



UNIVERSIDAD DE GUADALAJARA
CENTRO UNIVERSITARIO DE CIENCIAS EXACTAS E INGENIERÍAS

DIVISIÓN DE ELECTRÓNICA Y COMPUTACIÓN

DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

INGENIERÍA EN COMPUTACIÓN

PROYECTO MODULAR

THE CRASH PREDICTIT

PARA ACREDITAR LOS MÓDULOS

MÓDULO I. ARQUITECTURA Y PROGRAMACIÓN DE SISTEMAS

MODULO II. SISTEMAS INTELIGENTES

MODULO III. SISTEMAS DISTRIBUIDOS

P R E S E N T A

ISAAC GONZÁLEZ GUTIÉRREZ

OSWALDO LUNA GRADOS

ALAM ISAAC DURAN PLAZOLA

A S E S O R

ALEJANDRA SANTOYO SANCHEZ

Guadalajara, Jalisco, Noviembre de 2019

Contenido

I.	Introducción.	3
	Planteamiento del problema.	3
	Antecedente estado del arte.	3
	Alcance.	4
	Hipótesis.	4
	Justificación.	4
	Objetivos.	6
	A. Objetivo general.	6
	B. Objetivos específicos.	6
	Metodología.	7
II.	Diseño y Fase de Prototipo del proyecto Modular	9
	Cronograma de actividades	9
	Matrices de confusión de los modelos de Machine Learning	10
	Variables con mayor relevancia para el modelo de Random Forest	12
	Actividades y características de datos.	12
	Diagrama de despliegue	14
	Modelo vista controlador.	15
	Diccionario de datos.	15
III.	Manual de usuario.	22
	Inicio	22
	Análisis	24
	Formulario	28
IV.	Manual de administrador.	30
	Iniciar sesión	30
	Base de datos	31
V.	Conclusiones.	31
VI.	Referencia bibliográfica.	32

I. Introducción

Los choques son accidentes viales que desean evitarse porque causan muchos inconvenientes, como lo son personas heridas, muertes, daños materiales, tráfico, entre otros. Estos accidentes viales pueden ser registrados en una base de datos que se encargue de analizar horas frecuentes de accidentes viales, los accidentes, parámetros de edades propensas, las causas de los choques, etc. La base de datos que se propone necesita de cierto proceso para construir un modelo de Machine Learning: la primera fase se enfoca en una limpieza de datos, la segunda fase es la aplicación de los modelos predictivos de clasificación de accidentes Fatales y No fatales, la tercera fase es la propuesta de nuevas variables que mejoren el análisis y la cuarta fase es la recabación de datos con la cual se espera analizar los datos que se vayan capturando para poder prevenir dichos incidentes con la información obtenida. La base de datos se sustenta en modelos predictivos que ayudarán en el soporte de decisiones. Para la alimentación del software se implementará un apartado para cuando un usuario tenga un accidente: el mismo usuario o algún otro podrá introducir los datos del choque, por ejemplo, datos como las calles donde chocó, el tipo de vehículo, la iluminación, clima, el motivo, además de otros datos que ingresarán de forma automática, como la hora, un identificador, etc.

Planteamiento del problema

Mientras la ciudad crece, aumentan los problemas de tránsito, que usualmente son los accidentes viales que generan tráfico y otros conflictos. Y aunque los choques ocurren por diversas razones, con el uso de la tecnología y su implementación en los medios de comunicación se facilita que los accidentes viales sean publicados por diferentes medios, como la radio, la televisión, internet y redes sociales; lo que ayuda a evitar la zona donde ocurrió el accidente y las rutas afectadas, además de dar a conocer datos relevantes como daños colaterales, etc. Sin embargo, dichos medios no tienen una organización precisa ni específica para la modalidad de solo accidentes viales, son elementos separados que no trabajan en conjunto, por ello se ofrece una alternativa de solución para la prevención y estadística de los accidentes viales.

Antecedente estado del Arte

Actualmente hay mucha información acerca de choques automovilísticos en internet, pero solo son notas o esa información no está estructurada y algunas veces no se toman datos relevantes. INEGI tiene una gran cantidad de datos útiles acerca de todos los accidentes

automovilísticos, pero en México no existe un software (no se encontró ningún software en la investigación) que permita capturar la información de dicho choque de una manera eficaz y precisa. Por esto se implementará *the crash predictit*, para que cualquier usuario pueda registrar, con información relevante, accidentes viales desde algún dispositivo.

En una recopilación bibliográfica de algunos estudios realizados acerca de información de accidentes de tráfico se observa el uso de métodos de estimación, clasificación y predicción de accidentes. Por ejemplo, se usan modelos para predecir la gravedad de las lesiones según Chon y Paprzycki (2003), donde explican que su exactitud depende del tipo de modelo usado: Híbridos, Máquinas de Soporte Vectorial, Árboles de decisión. Chong (2004) realizó una clasificación de accidentes según la severidad de estos (no lesión, posible lesión, lesiones no incapacitantes, lesiones incapacitantes y lesiones fatales) y se compararon los modelos de árbol de decisión y red neuronal usando la función de base radial para el kernel, se concluyó que en todas las pruebas el árbol de decisión fue más preciso. En cambio, Kunt (2011), obtuvo buenos resultados para la predicción con las redes neuronales artificiales donde se desarrollaron modelos con mayor información sobre parámetros como la edad, género del conductor, el uso de cinturón de seguridad, el tipo y la seguridad del vehículo, las condiciones meteorológicas, la superficie de la carretera, el tipo de accidente, tipo de colisión y el flujo de tráfico.

Alcance

El alcance de este proyecto pretende aportar una herramienta donde se analice y se conozca la principal razón para un accidente fatal en relación a las otras variables recibidas, es decir, se busca una correlación con los accidentes fatales y los factores que pueden causarlos.

Hipótesis

Se plantea la posibilidad de que los accidentes viales son causados no por un solo factor, como lo es el conductor, sino que depende de otros como el medio, la hora, la edad, el género, tipo de carro, medidas de seguridad, etc. Se propone el uso de un sistema que reciba, procese y almacene datos de accidentes viales. Con la base de datos que se obtenga, se podrán realizar estadísticas, análisis de predicción y modelos de Machine Learning que podrán prevenir los accidentes viales alertando a los usuarios sobre los factores potenciales.

Justificación

Debido a la frecuencia constante con que ocurren choques en ciertas horas y en ciertos lugares, unos más que en otros, se planea localizar parámetros como hora-lugar, entre otros, y

guardar la información para establecer las causas que provocan dichos accidentes. Con dichos datos y parámetros se pretende crear un nuevo conjunto de datos sobre accidentes viales que mejore a los ya existentes como lo es el que proporcionó INEGI (2011-2017) y el cual se utiliza en este proyecto con algoritmos de Machine Learning.

Arquitectura y Programación de sistemas

Para el módulo se implementará el proceso de software *RUP (Proceso Unificado Racional)*. Se busca implementar la arquitectura en aplicación web con bases de datos MySQL. La implementación se basará en *Sistemas de Información de Gestión o MIS (Management Information System)*.

Se pretende demostrar los datos con relación a accidentes fatales por medio de gráficas a las que el usuario podrá ingresar, sin iniciar sesión, y podrá registrar un evento de choque para una nueva base de datos. El software dará información sobre los resultados de los modelos de Machine Learning.

Sistemas Inteligentes

En el apartado de sistemas inteligentes se realizará un modelo *de Machine Learning* utilizando repositorios recuperados del *INEGI* como *dataset* de entrenamiento y prueba. El modelo se creará a base de algoritmos de regresión tales como *SVM*, *Redes neuronales* y *Random forest*. Para la implementación y visualización del modelo, así como la manipulación de los datos, se van a utilizar las librerías de *Python: Sckit-Learn, Matplotlib, Numpy y Pandas*, entre otras librerías.

Sistemas distribuidos

Para el módulo referente a las bases de datos, se van a usar bases de datos relacionadas (*SQL*) y para gestionarla utilizaremos *MySQL* como también su sintaxis. La configuración que se implementará será la *master/slave* la cual permite tener bases de datos en distintos servidores con sus respectivos respaldos de seguridad. Para el mecanismo de seguridad se utilizarán dos métodos que trabajarán en conjunto, uno de ellos será el control de accesos, el cual nos permite controlar los privilegios de cada usuario, y el control de flujo con sus diferentes comandos, el cual nos ayudará a validar los respectivos usuarios y a tener un mejor control sobre el código.

Objetivos

Objetivo general

The crash predictit es un proyecto que aplicará modelos de Machine Learning (Logical Regresion, Support Vector Machines, Multilayer Perceptron, Random Forest) para analizar una gran cantidad de datos de accidentes viales que nos proporciona INEGI (2011 a 2017) para valorar diferentes variables, por ejemplo, a qué hora son más frecuentes los accidentes, que edades son las más propensas a chocar, las causas de los choques, entre otras para ofrecer una respuesta a qué factor o factores causan los accidentes. Con esta información se construyen modelos encargados de realizar predicciones de tipos de accidente viales enfocados en No fatal (no hubo muertos) y Fatal (Hubo al menos un muerto en el accidente). Además, con todos estos datos se espera ayudar a prevenir accidentes con la información obtenida por medio de modelos que ayudarán en el soporte de decisiones. Para poder alimentar al software de entorno web se implementará un apartado para cuando un usuario tenga un accidente, el mismo usuario o algún otro pueda introducir los datos del choque: las calles donde chocó, el estado de la calle, la iluminación, clima, el motivo, etc. Hay otros datos que se ingresan en automático, como la hora, un identificador, etc.

Objetivos específicos

Módulo para mostrar información:

- Identificar cuáles datos son más relevantes que otros
- Interpretar los datos de una manera simple y sencilla
- Brindar un apoyo preventivo con la interpretación de los datos
- Brindar un sistema de apoyo de decisiones

Módulo de captura de datos:

- Registrar accidentes viales de una manera sencilla
- Recolectar datos relevantes para el análisis a partir de diversas fuentes

Implementación de modelos de Machine Learning enfocados a Fatal y No fatal:

- Implementación Logistic Regresion.
- Implementación Support Vector Machines (Classifier).
- Implementación Multilayer Perceptron (Classifier).
- Implementación Random Forest (Classifier).

