Data Science Course

Proyecto Final
Communities and Crime Data

Equipo #16

Introducción

El objetivo de esté proyecto es aplicar la metodología *CRISP* para poder analizar, modelar y predecir la cantidad de crímenes violentos por población existente en diferentes comunidades en USA durante el año de 1990.

Para esto se va a recaer en el uso de diferentes técnicas de *Machine Learning* para regresión así como el uso de una técnica no supervisada, *K-Means*, para hallar patrones en las comunidades mediante los datos y poder brindar conclusiones adecuadas sobre las predicciones.

Business Understanding

Organizaciones

Se han proporcionado datos del año 1990 por tres organizaciones de USA:

- Censo de US
- Censo del LEMAS US sobre aplicación de la ley
- Información criminal dada por el FBI en 1995

Planteo del problema

- ¿Qué parámetros sociales/económicos influyen mayormente en los crímenes violentos?
- ¿Existirá un sesgo intrínseco en los datos?

Data Understanding

La base de datos fue obtenida a través del repositorio de *Machine Learning* del *Center for Machine Learning and Intelligent Systems* de la Universidad de California.

Dicho conjunto de datos están bajo el nombre de *Communities and Crime Data* bajo la etiqueta *Social* y posee las siguientes características:

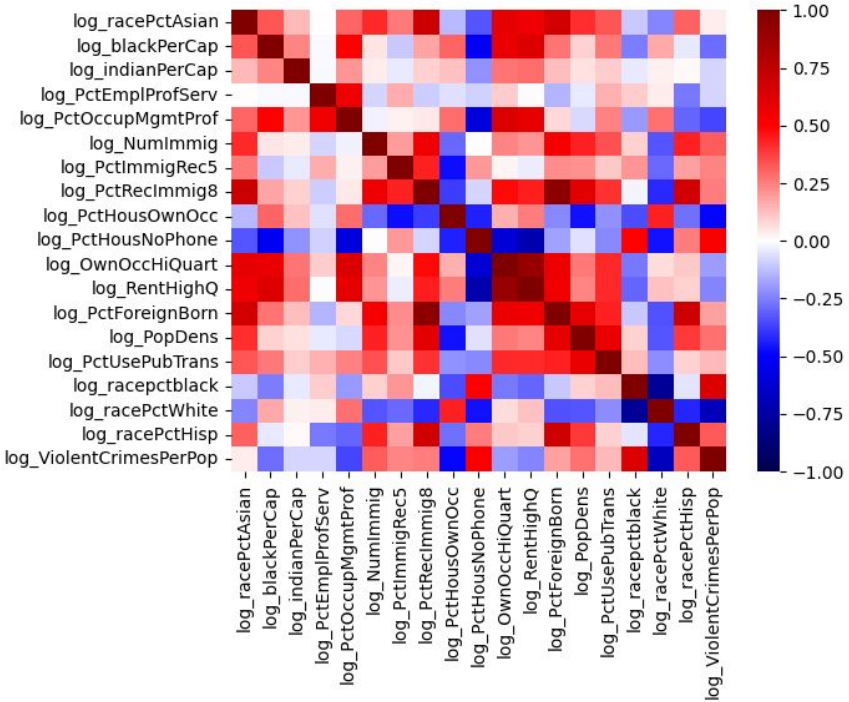
- Los datos son multivariados y reales
- Se cuenta con 1994 instancias, que representan el total de comunidades consideradas para el estudio , con 128 atributos para describir cada instancia
- Habrá instancias con atributos no disponibles
- Todos los atributos numéricos fueron transformados a una escala entre 0 y 1

El último atributo de nombre *ViolentCrimesPerPop* representa el número total de crímenes violentos por cada 100,000 habitantes y es el atributo objetivo.

Data Preparation

Los pasos seguidos para preparar la base de datos para su posterior análisis fueron los siguientes:

1. Remoción de atributos en los que no se contarán con suficientes datos disponibles.
2. Se buscaron datos no disponibles anómalos en los atributos restantes.
3. Se aplicó una transformación logarítmica dada la dispersión de los datos.
4. La técnica *PCA* se implementó para hallar los atributos más representativos de la base de datos actual.
5. Se buscaron *outliers*, bajo el uso de la distancia *Mahalanobis,* que toma en cuenta las correlaciones entre los atributos de datos.



Modeling

Para analizar los datos se utilizaron fracciones de 70% y 30% para conjuntos de entrenamiento y prueba, respectivamente; Y para predecir el atributo objetivo se propusieron tres modelos diferentes para aplicar regresión:

1. Modelo lineal
2. *Decision tree*
3. *Random Forest*

Para el modelo lineal se consideraron los pesos de cada atributo como positivos y constante nula. Con el resto, se optimizaron hiper parámetros para hallar aquella configuración que predijera mejor el atributo de crímenes violentos.

Además, para la técnica *K-means* se buscó el número óptimo de *clusters* utilizando el conjunto de datos completo.

Evaluation

Además de la cantidad R^2, se calcularon los errores cuadráticos y absolutos, *MS2* y *MA* respectivamente, de cada modelo de regresión para comparar y evaluar.

Se notó que en las tres cantidades más óptimas pertenecen al modelo *Random Forest*, logrando mejores predicciones en un 10% a comparación del modelo *Decision Tree*.

