Análisis de las víctimas reportadas por la Policía Nacional Civil en Guatemala durante el año 2017 usando varios modelos supervisados de clasificación de Machine Learning

Nathalia Morales
Universidad Francisco Marroquín
Ciudad de Guatemala, Guatemala
nmoralesrojas@ufm.edu

Abstracto

Este informe consta del análisis hecho a la base de datos de crímenes reportados por la PNC en Guatemala durante el año 2017. Los modelos de *Machine Learning* utilizados fueron modelos supervisados de clasificación con el principal propósito de encontrar un subconjunto y modelo con la habilidad de poder predecir el grupo de delito cometido reportado a la policía. Existió una modificación en cuanto a los grupos de delitos a considerar, los cuales no son los que originalmente fueron creados por la PNC. Así también se creó la subdivisión de la data en un subconjunto específico al departamento de Guatemala para observar si el desempeño de los modelos seleccionados incrementaba conforme a un área específica. El informe toma en cuenta la data solamente del año 2017 y cuenta con 35,169 registros después de la depuración de valores nulos, inconclusos o repetitivos. Es la espera del autor que esta investigación pueda ser usada como base para un análisis más profundo del tema en un futuro.

Palabras clave

Hurto, Robo, Delictivo, PNC, LDA,
QDA, Random Forest, INE, dataset,
crímenes, algoritmo de clasificación,
Machine Learning, algoritmo, modelo.

1 Introducción

Es de conocimiento común la alta tasa de criminalidad dentro de Guatemala, solamente en el 2017 se reportaron 35,169 crímenes a la PNC. Dentro de estos están los robos a mano armada, crímenes violentos y crímenes de conducta bajo la influencia de alcohol. Durante el 2019 varios departamentos de Guatemala incluyendo la ciudad de Guatemala, Escuintla, Chiquimula, Quetzaltenango, Izabal y Petén, fueron clasificados como nivel 3 ante la embajada estadounidense para los viajeros de esa nacionalidad. Las razones comunes para estos comportamientos son normalmente atribuidos por la falta de educación, trabajo, entre otros, existentes en el país que conllevan a actos violentos que atentan contra la persona o contra la propiedad privada. Con el propósito de encontrar una manera de anticipar estos comportamientos se ha creado este análisis.

La violencia conforma una de las mayores características por las cuales el país es reconocido mundialmente sin embargo, con la ayuda de análisis estadísticos se podrían anticipar características específicas de las personas más inclinadas a la violencia para poder optar por programas de rehabilitación, inclusión social y responsabilidad ciudadana con el propósito de prevenir a esas personas en mayor peligro y eliminarlas como posibles delincuentes de la ley. Mientras que en este reporte no se ha creado un plan específico para tales implementaciones, el propósito de este si ha de ser su uso para el futuro análisis de las personas que ya se encuentran dentro del sistema penitenciario o preventivo con el objetivo de prevención, más que corrección.

Las principales preguntas por tomar en cuenta en cuanto al procesamiento y análisis de la data con los modelos fueron las siguientes:

Las variables del mes, día del mes, día de la semana, grupo de hora, área geográfica, departamento, municipio, sexo, grupo de edad o si la persona es mayor o menor de edad tiene relación con el grupo de delito cometido. De tal forma que ya sean las variables individualmente o en un subconjunto de sus combinaciones tuvieran una influencia positiva en la predicción de la variable objetiva: el grupo del delito cometido.

Así también se creó una lista con las diferentes variables a predecir además de la variable clave, tales como, pero no limitando a la predicción del sexo de persona, grupo de edad, si la persona es mayor o menor de edad y el departamento en el que se cometió el crimen.

2 Adquisición de la data

La adquisición de la data provino directamente del el Instituto Nacional de Estadística de Guatemala. Su descripción oficial es:
"El Instituto Nacional de Estadística -INE-, es un organismo descentralizado del Estado, semiautónomo, con personalidad jurídica, patrimonio propio y plena capacidad para adquirir derechos y contraer obligaciones, cuyo principal fin es ejecutar la política estadística nacional.

El Instituto tiene dentro de sus principales funciones: recolectar, elaborar y publicar estadísticas oficiales, impulsar el Sistema Estadístico Nacional -SEN-, coordinar con otras entidades la realización de

investigaciones, encuestas generales y especiales, promover la capacitación y asistencia técnica en materia estadística, impulsar la aplicación uniforme de procedimientos estadísticos, entre otros. Todo esto en cumplimiento de la Ley Orgánica del INE, Decreto Ley 3-85."