Metodología

Para el análisis y el modelo se tomó información del INEGI, con la cual se extraerán los datos durante el 2011 al 2017.

Para el entrenamiento y prueba de nuestros modelos, se utilizarán dos datasets los cuales ya fueron limpiados y procesados. El dataset de entrenamiento contiene datos de accidentes viales desde 2011 hasta 2017 (repositorio balanceados con misma cantidad de registros para cada clase a predecir). El total de registros son de 13,529, el total de columnas son de 70. Mientras que el dataset de prueba contiene datos de accidentes viales desde 2011 hasta 2017. El total de registros son de 829,040 y el total de columnas son 70.

Se define la variable objetivo "CLASACC" donde se declara si es Fatal o No fatal, así como se eliminan las variables que no aportan información relevante al modelo. Posteriormente se asignan los registros a sus respectivas variables que utilizará el modelo.

Para la creación de los algoritmos de Machine Learning, se utilizaron las clases brindadas por la librería Sklearn. Los modelos se ajustaron con sus respectivos hiperparámetros obtenidos mediante la clase GridSearchCV, incluida en la librería de sklearn.

Se utiliza Random Forest un bosque aleatorio, que estima los ajustes a varios clasificadores de árbol de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste. El tamaño de la submuestra siempre es el mismo que el tamaño de la muestra de entrada original, logrando extraer la muestra con reemplazo o sin reemplazo. Los parámetros son: número de árboles generados por el bosque aleatorio, la función para medir la calidad de la división, profundidad máxima de los árboles, el número mínimo de muestras requeridas para dividir un nodo interno, el número mínimo de muestras requeridas para ser consideradas una hoja, número máximo de características a utilizar en un árbol.

Se implementa Support Vector Machines, el SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a dos espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Los parámetros son: función kernel utilizada en el algoritmo, parámetro de penalización para el término del error, coeficiente de los kernel 'rbf', 'poly' y 'sigmoid'.

Se utiliza Multilayer Perceptron, el perceptrón, una red neuronal artificial formada por múltiples capas con capacidad para resolver problemas que no son linealmente separables. Los parámetros son: función de activación para las capas ocultas, número máximo de iteraciones durante la optimización de los pesos, tamaño de mini lotes para optimizadores estocásticos, parámetro de penalización L2 (término de regularización), tasa de aprendizaje programada para la actualización de los pesos, tasa de aprendizaje inicial, solver para la optimización de los pesos.

Logistical Regresion, este análisis es utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Utiliza como función de enlace la función logística ($1 \frac{1}{1+e^{-z}}$). Los parámetros son: Inverso de la fuerza de regularización (función de coste), número máximo de iteraciones tomadas para que el optimizador converja, solver para la optimización de los pesos.

Para las métricas de evaluación se implementa Accuracy Classification score que busca la exactitud media de las predicciones con respecto al valor verdadero: $\frac{vp+vn}{vp+fp+vn+fn}$, donde vp es el número de verdaderos positivos y fp el número de falsos positivos.

Para hacer el reporte de clasificación se compuso con las principales métricas de clasificación que son:

Precisión, es la relación de $\frac{vp}{vp+fp}$, donde vp es el número de verdaderos positivos y fp el número de falsos positivos. La precisión es intuitivamente la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa. El mejor valor es 1 y el peor es 0.

Sensibilidad: es la relación $\frac{vp}{vp+fn}$, donde vp es el número de verdaderos positivos y fn el número de falsos negativos. La sensibilidad es intuitivamente la capacidad del clasificador para encontrar todas las muestras positivas. El mejor valor es 1 y el peor es 0.

F1_score: El puntaje F1 se puede interpretar como un promedio ponderado de la precisión y la sensibilidad, donde un puntaje F1 alcanza su mejor valor en 1 y el peor puntaje en 0. La contribución relativa de precisión y recuerdo al puntaje F1 es igual. La fórmula para el puntaje F1 es: $F1 = \frac{2*(precisión*sensibilidad)}{(precisión+sensibilidad)}$.

Para seguir alimentando a la base se pretende lograr la recolección de datos de accidentes viales, se quiere incentivar al usuario con un software sencillo y rápido para que registre toda la información necesaria acerca de cualquier accidente que vea, por lo que, si más de una persona

registra el mismo choque, la variable “hora” y “cruce de calles” nos ayudará a limpiar la base para el posterior análisis. Se utilizará un servidor web, en este caso *apache*, para que el usuario se pueda conectar desde cualquier lugar solo necesitando conexión a internet, se implementará un diseño responsivo para que se adapte a cualquier dispositivo con la librería de *Bootstrap*, por último, habrá un administrador que tendrá un acceso único y servirá de mediador en el software.

II. Diseño y Fase de prototipo del Proyecto Modular

Cronograma de actividades

Definición y modelado del sistema

Se definirá las herramientas, algoritmos y técnicas que se utilizarán durante el desarrollo del proyecto, los recursos necesarios y el alcance que tendrá el proyecto. Así como también, se precisarán las funcionalidades del sistema mediante el diseño de diagramas UML.

Agosto 19 - Septiembre 6

(Responsable: Todo el equipo).

Modelado de la estructura de la base de datos

Se identificarán los datos necesarios para su resguardo, se definirán y desarrollarán el modelo de las bases de datos, el diagrama de base de datos, diccionario de datos y todos los requisitos necesarios para un buen diseño de la base de datos y la correcta función de sistema.

Septiembre 9 – Octubre 14

(Responsable: Todo el equipo).

Modelado y diseño de la interfaz web

Se definirá y se diseñará los elementos gráficos que la aplicación mostrará al usuario, la funcionalidad de los componentes en dicha interfaz, el flujo de trabajo en la aplicación y se realizará la conexión a la base de datos.

Septiembre 9 – Octubre 14

(Responsable: Todo el equipo).

Desarrollo e implementación del sistema inteligente

Se desarrollará un sistema de soporte a la toma de decisiones para el procesamiento de los datos recolectados de incidentes viales de México, el cual nos permitirá mostrar los resultados en gráficas.

Septiembre 9 – Octubre 18

(Responsable: Todo el equipo).

Elaboración de pruebas

Se realizarán las pruebas necesarias para garantizar la seguridad e integridad del sistema de las primeras versiones de la aplicación. Se les dará seguimiento a los problemas encontrados y seguirá el plan de mantenimiento correspondiente.

Octubre 21 – Noviembre 22

(Responsable: Todo el equipo)

Matrices de confusión de los modelos de Machine Learning

Para demostrar cada algoritmo se implementó una matriz de confusión en la cual cada columna representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

Como se observa en la siguiente imagen 1 la parte superior izquierda representa los no fatales mientras que la parte inferior derecha los fatales donde las esquinas no nombradas son los falsos positivos.

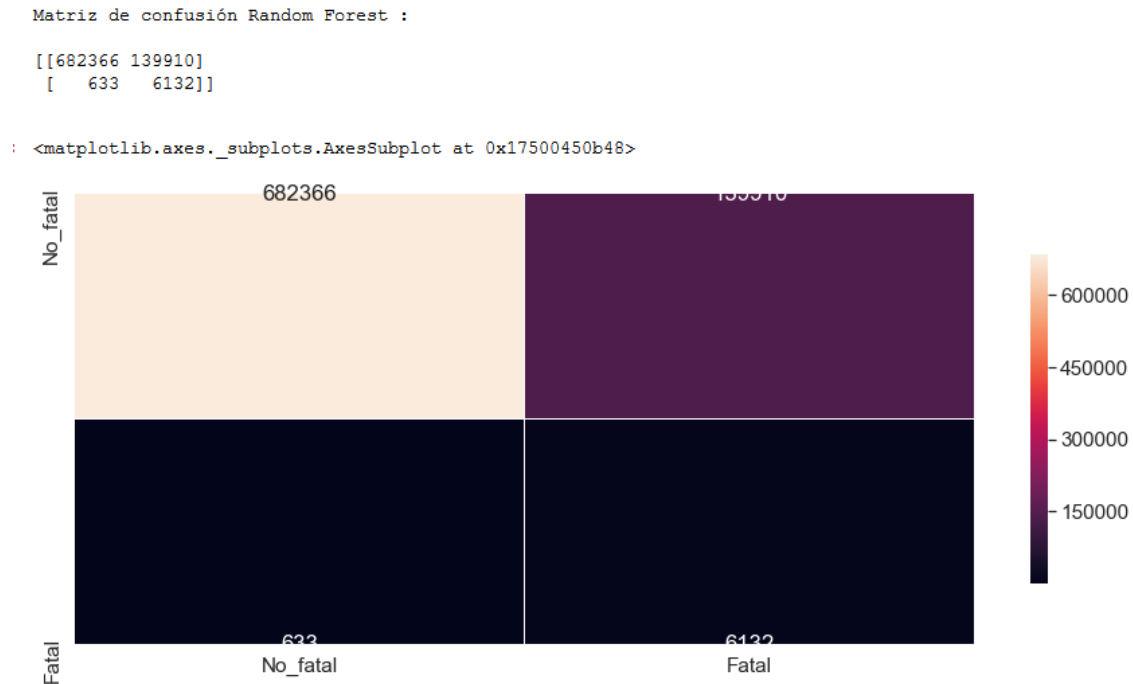


Imagen 1. Matriz de confusión de Random Forest.

La imagen 2 es la matriz de confusión de las máquinas de vectores de soportes donde en la parte superior izquierda es la parte no fatal y la parte fatal es la inferior derecha.

