Accidentes viales en la CDMX

Llamado oportuno

Yusuri Arciga
Diego Villegas
Yedam Fortiz

GOBIERNO DE LA CIUDAD DE MÉXICO
CENTRO DE COMANDO, CONTROL, CÓMPUTO, COMUNICACIONES Y CONTACTO CIUDADANO DE LA CIUDAD DE MÉXICO

Agenda

1. Objetivo
2. Características generales
3. Perfilamiento de los datos
4. Modelaje
5. Conclusiones

Objetivo

Predecir si una llamada al C5* para reportar un incidente vial es **Falsa** o no

Restricciones

Solo tenemos 20 ambulancias para enviar en caso de un incidente

*Centro de Comando, Control, Cómputo, Comunicaciones y Contacto Ciudadano de la CDMX

Caracteristicas Generales

Tenemos 1,383,138 registros

18 variables

	número de elementos únicos
folio	1,383,138
fecha_creacion	2,497
hora_creacion	86,299
dia_semana	7
codigo_cierre	5
fecha_cierre	2,496
año_cierre	7
mes_cierre	12
hora_cierre	86,375
delegacion_inicio	16
incidente_c4	26
latitud	82,501
longitud	78,984
clas_con_f_alarma	4
tipo_entrada	9
delegacion_cierre	16
geopoint	134,288
mes	12
lat_geo	82,501
long_geo	78,984
label	2
incidente_falso	2
incidente_falso_color	2
codigo_cierre_	5

Caracteristicas Generales

4 variables geoespaciales

	Tipo de dato
latitud	float64
longitud	float64
lat_geo	float64
long_geo	float64

4 variables de fecha

	Tipo de dato
fecha_creacion	datetime64[ns]
hora_creacion	datetime64[ns]
fecha_cierre	datetime64[ns]
hora_cierre	datetime64[ns]

9 variables categoricas

	Tipo de dato
dia_semana	category
codigo_cierre	category
año_cierre	category
mes_cierre	category
delegacion_inicio	category
clas_con_f_alarma	category
tipo_entrada	category
delegacion_cierre	category
mes	category

3 variables cadena

	Tipo de dato
folio	object
incidente_c4	object
geopoint	object

Caracteristicas Generales

Contamos con información del año 2014 al 2020

Las variables con faltantes

Variable	Proporcion faltante
delegacion_inicio	0.01%
delegacion_cierre	0.01%
latitud	0.03%
longitud	0.03%

Al hacer la conversión de formato

Variable	Proporcion faltante
hora_creacion	0.1%
hora_cierre	0.1%

Perfilamiento de los datos

Variables categóricas

metric	dia_semana	codigo_cierre	año_cierre	mes_cierre	delegacion_inicio	clas_con_f_alarma	tipo_entrada	delegacion_cierre	mes
num_categories	7	5	7	12	16	4	9	16	12
uniques	7	5	7	12	16	4	9	16	12
prop_missings%	0	0	0	0	0.01	0	0	0.01	0
num_na	0	0	0	0	158	0	0	140	0
top1_repeated	Viernes	(A) La unidad de atención a emergencias fue de...	2018	Octubre	IZTAPALAPA	EMERGENCIA	LLAMADA DEL 911	IZTAPALAPA	10
top2_repeated	Sábado	(D) El incidente reportado se registró en dos ...	2019	Agosto	GUSTAVO A. MADERO	URGENCIAS MEDICAS	LLAMADA DEL 066	GUSTAVO A. MADERO	8
top3_repeated	Jueves	(N) La unidad de atención a emergencias fue de...	2017	Septiembre	CUAUHTEMOC	FALSA ALARMA	BOTÓN DE AUXILIO	CUAUHTEMOC	9

Las únicas variables que hay datos faltantes son: delegación de inicio y cierre y representan 0.01%

El código de cierre que más se repite es el A=“Afirmativo”: Una unidad de atención a emergencias fue despachada, llegó al lugar de los hechos y confirmó la emergencia reportada

Iztapalapa es la delegación con más reportes y la forma más solicitado de apoyo es la llamada al 911 seguida de la llamada al 066

Perfilamiento de los datos

Variables fecha y hora

metric	fecha_creacion	hora_creacion	fecha_cierre	hora_cierre
num_categories	2497	86299	2496	86375
max	2020-12-10 00:00:00	1900-01-01 23:59:59	2020-12-10 00:00:00	1900-01-01 23:59:59
min	2013-12-31 00:00:00	1900-01-01 00:00:00	2014-01-01 00:00:00	1900-01-01 00:00:00
uniques	2497	86299	2496	86375
prop_missings%	0	0.1	0	0.1
num_na	0	1544	0	1542
top1_repeated	2020-02-14 00:00:00	1900-01-01 20:44:00	2020-02-14 00:00:00	1900-01-01 21:59:00
top2_repeated	2018-10-26 00:00:00	1900-01-01 19:16:00	2017-08-12 00:00:00	1900-01-01 21:52:00
top3_repeated	2019-11-30 00:00:00	1900-01-01 18:38:00	2017-02-12 00:00:00	1900-01-01 21:58:00

