

## Predicción del día, hora y municipio de un delito sexual:



- Eduard Alfonso Caballero Ponguta - 2151099
- Sergio Andrés Dulcey Lizarazo - 2151830
- Juan Pablo Viviescas Romero - 2152499

### Objetivos:

#### Objetivo general:

- Realizar una predicción del día de la semana, hora y municipio en que ocurre un posible delito sexual, por medio de un análisis de datos proporcionados por la policía nacional de Colombia

#### Objetivos específicos:

- Implementar diferentes modelos, tanto clasificadores como regresores para las características [ Dia(type: string) y Hora(type: timestamp)]
- Determinar los modelos que permitan una mejor predicción del municipio, día de la semana y la hora a partir de las demás características

### Motivación

Los casos de crímenes sexuales van en aumento, y proporcionalmente la inseguridad de algunos sitios también han incrementado, pero si hubiera una forma de predecir la hora y el lugar donde un posible delito sexual, con ello se podría reforzar la seguridad en dicho lugar y con ello el índice de delitos sexuales disminuye notablemente y así mejorando la calidad de vida de los ciudadanos.

## Documentación del problema

Los delitos sexuales ha sido uno de los grandes problemas que han asolado al país en los últimos años, pero y si hubiera una forma de predecirlo, así se podría tratar de disminuir los factores que lo causan y poder evitar que se genere más casos, lo cual se va a investigar en este proyecto.

Con la implementación de modelos de predicción, tanto clasificadores como regresores para el dataset proporcionado por la policía, y con ellos hallar una pauta que permita predecir la hora, fecha y lugar de un posible delito sexual, con el fin de mejorar la seguridad de esos sitios, mejorando la calidad de vida de los ciudadanos.

## Estado del arte

Se han realizado varios intentos en tratar de predecir un delito sexual en Colombia, pero más enfocado en que si un agresor reincide a cometer de nuevo el delito, ya que un estudio hecho por la policía nacional colombiana determinó que la gran mayoría de los agresores volvieron a cometer delitos sexuales, y en especial contra menores de edad, por ello endurecieron más el castigo a los agresores.

## Dataset sin tratar

Delitos sexuales 2019  
Delitos sexuales (del 01 de enero al 31 de diciembre año 2019).  
Fuente: Policía Nacional. Datos extraídos el día 12 de enero del año 2020 a las 14:00 horas.

Más vistas Filtrar Visualizar Exportar Debatir Incrustar Acerca de

MINISTE...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...	Unname...
2019 Jan 01 1...	AMAZONAS	LETICIA (CT)	ARTÍCULO 20...	Martes	18:00:00	NO REPORTA...	RURAL	CASAS DE AP...	SIN EMPLEO ...	A PIE	A PIE	7	FEMEN
2019 Jan 01 1...	AMAZONAS	LETICIA (CT)	ARTÍCULO 20...	Martes	18:00:00	NO REPORTA...	RURAL	CASAS DE AP...	SIN EMPLEO ...	A PIE	A PIE	21	FEMEN
2019 Jan 01 1...	ANTIOQUIA	ABRIQUÍ	ARTÍCULO 20...	Martes	0:00:00	SAN JOSE	RURAL	FINCAS Y SIM...	SIN EMPLEO ...	A PIE	A PIE	16	FEMEN
2019 Jan 01 1...	ANTIOQUIA	AMAGÁ	ARTÍCULO 20...	Martes	10:00:00	LA GUALI	RURAL	FINCAS Y SIM...	SIN EMPLEO ...	A PIE	A PIE	4	FEMEN
2019 Jan 01 1...	ANTIOQUIA	AMAGÁ	ARTÍCULO 20...	Martes	10:00:00	LA GUALI	RURAL	FINCAS Y SIM...	SIN EMPLEO ...	A PIE	A PIE	8	FEMEN
2019 Jan 01 1...	ANTIOQUIA	AMAGÁ	ARTÍCULO 20...	Martes	10:00:00	LA GUALI	RURAL	FINCAS Y SIM...	SIN EMPLEO ...	A PIE	A PIE	28	FEMEN