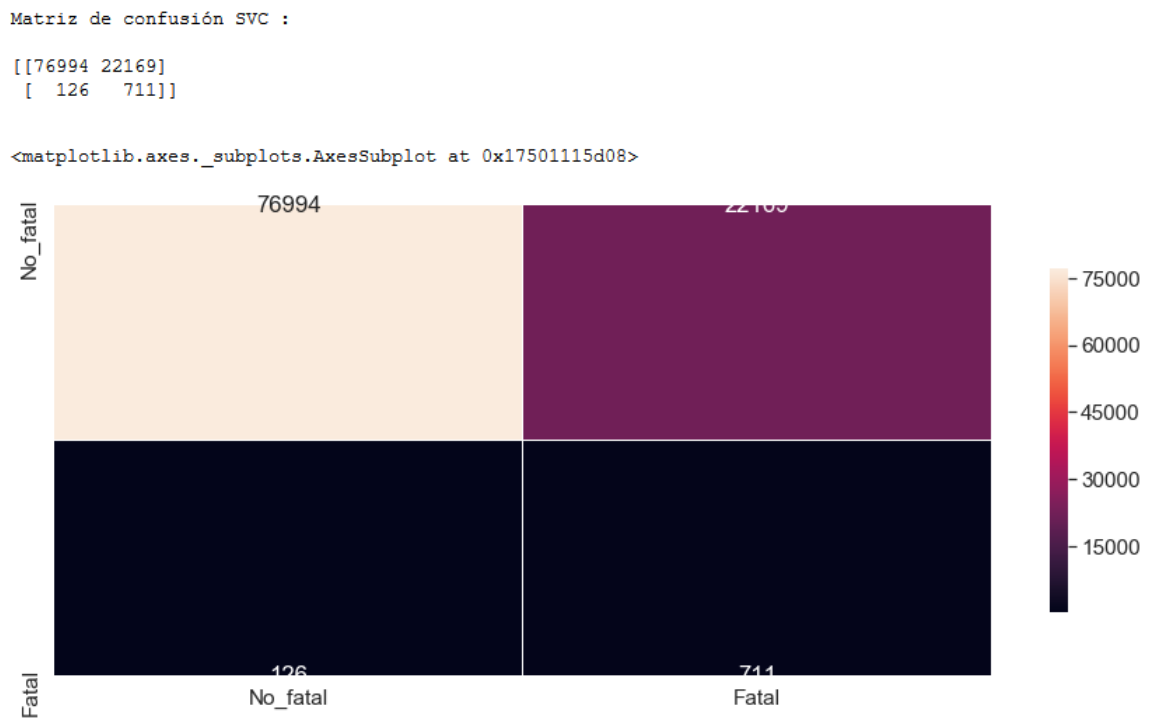


Imagen 2. Matriz de confusión de SVC.

La matriz de confusión de la red neuronal es como se observa en la imagen 3.

Matriz de confusión MLP :

```
[[79234 19929]
 [  123   714]]
```

<matplotlib.axes._subplots.AxesSubplot at 0x175003bea88>

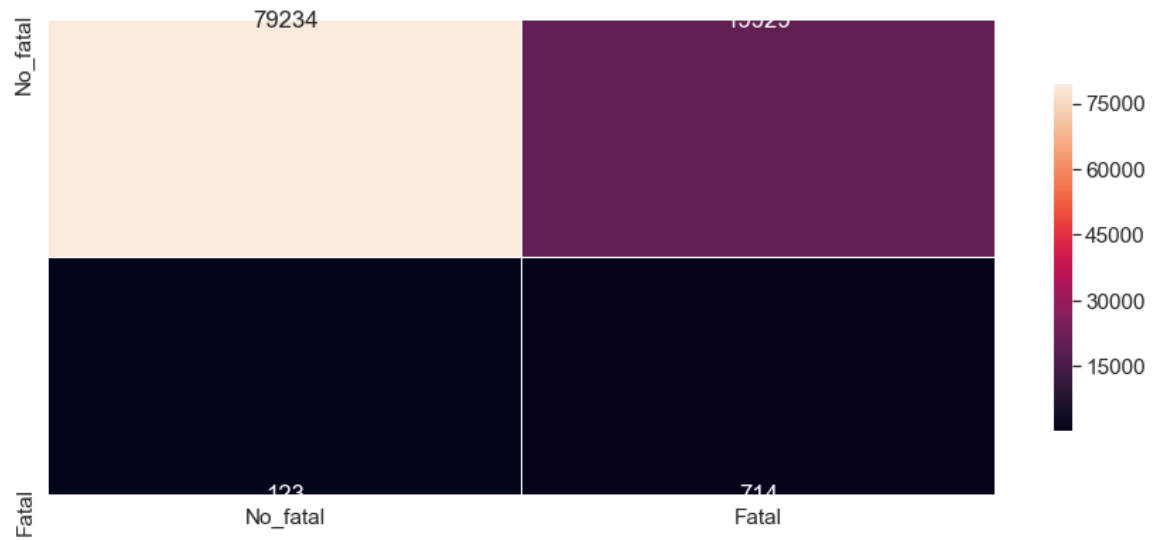


Imagen 3. Matriz de confusión de MLP.

Dentro de la imagen 4 se ve la matriz de confusión de la regresión logística.

Matriz de confusión Logistic Regression :

```
[[645381 176895]
 [ 1581   5184]]
```

: <matplotlib.axes._subplots.AxesSubplot at 0x175013994c8>

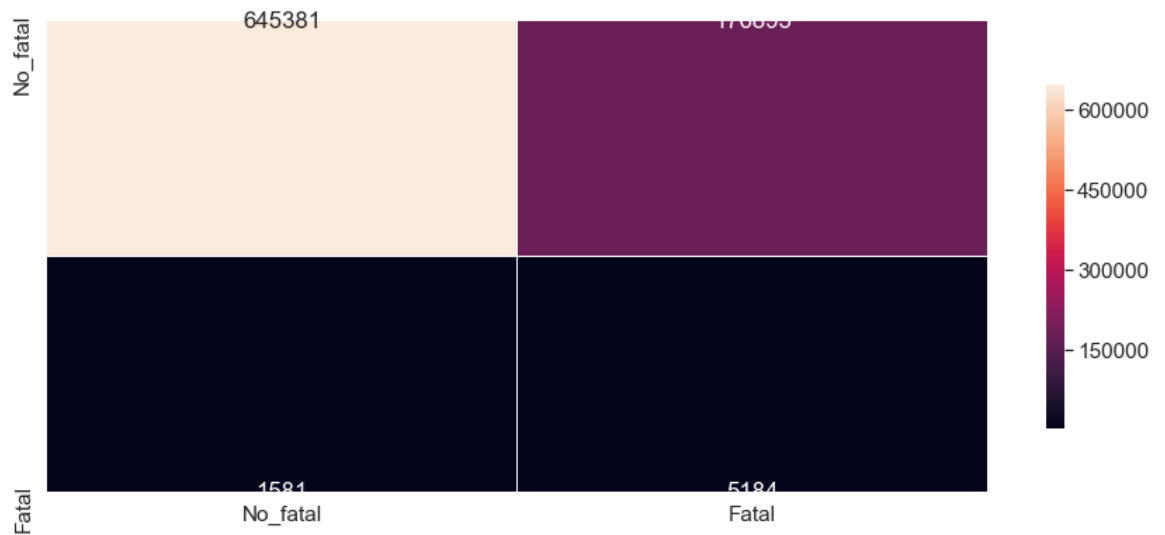


Imagen 4. Matriz de confusión de Logistic Regression.

Se concluyó que el mejor modelo fue el Random Forest para este caso. Ya que dio el mejor score.

Variables con mayor relevancia para el modelo de Random Forest

Variables con más relevancia para el modelo en porcentaje, ordenadas de forma descendiente.

NOMBRE	COEFICIENTE
ID_ENTIDAD	8.871210
ID_HORA	7.988062
AUTOMOVIL	4.313406
MOTOCICLETA	2.021481
ALIENTO_SI	1.837024
DIASEMANA_DOMINGO	1.163235
SEXO_HOMBRE	0.902867

Tabla 1. Variables relevantes en el modelo de Random Forest.

Actividades y características de datos

El sistema tiene como función mostrar análisis previos y el desarrollo de una aplicación para recibir datos y analizarlos.

Las funciones básicas son:

- Soporte de una plataforma web que realice los servicios requeridos.
- Permitir la interacción entre el usuario y el sistema.

Usuario

-Entrar sin iniciar sesión (en el caso que los datos no serán tan específicos) o iniciando sesión (si se necesita datos que solo un agente vial pueda obtener y registrar).

-Registrar accidente vial.

-Consulta de análisis de datos.

Administrador

-Revisión de datos registrados por usuarios.

-Procesar datos revisados para procesar.

Dentro de las Actividades que cada uno de los usuarios puede realizar se encuentran algunas limitantes para algunos.

Usuario

El usuario podrá registrar los accidentes viales ingresando las calles momento del día donde se observa el tipo de auto y si fue caravana o no. El usuario podrá consultar en una pestaña los análisis de los datos capturados anteriormente.

Administrador

El administrador podrá realizar todo lo que el usuario en tránsito puede realizar. Podrá modificar la base de datos y gestionar la misma. Revisará si los datos registrados de los usuarios tienen coherencia. Revisar, modificar o eliminar en la base de datos donde se ingresan los nuevos valores. De estos datos ya revisados, se envían a procesar junto a los datos ya procesados.

Diagrama de despliegue

Para uso del software es necesario que el usuario cuente con conexión de internet para poder navegar por la página web, se mostrará un diagrama de despliegue para demostrar los módulos que lo conforman en la imagen 5.