Considerando el número de años y días, corresponde a que casi diario se hacen llamadas pues los 2497 datos únicos se encuentran en el rango de fechas

En las variables de horas el porcentaje de faltantes es de 0.1%

Las horas de creación del reporte que más se repiten se realizan en la tarde-noche. Y en las que se cierran tienen un rango muy similar, solo se diferencían por minutos

Perfilamiento de los datos

Variables geoespaciales

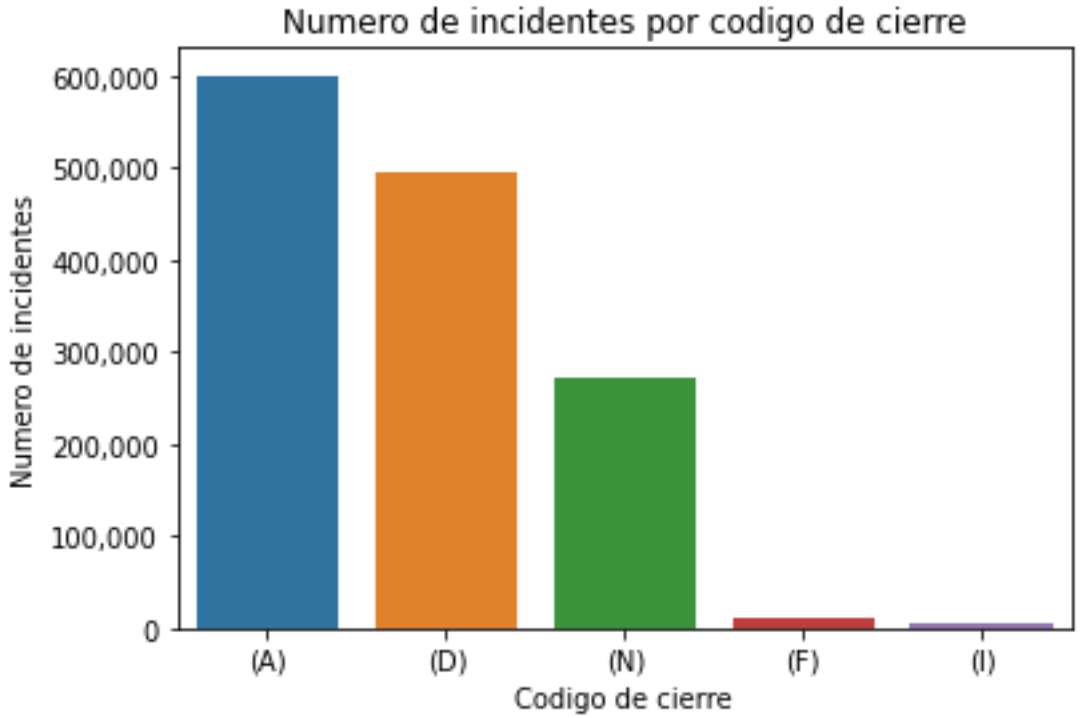
metric	latitud	longitud	lat_geo	long_geo
uniques	82501.	78984	82501.	78984
prop_missings%	0.03	0.03	0.03	0.03
num_na	443.	435.	443.	435.
top1_repeated	19.304320	-99.080240	19.304320	-99.080240
top2_repeated	19.371680	-99.087140	19.371680	-99.087140
top3_repeated	19.347021	-99.180646	19.347021	-99.180646

Solo tenemos las variables geoespaciales como numéricas pues las demás consideramos pueden ser categóricas