Específicamente para esta publicación se utilizó la data de hechos delictivos sobre las víctimas reportadas por la PNC (Policía Nacional Civil) durante el año 2017. Su descripción exacta en el sitio es: *"Las estadísticas de hechos delictivos incluyen la información relacionada con los hechos, víctimas, sindicados y sentencias de los delitos cometidos en la República de Guatemala. También se incluyen datos relaciones con los exámenes médico-forenses que se realizan en virtud o sospecha de un hecho delictivo."*

Dentro de las categorías encontradas de la data estaban los grupos por delito categorizados directamente por la PNC sin embargo durante el pre-procesamiento de la data se encontraron incoherencias de categorización en cuanto a la gravedad de los hechos delictivos descritos. Por lo tanto, se creo una nueva categorización de la data acorde a su gravedad y grupo delictivo, la descripción exacta de los grupos creados puede ser encontrada en el documento de "Data Wrangling".

Algunas modificaciones hechas a la data fueron:

- La zona de ocurrencia tuvo que ser eliminada dado que casi nunca es registrada por la policía.
- Existen tres diferentes categorizaciones de grupos de edad, por lo cual dos de ellos fueron eliminados por redundancia.
- El año de ocurrencia y el número de correlativo son irrelevantes, también fueron eliminados.

Algunas genralidades de la data, encontradas durante el pre-procesamiento son estas:

- Los días 1, 15 y 30 del mes tienen una cantidad mayor de delitos cometidos.
- Los meses con mayor cantidad de crímenes reportados son enero y octubre.
- Los días lunes, viernes y domingo contienen un pequeño margen de incremento en los crímenes reportados.
- Los grupos de horas con mayores crímenes reportados son de las 6 p.m. en adelante.
- El departamento con más crímenes reportados es la ciudad de Guatemala.
- La cantidad de crímenes reportados en personas del sexo masculino son más que el doble en personas del sexo femenino.
- El grupo de edad de las personas que más son víctimas/victimizan tienen entre 15 a 39 años.
- Los mayores delitos cometidos reportados son:
    - Extorsión a residencias
    - Homicidio por arma de fuego
    - Lesiones por arma de fuego
    - Hurto de motocicletas
    - Desaparecidos
    - Hurto de vehículos

La data utilizada esta disponible públicamente en el sitio: https://www.ine.gob.gt/index.php/estadisticas-continuas/hechos-delictivos
Así también como otras bases de datos con datos de diferentes categorías. Para las

visualizaciones del prepocesamiento se puede referir al documento de "Data Wrangling".

3 Algoritmos

Los algoritmos utilizados en el análisis primario de la data fueron una combinación entre algoritmos de clasificación y algoritmos de regresión sin embargo, por la naturaleza de la data se decidió en el análisis secundario de la data, utilizar solamente algoritmos de clasificación.

La métrica utilizada para la evaluación de cada modelo utilizado se utilizó el atributo de *score* el cuál devuelve la precisión (*accuracy*) media en los datos y etiquetas de prueba dados.

Los algoritmos de clasificación utilizados fueron:

- Regresión logística: utilizado de forma simple y múltiple para categorizar la variable a predecir y sus categorías adecuadamente. La regresión logística se inclinó a categorizar la categoría con el mayor número de registros lo que causo puntajes en la precisión muy altos pero la precisión real muy baja.

- Análisis Discriminatorio Linear (LDA): es una técnica de *reducción de dimensionalidad supervisada*. El objetivo es proyectar un conjunto de datos en un espacio de dimensiones inferiores con una buena capacidad de separación de clases para evitar el sobreajuste y también reducir los costos computacionales.

- Análisis Discriminatorio Cuadrático: es un clasificador con un límite de decisión cuadrático, generado por el ajuste de densidades condicionales de clase a los datos y utilizando la regla de Bayes.

- K-Vecinos Cercanos (KNN): es un tipo de algoritmo supervisado de aprendizaje automático. KNN es extremadamente fácil de implementar en su forma más básica y, sin embargo, realiza tareas de clasificación bastante complejas. Utiliza todos los datos para el entrenamiento mientras clasifica un nuevo punto de datos o instancia. KNN es un algoritmo de aprendizaje no paramétrico, lo que significa que no asume nada sobre los datos provistos.

- Random Forest Classifier: es un algoritmo de aprendizaje supervisado. Puede ser utilizado tanto para clasificación como para regresión. También es el algoritmo más flexible y fácil de usar. Los bosques aleatorios crean árboles de decisión en muestras de datos seleccionadas al azar, obtienen predicciones de cada árbol y seleccionan la mejor solución por medio de la votación. También proporciona un buen indicador de la importancia de la característica.