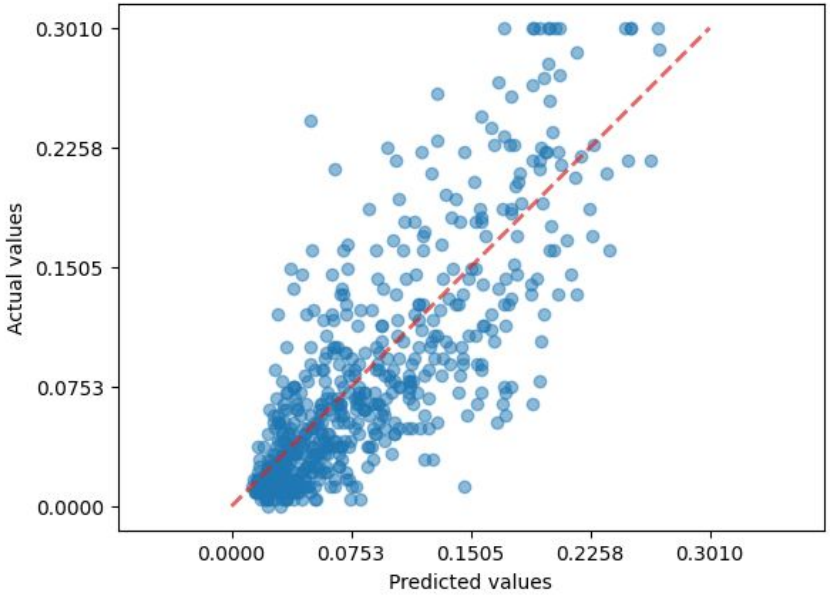
Respecto al uso de *K-Means*, se estableció que el número óptimo de *clusters* es de 3 y con la cantidad de comunidades ilustrada.

Módelo	MS2	MA	R2
Linear	0.0460	0.0347	0.5723
Decision Tree	0.0479	0.0345	0.5373
Random Forest	0.0428	0.0317	0.6395

Número de *cluster*	Número de comunidades por *cluster*
0	875
1	814
2	305

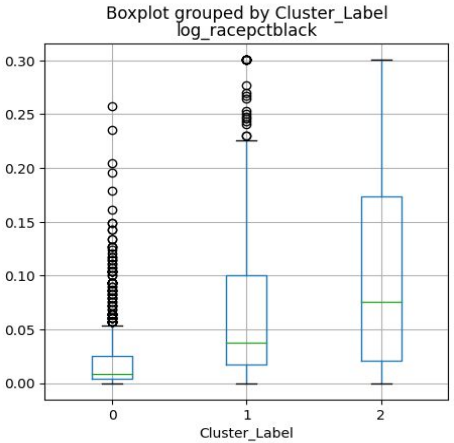
Deployment

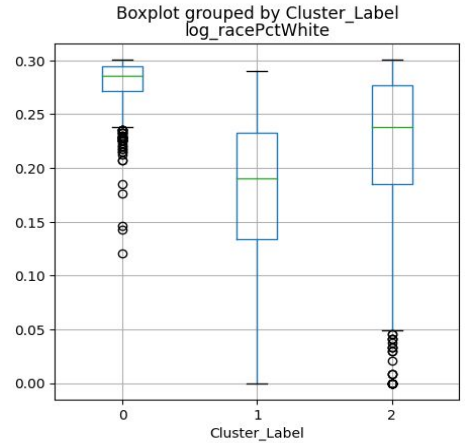
Aunque la R^2 de 0.6395 del mejor modelo, el *Random Forest*, no sea suficientemente alta para prometer excelentes predicciones; cabe resaltar que el estudio es de ámbito social por lo que deberíamos esperar influencias por comportamiento humano y sesgos sociales.

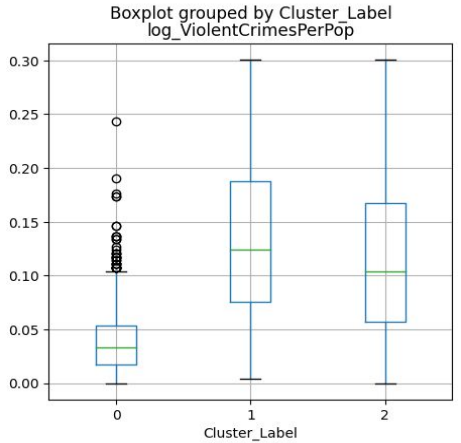


Dichas influencias se visualizan mayormente al comparar los valores reales contra los predichos por el mismo modelo. Es fácil de observar que hay una gran concentración de comunidades con crímenes violentos poco frecuentes.

Un par de variables muy resaltantes en los *clusters* obtenidos fueron los porcentajes de raza negra y blanca así como la tasa de crímenes violentos. En base a los gráficos de cajas, podemos dar descripciones cualitativas







# Cluster	% Negra	% Blanca	Crímenes Violentos
0	Baja y dispersa	Alta y levemente dispersa	Baja y levemente dispersa
2	Alta y dispersa	Alta y dispersa	Completamente dispersa

Conclusión

Tras concluir con esté proyecto, se pudieron aplicar y fortalecer enormemente las teoría y técnica detrás del uso métodos computacionales del *Machine Learning* para estudiar datos y hallar tanto patrones como poder predictivo.

En particular, la tarea principal fue de poder predecir la cantidad de crímenes violentos por población entre diferentes comunidades en USA y, aunque se logró bajo estándares de estudios sociales, cabe recalcar que la actividad humana podría estar sesgando nuestras conclusiones; por esta razón se vió viable un análisis por medio de *clusters*.

Y afortunadamente, además de la predicción, se pudo notar que efectivamente existen sesgos sociales que no se reflejan en la regresión; cuestión que podría ser delicada de no tratarse en esté tipo de información.