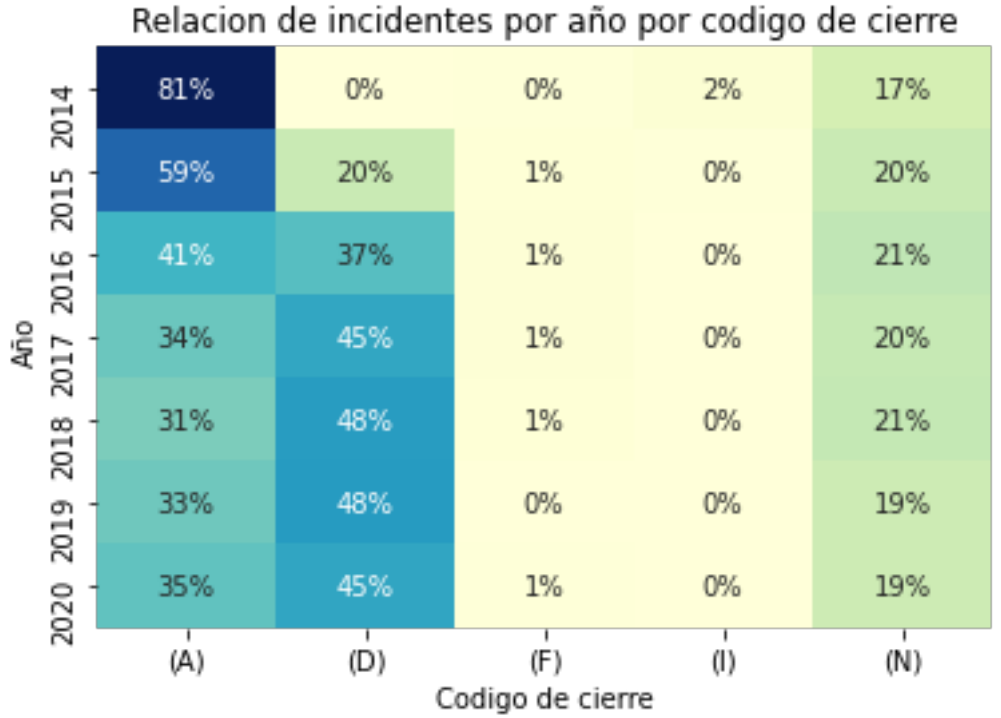
Hicimos la comparación de la columna geopoint y latitud, longitud y observamos que son iguales al hacer la separación correspondiente

El punto geoespacial que más se repite se encuentra en Iztapalapa

Perfilamiento de los datos



Los codigos de atencion despachada y confirmada (A) e incidente reportado en varias ocasiones (D) representan el 43% y 36% de los incidentes respectivamente (79% del total)

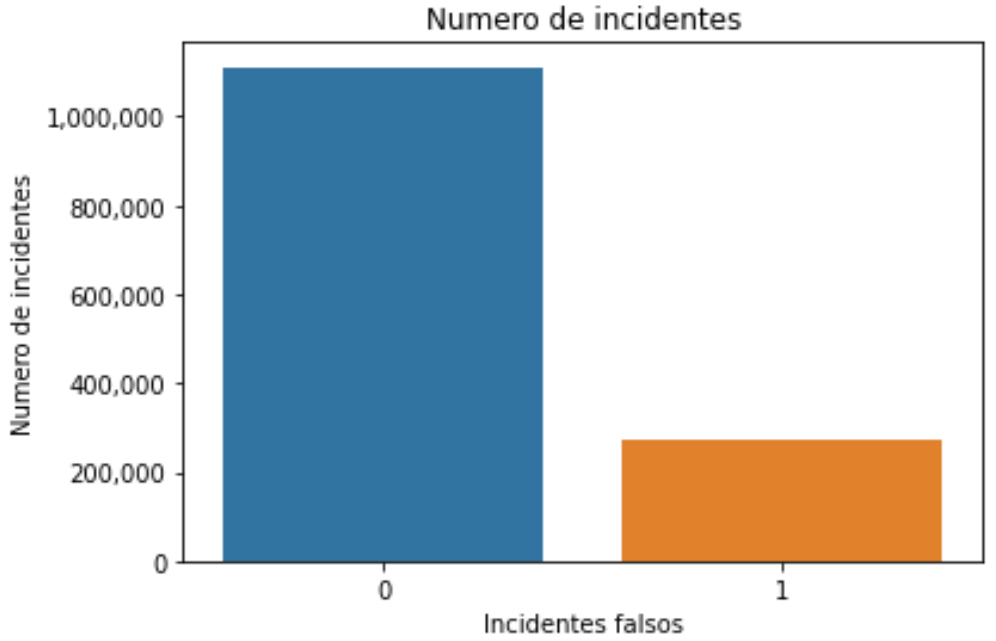


Desde 2014 las categorias A y D representan entre 78 y 81 por ciento.
La siguiente categoría (N) es atención despachada pero en el lugar nadie habia solicitado el servicio

