



Universidad  
Católica del  
Uruguay

---

**Ingeniería en informática**  
**Introducción a los métodos de aprendizaje automático**  
**(Inteligencia Artificial I)**

*Profesor : Ernesto Ocampo*  
*Profesores Asistentes: Juan Francisco Kurucz, Jose Vilaseca*

**05/12/2023**

<b>1. Abstract</b>	<b>3</b>
<b>2. Introducción</b>	<b>4</b>
<b>3. Comprensión del negocio</b>	<b>4</b>
PREDICTIVE POLICING	4
FACTORES QUE INFLUYEN EN LA INCIDENCIA DE UN CRIMEN	5
TIPO DE PROBLEMA	6
<b>4. Comprensión de los datos</b>	<b>6</b>
VALORES FALTANTES	7
VALORES DUPLICADOS	7
ANÁLISIS DE LOS ATRIBUTOS	7
ATRIBUTOS CORRELACIONADOS, OUTLIERS, NORMALIZACIÓN	10
VISUALIZACIÓN DE LOS DATOS	10
<b>5. Preparación de los datos</b>	<b>15</b>
TRABAJO SOBRE LOS ATRIBUTOS	15
<b>6. Modelado</b>	<b>15</b>
PROTOTIPO Y PRUEBA DE CONFIGURACIONES	16
PROCESO UTILIZADO	17
MODELO EN PYTHON:	17
MODELO EN RAPIDMINER:	18
FEATURE SELECTION:	19
FORWARD SELECTION	19
BACKWARD ELIMINATION:	20
POSIBLES MEJORAS AL MODELO	21
<b>7. Evaluación</b>	<b>21</b>
COMPARACIÓN DE PERFORMANCE ENTRE LOS MODELOS	21
CONCLUSIONES FEATURE SELECTION	22
CONCLUSIONES EVALUACIÓN	23
<b>8. Deploy</b>	<b>23</b>
<b>9. Referencias</b>	<b>23</b>
<b>10. Anexos</b>	<b>25</b>

## **1. Abstract**

El siguiente trabajo fue propuesto en el marco de la materia Introducción a los Métodos de Aprendizaje Automático (Inteligencia Artificial I). Se trata de un trabajo de aplicación en el cual se verán reflejados los diferentes temas trabajados a lo largo del curso a través de la metodología CRISP-DM.

## 2. Introducción

De 1934 a 1963, San Francisco fue tristemente famoso por albergar a algunos de los criminales más célebres del mundo en la ineludible isla de Alcatraz. Hoy, la ciudad es más conocida por su escena tecnológica que por su pasado criminal. Pero, con el aumento de la desigualdad económica y la escasez de vivienda, la delincuencia no escasea en la zona de la bahía. Este dataset ofrece casi 12 años de informes sobre delitos en todos los barrios de San Francisco. (Ocampo, 2023)

Dada la hora y la ubicación, se debe predecir la categoría de delito que se ha producido. Para esto se dispone del siguiente dataset extraído de kaggle: San Francisco Crime Classification | Kaggle , <https://www.kaggle.com/competitions/sf-crime/overview>.

Este hace referencia a una competición de Machine Learning ocurrida entre 2015 y 2016.

Para realizar el estudio de este problema se utilizará el proceso CRISP-DM (Cross Industry Standard Process for Data Mining). (Ver ilustración anexo 17)

Este proceso se verá aplicado tanto en RapidMiner como en Python mediante el uso de las diferentes librerías.

Se adjuntan los archivos generados por RapidMiner, un archivo .ipynb, un anexo, un archivo excel con los resultados de rapidminer.

## 3. Comprensión del negocio

### PREDICTIVE POLICING

En los últimos años se ha visto cómo el aprendizaje automático ha ganado terreno en el ámbito policial. Esto debido a una necesidad frente al aumento de crímenes y la urgencia de establecer planes de acción contra los mismos. Incluso hay una rama de la ciencia de datos que se encarga de este tipo de problemas conocido como Predictive policing; el cual según Wikipedia (2023) consiste en utilizar matemática, análisis predictivo, ML y otras técnicas de análisis y predicción de datos con el fin de identificar potenciales actividades criminales. Existen cuatro categorías: predicción de delitos, métodos de predicción de delincuentes, métodos de predicción de la identidad de los autores y métodos de predicción de las víctimas de delitos. En el caso del presente informe nos encontramos ante la primera categoría: predicción de delitos.

El objetivo de esta disciplina es que mediante el uso de datos de tiempo, localización y naturaleza de crímenes pasados, se obtenga una mirada y brinde diferentes estrategias a la policía para conocer dónde y a qué hora deben patrullar los móviles policiales, o mantener una presencia, para hacer el mejor uso posible de los recursos para disuadir o prevenir futuros delitos.

Las predicciones generadas por el algoritmo deben ir acompañadas de una estrategia de prevención o anticipación, que suele consistir en enviar a un agente a la hora y el lugar posible del delito. El uso de la policía predictiva automatizada

proporciona un proceso más preciso y eficaz a la hora de analizar futuros delitos, ya que existen datos que respaldan las decisiones, en lugar de limitarse a los instintos de los agentes. Si la policía utiliza la información procedente de la vigilancia predictiva, puede anticiparse a las preocupaciones de las comunidades, asignar sabiamente los recursos a los momentos y lugares y prevenir la victimización.

En la actualidad muchos países aplican esta metodología, entre ellos la UE, China, Estados Unidos. Concretamente en Estados Unidos este tipo de tecnologías se aplica en numerosos estados. En Nueva York se implementan modelos que ayudan a los agentes de policía a identificar puntos en común en los delitos cometidos por los mismos delincuentes o el mismo grupo de delincuentes. Los agentes pueden ahorrar tiempo y ser más eficientes, ya que el programa genera el posible "patrón" de diferentes delitos. A continuación, el agente tiene que buscar manualmente entre los posibles patrones para ver si los delitos generados están relacionados con el sospechoso actual. Si los delitos coinciden, el agente iniciará una investigación más profunda de los delitos del patrón. Para San Francisco no hay antecedentes, pero sí se utilizó un modelo de predicción en LA, este consistía en generar mapas de calor de zonas donde podía ocurrir un crimen.

El principal problema que surge del Predictive policing es: se basa en la aportación humana para determinar patrones. Los datos erróneos pueden conducir a resultados sesgados y posiblemente racistas. La tecnología no puede predecir la delincuencia, sólo puede convertir la proximidad en un arma policial. Aunque se afirma que los datos son imparciales, las comunidades de color y de bajos ingresos son las más afectadas. También hay que señalar que no todos los delitos se denuncian, lo que hace que los datos sean erróneos e inexactos.

Se considera de suma importancia lograr entrenar modelos claros y sin sesgo alguno. Esto debido a que se está tratando con personas y un asunto sensible como lo es la seguridad de las mismas. Por lo tanto la evaluación de los modelos, el procesado de datos y la interpretación de los datos debe ser tomada en cuenta junto a un análisis humano.

En resumen:

Los métodos policiales predictivos no son una bola de cristal: no pueden predecir el futuro. Sólo pueden identificar a las personas y los lugares con mayor riesgo de delincuencia. Los enfoques policiales predictivos más eficaces son elementos de estrategias proactivas más amplias que establecen relaciones sólidas entre los departamentos de policía y sus comunidades para resolver los problemas de delincuencia (Perry et al,2013).

## FACTORES QUE INFLUYEN EN LA INCIDENCIA DE UN CRIMEN

Si bien existen diferentes estudios sobre el tema, muchos de estos coinciden que los factores principalmente son: Fuerte expansión demográfica, Bajo nivel educacional, Deficiente situación sanitaria, Escaso nivel de vida, Condiciones de trabajo inadecuadas, Estructuras sociales atrasadas, Desarrollo de la clase media, Deficiente integración nacional esencialmente en el plano económico, Toma de conciencia en su realidad social, Reducida industrialización, Escasa renta nacional,

Débil desarrollo agrícola, Bajo nivel de consumo de energía mecánica, Hipertrofia del sector comercial.

Tener en cuenta estos factores facilitan la comprensión de la data y poder entender mejor los indicadores de criminalidad a lo largo del modelo.

## TIPO DE PROBLEMA

Nos encontramos frente a un problema de clasificación (aprendizaje supervisado) esto es debido a que se posee previamente los datos etiquetados con una variable objetivo, por lo que sabemos qué esperar del modelo y evaluarlo en base al mismo. La variable objetivo es la categoría.

## 4. Comprensión de los datos

El dataset de entrenamiento de la competencia posee 878049 entradas. Las cuales son una reducción bastante considerable del dataset que se encuentra disponible en Police Department Incident Reports: Historical 2003 to May 2018 con un total de 2.13 millones de entradas. Además de que se redujo la dimensionalidad de 14 originalmente a 9.

El gran tamaño del dataset nos da un indicador claro: se requerirá una gran capacidad de cómputo. Por lo que la optimización, la feature selection así como el manejo inicial de muestras representativas serán claves para lograr un mejor trabajo.

El dataset provisto, según la página de kaggle y luego de realizar una breve lectura a los datos de entrenamiento, contiene los siguientes atributos:

Nom Atributo	Tipo_dato	Descripción	¿Faltan valores?
<b>Dates</b> (Fecha)	Categórico	Fecha y hora en la que ocurrió el crimen	No
<b>Category</b> (Categoría crimen)	Categórico - variable objetivo	Categoría del crimen	No
<b>Descript</b> (Descripción)	Categórico	Descripción detallada del crimen	No
<b>DayOfWeek</b> (Dia de la semana)	Categórico	Dia de la semana	No
<b>PdDistrict</b> (Distrito policial)	Categórico	Nombre del departamento de policía	No
<b>Resolution</b> (Resolución)	Categórico	Como el crimen fue resuelto	No

Address (Dirección)	Categórico	Representa aproximadamente la calle donde se resolvió el crimen	No
X	Real	Representa la longitud	No
Y	Real	Representa la latitud	No

Por otro lado se posee el dataset de prueba con un total de 884262 filas. Este posee todos los atributos del dataset mencionado anteriormente a excepción de descripción y resolución; además de agregarse un atributo de id.

	Entrenamiento	Prueba
Elementos	878049	884262
Dimensionalidad	9	7

## VALORES FALTANTES

Ningún atributo posee valores faltantes (Ver anexo 1).

## VALORES DUPLICADOS

Tras aplicar un filtro de datos duplicados se obtiene que existen 2,323 filas duplicadas. Es imposible determinar si efectivamente se trata de un error o un mismo crimen ocurrió en el mismo lugar a la misma hora.

## ANÁLISIS DE LOS ATRIBUTOS

**FECHA:** Viene con el formato yyyy-MM-dd HH:mm:ss. Es decir que engloba dos datos en un campo solo por un lado la fecha como tal y por otro la hora. El campo es importante debido a que se da que ciertos crímenes son cometidos en fechas más concretas (Ejemplo aumentos de ciertos crímenes en navidad), así como también en fechas del año y hasta en determinados rangos horarios. De manera estadística en el formato que se encuentra es muy difícil de trabajar, pues es mucha información junta.