Los siguientes modelos fueron utilizados dentro de las funciones para validar el grupo de *testing* y *training:*

- Validación cruzada de K iteraciones: es un método estadístico utilizado para estimar la habilidad de los modelos de aprendizaje automático. Se usa comúnmente en el aprendizaje aplicado para comparar y seleccionar un modelo para un problema de modelado

predictivo porque es fácil de entender, fácil de implementar y da como resultado estimaciones de habilidades que generalmente tienen un sesgo más bajo que otros métodos.

- Matriz de confusión: es una tabla que se usa para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba para los cuales se conocen los valores reales.

4 Utilización de la data

Para la mayor eficiencia en el análisis de datos se creó una nueva función que evaluaba los modelos contra el subconjunto de datos provistos por el usuario. Mientras que la función original tomaba en cuenta todos los algoritmos de clasificación y regresión para el propósito de este reporte se creo otra función aparte para evaluar solamente los modelos de clasificación. La tabla que la función devolvía era como la siguiente:

	Modelo	Score/Resultado
0	Regresión Logística	0.256545
1	Análisis Discriminatorio Linear	0.256545
2	Análisis Discriminatorio Cuadratico	0.239873
3	K Vecinos Cercanos	0.253238
4	Random Forest Classifier	0.304009

Para el testeo validación de la data, ésta fue descompuesta en dos conjuntos separados con la función de *Train Test Split* con el propósito de probar el modelo contra resultados reales dentro de la misma base de datos.

Si la evaluación del subconjunto de variables obtenía un resultado mayor a un 0.35 de precisión se utilizaban las funciones de gráficas generadas por los autores de este reporte con el propósito de visualizar la precisión del modelo ya fuera en una matriz de confusión o a través de gráficas de *performance* en el tiempo dependiendo también de cuál fuera el modelo con el cual se había obtenido el alto puntaje.

Se crearon dos subconjuntos para el análisis de la data. Primero la data fue analizada en su totalidad, y luego fue analizada solamente en el departamento de la ciudad de Guatemala con el objetivo de encontrar relaciones más intrínsecas en este departamento al ser el departamento con más crímenes reportados del país.

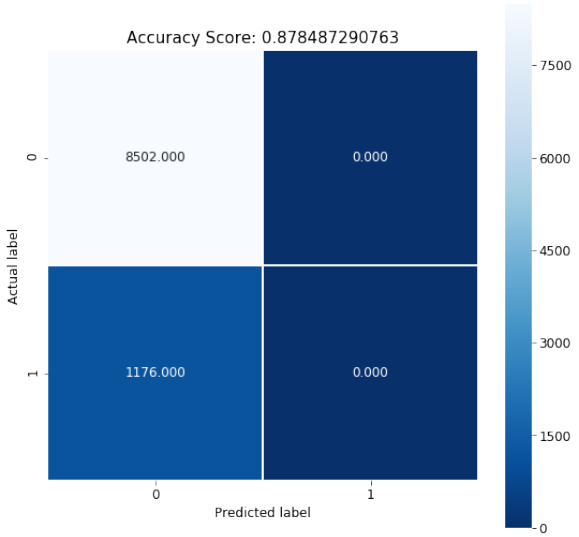
5 Resultados

El rendimiento de los modelos de clasificación en general fue aproximadamente el mismo, siendo el modelo de Random Forest el único en sobrepasar a todos los demás modelos en el testeo de subconjuntos entre variables con un puntaje más alto por 0.05, aún así no se reconoció que ningún modelo, siendo moderadamente preciso obtuviera una precisión mayor al 45%.

En los análisis hechos, si el modelo sacaba un puntaje mayor al descrito era porque probablemente estaba sobre-estimando la

suma de datos a la variable con mayor cantidad. Una prueba que puede ejemplificar este comportamiento es la evaluación de los modelos con el subconjunto del mes, día del mes, día de la semana, hora, área, municipio, sexo, y delito específico para poder predecir si la persona era mayor o menor de edad.

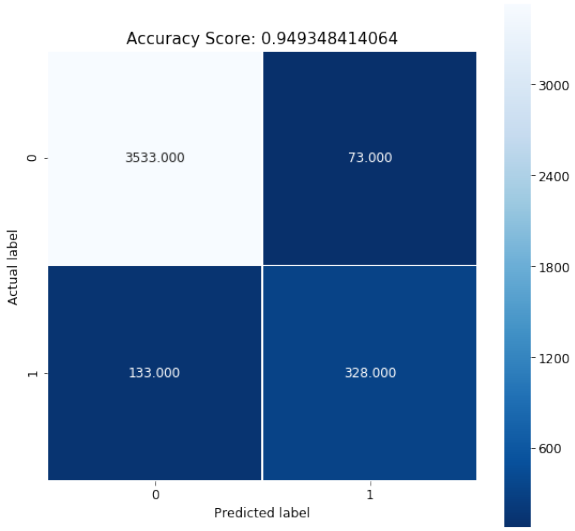
La matriz de confusión que se obtuvo fue:



Y tal como se ve en la imagen este modelo obtuvo una precisión del 87%.