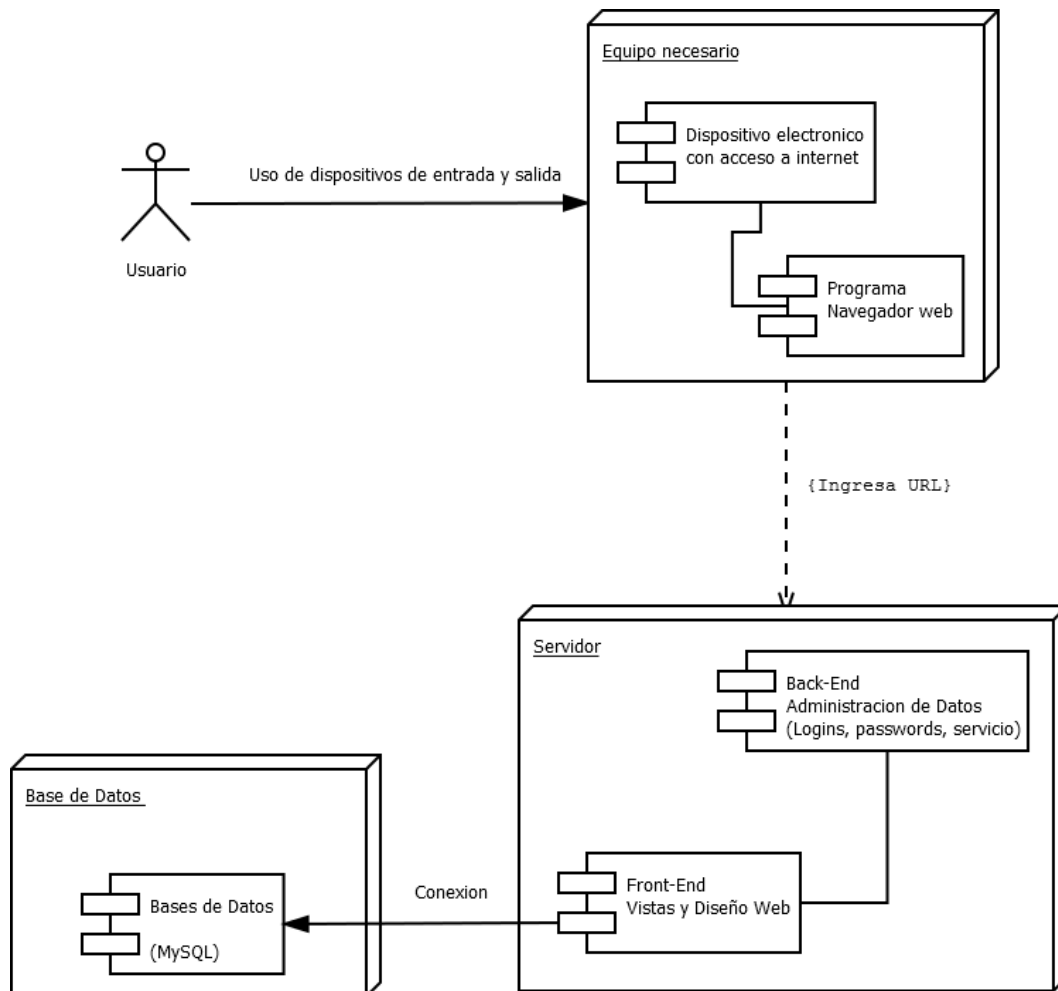


Imagen 5. Diagrama de despliegue.

Modelo vista controlador.

Teniendo esta estructura establecida pasaremos a destacar con el modelo vista controlador dicha estructura en la imagen 6.

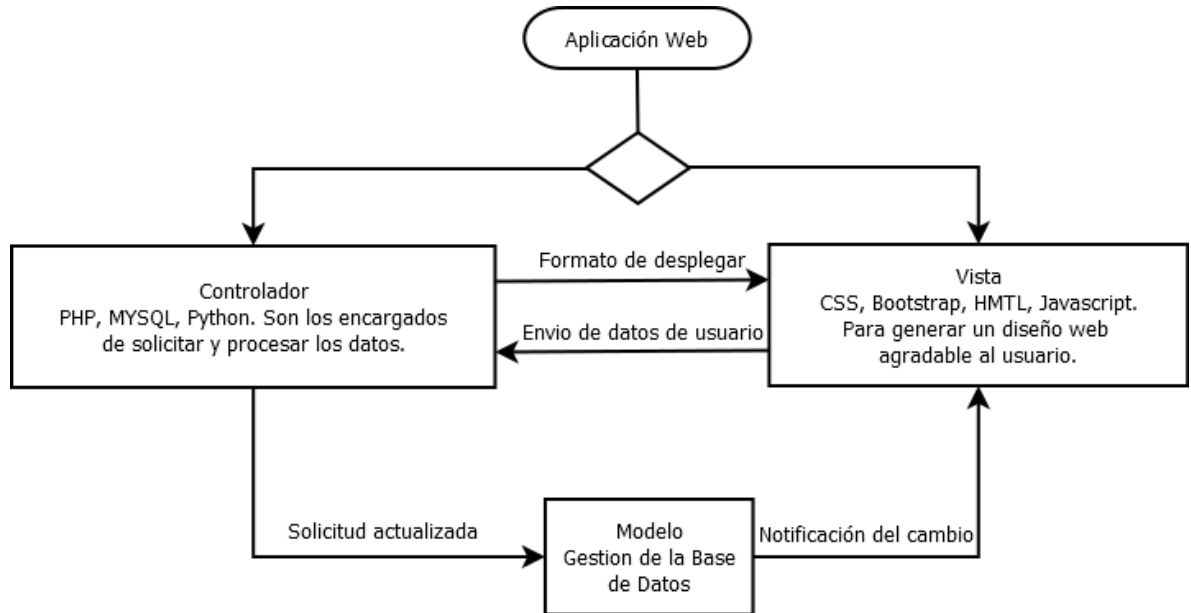


Imagen 6. Modelo vista controlador.

Existe una estructura en la base de datos para el funcionamiento de la página web, logrando el procesado y guardado de toda la información, tanto la almacenada como la que se espera recibir. En esta fase se dividirá en dos sectores: los datos no analizados por el administrador y los datos para procesar.

Diccionario de datos

El diccionario de datos se creó con base en variables ya analizadas por medio de estudios previos que señalaron los antecedentes, los cuales ofrecen información para saber cuáles variables están implicadas en los accidentes viales.

COLUMNA	TIPO_DATO	LONGITUD	CODIGO_VALIDO	DESCRIPCION
MES	Varchar	5	01,02,03,04,05,06, 07,08,09,10,11,12	Correspondiente al mes de referencia en que ocurrió el accidente
ID_HORA	Int		0-23	La hora (sin los minutos) en que ocurrió el accidente, con rango: 00-23 horas. Clave 99 Hora no especificada

ID_MINUTO	Int		0-59	Los minutos en que ocurrió el accidente, con rango: 00-59. Clave 99 Minutos no especificados
ID_DIA	Int		01-31	Número correspondiente al día del mes en que ocurrió el accidente, con rango: 01 a 28, 30 o 31 según corresponda al mes de referencia. Clave 32 Día no especificado.
DIASEMANA	Varchar	50	Lunes-Domingo	El día de la semana en que ocurrió el accidente
URBANA	varchar	150		Es el área habitada o urbanizada que, partiendo de un núcleo central, presenta continuidad física en todas direcciones hasta ser interrumpida, en forma notoria, por terrenos de uso no urbano como bosques, sembradíos o cuerpos de agua. Se caracteriza por presentar asentamientos humanos concentrados de más de 15,000 habitantes. En estas áreas, se asienta la administración pública, el comercio organizado y la industria. Cuenta con infraestructura, equipamiento y servicios urbanos, tales como drenaje, energía eléctrica, red de agua potable, escuelas, hospitales, áreas verdes y de diversión, etc.
SUBURBANA	Varchar	150		Son aquellas zonas donde la población es de 2,500 a 14,999 habitantes, las viviendas se encuentran dispersas y en algunas ocasiones carecen de algunos servicios.

TIPACCID	varchar	150		<p>Corresponda al tipo de accidente de tránsito, de acuerdo con las siguientes descripciones:</p> <p>1) Colisión con vehículo automotor Encuentro violento, accidental o imprevisto de dos o más vehículos en una vía de circulación, del cual resultan averías, daños, pérdida parcial o total de vehículos o propiedades, así como lesiones leves y/o fatales a personas. Puede ser lateral, frontal o por alcance. 2) Colisión con peatón Evento vial donde un vehículo de motor arrolla o golpea a una persona que transita o que se encuentra en alguna vía pública, provocando lesiones leves o fatales. 3) Colisión con animal Es aquel accidente en el que un vehículo de motor arrolla a cualquier tipo de animal provocando daños materiales, inclusive lesiones leves o fatales a personas ocupantes o no del vehículo. 4) Colisión con objeto fijo Encuentro violento de un vehículo de motor con cualquier tipo de objeto, que por sus características se encuentre sujeto al piso o asentado en él, tales como postes, guarniciones, señales de tránsito, árboles, contenedores de basura, etc. También se incluye en este tipo de colisión, el percance de un automotor en movimiento contra otro estacionado. 5) Volcadura Es el tipo de accidente que debido a las circunstancias que lo originan, provocan que el vehículo pierda su posición normal, incluso dé una o varias volteretas. 6) Caída de pasajero Accidente donde una o más personas que viajan en el vehículo, (excluyendo al conductor), caen fuera del mismo. No se considera este tipo de accidente si la caída fue por consecuencia de otro tipo de accidente. 7) Salida del camino Evento en donde el vehículo, por causas circunstanciales, abandona de manera violenta e</p>
----------	---------	-----	--	---