**CATEGORÍA:** Es la variable objetivo, la cual es categórica. Esta tiene un total de 39 clases, las cuales son:

Larceny/theft, other offenses, non-criminal, assault, drug/narcotic, vehicle theft, vandalism, warrants, burglary, suspicious occ, missing person, robbery, fraud, forgery/counterfeiting, secondary codes, weapon laws, prostitution, trespass, stolen property, sex offenses forcible, disorderly conduct, drunkenness, recovered vehicle, kidnapping, driving under the influence, runaway, liquor laws, arson, loitering, embezzlement, suicide, family offenses, bad checks, bribery, extortion, sex offenses non-forcible, gambling, pornography/obscene mat, and trea.

Donde las tres más recurrentes (y mediante una gran diferencia) son LARCENY/THEFT, OTHER OFFENSES y NON-CRIMINAL. En cuarto y quinto puesto se encuentra ASSAULT y DRUG/NARCOTIC (Para gráficas y estadísticas relevantes ver anexo 2).

**DESCRIPCIÓN:** Funciona como una subcategoría de Categoría misma. Encontramos un total de 821 clases. Este atributo da una descripción más detallada del crimen. Por ejemplo si vemos la categoría de vandalismo encontramos que tenemos 25 sub categorías de vandalismo: (Para gráficas y estadísticas relevantes ver anexo 3).

Este campo no se encuentra en el dataset de prueba, por lo que se excluirá del dataset para entrenar el modelo.

**DIA DE LA SEMANA:** Corresponde al día de la semana en la que ocurrió el crimen, están los 7 días de la semana. Es importante porque se da que ciertos crímenes se dan más en un día que en otros (Para gráficas y estadísticas relevantes ver anexo 4).

Se puede ver que los crímenes están bastante balanceados en términos de días, donde la diferencia entre la mayor diferencia entre el día que se cometen más crímenes (Viernes) y el día donde se cometen menor cantidad de crímenes (Domingo) es de 2%.

**DISTRITO POLICIAL:** Según Wikipedia (2023), la SFPD (San Francisco Police Department) se divide en 10 distritos (Para gráficas y estadísticas relevantes ver anexo 6).

Podemos observar que en el dataset se encuentran los 10 distritos mencionados. A mencionar que los dos lugares donde se registraron mayor cantidad de crímenes fue en SOUTHERN Y MISSION mientras que donde se registró la menor fue en PARK y RICHMOND.

**SOUTHERN:** abarca el área Sur de Market, el Embarcadero, China Basin, Mission Bay, Treasure Island y Yerba Buena Island.

**MISSION:** cubre el área al este de Twin Peaks hasta la autopista James Lick Freeway y al sur de Market Street hasta Cesar Chavez Street. Aislados de la niebla y el viento que afectan a gran parte del resto de la ciudad, los residentes disfrutan de una rara cantidad de días soleados. En los últimos años, la Misión se ha convertido en un centro de restaurantes especializados, tiendas y clubes nocturnos.

**PARK:** El distrito de parques incluye el vasto extremo oriental del parque Golden Gate y los barrios mundialmente famosos de Haight-Ashbury y Castro. Está rodeado por Geary Boulevard, Steiner, Market, Upper Market, 7th Avenue y el extremo este del parque.

RICHMOND: la esquina noroeste de la ciudad. Aunque es principalmente residencial, también incluye los largos corredores comerciales, de compras y de restaurantes de Geary Boulevard y Clement Street, además de la mayor parte del Golden Gate Park, el Área Recreativa Nacional Golden Gate y el campus de la Universidad de San Francisco.

Sobre Mission y South se puede entender que tengan la mayor cantidad de ocurrencias debido a las razones socio-culturales y económicas que hay en la zona. Estos dos distritos han sufrido últimamente una fuerte gentrificación y por ende un mayor aumento de las diferencias sociales y económicas así como también una inestabilidad poblacional generando una fuerte expansión demográfica, que como se vio en la compresión es un disparador de incidencias criminales.

Este dato es importante puesto que nos da la jurisprudencia de donde se produjo el crimen y nos da una forma de englobar por distritos para su posterior análisis.

Algo interesante a mostrar es si hacemos un scatter plot de x e y (ver gráfica en anexo 9) donde el color es el distrito, se aprecia que estos datos son en su mayoría correctos. Sin embargo, hay algunos datos pertenecientes a determinado distrito que dadas las coordenadas se encuentran en jurisprudencia de otro.

Se pueden hacer dos cosas frente este problema: la primera es sustituir todos los puntos que están por fuera con las coordenadas del distrito de policía. Para esto se debería buscar los límites de cada uno y si se encuentra un x e y por fuera o en otra categoría cambiarlo. Esta opción es sumamente costosa porque tenemos que recorrer todo el dataset varias veces con el fin de poder encontrar aquellos que están fuera de los límites.

La segunda es ignorarlo, pues al ser relativamente pocos puntos en comparación el ruido que generan no es mucho. Esta opción es la más viable puesto que ahorra una gran cantidad de recursos computacionales.

**RESOLUTION:** Es el fin en el cual terminó el reporte de un crimen. Son 17 clases (Para gráficas y estadísticas relevantes ver anexo 6).

Al ver los valores se puede destacar que más de un 60% de los casos no condujeron a ninguna resolución. Lo cual puede ser entendible puesto que un análisis de los crímenes nos da que muchos son crímenes que se denuncian tiempo después de cometido el delito por ejemplo: Robo, actos no criminales, vandalismo, etc. (Para gráficas y estadísticas relevantes ver anexo 7).

A su vez destacar que este campo no se encuentra en el dataset de prueba, por lo que se excluirá del dataset para entrenar el modelo.

**DIRECCIÓN:** Representa la dirección en la que ocurrió la categoría. No tienen un formato prefijado, pero se evidencia que las categorías más predominantes tienen la palabra BLOCK contenidas en su dirección correspondiente. Se puede evaluar representar BLOCK como un atributo aparte.

**X:** Valor numérico que representa la longitud. La mínima es -122.514 y la máxima -120.500. Tiene una desviación estándar de 0.03.

**Y:** Valor numérico que representa la latitud. La mínima es 37.798 y la máxima 90. Tiene una desviación estándar de 0.414.

Latitud y longitud de San Francisco:

- Longitud: -122.4194200
- Latitud: 37.7749300

Por lo tanto se espera que los valores de X e Y estén entre esos valores.

## ATRIBUTOS CORRELACIONADOS, OUTLIERS, NORMALIZACIÓN

Al ser la gran mayoría datos categóricos se vuelve complicado aplicar técnicas tanto como para detectar outliers o matrices de correlación entre los atributos. Sobre los únicos dos datos en los cuales es posible realizar una detección de outliers es X e Y pues son los únicos datos numéricos (Ver gráfica en anexo 8).

Se puede observar que tanto los valores de X como de Y se agrupan en una parte del gráfico a excepción de unos pocos los cuales corresponden a las coordenadas (-120.5, 90.0). Se puede apreciar mismo en la gráfica como estos datos son totalmente atípicos puesto que están totalmente alejados de la gran mayoría de coordenadas. Además cabe destacar que al ser coordenadas los decimales hacen una gran diferencia.

Es más, si buscamos estas coordenadas en un mapa podemos encontrar que estas corresponden al polo norte. Es decir que son valores que están fuera de los límites de nuestro problema (San Francisco). Claramente esto se trata de un error de ingreso de los datos (Ver anexo 8).

Se cuentan cuantos (-120.5, 90.0) hay en el modelo, nos da un total de 67.

Se pueden pensar dos soluciones posibles para solucionar esto. La primera es a partir de los otros datos geográficos que se poseen en el dataset reemplazar las columnas X e Y. La segunda es eliminar todos los registros donde se encuentre este valor, los cuales no son muchos y representan solo el 0.008% de los datos.

Por último destacar que la normalización de la longitud y latitud se evalúa en función del modelo a utilizar. La desviación es muy cercana a 0 para el caso de X, esto puede perjudicar algún algoritmo que funcione en función del cálculo de distancia. Sin embargo esto se evaluará en el modelo mismo para ver cómo afecta la performance del mismo.

## VISUALIZACIÓN DE LOS DATOS

El objetivo de esta sección es realizar y tratar de entender un poco más sobre los datos que tenemos y cómo se relacionan entre ellos (Ver gráfica anexo 10).

Se puede ver como en todos los distritos predomina el jueves como el día de más denuncias para casi todos los crímenes.

### Relación categoría y año

No hay un crecimiento fijo ni una tendencia marcada, los datos son bastante parejos entre sí, aunque se observa que el año 2015 tiene menos de la mitad de registros que todos. Además de contener registros solo hasta el mes 5 (mayo). Esto

puede afectar los resultados del modelo por lo cual se evalúa eliminar los registros (son un total de 27 mil, lo cual no es una gran cantidad considerable del dataset) (Ver gráfica Anexo 16).

### **Relación categoría y mes**

Se observa que la distribución de la categoría según el mes es bastante uniforme sin presentar picos ni cambios grandes (Ver gráfica Anexo 15).

### **Relación categoría hora**

Se puede ver que las horas principales con mayor ocurrencia de categoría son 18, 17, 12, 16, 19, 22, 0, 20, 14, 21, 13, 23 (Ver gráfica Anexo 16). Se podría desprender que hay una mayor cantidad de categorías correspondiente a la Noche que al día (Definiendo Noche entre las 18 hs y las 06 hs). Lo cual se contrarresta con la evidencia puesto que durante la noche se obtiene un total de 405939 ocurrencias contra las 469787 del día. A pesar de esto la diferencia no es grande en comparación al tamaño del dataset, representa menos de un 10% de diferencia aprox. (Ver gráfica Anexo 17).

### **Categoría más recurrente por distrito**

PdDistrict	Categoría
BAYVIEW	OTHER OFFENSES
CENTRAL	LARCENY/THEFT
INGLESIDE	OTHER OFFENSES
MISSION	OTHER OFFENSES
NORTHERN	LARCENY/THEFT
PARK	LARCENY/THEFT
RICHMOND	LARCENY/THEFT
SOUTHERN	LARCENY/THEFT
TARAVAL	LARCENY/THEFT
TENDERLOIN	DRUG/NARCOTIC

### Crimen mas cometido por dia

DayOfWeek	Category	count
Friday	LARCENY/THEFT	27104
Monday	LARCENY/THEFT	23570
Saturday	LARCENY/THEFT	27217
Sunday	LARCENY/THEFT	24150
Thursday	LARCENY/THEFT	24415
Tuesday	LARCENY/THEFT	23957
Wednesday	LARCENY/THEFT	24487

Se puede observar que mientras más al este de la ciudad aumenta Recovered Vehicles (Ver gráficas Anexo 9, 10, 11).

Se confirma al ver las estadísticas de Recovered Vehicles como en donde hay mayor cantidad es BAYVIEW, INGLESIDE, SOUTHERN Y MISSION que son los distritos más al este.

### Categoría más recurrente de dia y noche segun distrito

Distrito	Categoría Dia	Categoría Noche
BAYVIEW	OTHER OFFENSES	OTHER OFFENSES
CENTRAL	LARCENY/THEFT	LARCENY/THEFT
INGLESIDE	OTHER OFFENSES	OTHER OFFENSES
MISSION	OTHER OFFENSES	LARCENY/THEFT
NORTHERN	LARCENY/THEFT	LARCENY/THEFT
PARK	LARCENY/THEFT	LARCENY/THEFT
RICHMOND	LARCENY/THEFT	LARCENY/THEFT
SOUTHERN	LARCENY/THEFT	LARCENY/THEFT
TARAVAL	DRUG/NARCOTIC	DRUG/NARCOTIC

Se observa que salvo el distrito de MISSION no cambian las categorías dependiendo de si es día o noche.