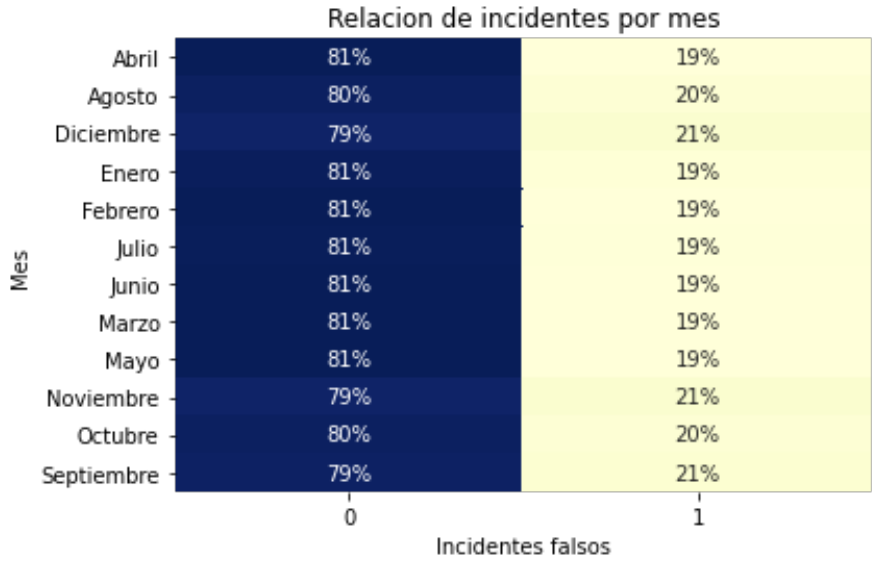
Perfilamiento de los datos

Para enfocar el analisis consideraremos los codigos de atencion despachada pero nadie en el lugar nadie habia solictado el servicio (N) y emergencia falsa (F) como incidentes falsos (1), mientras que las demas claves serán considerados incidentes verdaderos (0)

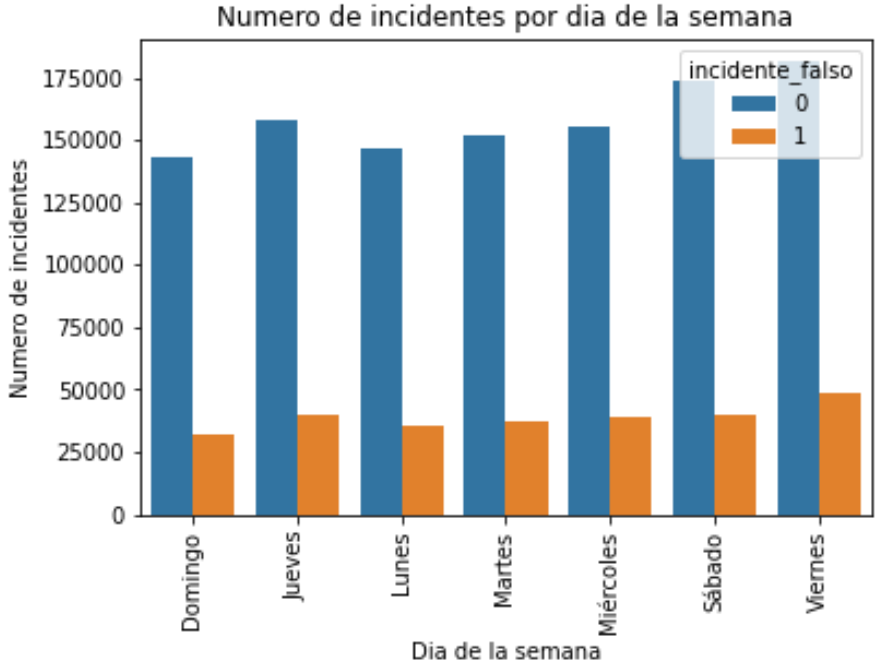
Ochenta por ciento de los casos se consideran como incidentes verdaderos mientras que veinte por cierto son falsos para el periodo de 2014 a 2020