				<p>imprevista la vía de circulación por la cual transita. Incluso si por la acción del vehículo cae a una zanja, cuneta, barranca, etc. 8) Incendio Es el accidente ocasionado por un corto circuito, derrame de combustible o cuestiones desconocidas, que propician la generación de fuego mediante el cual se consume parcial o totalmente el vehículo automotor.</p> <p>Nota: No se clasifique este accidente en este tipo, si el incendio es resultado de una colisión con otro vehículo automotor en circulación, o si el fuego se produce después de una colisión, volcadura o salida del camino. 9) Colisión con ferrocarril Choque de un vehículo automotor con una locomotora, vagón, góndola o cualquier otro vehículo clasificado como transporte ferroviario. 10) Colisión con motocicleta Percance vial en donde un vehículo automotor de cualquier tipo, tiene un encuentro violento, accidental o imprevisto con una motocicleta. Incluso se puede dar el caso de que sea entre dos motocicletas. 11) Colisión con ciclista Hecho en el cual un vehículo automotor de cualquier tipo, arrolla a un ciclista sobre la vía de circulación o en un cruce vial. 12) Otro Cualquier otro tipo de accidente que no pueda ser clasificado en los 11 incisos descritos anteriormente, tales como derrumbes, deslaves o cualquier otro objeto que caiga sobre los vehículos en circulación y como consecuencia se produzca algún accidente vial.</p>
AUTOMOVIL	Int		0-9	Automóvil. Comprende los vehículos de motor destinados principalmente al transporte de personas, que cuentan hasta con 7 asientos (incluyendo el del conductor).
CAMPASAJ	Int		0-9	Camioneta de pasajeros. Comprende todos los vehículos de motor destinados primordialmente al transporte de

				personas y que tengan de 8 a 15 asientos (incluyendo el del conductor).
MICROBUS	Int		0-9	Microbús. Autobús de menor tamaño usado por lo general en el transporte urbano de pasajeros y que tenga de 16 a 20 asientos (incluyendo el del conductor).
PASCAMION	Int		0-9	Camión urbano de pasajeros. Comprende los autobuses urbanos y suburbanos y en general los vehículos que tengan de 21 a 29 asientos, destinados al transporte público y privado de personas, los cuales cuentan con rutas fijas.
OMNIBUS	Int		0-9	Ómnibus. Comprende los vehículos automotores con 30 o más asientos, destinados al transporte público y privado de personas, con destinos establecidos, así como horarios de llegada y salida.
TRANVIA	Int		0-9	Tren eléctrico o trolebús. Comprende los vehículos de motor destinados al transporte de personas, propulsados por energía eléctrica captada de cables aéreos, que no circulan sobre rieles (este tipo de transporte sólo se encuentra registrado en el Ciudad de México y Guadalajara).
CAMIONETA	Int		0-9	Camioneta de carga. Son aquellas que están destinadas exclusivamente al transporte de carga; se identifican de acuerdo al tamaño y a la capacidad de hasta 999 kilogramos.
CAMION	Int		0-9	Camión de carga. Comprende los vehículos de propulsión mecánica propia, destinados exclusiva o principalmente al transporte de carga, con capacidad de 1,000 hasta 5,000 kilogramos.
TRACTOR	Int		0-9	Tractor con o sin remolque. Comprende los vehículos de propulsión mecánica propia, diseñados exclusiva o

				principalmente para remolcar otros vehículos (excluye los tractores agrícolas, industriales y de construcción).
FERROCARRI	Int		0-9	Ferrocarril. Medio de transporte sobre rieles para el transporte de pasajeros y carga, que recorre distancias relativamente largas a velocidades medias.
MOTOCICLET	Int		0-9	Motocicleta. Vehículo automotor de dos, tres o cuatro ruedas, cuyo peso no excede los 400 kilogramos.
BICICLETA	Int		0-9	Bicicleta. Vehículo de dos o tres ruedas generalmente iguales, movidas por pedales y una cadena, el cual es propulsado por el esfuerzo humano.
OTROVEHIC	Int		0-9	Otro. Considérese cualquier otro tipo de vehículo no descrito en la clasificación anterior; por ejemplo las ambulancias, grúas, vehículos de tracción animal o humana, carro de bomberos, tractores agrícolas, industriales y de construcción, etc.
CAUSAACCI	Varchar	100		La causa presunta o determinante puede considerarse como: El motivo principal que causó el accidente, ya sea por condiciones inseguras o actos irresponsables potencialmente prevenibles, atribuidos a conductores de vehículos, así como a peatones o pasajeros, falla de vehículos, condiciones del camino, circunstancias climatológicas, etc.
CAPAROD	Varchar	100		Superficie de rodamiento en donde ocurrió el accidente de tránsito. Pavimentada Conjunto de capas de material rígido (concreto hidráulico) o flexible (carpeta asfáltica) compactado sobre el subsuelo, que permite el tránsito adecuado de vehículos y su carga. No pavimentada. Camino acondicionado con materiales naturales (piedra, bola, tezontle, etc.), para el tránsito de vehículos y/o personas.

SEXO	Varchar	50		Genero del conductor presunto responsable de ocasionar el accidente.
ALIENTO	Varchar	50		Sobriedad del conductor presunto responsable del accidente
CINTURON	Varchar	50		El Conductor presunto responsable del accidentes usaba el Cinturón de seguridad
ID_EDAD	Varchar	100	0-99	La edad del conductor presunto responsable, la cual debe estar anotada con un número arábigo de 12 a 98. La clave 0 se refiere a los registros en donde el conductor se fugó y los registros con clave 99 se refiere Se ignora la edad del conductor
CLASACC	Varchar	50		Los accidentes se clasifican en Fatales: Se refiere a todo accidente de tránsito en el cual una o más personas fallecen en el lugar del evento; No fatales: Se refiere a todo accidente de tránsito en el cual una o más personas resultan con lesiones con o sin consecuencia de muerte y Sólo daños: Se refiere a todo accidente en el que se ocasionaron daños materiales a vehículos automotores, propiedad del estado, inmueble particular y otros.
ESTATUS	Varchar	100		Estatus de las cifras conforme a los Lineamientos de Cambios a la información divulgada en las publicaciones estadísticas y geográficas del INEGI
CLIMA	Varchar	20		Muestra las condiciones del clima (lluvia, tormenta eléctrica, neblina, granizo, condiciones regulares, etc.)
ILUMINACION	Varchar	20		Muestra las condiciones de iluminación (oscuro: sin lámparas, oscuro: sin lámparas, luz de día)
LIMI_VELOCI	Int	11	0-9	
TIPO_CALLE1	Varchar	20		ejemplo: avenida, calle, callejón, etc.
CALLE1	Varchar	50		nombre de calle 1
TIPO_CALLE2	Varchar	20		ejemplo: avenida, calle, callejón, etc.

CALLE21	Varchar	50		nombre de calle 2
CONDI_CALLE	Varchar	20		Muestra las condiciones de la calle (Con baches, empedrado, etc.)

Tabla 2. Diccionario de datos.

III. Manual de usuario

La plataforma web está dividida en diferentes funciones donde se explicará de manera detallada el funcionamiento de la aplicación web vista por el usuario:

Inicio

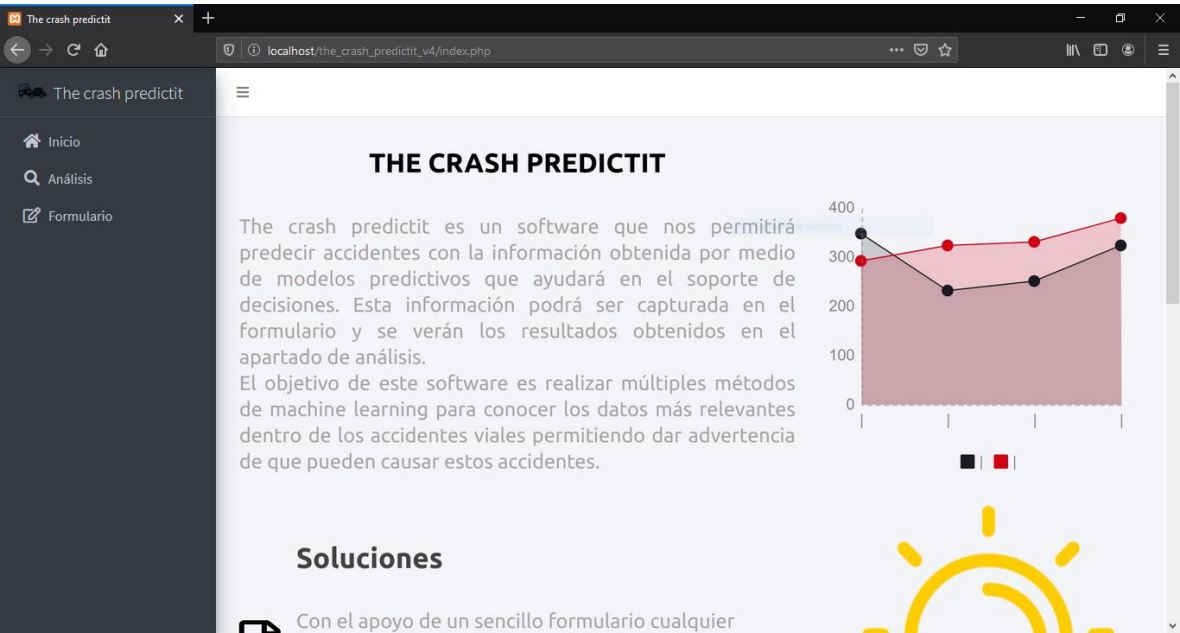


Imagen 7. Inicio de plataforma.

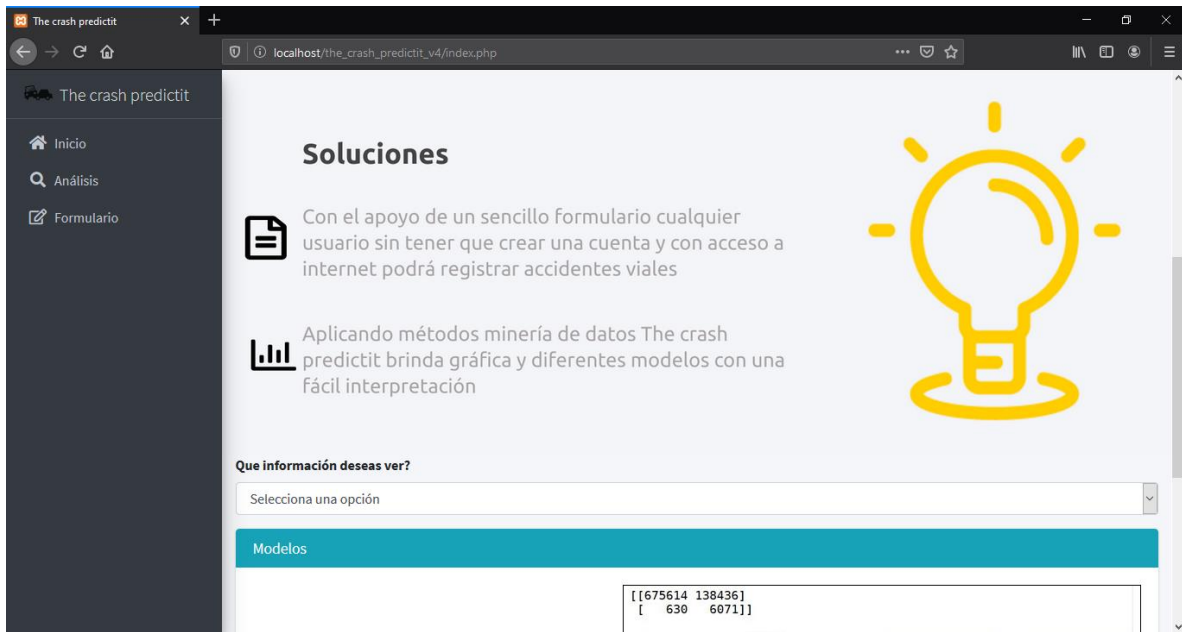


Imagen 8. Inicio de la plataforma parte dos.