#### **Categorías mas comunes segun si es fin de semana o dia de semana**

Fin de semana	Ocurrencias	Dia de semana	Ocurrencias
LARCENY/THEFT	78193	LARCENY/THEFT	96112
OTHER OFFENSES	51086	OTHER OFFENSES	74857
NON-CRIMINAL	40773	NON-CRIMINAL	51138
ASSAULT	35209	ASSAULT	41602
VEHICLE THEFT	24198	DRUG/NARCOTIC	33981
VANDALISM	20949	VEHICLE THEFT	29499
DRUG/NARCOTIC	19938	WARRANTS	25583
WARRANTS	16554	VANDALISM	23631
BURGLARY	15238	BURGLARY	21361
SUSPICIOUS OCC	13082	SUSPICIOUS OCC	18310

Si bien se puede observar que la categoría depende de si es un dia de semana o fin de semana.

#### **Delitos más comunes según si es fin de semana o dia de semana por distrito**

District	Categoría Fin de semana	Categoría Semana
BAYVIEW	OTHER OFFENSES	OTHER OFFENSES
CENTRAL	LARCENY/THEFT	LARCENY/THEFT
INGLESIDE	OTHER OFFENSES	OTHER OFFENSES
MISSION	LARCENY/THEFT	OTHER OFFENSES
NORTHERN	LARCENY/THEFT	LARCENY/THEFT
PARK	LARCENY/THEFT	LARCENY/THEFT
RICHMOND	LARCENY/THEFT	LARCENY/THEFT
SOUTHERN	LARCENY/THEFT	LARCENY/THEFT

TARAVAL	DRUG/NARCOTIC	DRUG/NARCOTIC
TENDERLOIN		

Como con

**Días con mayores ocurrencias:**

2011-01-01	650, 2013-11-01	626, 2006-01-01	613,
2004-04-01	562, 2013-10-04	555, 2012-10-01	548,
2012-01-01	526, 2004-09-01	523, 2003-04-01	523,
2003-01-08	523		

Algo que se puede destacar en como las 10 fechas con mayor ocurrencias se repite en dos ocasiones año nuevo, dándose en primera posición y en segunda posición. También se puede destacar como las restantes ocurren en los primeros días del mes.

**Categorías según fechas especiales:**

Cantidad total de crímenes en Navidad: 1967	Cantidad total de crímenes en Año Nuevo: 1789	Cantidad total de crímenes en el Cuatro de Julio: 347
5 crímenes más comunes en Navidad:	5 crímenes más comunes en Año Nuevo:	5 crímenes más comunes en el Cuatro de Julio:
Category count	Category count	Category count
-LARCENY/THEFT 368	-OTHER OFFENSES 339	-LARCENY/THEFT 57
-NON-CRIMINAL 241	-NON-CRIMINAL 261	-OTHER OFFENSES 46
-OTHER OFFENSES 206	-LARCENY/THEFT 241	-ASSAULT 42
-ASSAULT 201	-ASSAULT 198	-NON-CRIMINAL 34
-VEHICLE THEFT 153	-SUSPICIOUS OCC 122	-SUSPICIOUS OCC 19

Para realizar las estadísticas en fechas especiales se decantó por tres opciones, todas feriados en Estados Unidos: Navidad, Cuatro de Julio y año nuevo.

Cabe destacar que es igual de interesante (e incluso hasta más) ver el comportamiento de la variable Category en los días de acción de gracias, black friday o el labor day. El problema radica en que estas fechas no son fijas, por lo tanto para lograr filtrarlas requiere de una complejidad computacional extra, la cual en el contexto del informe no se puede dar. A pesar de esto, con las tres fechas definidas inicialmente sirve para tener un buen indicador de cómo se comporta.

Fechas que no son días especiales:

Cantidad total de crímenes en otras fechas: 871550

Category count

LARCENY/THEFT	173639
OTHER OFFENSES	125352
NON-CRIMINAL	91375
ASSAULT	76370
DRUG/NARCOTIC	53789

## 5. Preparación de los datos

Lo primero que se va a realizar es quitar los atributos Descripción y Resolución puestos que estos no están en el dataset de prueba.

Se eliminan las filas con coordenadas en el polo norte.

### TRABAJO SOBRE LOS ATRIBUTOS

FECHA: Se va a subdividir en varias partes para un mejor manejo de los datos relacionados a la fecha.

Nuevas clases:

- Dia
- Mes
- Año
- Hora

A su vez se generarán nuevos atributos como lo son noche (Si la hora está entre las 18 hs y las 06 hs) y si es fin de semana (Sábado y Domingo).

Se evalúa eliminar todo registro del 2015.

DIRECCIÓN: Se crea la variable binaria BLOCK, en donde se representa si la dirección contiene un BLOCK.

A destacar que estos nuevos atributos se verán en la evaluación y en el modelado como afectan el modelo.

## 6. Modelado

Al ser un problema de clasificación nos limitamos a ciertos algoritmos. Es decir algoritmos pertinentes a este tipo de problemas. El segundo problema surge de que se trata de una clasificación no binaria. Al ser no binaria muchos de los algoritmos quedan descartados puesto que estos solo sirven para la clasificación binaria. Para este caso se utilizarán los siguientes algoritmos:

- Árboles de Decisión
- RandomForest
- Naive Bayes
- AdaBoost

Otro problema propio de los algoritmos en sí es debido al tamaño del dataset en cuestión. Para solucionar esto y así obtener un mejor uso del tiempo en ajustes y

demás, se realizará un prototipo en Rapidminer con una muestra pequeña pero significativa del dataset. Tras esto y de probar diferentes ajustes se realizarán los modelos sobre el dataset completo.

## PROTOTIPO Y PRUEBA DE CONFIGURACIONES

En esta parte se crearon cuatro dataset diferentes con características distintas para trabajar y cubrir las diferentes configuraciones sobre los datos discutidas en los puntos anteriores. A cada uno de los mismos se les generó una muestra significativa para disminuir la cantidad de cómputo y a su vez agilizar la prueba de distintos parámetros a los algoritmos. (El proceso para ver cómo se generaron los samples se encuentra disponible en los archivos de rapidminer como: GENERAR\_SAMPLES\_DATASET).

Los dataset en cuestión son:

- sample\_procesed\_train\_1: Sample del dataset procesed\_train\_1 el cual cuenta con todas las clases generadas en la preparación de datos (Year, Month, Day, Hour, isNight, Weekend, Block)
- sample\_procesed\_train\_2: Sample del dataset procesed\_train\_2 en el cual se poseen todas las clases creadas mencionadas en preparación de los datos menos BLOCK.
- sample\_procesed\_train\_1\_encoded: Sample del dataset con los atributos encoded tal como lo utiliza python para aplicar los modelos).
- sample\_train: Sample del dataset sin modificar.

A cada uno de estos se le aplicó un cross validation para cada algoritmo. Con los siguientes parámetros: (Se puede ver el proceso en el archivo RapidMiner PROTOTIPOS\_SAMPLES)

**Los parámetros fueron los siguientes:**

- **Cross Validation:**

Parámetros por defecto

Seed: 1992

Algoritmos:

- **ARBOL DE DECISION:**

criterion: gini\_index

El resto por defecto

- **RANDOM FOREST:**

numero de arboles: 10

criterion: gini\_index

max\_depth: 10

El resto por defecto

- **NAIVE BAYES:**

Laplace Correction: activada

- **ADA BOOST:**

Iterations: 10

Algoritmo utilizado: Decision Tree (Esto debido a que su contraparte en python utiliza una versión de Decision tree).

Tras correr este proceso obtenemos que la mejor performance está en el dataset sample\_procesed\_2, sample\_procesed\_1. La performance del sample\_train fue bastante baja por lo cual se descarta trabajar con el dataset sin modificar. (Ver archivo excel adjunto resultados\_prototipos.xlsx).

Otra configuración que se probó como prototipo fue cambiar el criterio del árbol de decisión y random forest a information gain (entropy).

Esto nos dio como resultado una bajada de 5 puntos en performance.

Si se utiliza un criterio de gain\_ratio la performance aumenta con respecto a gini\_index.

DATOS DE 2015: Se realizan dos mediciones utilizando como benchmark los árboles de decisión. El resultado que se obtiene es que al sacar las filas del 2015 el modelo disminuye un 1% su performance.

En resumen se optó por trabajar con la siguiente configuración:

- Dataset: sample\_procesed\_2 (Es el que nos dio la mayor performance)
- Filas del año 2015: Se deja (Sin él baja el rendimiento y en el entorno que nos encontramos un 1% hace la diferencia)
- Configuración de parámetros en RapidMiner (Luego en python se homologara): se utilizaran los mismos que anteriormente. El algoritmo elegido para los AD es gain\_ratio (es el que tiene mas performance)

## PROCESO UTILIZADO

Tanto en python como en RapidMiner lo que se realiza es un cross validation del modelo para luego obtener su accuracy, su error y su matriz de confusión y las estadísticas de la variable objetivo (precision recall f1-score support).

## MODELO EN PYTHON:

### ÁRBOLES DE DECISIÓN:

Parámetros: `parameters_dt = {"criterion": ["gini", "entropy"], "splitter": ["best", "random"], "max_features": [None, "sqrt", "log2"], "max_depth": [10], "min_samples_leaf": [2]}`

Precisión: 0.2655554984554419 aprox 26,5%

RECALL y PRECISIÓN DE CLASE: Muchos 0, las únicas clases que tienen recall son: OTHER OFFENSES 0.31 0.30, LARCENY/THEFT 0.28 0.73, DRUG/NARCOTIC 0.30 0.38, ASSAULT 0.16 0.20

## **RANDOM FOREST:**

Parámetros: `parameters_rf = {"n_estimators": [10], "criterion": ["gini", "entropy"], "max_features": [None, "sqrt", "log2"], "max_depth": [10], "min_samples_split": [5]}`

Precisión: 0.27356664439762235 aprox 27,3%

RECALL y PRECISIÓN DE CLASE: Las que poseen mayor son bastantes bajas las que poseen más son: ASSAULT 0.16 0.21, DRUG/NARCOTIC 0.30 0.41, LARCENY/THEFT 0.28 0.76, MISSING PERSON 0.56 0.17

## **NAIVE BAYES:**

Parámetros: Defecto

Precisión: 0.18854457520370466 aprox un 18,8%

RECALL y PRECISIÓN DE CLASE: muy mal recall y precision a destacar: LARCENY/THEFT 0.23 0.55, OTHER OFFENSES 0.25 0.37

## **ADA BOOST:**

Parámetros: `parameters_ada = {"n_estimators": [10]}`

Precisión: 0.23210625189144127 aprox un 23,2%

RECALL y PRECISIÓN DE CLASE: Malos resultados tanto en recall como class precisión solo a destacar OTHER OFFENSES 0.25 0.37 0.30

(Se pueden observar las diferentes métricas en el archivo parcial2.ipynb)

## **MODELO EN RAPIDMINER:**

Los parámetros de los algoritmos son los definidos previamente en el prototipado.