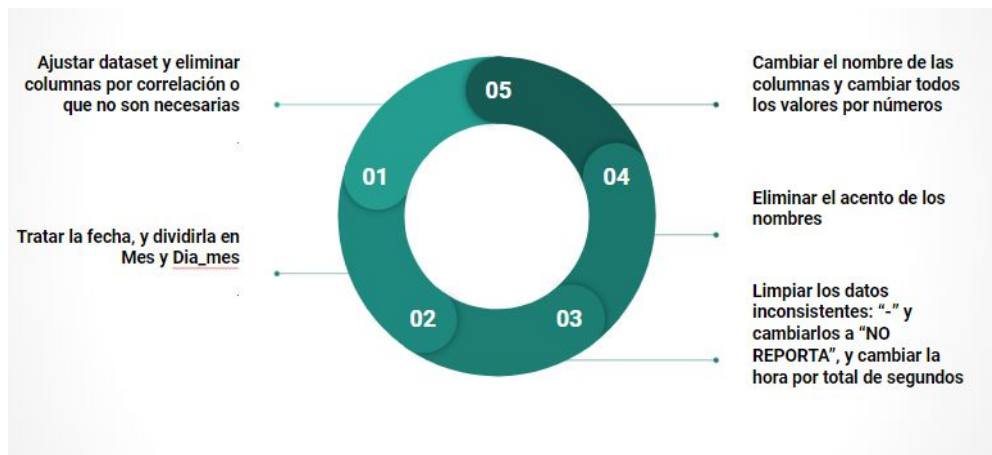
Gráfica 1. Dataset sin tratar

Fuente: <https://www.datos.gov.co/Seguridad-y-Defensa/Delitos-sexuales-2019/tdf6-87g7/data>

Tamaño: 140.400

Columnas: 21

En la gráfica 1 se evidencia el conjunto de datos sin tratar, este dataset se encuentra en el link publicado en fuente, inicialmente su tamaño es de 140 mil datos, y 21 columnas.



Gráfica 2. Pasos a usar para tener el dataset tratado.

En la gráfica 2 se evidencia los pasos que se usó para tratar los datos, se puede evidenciar de mejor manera en el notebook.

### Dataset tratado

	Día	Hora	Barrio	Zona	Clase_de_sitio	Arma_empleada	Móvil_Agresor	Móvil_Victima	Edad	Sexo	Estado_civil	Clase_de_empleado	Mes	Dia_mes	Mun_Dep
0	3	64800	3514	2	25	7	0	0	67	0	2	19	1	1	472
1	3	64800	3514	2	25	7	0	0	14	0	5	17	1	1	472
2	3	0	4172	2	59	7	0	0	8	0	4	15	1	1	2
3	3	36000	2719	2	59	7	0	0	34	0	4	11	1	1	31
4	3	36000	2719	2	59	7	0	0	78	0	4	11	1	1	31

Gráfica 4. Dataset codificado

Tamaño:33790

Columna: 15

En la gráfica 4 se evidencia el dataset codificado, se pasa de string a números con labelencoder.

### Exploración de los datos:

- Fecha: MAYO-1-2019 3442 Casos
- Municipios y departamentos: BOGOTÁ D.C- CUNDINAMARCA 5419 16%
- Género: FEMENINO 29.003, 85.8% - MASCULINO 4707, 13.9% - NO REPORTA 81 , 0.24%
- Edad: 13 años, 3829, 11.3%

En la exploración de los datos, se grafica de acuerdo a los datos tratados sin codificación, y lo que se busca es analizar los datos para evidenciar que datos pueden ser importantes a la hora de usar la implementación con el modelo.