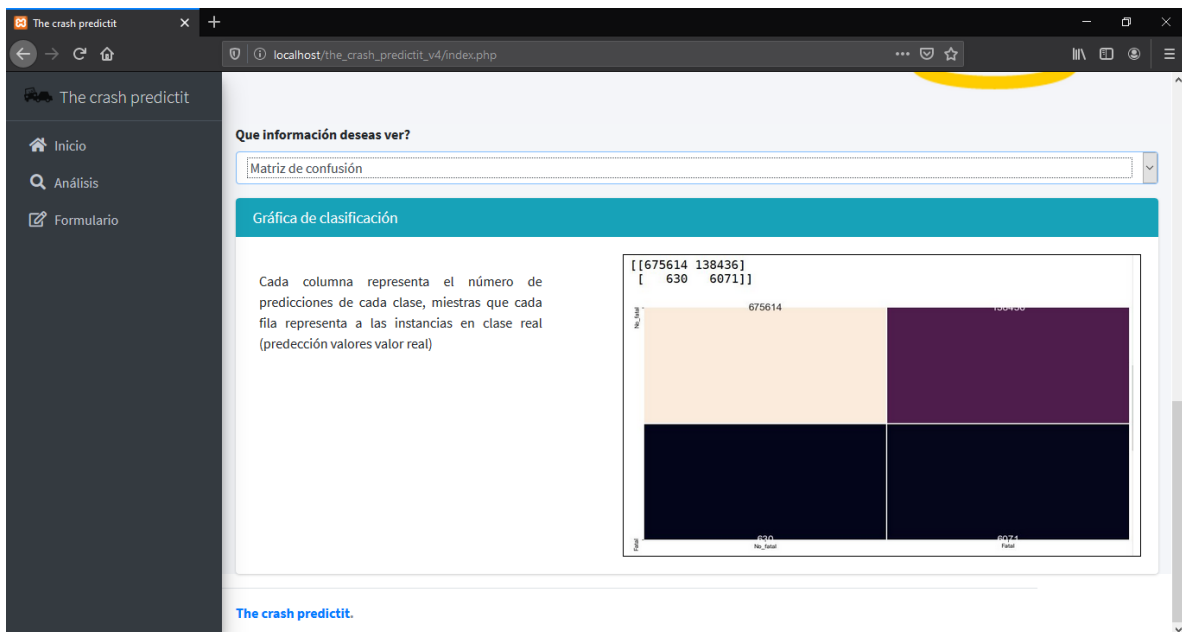


Imagen 9. Inicio de la plataforma parte tres.

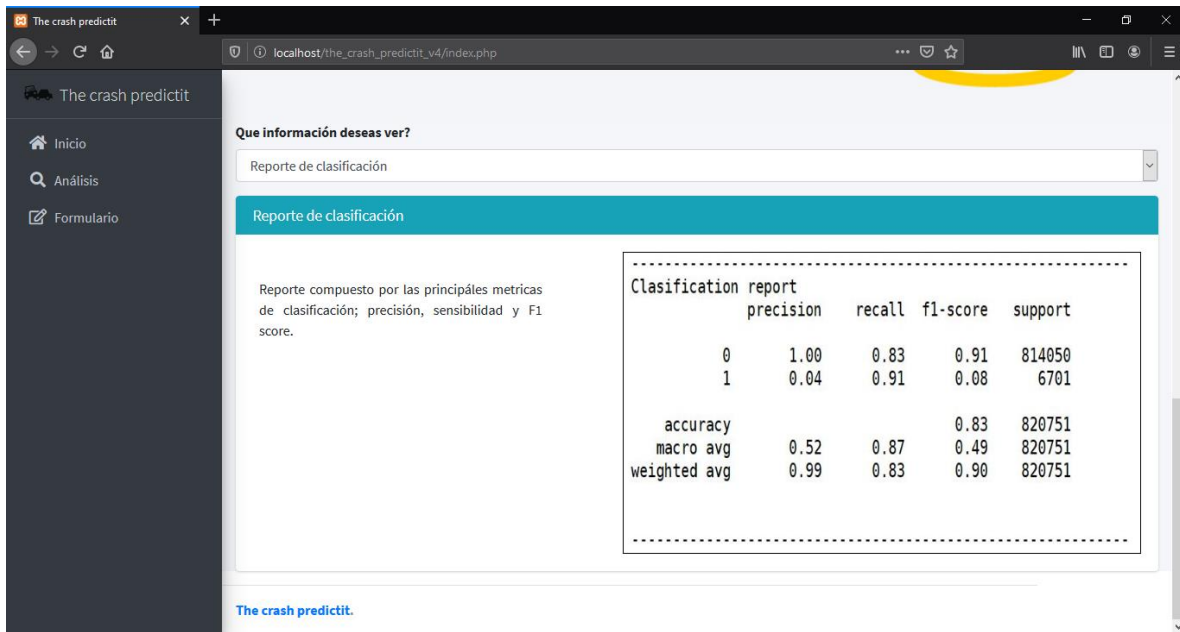


Imagen 10. Inicio de plataforma parte cuatro.

Como se observa en las anteriores imágenes se da una sencilla introducción del software y se muestra los modelos de confusión, y los reportes de clasificación de los datos ya analizados.

Análisis



Imagen 11. Pestaña de análisis frecuencia de accidentes fatales de acuerdo al día.



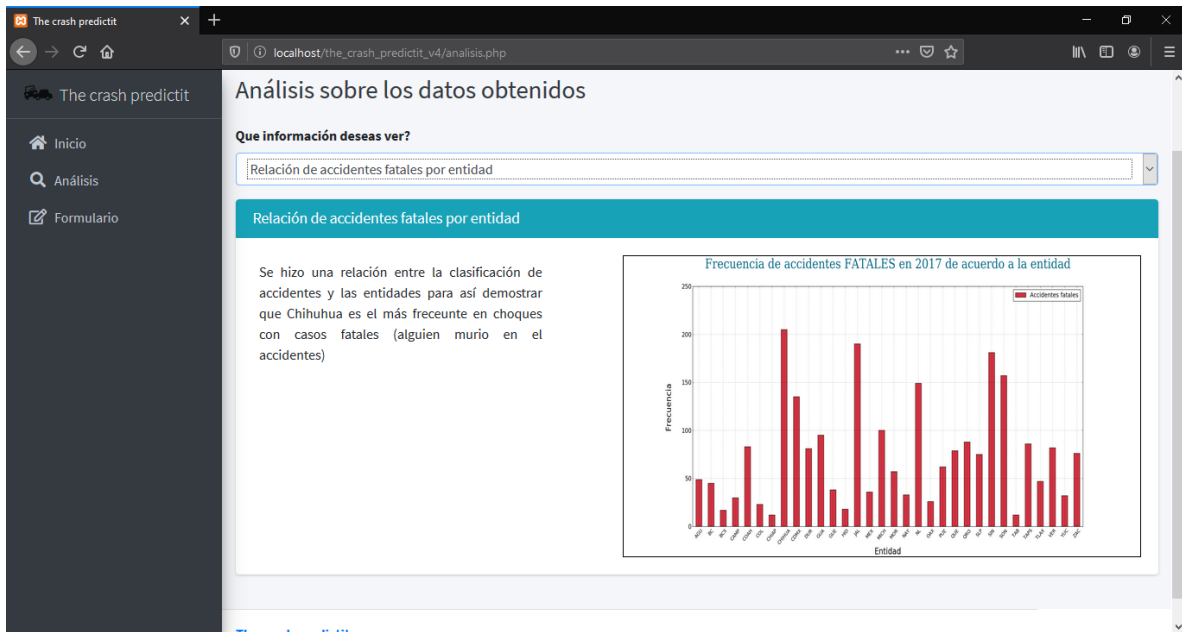


Imagen 14. Pestaña de análisis, frecuencia de accidentes fatales de acuerdo a la entidad.

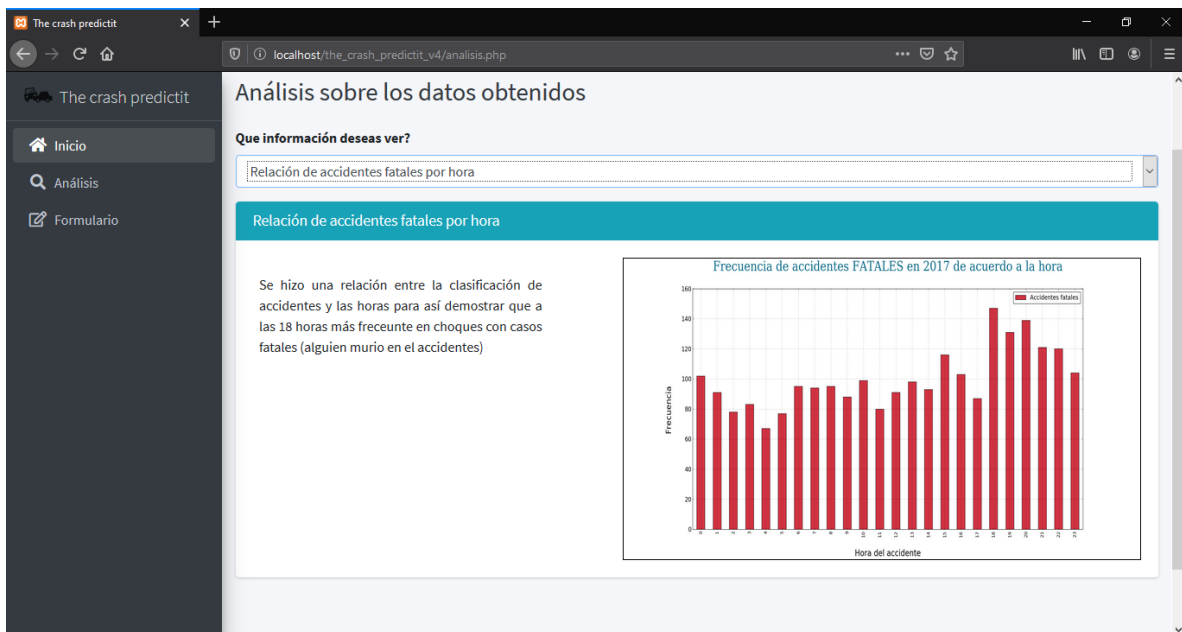


Imagen 15. Pestaña de análisis, frecuencia de accidentes fatales de acuerdo a la hora.