## **ÁRBOLES DE DECISIÓN:**

Precisión: 21.96% +/- 0.80% (micro average: 21.96%)

logistic\_loss: 0.640 +/- 0.002 (micro average: 0.640)

RECALL y PRECISIÓN DE CLASE: Tiene un recall en varias clases

## **RANDOM FOREST:**

Precisión: 23.22% +/- 0.43% (micro average: 23.22%)

logistic\_loss: 0.639 +/- 0.001 (micro average: 0.639)

RECALL y PRECISIÓN DE CLASE: El recall disminuye a medida de que hay menos

### **NAIVE BAYES:**

Precisión: 28.37% +/- 0.10% (micro average: 28.37%)

logistic\_loss: 0.602 +/- 0.000 (micro average: 0.602)

RECALL y PRECISIÓN DE CLASE: Tanto el recall como la precisión está bien distribuida a pesar de ser baja. Todas tienen un recall

### **ADA BOOST:**

Precisión: 4.81% +/- 0.00% (micro average: 4.81%)

logistic\_loss: 0.680 +/- 0.000 (micro average: 0.680)

RECALL y PRECISIÓN DE CLASE: Predijo solo una clase (warrants) con el 100%. Warrants es la única que tiene una precisión. esta es el 4.81% (total de warrants en el dataset).

(Para ver los resultados completos de los modelos se puede revisar el archivo adjunto: resultadosModelosRM)

## FEATURE SELECTION:

Se realizarán dos diferentes técnicas de feature selection:

- Forward Selection
- Backward Elimination

Se aplicarán estos dos diferentes algoritmos heurísticos de optimización con el fin de poder determinar si aumenta la performance de los distintos algoritmos al aplicar unos de estos métodos. Se aplicará sobre los algoritmos Naive Bayes y Árboles de decisión. Esto pues son los que dieron mejores resultados en RapidMiner.

### FORWARD SELECTION

Parámetros: Configuración por default RapidMiner

### **NAIVE BAYES:**

Precisión: 30.50% +/- 0.16% (micro average: 30.50%)

Logistic\_loss: 0.606 +/- 0.000 (micro average: 0.606)

Class recall y Class Precision:

Bastante distribuida con mucha precision y recall en OTHER OFFENSES y LARCENY/THEFT, el resto tiene aproximadamente un 20%.

Atributos eliminados: PdDistrict, X,Y,Month,isNight,Weekend, Block

Atributos conservados: DayOfWeek, Adress, Year, Day,Hour

### **ÁRBOLES DE DECISIÓN:**

Precisión: 23.47% +/- 0.07% (micro average: 23.47%)

Logistic\_loss: 0.636 +/- 0.000 (micro average: 0.636)

Class recall y Class Precision:

Bastante distribuida con mucha precision y recall en OTHER OFFENSES y ROBERY, el resto tiene aproximadamente menos de un 20% y hay muchos con 0%

Atributos eliminados: DayOfWeek, X,Y,Month,Day,Hour

Atributos conservados: PdDistrict, Adress,isNight,Weekend,Block

(Para ver los resultados completos de los modelos se puede revisar el archivo adjunto: resultadosModelosRM)

(Ver gráficas de pesos en anexo 18,19)

### **BACKWARD ELIMINATION:**

Parámetros: Configuración por default RapidMiner

### **NAIVE BAYES:**

Precisión: 30.48% +/- 0.11% (micro average: 30.48%)

Logistic\_loss: logistic\_loss: 0.606 +/- 0.000 (micro average: 0.606)

Class recall y Class Precisión: Ocurre lo mismo que con foward

Atributos eliminados: DayOfweekm pdDistrict, X,Y, isNight, Weekend, Block

Atributos conservados: Adress, Year, Month, Day, Hour

### **ÁRBOLES DE DECISIÓN:**

Precisión: 23.32% +/- 0.14% (micro average: 23.32%)

Logistic\_loss: 0.636 +/- 0.001 (micro average: 0.636)

Class recall y Class Precisión: Ocurre lo mismo que con foward

Atributos eliminados: DayOfWEEK, Y, Day, Hour

Atributos conservados: PdDistrict, Adress, Year, Month, IsNight, Block

(Para ver los resultados completos de los modelos se puede revisar el archivo adjunto: resultadosModelosRM)

(Ver gráficas de pesos en anexo 20,21)

## POSIBLES MEJORAS AL MODELO

Menos clases mejor análisis separar contenido probar otros feature selection utilizar PCA etc mejorar parámetros de evaluación.

Para aumentar el rendimiento del modelo se podrían tomar varias acciones sobre el mismo.

Primero continuar con el análisis exhaustivo de los datos para esto se podría integrar otras técnicas como PCA, probar diferentes enfoques de atributos en los atributos. Por ejemplo quitar Hora, separar Distritos en menos clases.

El segundo es reducir el número de clases de la variable objetivo. Actualmente se poseen 31 diferentes clases, lo cual es bastante. Se proponen diferentes formas de agruparlas. Por ejemplo se podría agrupar por la gravedad de los crímenes o por si el crimen computa en cárcel o no (se podría utilizar el atributo de resolución como guía).

El tercero es utilizar otros algoritmos de Feature Selection como Evolutionary entre otros.

## 7. Evaluación

### COMPARACIÓN DE PERFORMANCE ENTRE LOS MODELOS

Se descarta ADAboost, puesto que en RapidMiner no se obtuvo un resultado coherente y para nada confiable. Esto debido a que solo predice una clase. Se presume que este error puede ser debido a errores de configuración y/o Rapidminer al ser de alto nivel cubre muchas configuraciones que al no poder conocerlas termina ocurriendo esta baja performance.

Cabe destacar que ADAboost en Python nos da una performance de 23.3%. A pesar de esto la evaluación se va a continuar con los tres algoritmos restantes debido a la baja performance en rm.

Algoritmo	Performance python	Performance rm
Random Forest	27,3%	23.22%

Decision Tree	26,5%	21.96%
Naive Bayes	18,8%	28.37%

Lo primero que se puede destacar es que en ambos se posee la misma tendencia de que los random forest son mejores que los árboles de decisión, esto por aproximadamente un 2%. Lo segundo es notar que si bien existen diferencias entre plataformas, estas no son abismales en relación de una con la otra, todas se encuentran en un rango de lo esperable (posiblemente debido a la implementación interna que ambos realizan). El gran cambio está en Naive Bayes, donde nos da un 10% de diferencia. Una posible explicación a esto es que en Python las variables tienen que ser encoded para poder ser procesadas por el modelo, algo que en rm no es necesario. Por lo tanto esto puede afectar. Esto se puede confirmar si vemos en los prototipos naive bayes con el dataset encoded se obtenía resultados alrededor del 15 - 17%.

### **CLASS RECALL, CLASS PRECISION:**

#### **CONCLUSIONES FEATURE SELECTION**

Se observa que en las dos se obtiene mejoras de rendimiento

	Forward	Backward
Decision Tree	23.47%	23.32%
Naive Bayes	30.50%	30.48%

Primero se puede ver que Decision Tree en ambos no tuvo una mejora, solo con un 0,10% en el mejor caso. Por el otro lado, naive bayes logró llegar a un 30.50% y 30.48% en forward y backward respectivamente. En todo caso Forward Selection se puede decir que se obtiene una mejor performance.

Por el lado del peso de los atributos se observa que en 3 de los 4 casos se elimina DayOfWeek, mismo con la X e Y (este se elimina siempre). Algo a destacar es que también se conservan las clases creadas entre ellas: IsNight, Block, Weekend, en la mayoría de los casos.

Sobre la class recall y las clases tienen una buena distribución donde todas tienen más de 0% es decir que aunque sea mínimo poseen un recall y una precisión para todos y si bien hay clases que poseen más recall como: WARRANTS, OTHER OFFENSES, ARDENY/THEFT, VEHICLE THEFT, VANDALISM, NON-CRIMINAL.

## CONCLUSIONES EVALUACIÓN

Se puede decir que basándonos en el análisis previo de la evaluación de los modelos, el mejor es: Naive Bayes pues es el que tiene la mejor performance y el recall entre las clases está más distribuido, donde si bien hay algunos que la estadísticas de confianza no son los mejores, casi todas las clases tienen estos defectos. Pero no es tan evidente como en Random Forest y Árboles de decisión.

Con un porcentaje de performance de 30.50% (en caso de utilizar Forward Selection) nos da una estadística de predecir 1 de cada 3 casos de pruebas aproximadamente. Esto si bien parece poco, es un aumento de 12.2 veces más de que si elegimos una categoría al azar ( 100 / 39 categorías) la cual es 2,5%.

Sin embargo sigue siendo una performance baja dentro de todo, tener un caso de 3 es desfavorable, y más si se utiliza como guía para la seguridad. Como se dijo antes en la comprensión del negocio: es muy difícil de predecir el crimen porque es difícil encontrar una forma de modelar. Si bien hay indicadores, los cambios sociales y económicos juegan el gran papel decisivo a la hora de tomarlos en cuenta.

Se puede hacer un análisis de los datos y establecer modelos que ayuden a generar acciones en conjunto con una analítica de datos mejor de la que se realizó en el informe, pero no se puede usar como algoritmo definitivo para predecir a futuro.

## 8. Deploy

El objetivo de este informe no es generar una herramienta de deploy de este modelo, puesto que este enfoque lleva más tiempo y una investigación paralela al modelo en sí. Sin embargo, se pueden proponer diferentes enfoques para un eventual deploy. El primero es generar aplicaciones de compañía para los diferentes agentes que patrullan por San Francisco. Otra cosa que se puede hacer es establecer períodos de fechas para generar estadísticas predictivas y establecer planes de acción. Se destaca que siempre estas formas de deploy van acompañadas de una mirada crítica por parte de un humano y no debe tomarse como una solución absoluta.

Estas no son más que meras guías generales y bosquejos rápidos para tomar como inspiración para un posible deploy. Se vuelve a destacar la necesidad de profundizar en estas ideas y delegar a un equipo encargado de desarrollar un buen soporte para el posible deploy del modelo entrenado.

