Predicción del tipo de delito que ocurre en la Ciudad Autónoma de Buenos Aires y busqueda de similaridad mediante Clustering

clusterAI

Alcuri Federico, Girola Federico, Pantoni Francisco
Ingenieria Industrial, Universidad Tecnologica Nacional, Facultad Regional Buenos Aires

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

Introducción

A partir del registro durante cuatro años (2016-2019) de delitos en Capital Federal, se procederá a realizar un análisis de los datos para entender cuáles son las variables que intervienen y cómo se comportan a la hora de registrar un hecho delictivo. El objetivo es poder clasificar los delitos una vez ocurridos en base al “Tipo” del cual se trate, mediante la utilización de diferentes modelos de clasificación (Machine Learning), como SVM, KNN y LR. Se buscará clasificar la ocurrencia del delito según sea Hurto, Robo con violencia, Lesiones y Homicidios.
Ademas obtener la similaridad de las muestras mediante clustering.

Datasets utilizados

Para el trabajo se utilizaron 6 dataset:

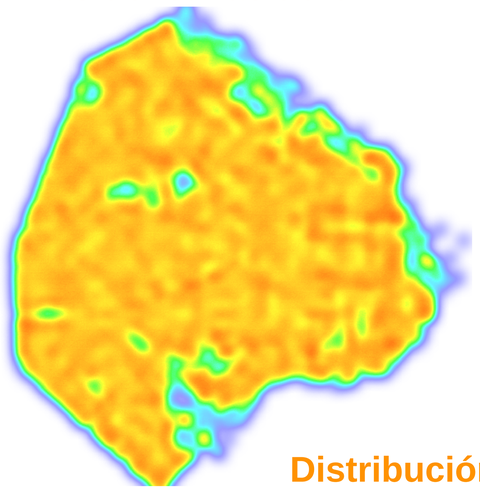
- 4 de delitos ocurridos en la Capital Federal entre 2016 y 2019 --> Datos abiertos CABA
- Características poblaciones --> Elaboración propia en base a datos abiertos CABA
- Ingresos por barrios --> Elaboración propia en base a datos abiertos CABA

Análisis Exploratorio de Datos

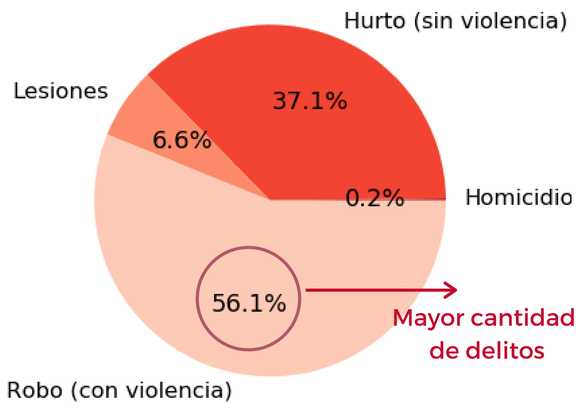
Pasos iniciales: Unificación y limpieza de datos

- Unión de los 4 datasets de delitos --> 10 features y 488.000 samples
- Eliminación de features irrelevantes (ID, Subtipo de delito)
- Eliminación de valores nulos --> Representan el 1.74% de los datos
- Incorporación de Ingresos por barrios al datasets --> 11 features y 488.000 samples
- Separación de la feature "Fecha" en "Dia" "Mes" y "Año"

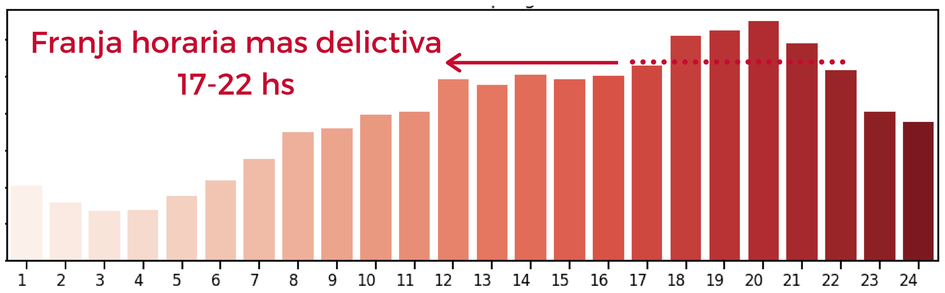
Visualización de las zonas más delictivas de CABA