Perfilamiento de los datos

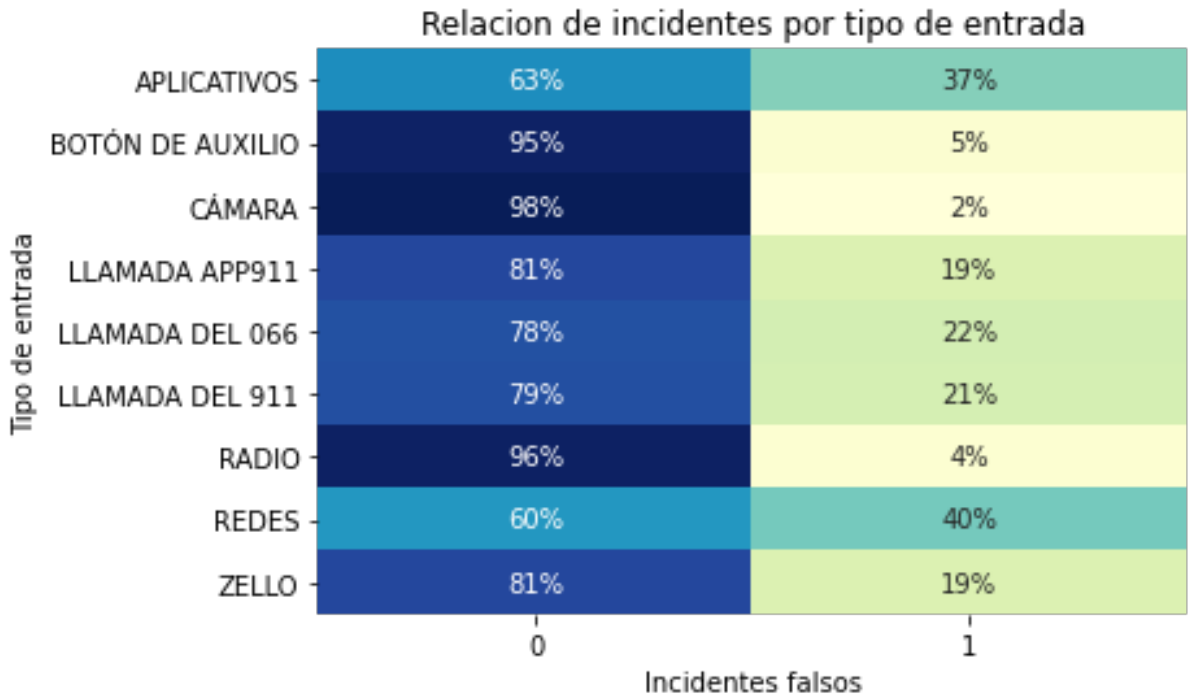


Existe un comportamiento estable entre todos los meses del año; existiendo únicamente una diferencia de 2% entre consultas en diferentes meses



Los dias de la semana donde se reportan mas incidentes son Sabado y Viernes. La proporcion de incidentes verdaderos ronda entre el 78% y 81%, volvemos a ver estabilidad entre periodos