## 9. Referencias

(n.d.). San Francisco Police Department. Retrieved December 3, 2023, from

<https://www.sanfranciscopolice.org/>

V. (2023, June 16). YouTube. Retrieved December 3, 2023, from  
<https://twitter.com/SFPD/status/622106119742173184/photo/1>

*Factores de riesgo que provocan la criminalidad.* (n.d.). Academia Mexicana de Ciencias. Retrieved December 3, 2023, from  
[https://www.amc.edu.mx/revistaciencia/images/revista/68\\_4/PDF/68\\_4\\_factores\\_riesgo.pdf](https://www.amc.edu.mx/revistaciencia/images/revista/68_4/PDF/68_4_factores_riesgo.pdf)

*Geopolitica.* (n.d.). Wikipedia. Retrieved December 3, 2023, from  
<https://en.wikipedia.org/wiki/Geopolitica>

*Predictive policing.* (n.d.). Wikipedia. Retrieved December 3, 2023, from  
[https://en.wikipedia.org/wiki/Predictive\\_policing](https://en.wikipedia.org/wiki/Predictive_policing)

*Predictive policing in the United States.* (n.d.). Wikipedia. Retrieved December 3, 2023, from  
[https://en.wikipedia.org/wiki/Predictive\\_policing\\_in\\_the\\_United\\_States](https://en.wikipedia.org/wiki/Predictive_policing_in_the_United_States)

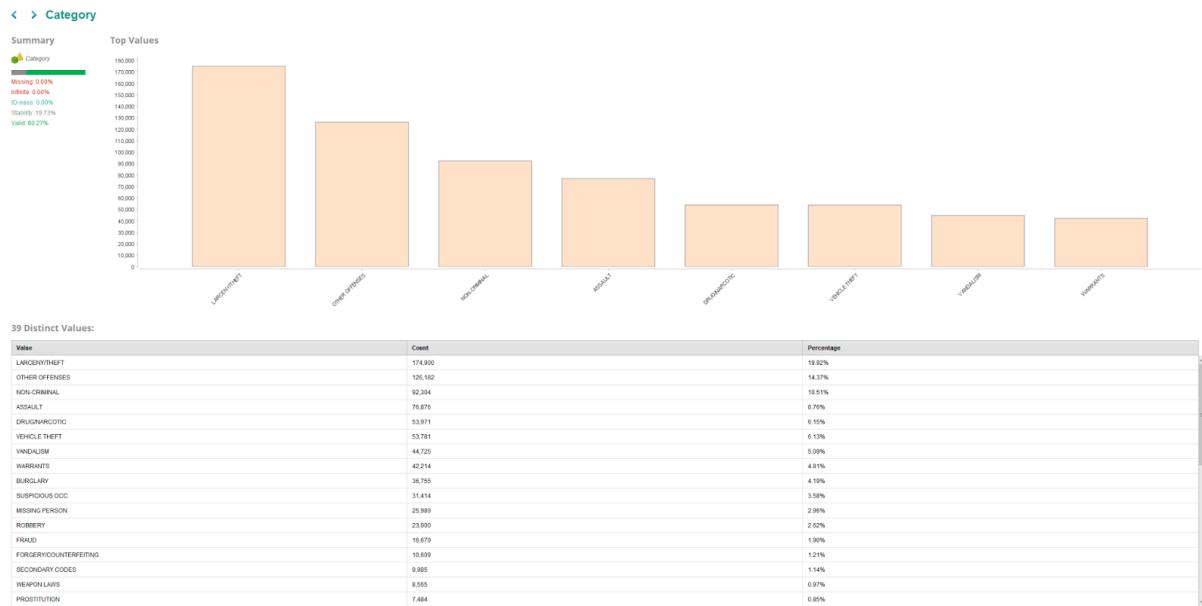
*San Francisco (California).* (n.d.). Wikipedia. Retrieved December 3, 2023, from  
[https://es.wikipedia.org/wiki/San\\_Francisco\\_\(California\)](https://es.wikipedia.org/wiki/San_Francisco_(California))

*San Francisco (California).* (n.d.). Wikipedia. Retrieved December 3, 2023, from  
[https://es.wikipedia.org/wiki/San\\_Francisco\\_\(California\)](https://es.wikipedia.org/wiki/San_Francisco_(California))

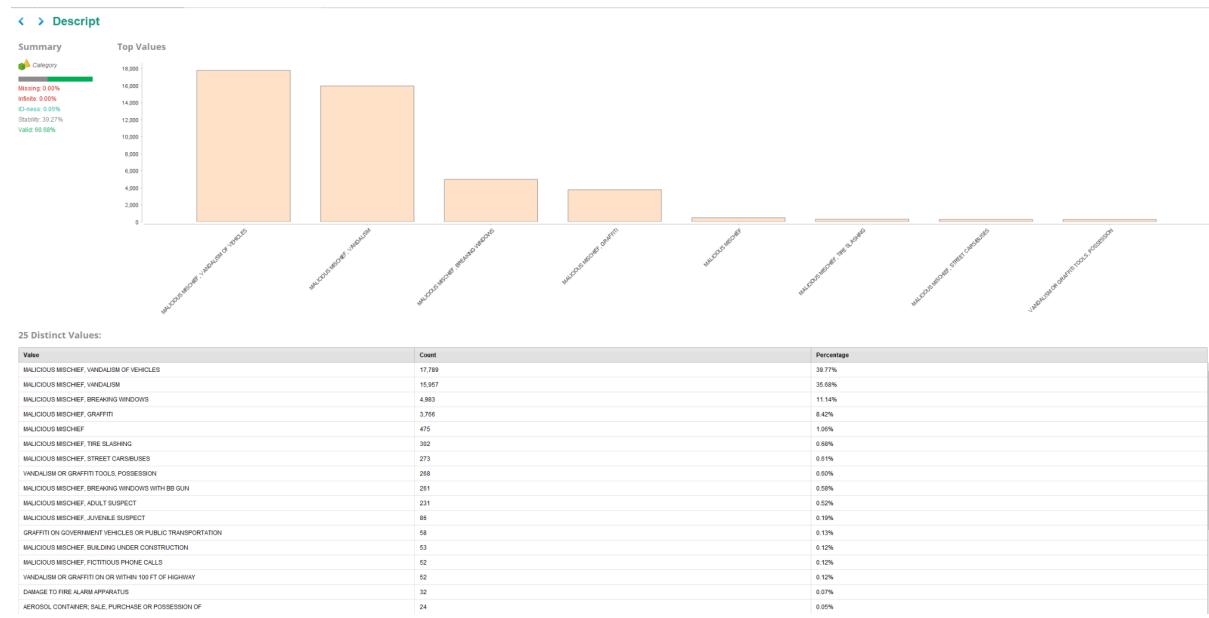
## 10. Anexos

Name	Type	Missing	Statistics		
▼ Dates	Date-time	0	Earliest date Jan 6, 2003 12:01 AM	Latest date May 13, 2015 11:53 PM	Duration 4510d 23h 52m 0s
▼ Category	Nominal	0	Least TREA (6)	Most LARCENY/THEFT (174...)	Values LARCENY/THEFT (174...), OTHER OFFENSES (126182), ...[3 more]
▼ Descript	Nominal	0	Least WEARING [...] CEIVE (1)	Most GRAND TH [...] O (60022)	Values GRAND TH [...] CKED AUTO (60022), LOST PROPERTY (31729), ...[877 more]
▼ DayOfWeek	Nominal	0	Least Sunday (116707)	Most Friday (133734)	Values Friday (133734), Wednesday (129211), ...[5 more]
▼ PdDistrict	Nominal	0	Least RICHMOND (45209)	Most SOUTHERN (157182)	Values SOUTHERN (157182), MISSION (119908), ...[8 more]
▼ Resolution	Nominal	0	Least PROSECUT [...] ENSE (...)	Most NONE (526790)	Values NONE (526790), ARREST, BOOKED (206403), ...[15 more]
▼ Address	Nominal	0	Least YUKON ST / 19TH ST (1)	Most 800 Bloc [...] T (26533)	Values 800 Block of BRYANT ST (26533), 800 Block of MARKET ST (6581), ...[23226 more]
▼ X	Real	0	Min -122.514	Max -120.500	Average -122.423
▼ Y	Real	0	Min 37.708	Max 90	Average 37.771

Anexo 1 - Estadísticas de RapidMiner. Se observa como no hay datos faltantes en todo el dataset.