**Implementación:**  
**Modelo:**

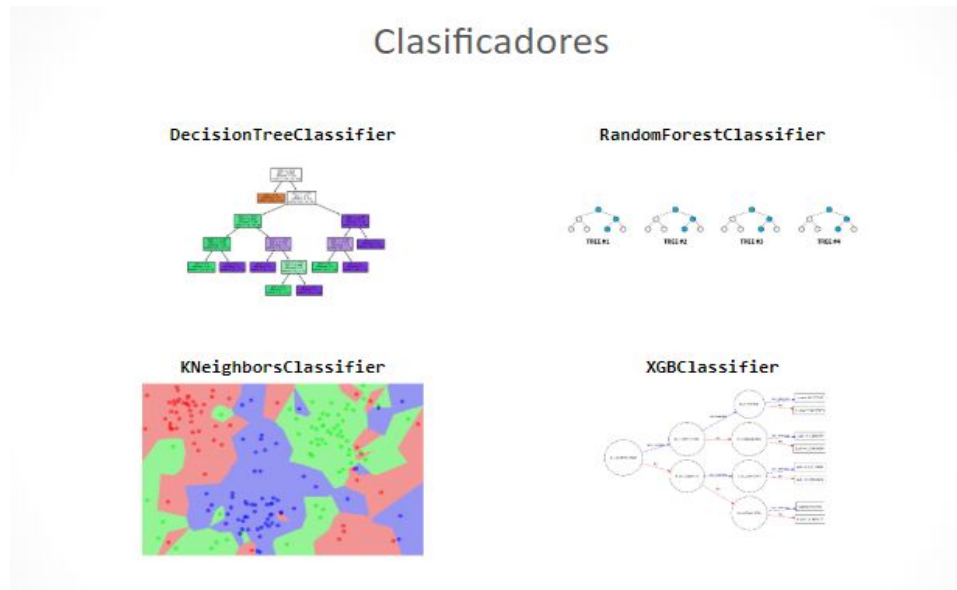
**Entrenamiento:** 80%

**Testeo:** 20%

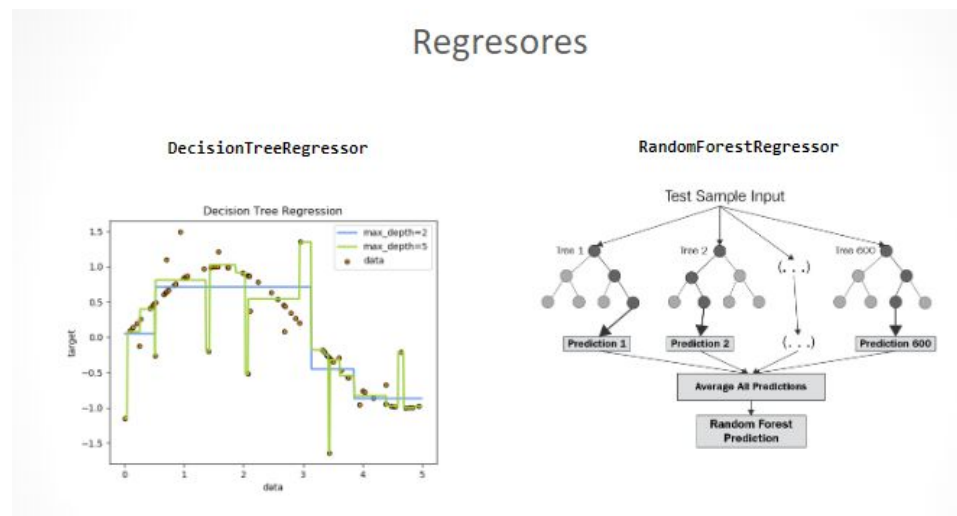
Estos son los porcentajes a usar para los clasificadores y regresores, se entrena el 80% y luego se testea el 20%, usando:

**Clasificador:** y[Dia,Municipio], se usa como y esas columnas, porque son datos discretos.

**Regresor:** y[Hora], se usa como y esa columna, porque son datos continuos.



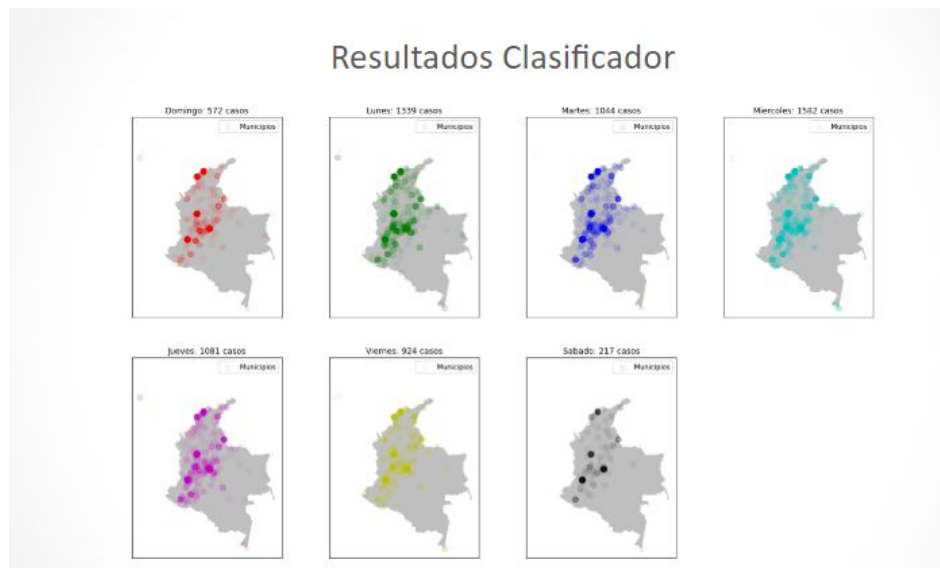
Gráfica 5.



Gráfica 6.

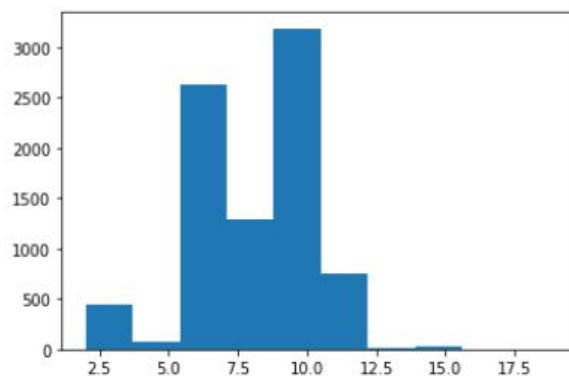
En la gráfica 5 y 6, se usaron los clasificadores y regresores respectivamente, los que se usaron fueron el XGBClassifier y RandomForestRegressor, el clasificador fue el de mejor accuracy y los dos tuvieron menor RMSE.

## Resultados:



Gráfica 7.Resultados clasificador

## Resultados Regresor



Gráfica 8. Resultados regresor

## Conclusiones:

- ❖ El clasificador con mejor accuracy, y evidentemente menor RMSE, es Multi Output Classifier con XGBClassifier como parámetro
- ❖ El regresor con mejor accuracy y evidentemente menor RMSE, es Decision Tree Regressor
- ❖ Con el clasificador, Se predice que el día miércoles, es el día que se produce más casos de delitos sexuales
- ❖ Con el regresor, se predice la hora donde ocurren más casos, se encuentra en el intervalo de la mañana y el medio día.

## Bibliografía:

- <https://www.datos.gov.co/Seguridad-y-Defensa/Delitos-sexuales-2019/tdf6-87gZ>
- <https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html>
- <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
- <https://gitlab.com/bivl2ab/academico/cursos-uis/ai/ai-uis-student.git>