En el análisis de la data hecha específicamente al departamento de Guatemala tampoco se logró encontrar una combinación adecuada para poder predecir el grupo de delito que fue cometido o cualquier otra variable dentro del *dataset*. Pero hubo una mejora en cuanto a la combinación de subconjunto para predecir si la persona reportada por la PNC era mayor o menor de edad.

Utilizando las variables de día de la semana, hora específica, área, grupo de edad, grupo de delito y sexo se obtuvieron los siguientes resultados, con una precisión del 94%.



6 Conclusiones y recomendaciones para futuros trabajos

A pesar de que no se pudieron obtener resultados relevantes para ser utilizados en futuras aplicaciones, se logra registrar por medio de este reporte un análisis no hecho anteriormente por nadie, con data y código públicamente disponible para que otras personas puedan hacer sus propios análisis y contribuciones a la investigación. Así también se conoce ahora que los modelos de clasificación son los mejores para poder predecir las categorías dentro del *dataset*, pues los modelos de regresión no son sólo inadecuados para la data sino también sus resultados son normalmente muy bajos o incluso negativos.

Mientras que la separación de la data por el departamento de Guatemala no fue tan efectiva como lo anticipado, aún así se lograron mayores puntajes de precisión en algunas áreas muy específicas.

En cuanto al desempeño de los modelos vemos que el modelo de Random Forest siempre tiene puntajes más altos a los demás. Así también el cambio entre el número de k-vecinos cercanos no tiene una gran repercusión en la precisión del modelo más que un punto o un punto y medio de diferencia. En promedio los modelos del *dataset* en general obtuvieron un puntaje entre el 20% y 30% de precisión, lo que indica que normalmente predicen correctamente uno de cada 3 a 5 resultados.

Las recomendaciones para futuros trabajos serían la exclusión del departamento de Guatemala para el análisis general del comportamiento de los demás departamentos, pues por la gran cantidad de crímenes reportados en el área, esto puede estar causando *bias* en cuanto al análisis y desempeño de los modelos. Así también la categorización como ciudad o no ciudad para su análisis personal dependiendo del departamento en cuestión o la población aproximada tomada de otras fuentes de información, para que de tal forma se pueda analizar si los departamentos de tipo ciudad tienden a tener un comportamiento parecido entre ellos o si fueran excluidos, si existe un comportamiento parecido entre todos los departamentos no considerados como ciudad.

Otra recomendación sería el basarse en otros análisis por otras instituciones, para analizar individualmente a departamentos con aumentos en hechos delictivos durante un tiempo específico, tales como el departamento de Chiquimula y los problemas de inmigración que tienen.

Si en el futuro pudiera ser posible una alianza entre el INE y la PNC sería una recomendación el recaudar otro tipo de variables que puedan describir más los crímenes tal como: zona en la que vive, en qué sector trabaja, número de personas que viven en su casa, salario promedio total que gana, si ha reportado o hecho crímenes en el pasado.

Por último, la recomendación para el procesamiento futuro de los datos sería el uso de Redes Neurales, para la detección de patrones o formas de los datos. Pues este tipo de procesamiento es conocido por saber encontrar patrones no visibles al ojo humano y con una capacidad de computo mucho mayor a cualquiera de los algoritmos utilizados en este reporte.

La data y procesamiento evaluado puede ser encontrado en:
https://github.com/nathsmo/Victimas_PNC_2017/

8 Referencias

https://www.prensalibre.com/tema/crimenes/

https://www.ine.gob.gt/ine/

https://scikit-learn.org/stable/modules/lda_qda.html

https://scikit-learn.org

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://lahora.gt/departamentos-alta-migracion-presentan-elevado-indice-creditos/

https://www.who.int/violence_injury_prevention/violence/national_activities/informe_estadistico_violencia_guatemala.pdf

http://www.dialogos.org.gt/wp-content/uploads/2018/07/Informe-SEMESTRAL-sobre-la-Violencia-Homicida-en-Guatemala-2018-ver-FINAL.pdf

https://www.elnuevodiario.com.ni/internacionales/centroamerica/451242-guatemala-reporta-4-410-homicidios-2017-mantiene-t/

https://www.insightcrime.org/wp-content/uploads/2018/08/Homicidios-de-pandillas-y-OTD-Guatemala-Informe-InSight-Crime.pdf

https://www.prensalibre.com/guatemala/comunitario/las-cifras-sobre-violencia-suicidios-y-complicaciones-en-embarazos-principales-causas-de-muertes-en-jovenes-guatemaltecos/

https://www.ine.gob.gt/index.php/historia