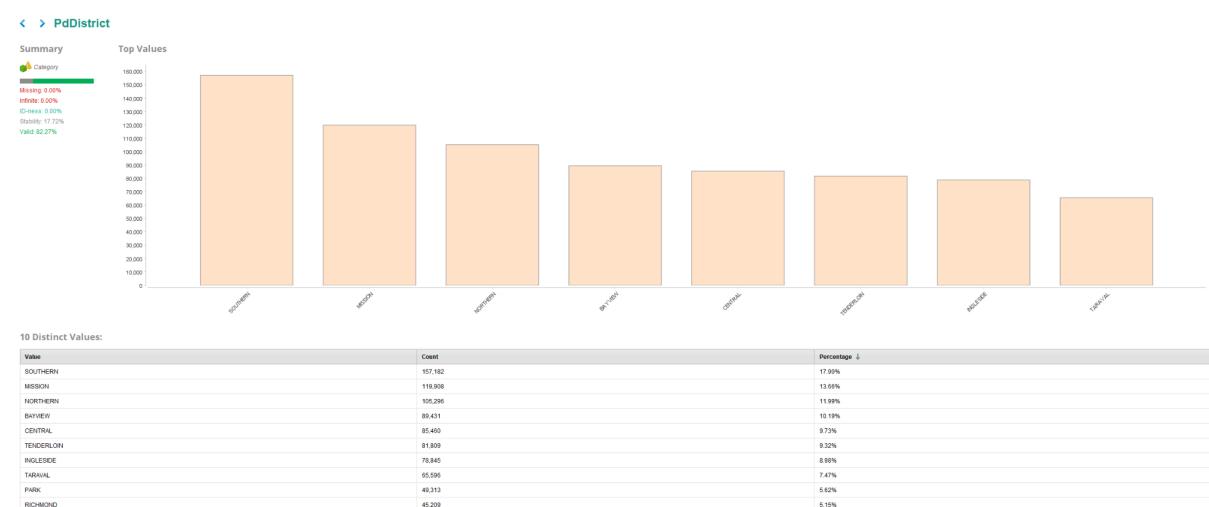
Anexo 2 - Estadísticas Categoría



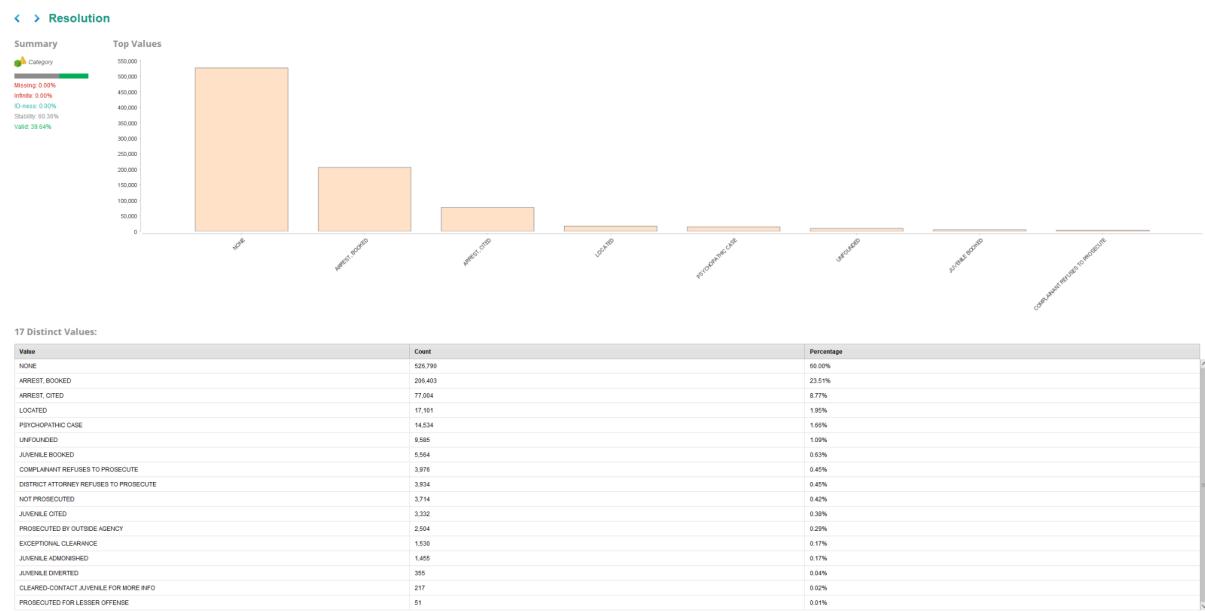
Anexo 3 - Estadísticas Descripción



Anexo 4 - Estadísticas Dia de la semana



Anexo 5 - Estadísticas Distrito Policial



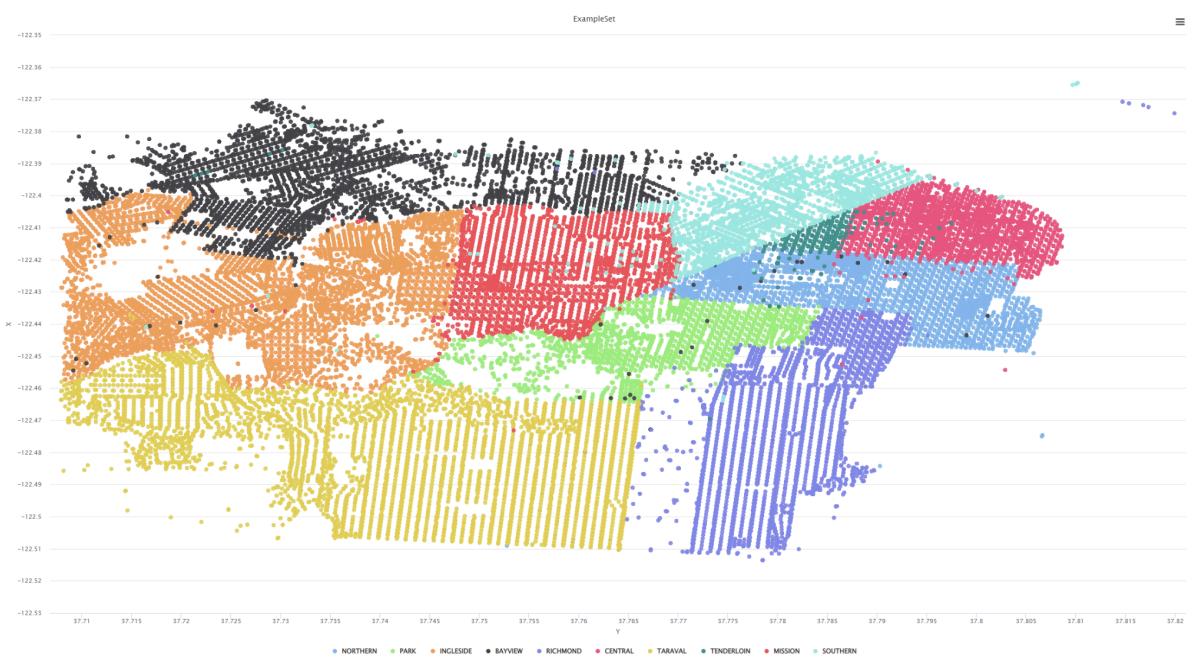
Anexo 6 - Estadísticas Resolución



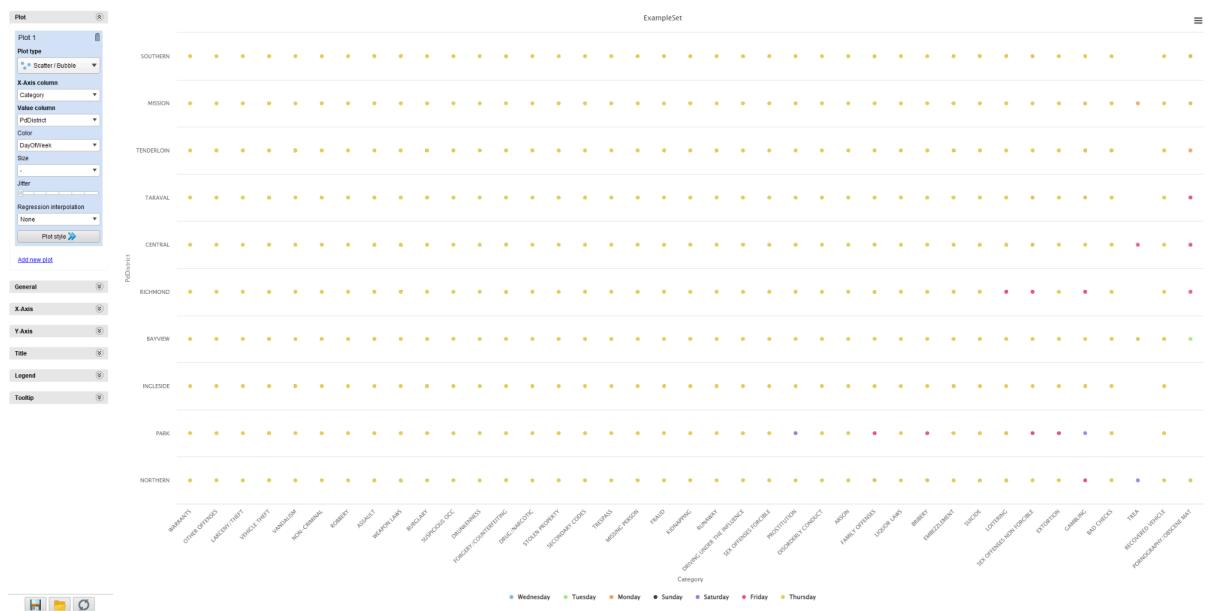
Anexo 7 - Categorías de los crímenes con resolución igual a None



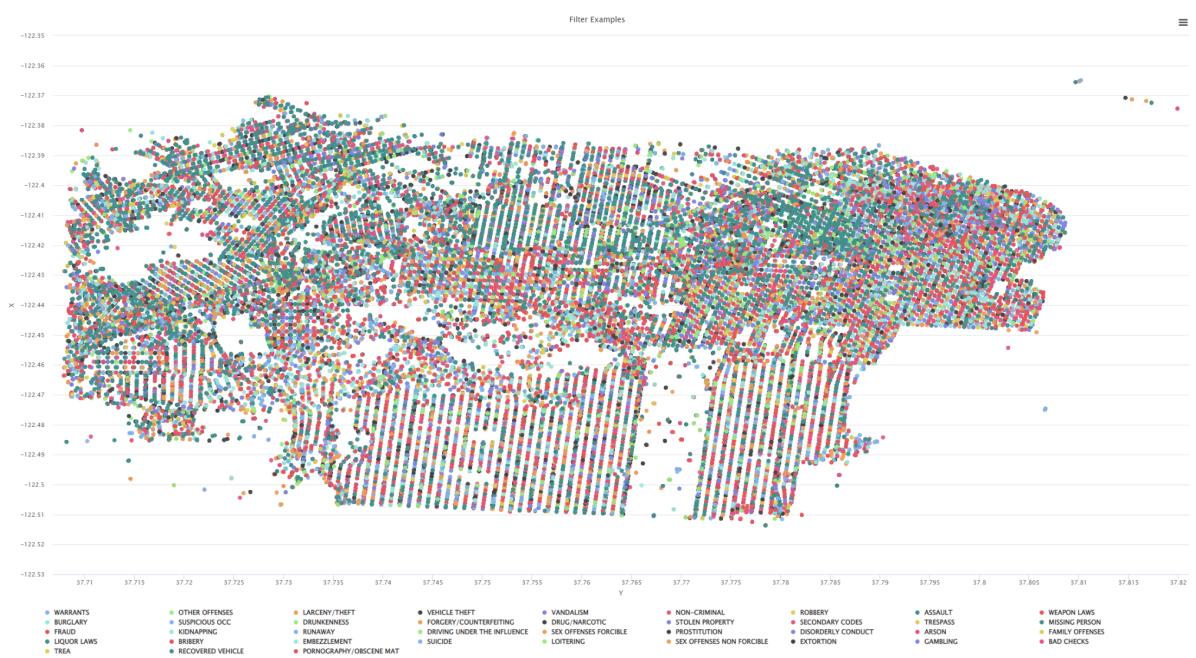
Anexo 8 - Gráfica de puntos X en función de Y con Y con color



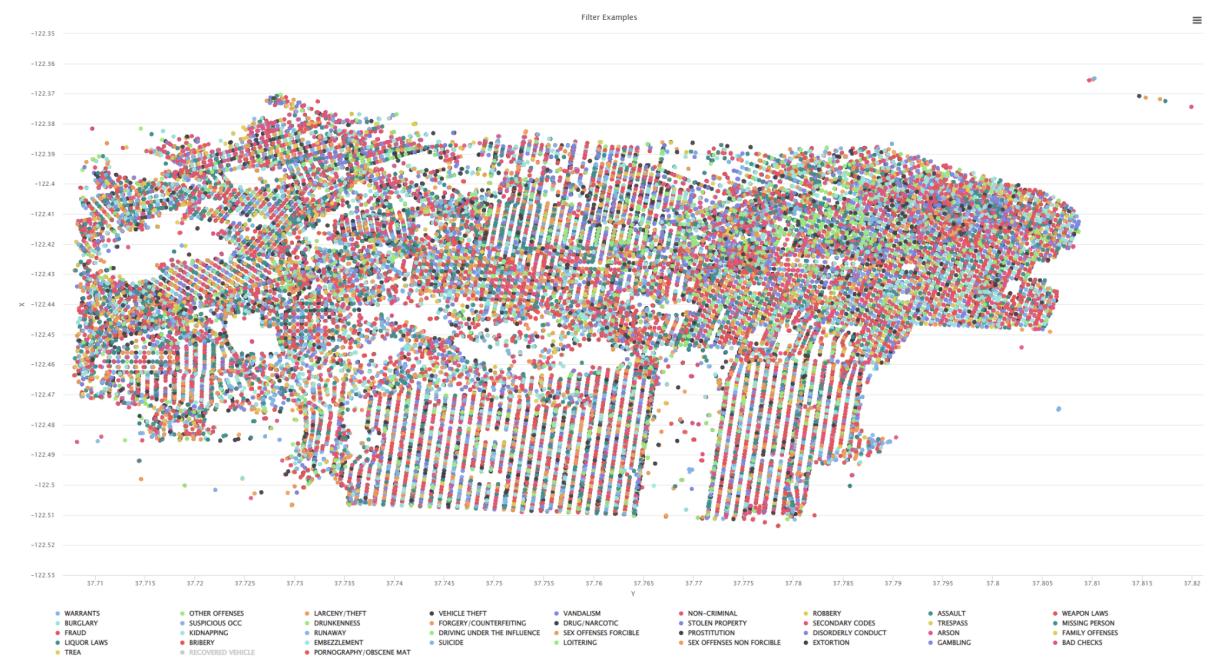
Anexo 9 - Gráfica de puntos Y en función de X con Distrito como color