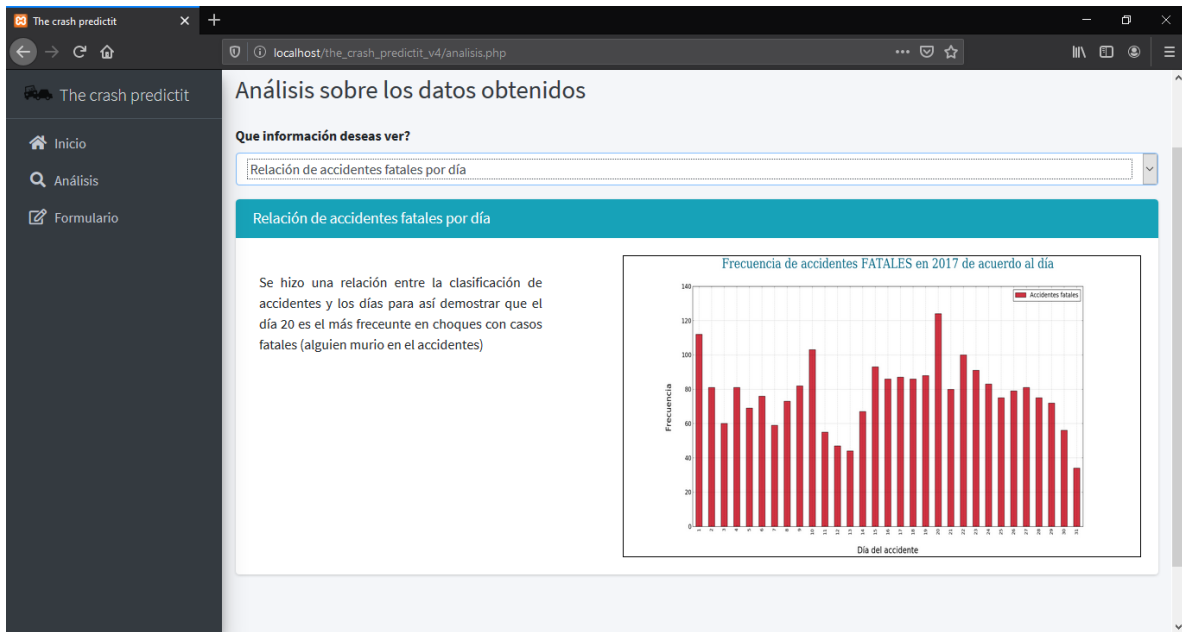


Imagen 16. Pestaña de análisis, frecuencia de accidentes fatales de acuerdo al día.

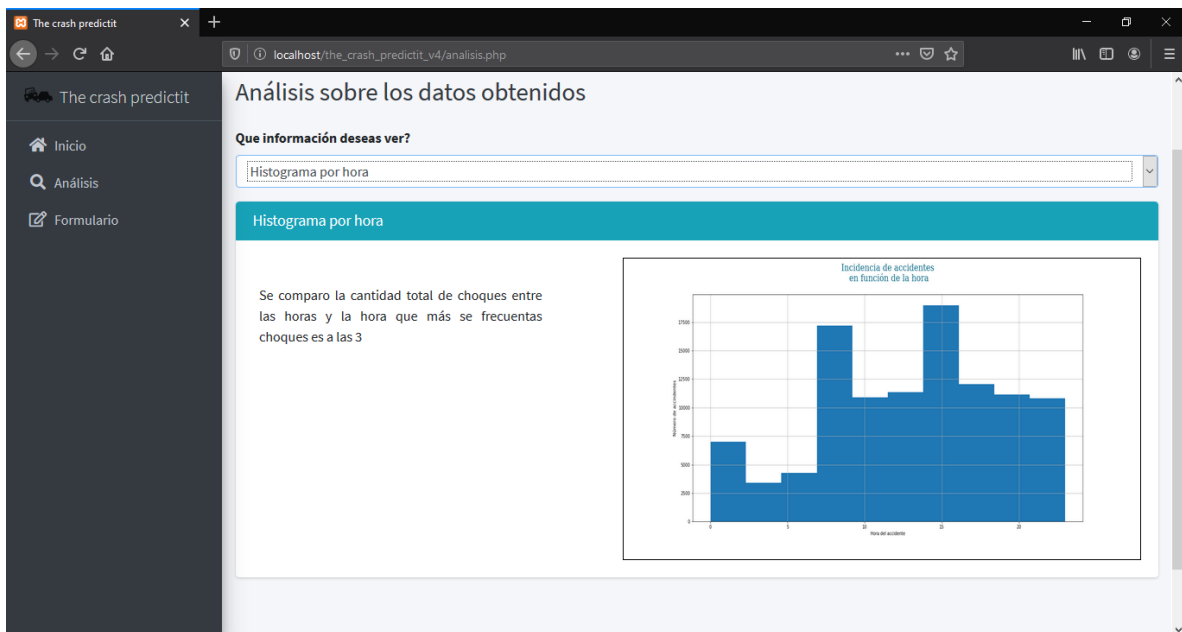


Imagen 17. Pestaña de análisis, histograma por hora.

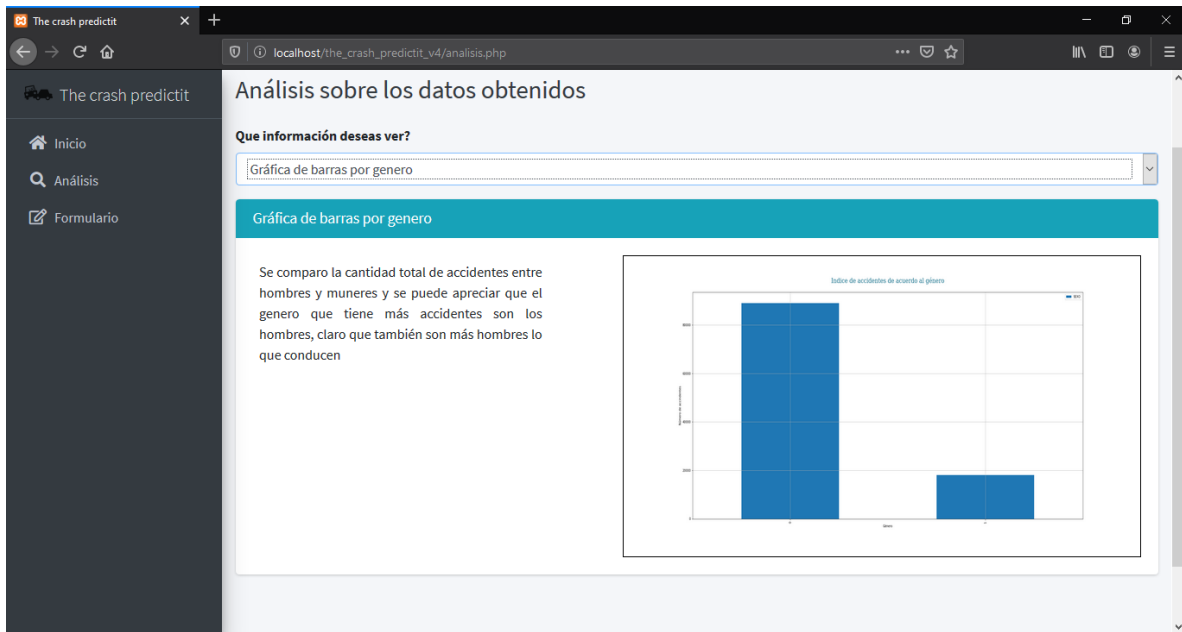


Imagen 18. Pestaña de análisis, gráficas de barras por género.

Como se observa en las anteriores imágenes son las gráficas mostradas de las variables analizadas y comparadas para ver la relevancia que tienen y si puede haber un impacto con estos a comparación de otras.

Formulario

The crash predictit

Inicio

Análisis

Formulario

Ingresa la información del accidente

Tipo de calle 1

Selecciona una opción

Nombre de calle 1

Tipo de calle 2

Selecciona una opción

Nombre de calle 2

Sexo

Selecciona una opción

Edad

Selecciona una opción

Imagen 19. Pestaña de formulario parte 1.

Imagen 20. Pestaña de formulario parte dos.

Imagen 21. Pestaña de formulario parte 3.

En las anteriores imágenes se observa el formulario y cada espacio para agregar, donde se autocompleta el nombre de la calle para que no exista incongruencia con el nombre.

IV. Manual de administrador

Iniciar sesión

En el manual de administrador solo se muestra la pestaña de datos dentro de la plataforma web. Para lograr acceder al modo administrador se necesita iniciar sesión. Como se ve en la imagen 22.

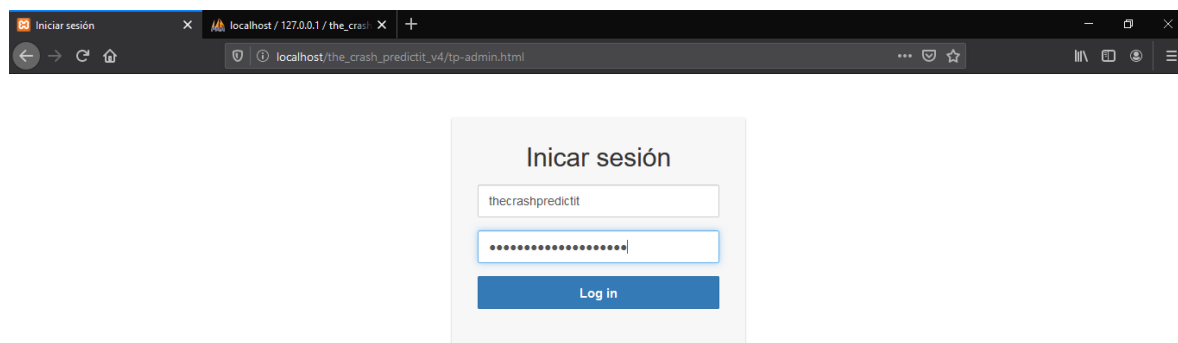


Imagen 22. Pestaña de inicio de sesión del administrador.