Distribución del tipo de delito

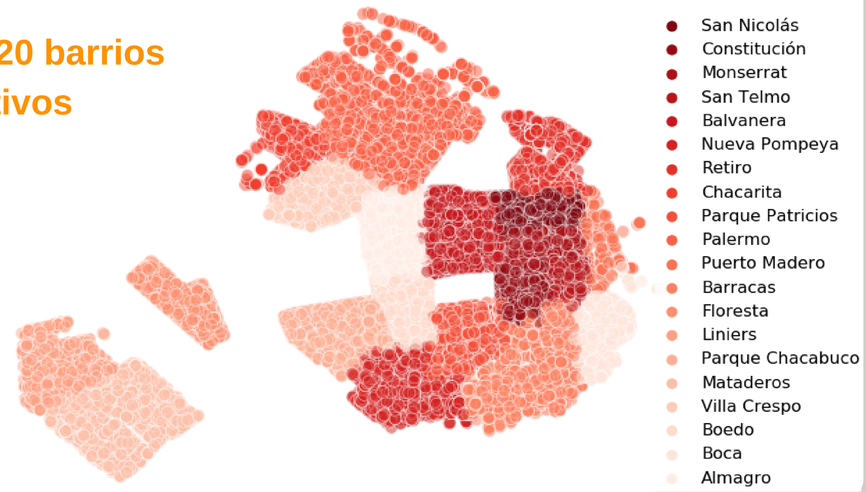


Distribución horaria de los delitos



Scaterplot Top 20 barrios más delictivos

Mediante el uso de las características poblaciones de los barrios según un Índice de delitos cada 100 Habitantes por barrios Sin afectar a los barrios por el índice, el barrio mas delictivo es Palermo, al afectarlo podemos ver que es San Nicoles en cantidad de delitos cada 100 habitantes

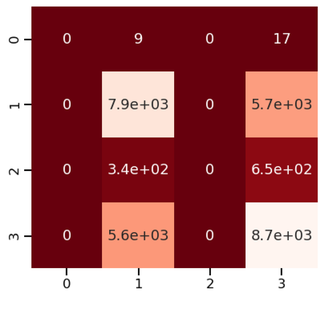


Métodos y Modelos

Clasificación: KNeighborsClassifier y Support Vector Machines

- Realizado sobre los 5 barrios más delictivos del 2019
- Generación de dummies para la feature "Barrio" --> Obtención de 5 features numéricas extras
- Transformación de la feature "Tipo" mediante Label Enconder --> Toma valores 0 , 1 , 2, 3
- Determinación de la matriz "X" con las variables independientes e "Y" con las etiquetas de "Tipo"
- Muestras de entrenamiento y testeo mediante Test Split --> Test size 0.8
- Auto-scaling de "xtrain" e "ytrain" mediante Standar Scaler

KNeighborsClassifier

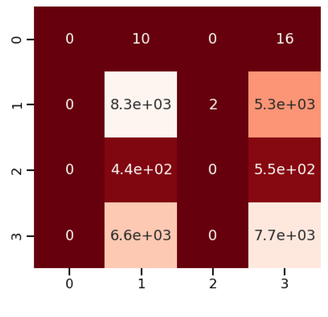


Grid Search & Cross Validation:

- Kneighbors:
  - Arrange (1,20,1)

Accuaracy 55.37%

Support Vector Machines



Grid Search & Cross Validation:

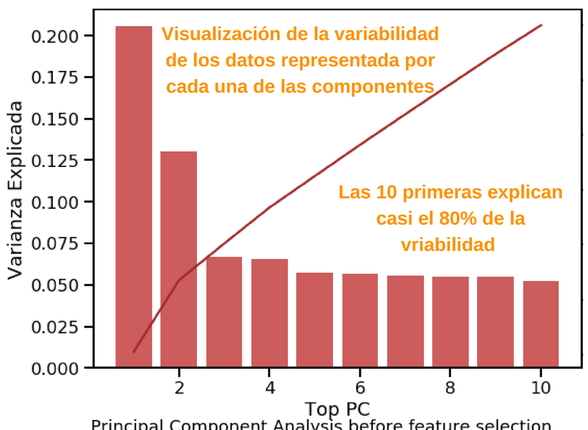
- C: 0.1, 1, 10
- Kernel : "RBF" "Linear"
- Gamma : 0.01, 0.1 , 1

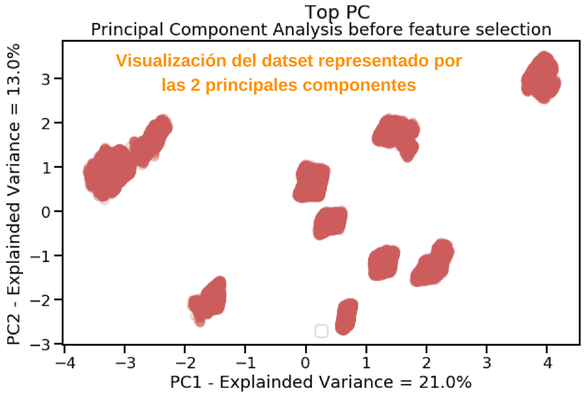
Accuaracy 57.36%

Clustering : Reducción de la dimensionalidad PCA & K-Means

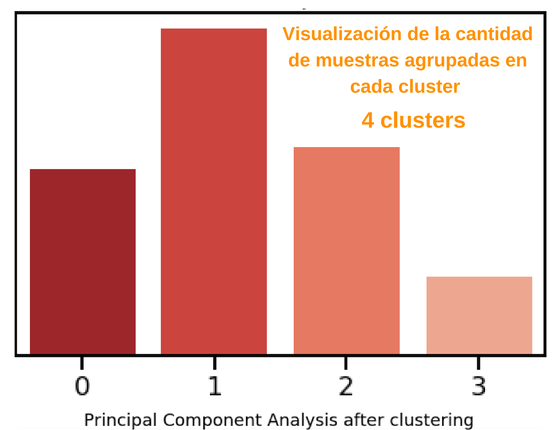
- Realizado sobre los 10 barrios más delictivos del 2019
- Generación de dummies para la feature "Barrio" --> Obtención de 10 features numéricas extras
- Transformación de la feature "Tipo" mediante Label Enconder --> Toma valores 0 , 1 , 2, 3
- Determinación de la matriz "X" con todas la variables y Auto-scaling

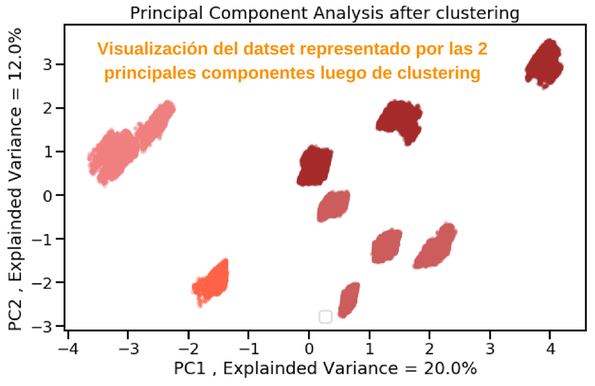
PCA





K-means





Resultados y Conclusiones

Clasificación

En cuanto al modelo de clasificación podemos observar que tuvo un resultado aceptable, por lo que puede clasificar muestras de delitos ocurridos en CABA y determinar cual es el "Tipo" del cual se trata con una exactitud del 56%

Pca + K-means

Como se pudo observan en el PCA, las 10 principales componentes explican casi el 80% de la variabilidad de los datos. Luego de la clusterización utilizando las 10 principales componentes pudimos observar que las muestras fueron agrupadas en cluster según la similaridad determinada por la comuna a la que pertenece cada muestras y a los disitntos barrios. Obteniendo un Silhouette score de 35%