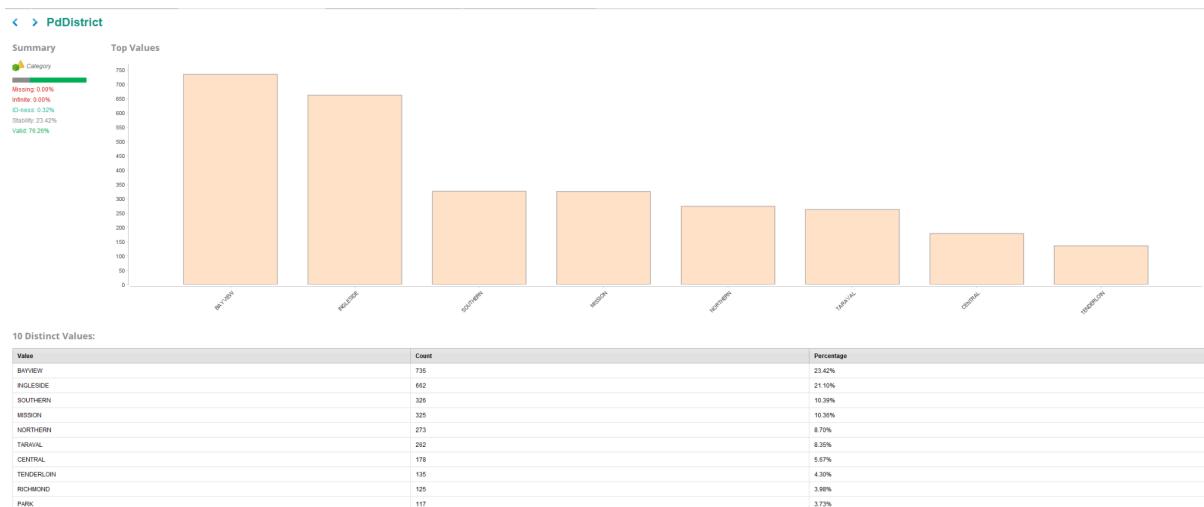
Anexo 10 - Gráfica de puntos clasificación en función de distrito con color en dia de la semana



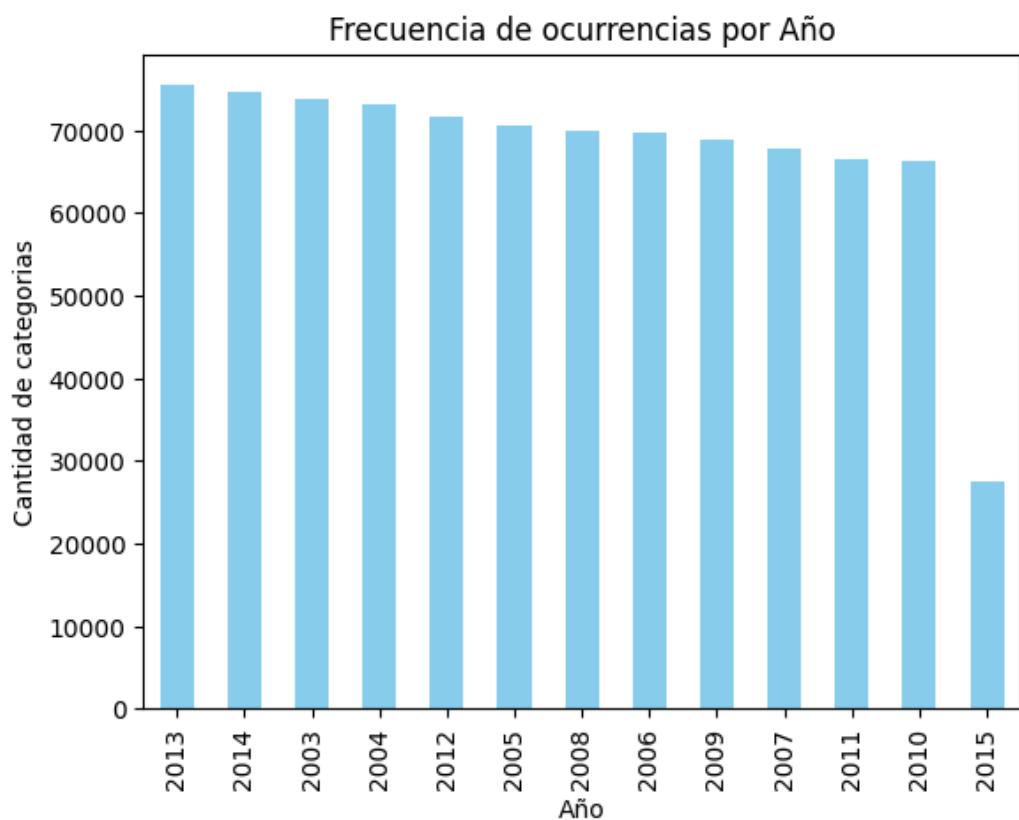
Anexo 11 - Gráfica de puntos X en función de Y distrito con color categoría



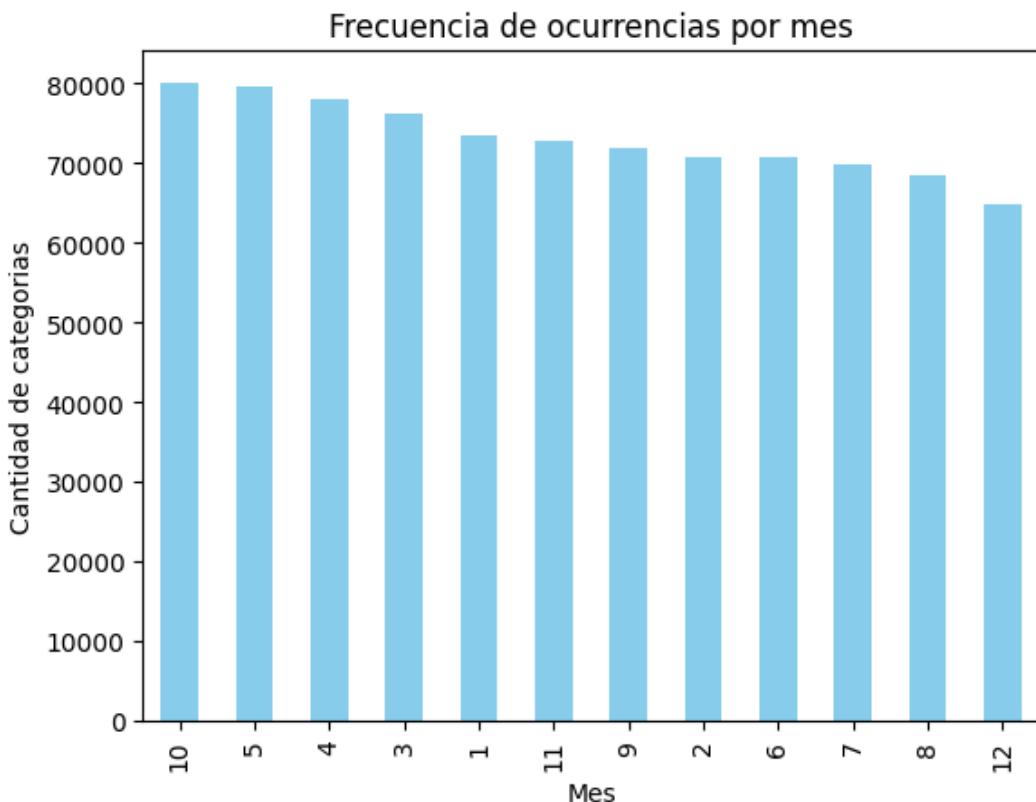
Anexo 12 - Gráfica de puntos X en función de Y distrito con color categoría con la variable Recovered Vehicles desactivada (el color verde petróleo disminuye)



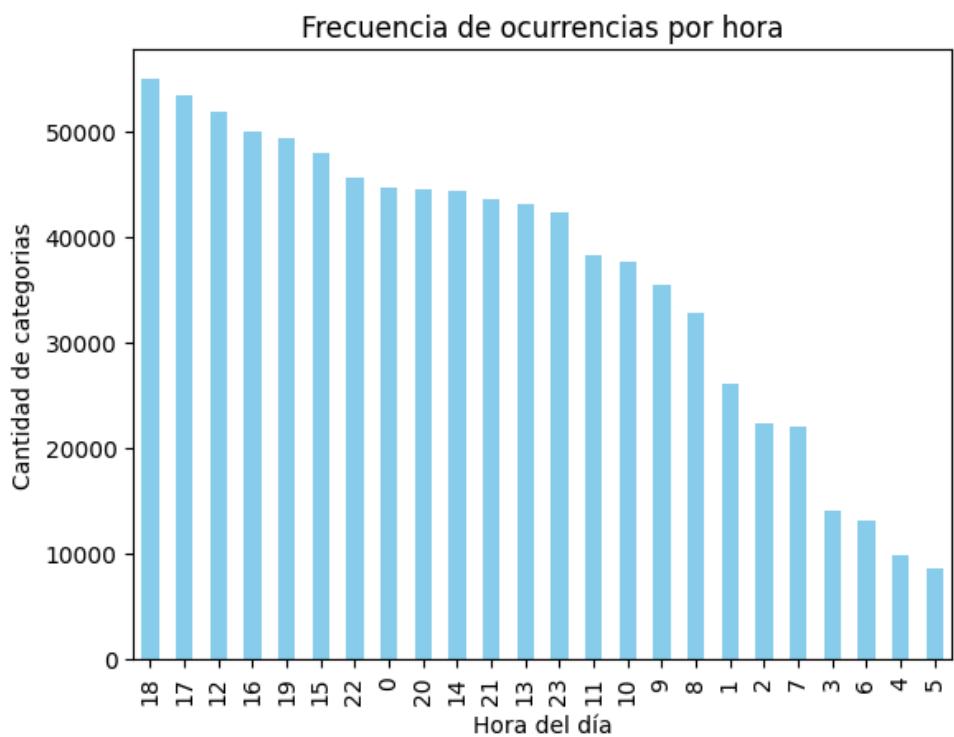
Anexo 13 - Gráfica de los distritos en función de Recovered Vehicles



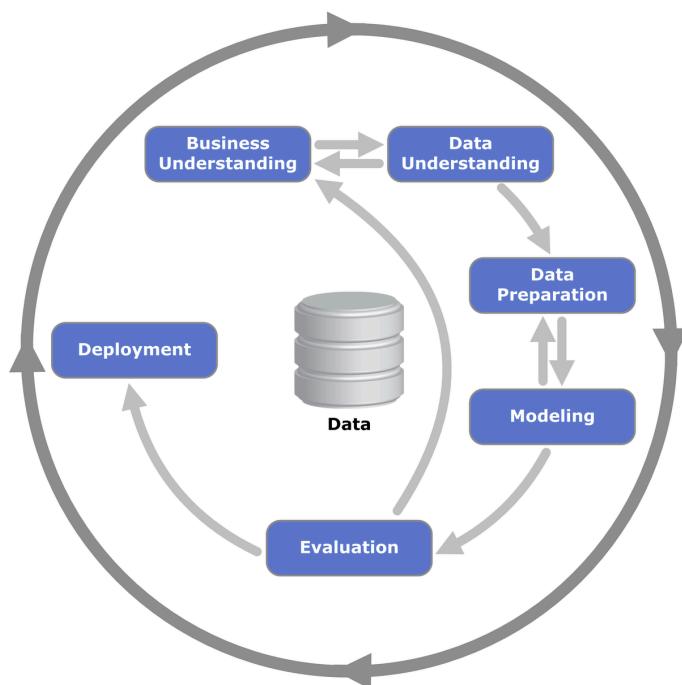
Anexo 14 - Gráfica cantidad de categorías por año