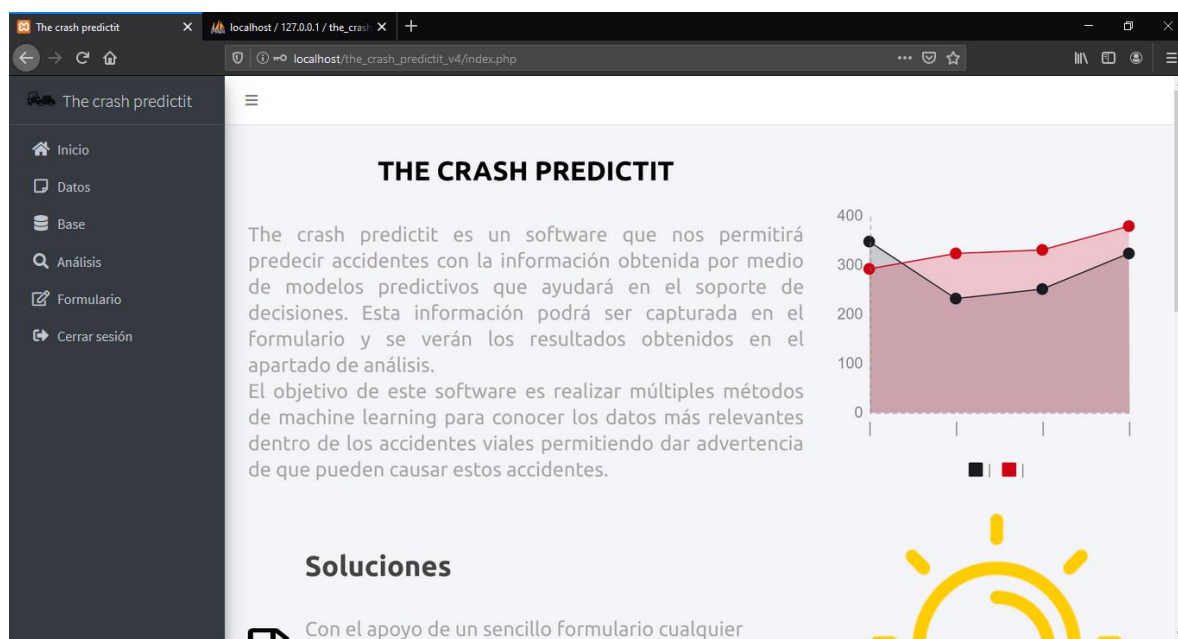
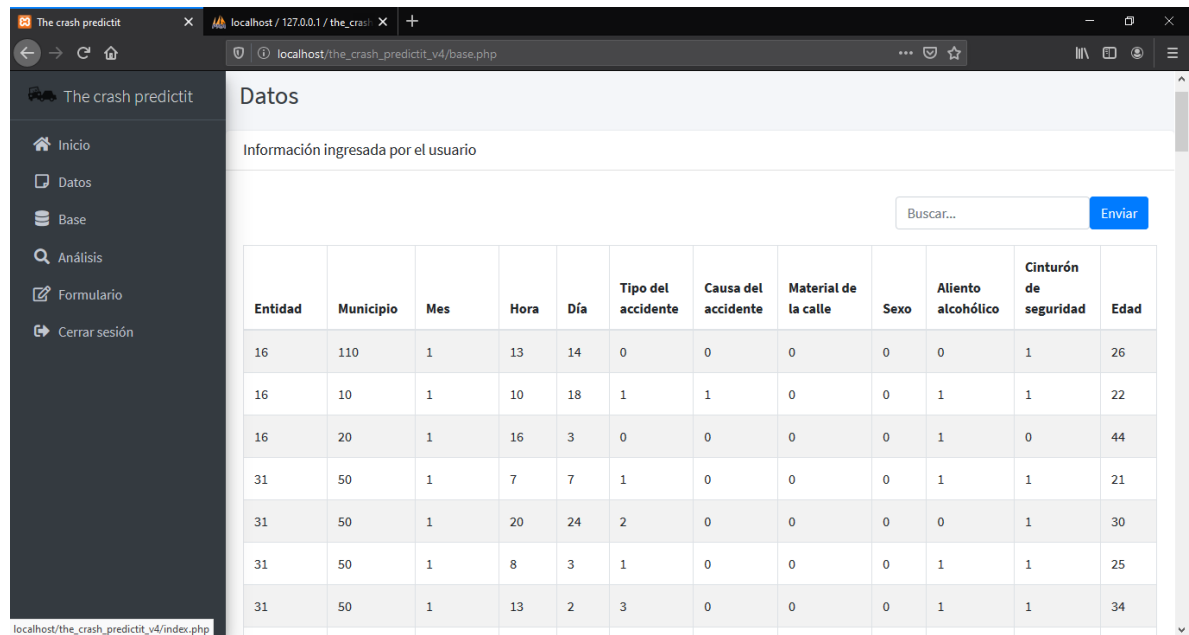


Imagen 23. Pestaña de inicio de administrador.

Base de datos

Como se ve en la imagen 23, se introducen las opciones de entrar a ver la base de datos y cerrar sesión. En la pestaña de base, como se ve en la imagen 24. Se puede editar el registro y eliminar datos que crea irrelevante.



The screenshot shows a web browser window with the URL `localhost/the_crash_predictit_v4/base.php`. The page title is "Datos". Below the title, there is a search bar with the placeholder text "Buscar..." and a blue "Enviar" button. The main content is a table with 12 columns: Entidad, Municipio, Mes, Hora, Día, Tipo del accidente, Causa del accidente, Material de la calle, Sexo, Aliento alcohólico, Cinturón de seguridad, and Edad. The table contains 8 rows of data.

Entidad	Municipio	Mes	Hora	Día	Tipo del accidente	Causa del accidente	Material de la calle	Sexo	Aliento alcohólico	Cinturón de seguridad	Edad
16	110	1	13	14	0	0	0	0	0	1	26
16	10	1	10	18	1	1	0	0	1	1	22
16	20	1	16	3	0	0	0	0	1	0	44
31	50	1	7	7	1	0	0	0	1	1	21
31	50	1	20	24	2	0	0	0	0	1	30
31	50	1	8	3	1	0	0	0	1	1	25
31	50	1	13	2	3	0	0	0	1	1	34

Imagen 24. Pestaña de base de datos.

V. Conclusiones

Durante el proceso de la elaboración del software se tuvieron diferentes dificultades para el procesamiento de los datos y para crear los modelos de Machine Learning, pero poco a poco se fueron resolviendo.

Dentro de los algoritmos de Machine Learning se pudo apreciar que el método de Random Forest tiene mayor exactitud en las predicciones con respecto al valor verdadero porque se obtuvo un score de 83 por ciento en el dataset de prueba. Mientras otros modelos, como las máquinas de vectores de soportes y la red neuronal solo utilizaron una cantidad de 100,000 del dataset, donde el total de datos en el conjunto de datos es de 829,040, por ello se obtuvo un score de 77 por ciento en máquinas de vector de soportes y 79 por ciento en la red neuronal. El score dentro de la regresión logística que utilizó todo el dataset se obtuvo como resultado un porcentaje de 78. Dentro de los reportes de clasificación el Random Forest tiene mayor

sensibilidad y mayor f1 score, por lo que se buscará modelar los datos que se obtengan en el software capturado de los usuarios.

Se eligió el tema de accidentes viales ya que es algo que se ve día tras día y el no ir en un automóvil no nos exenta de estar a salvo por la poca cultura vial por parte de los automovilistas y peatones. Por eso es necesario concientizar a las personas en este tema tan importante de la seguridad y prevención de accidentes.

Actualmente el proyecto se encuentra en una etapa funcional, pero aún se pueden hacer mejoras para que sea un sistema más óptimo que pueda llegar a crear nuevos modelos de Machine Learning que clasifiquen más variables clave, y que la base de datos pueda llegar a tener autonomía. Por ese motivo ya se contemplan ciertas mejoras en las que se trabajan porque será un excelente apoyo para las instituciones de movilidad. Ésta nueva manera en que se recolectarán los datos y la manera en la que se procesarán para tomar medidas pueden evitar accidentes antes de que sea demasiado tarde.

VI. Referencias Bibliográficas.

- SISCAV S.F (2019) Actas de accidentes. Quejas administrativas por choque y reporte de accidentes del transporte público de la dirección General Jurídico y Dirección General de Delegaciones Foráneas, Secretaria de Movilidad. Recuperado de:
[accidentes_viales_en_jalisco_2012-2019.0.pdf](#).
- INEGI (2017). Accidentes de tránsito terrestre en zonas urbanas y suburbanas. Recuperado de:
<https://www.inegi.org.mx/sistemas/olap/proyectos/bd/continuas/transporte/accidentes.asp>
- Oracle. (2019). MySQL. Recuperado de:
<https://www.oracle.com/database/technologies/mysql.html>
- Inc. Management Information Systems MIS (2019) de:
<https://www.inc.com/encyclopedia/management-information-systems-mis.html>.
- Python Software Foundation. (2019). Sobre aplicaciones para Python de:
<https://www.python.org/>

- Chong, M., Ajit, A. y Paprzycki, M. (2003). Accident Data Mining Using Machine Learning Paradigms. Computer Science Department. Oklahoma State University, USA : s.n.,. cities of Turkey, pp. 906-913.
- Kunt, M., Aghayan, I. y Noi, N. (2011). Prediction for traffic accident severity: Comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods Transport Volume 26, Issue 4, 1, Pages 353-366.