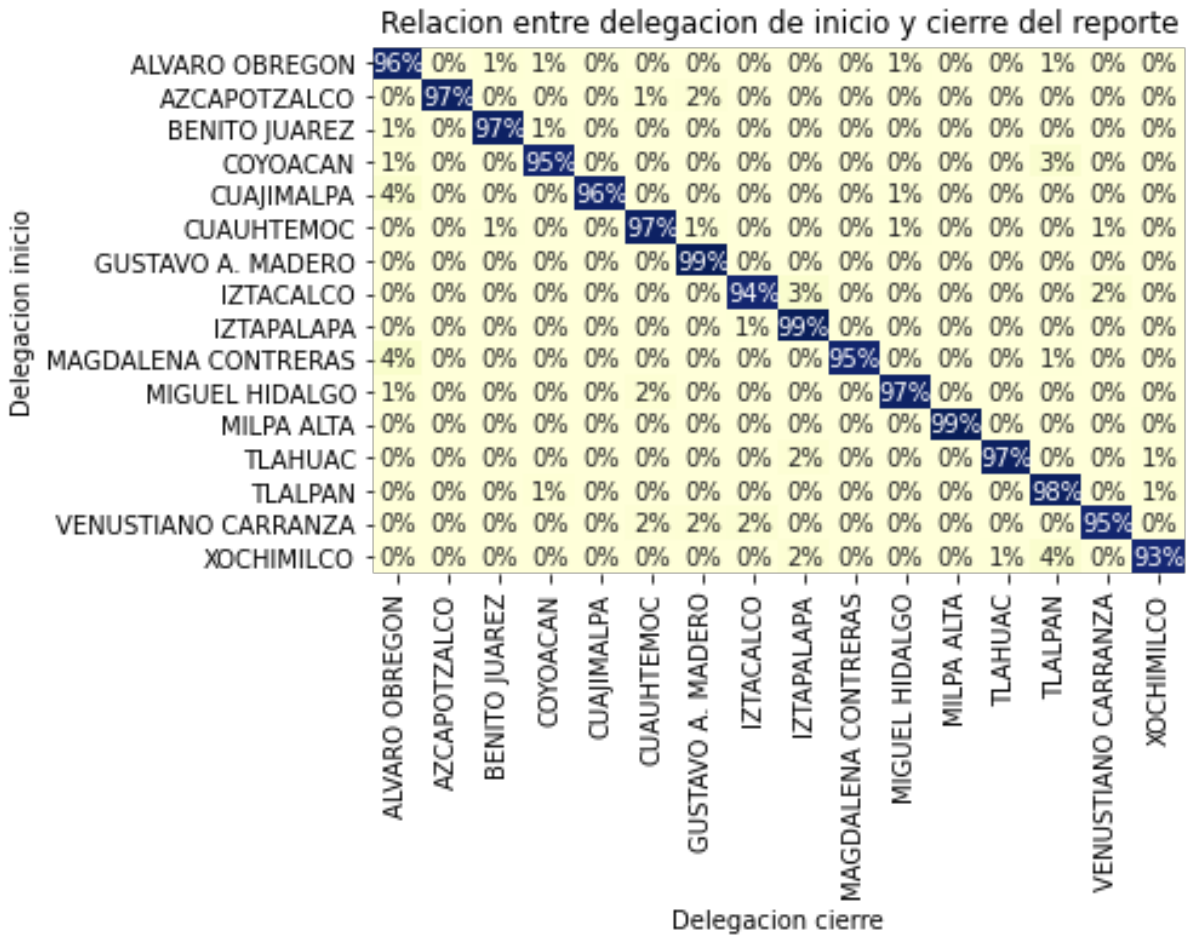
Perfilamiento de los datos



Podemos notar que dependiendo el tipo de entrada para reportar un incidente si existe una diferencia considerable entre que el caso sea real o falso

Para los casos de aplicativos y de redes la tasa de casos verdaderos se encuentra alrededor de 60% mientras que cuando se reporta a traves de camara o radio la efectivdad ronda 96%.

Perfilamiento de los datos



Verificamos si la delegacion donde se abre el incidente vial es la misma donde cierra o si existe alguna tendencia.

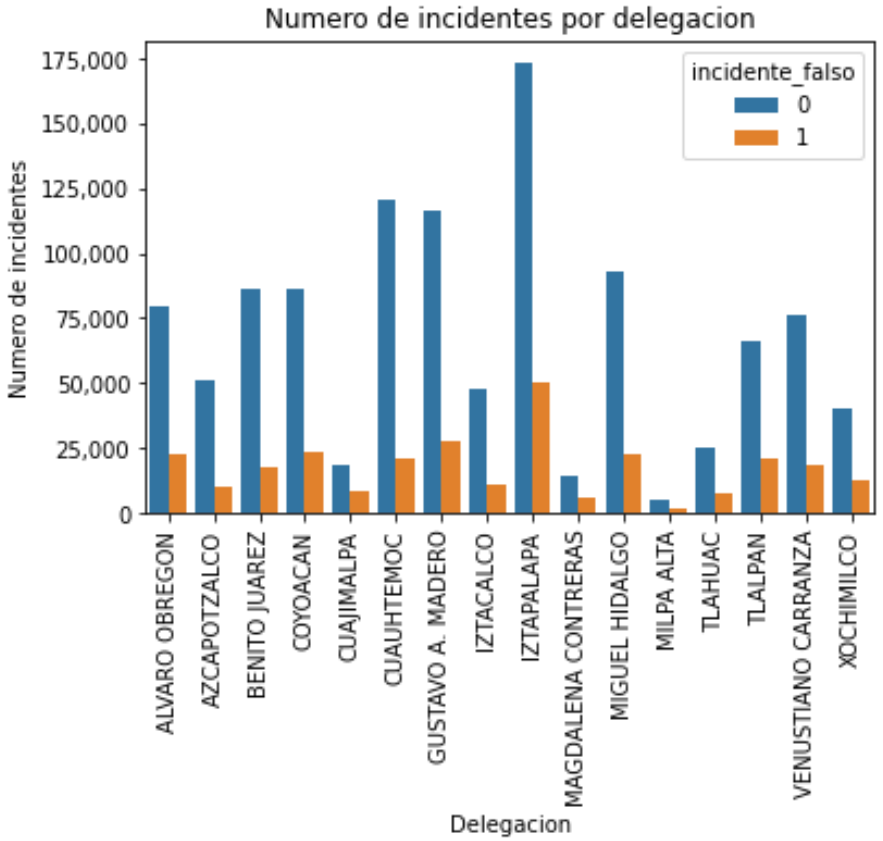
Cada renglon representa la delegacion donde se inicio y la columan donde se cerro. El color representa donde se encuentran los mayores casos.

En particular, se muestra la diagonal con mas incidencias lo que significa que la delegacion donde se abrio es donde se cerro.