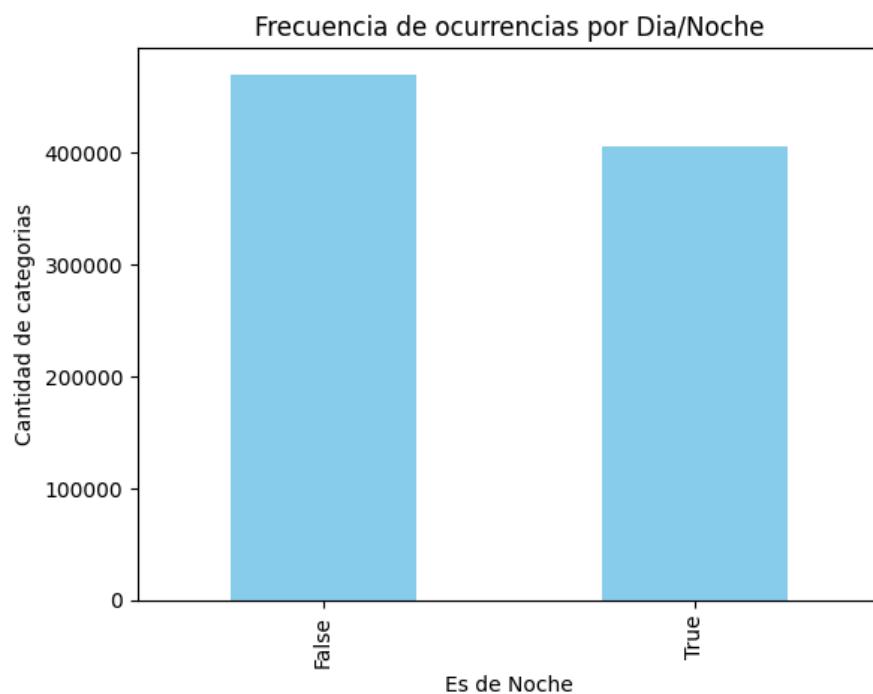
Anexo 15 - Gráfica cantidad de categorías por mes



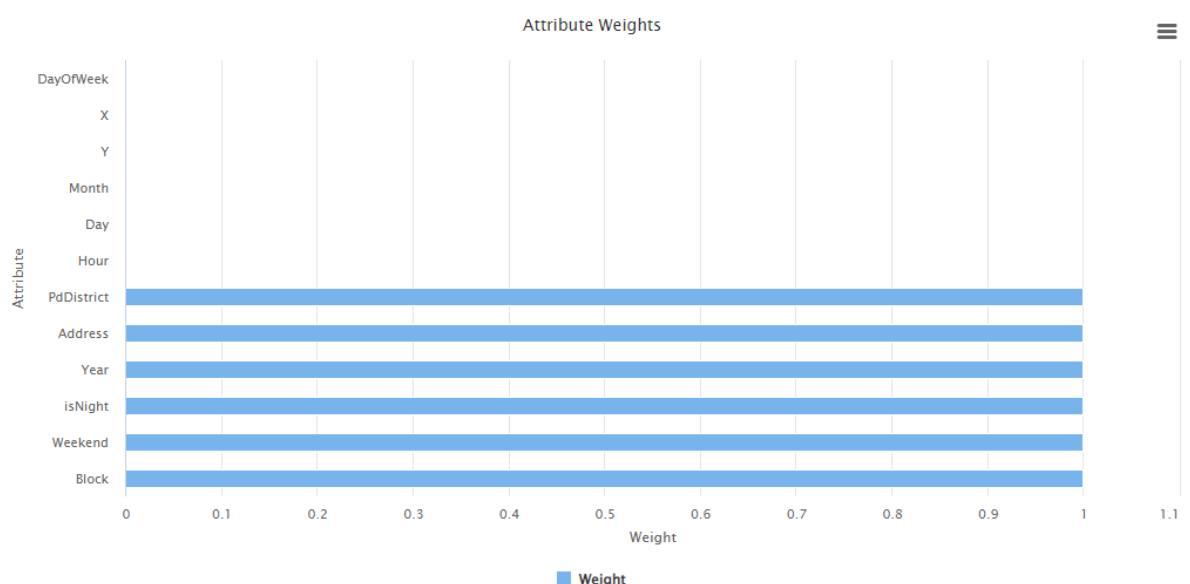
Anexo 16 - Cantidad de ocurrencias de categoría por hora



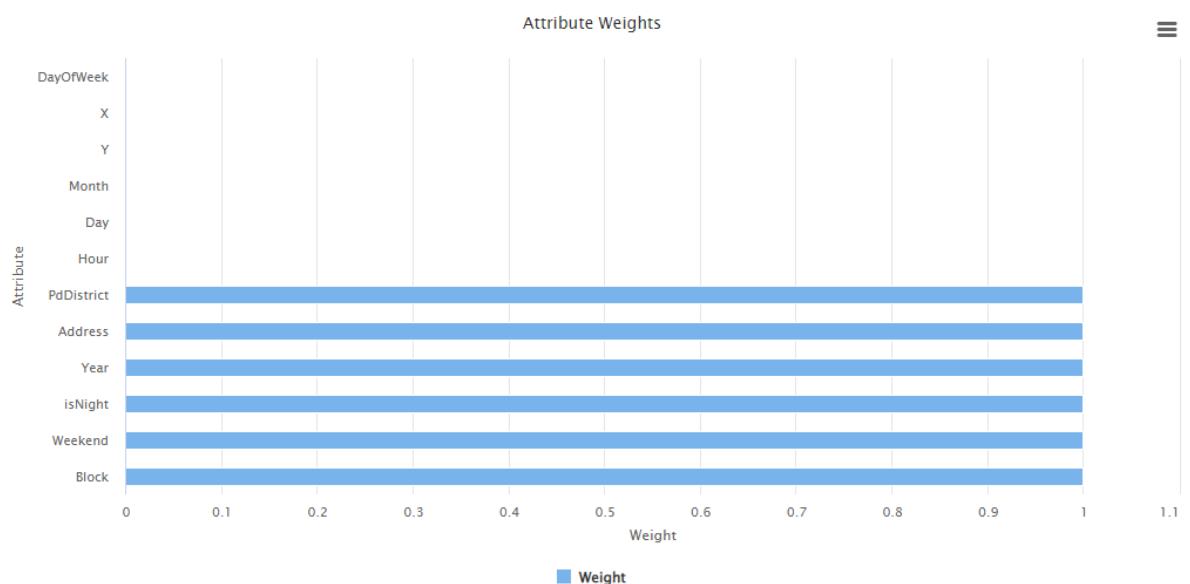
Anexo 17 - Proceso CRISP-DM



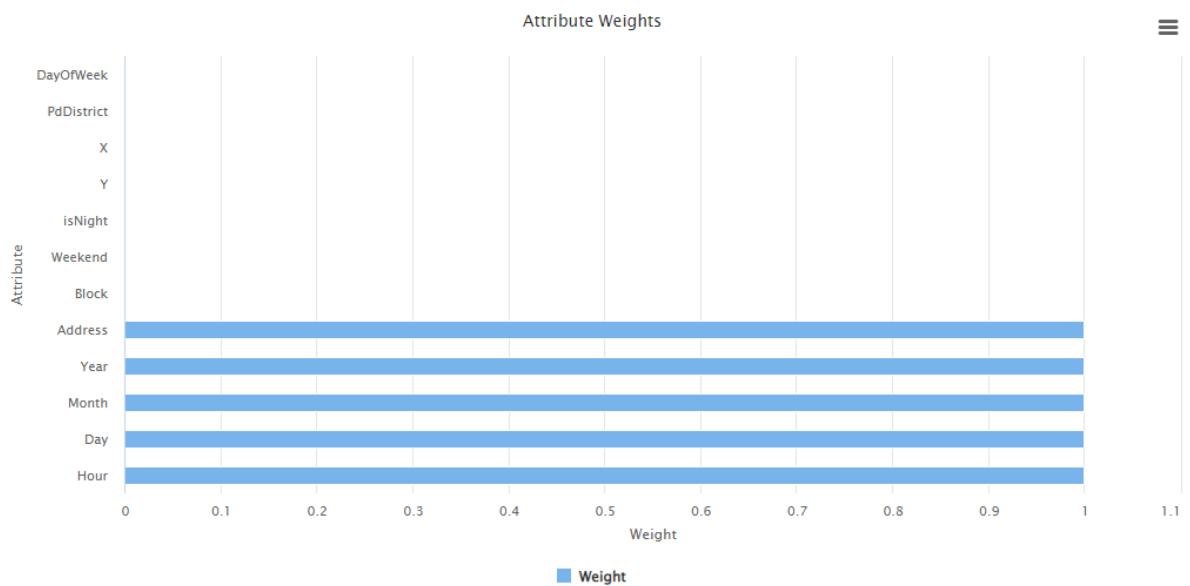
Anexo 18 - Gráfica recurrencia DIA/NOCHE



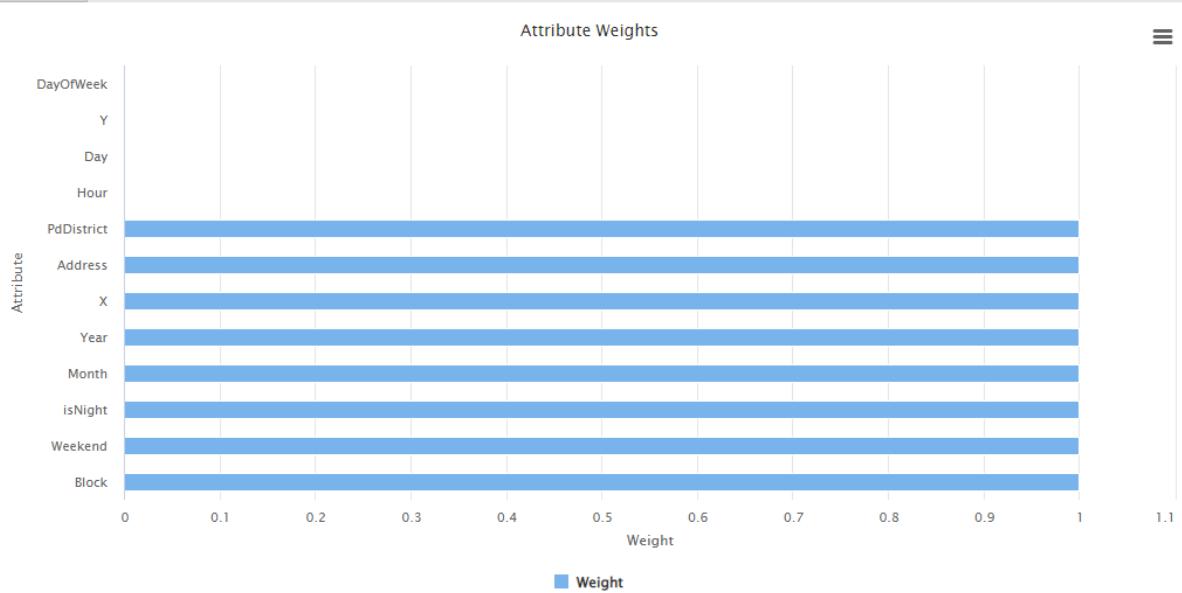
Anexo 19 - Gráfica pesos atributos Forward Selection Naive Bayes



Anexo 20 - Gráfica pesos atributos Forward Selection Árboles de decisión



Anexo 21 - Gráfica pesos atributos Backward Elimination NB



Anexo 22 - Gráfica pesos atributos Backward Elimination AD