes terrestres registrados.



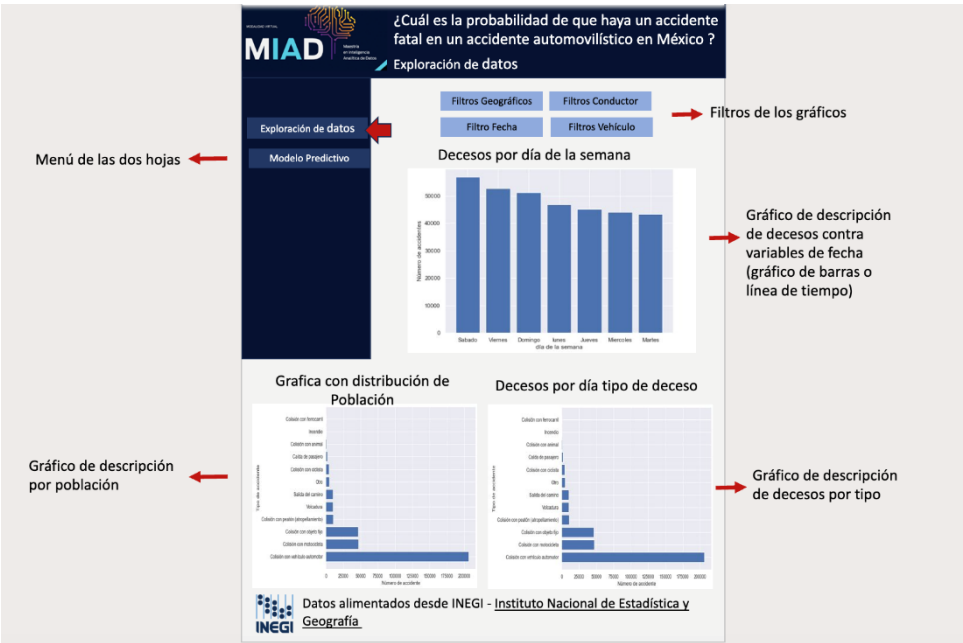
La mediana tanto para hombres como para mujeres se encuentra en los 35 años de edad, donde se observa una alta distribución de la edad de los conductores entre los 23 y 30 años.



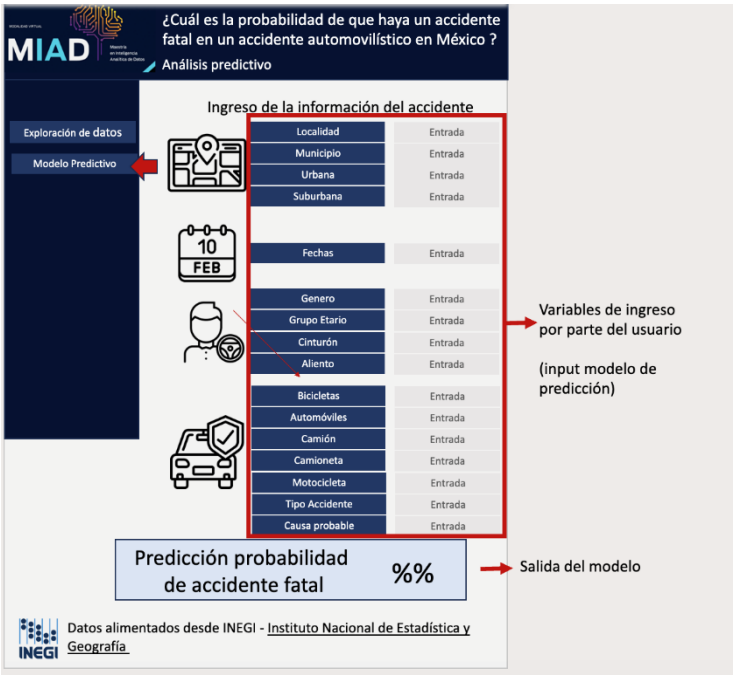
Por ahora, no se evidencian correlaciones significativas con la variable de respuesta.

5. **Maqueta del prototipo.**

La herramienta desplegada comprende dos secciones, desarrolladas por medio de Dash, la primera comprende un análisis de la exploración de datos, donde las gráficas representan el comportamiento de los datos según las tres clasificaciones de las variables.



La segunda hoja comprende el modelo predictivo, donde el usuario ingresa las variables de entrada del modelo y como resultado el modelo entregará la probabilidad de que el accidente sea fatal.



6. Reporte de trabajo en equipo

El entregable se dividió en partes iguales, donde cada miembro del equipo tomó dos puntos para agregar. Gloria trabajó la redacción del problema a abordar y del problema de negocio. Sergio, trabajó la parte que lanzar la instancia del grupo y realizar los repositorios de DVC. Cristian y Andrés, trabajaron la parte de exploración de los datos y de construir la maqueta. Al final, cada uno se unió a la instancia creada y clonó el repositorio subiendo cada uno su propia asignación.

Creando un commit cada uno de los integrantes dentro de la misma rama y actualizando la información en el repositorio de GitHub del proyecto.