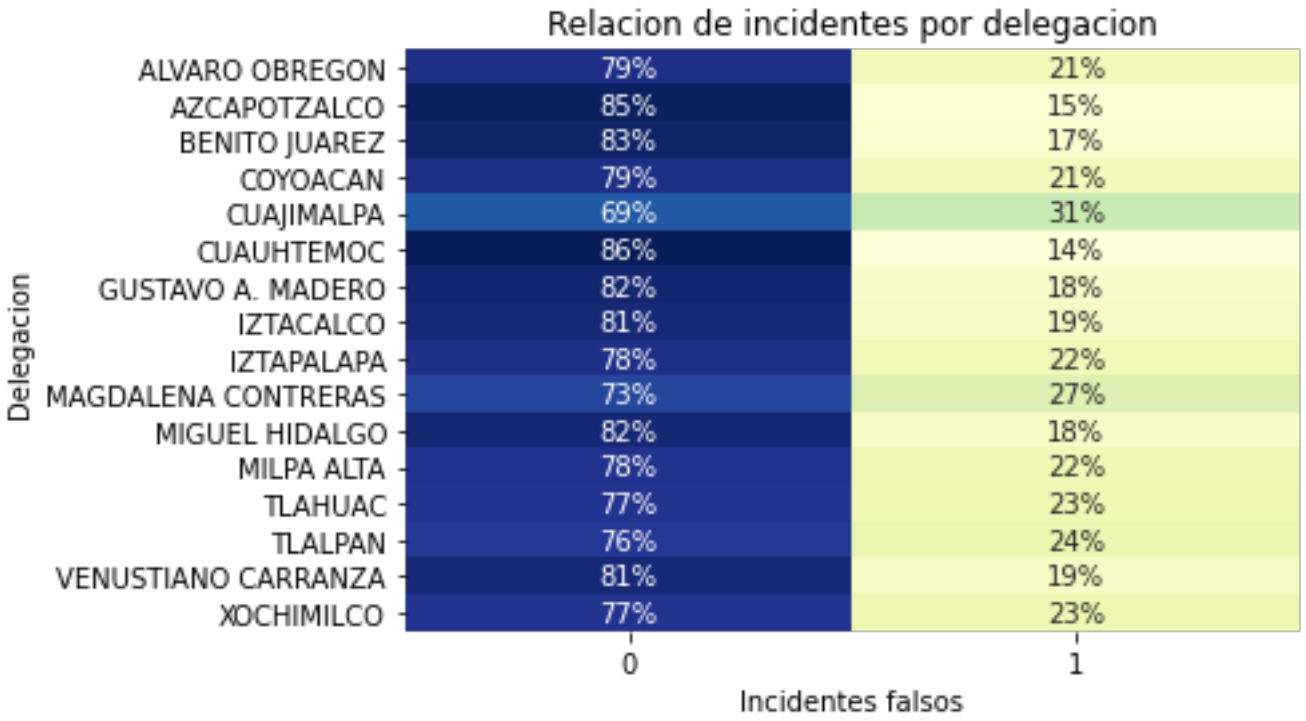
El rango nos indica que entre 94% y 99% se cerro donde se abrio el incidente.

Siendo Iztacalco donde solo 94% de los casos sucedió asi y 99% para la Gustavo A. Madero y Milpa Alta.

Perfilamiento de los datos



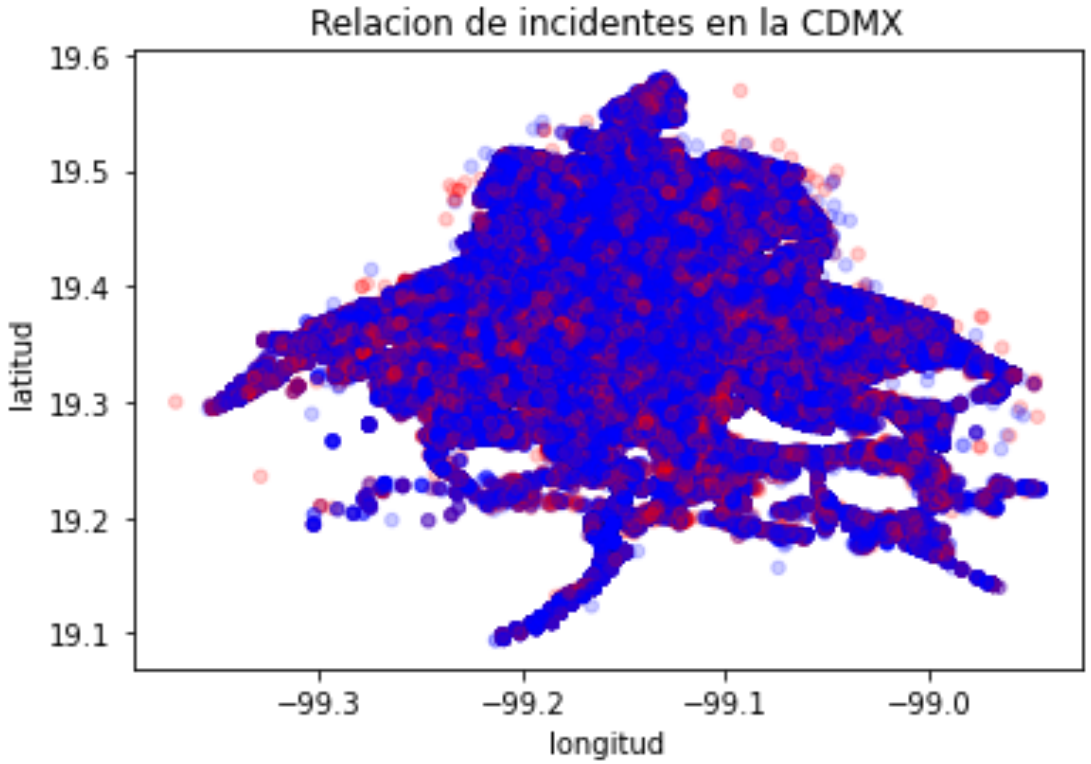
Las delegaciones con mayor numero de registros de incidentes son : Iztapalapa, Cuauhtemoc y Gustavo A. Madero.



La relacion de incidentes reportados que son verdaderos y falsos, varía considerablemnte entre delegaciones.
Podemos notar que Cuajimalpa y Magdalena contreras tienen el porcentaje mas alto de casos falsos.
Para las demas delegaciones ronda entre 14% y 24%

Perfilamiento de los datos

Podemos notar en el mapa de la ciudad de Mexico con color azul los casos reportados que son verdaderos y con rojo los falsos.



80% de los casos que se reportan son verdaderos por esta razon el mapa se visualiza mayormente de color azul. Notamos una concetracion de casos falsos en el Oeste de la CDMX, referente a Cuajimalpa y Magadalena Contreras asi como en el lado Este, en la zona de Iztapalapa.

Modelaje

Eliminacion de variables

Dentro de este análisis decidimos eliminar las variables: latitud, longitud, codigo_cierre, fecha_creacion, incidente_c4

Eliminamos las variables de cierre ya que esta información es provista cuando se cierra el incidente y se conocen las causas

Eliminamos la fecha de creación ya que ocuparemos la variable de año y mes referente a esa fecha

Modelaje

Creacion de variables

Creamos dos nuevas variables cíclicas para el análisis, que son el coseno y el seno de la hora del reporte del siniestro para reflejar la distancia entre los periodos de tiempo

Imputacion de variables

Al realizar el cambio de variable de la hora de creación a tipo fecha encontramos errores por lo que decidimos realizar una imputación simple de los datos de esta variable con la media de todas las horas, esto se ve reflejado después de hacer el cambio a seno y coseno del tiempo. Estos registros representan un porcentaje minusculo de la tabla por lo que decidimos hacer una imputacion sencilla que sea entendible para el negocio y que no tenga una afectacion a la variable

Modelaje

Selección de variables

En primera instancia aplicamos OneHotEncoder para transformar las variables categóricas a columnas obteniendo 40, aplicamos GridSearchCV para encontrar los mejores parámetros de nuestra selección inicial con lo que obtenemos los siguientes resultados:

Parametros de major estimador para Random Forest: {'max_depth’: 5, 'min_samples_split': 2, 'n_estimators’: 20}

Variables que aportan más del 7% al modelo (clas_con_f_alarma y tipo_entrada):

importance	col_name
27%	DELITO
27%	FALSA ALARMA
17%	URGENCIAS MEDICAS
9%	CÁMARA

Conclusiones

La informacion es **relevante** para predecir si una llamada al C5 para reportar un incidente vial es **Falsa** o no

Considerando que tenemos recursos limitados para atender las solicitudes, 20 ambulancias, se evaluó la **precision** del modelo para poder dar ayuda a los que estamos seguros que la necesitan

El modelo resultante es un **RandomForest** (Max_depth: 5, Min_sample_split: 2, N_estimators: 20) con los siguientes resultados:

- Precision en los datos de prueba de 80%

Las variables que aportan más del 7% son (**clas_con_f_alarma y tipo_entrada**):

importance	col_name
27%	DELITO
27%	FALSA ALARMA
17%	URGENCIAS MEDICAS
9%	CÁMARA