Estudio de Datos Poblacionales de Galicia

Autor: **Cuesta Calvo, Roberto**

Director: Rodero Lacruz, Miguel

Contacto: rcuesta8@hotmail.com

Resumen: Se ha pretendido establecer un estudio sobre la posibilidad de encontrar determinadas Características cuantitativas, objetivas y medibles que pueden predecir, en cierta manera, el comportamiento de una Población desde el punto de vista de la transgresión de leyes.

Dichas Características se han intentado provisionar, a partir de información tanto de entidades y organismos públicos como de información privada y reservada correspondiente con otros entes u organizaciones, con el objeto de analizar el grado de relación existente respecto del hecho final que se quiere medir.

El estudio se ha limitado a la comunidad autónoma de Galicia con el objeto de analizar si esas características pueden adivinar patrones conductuales desde el punto de vista de conjunto, Población, en cuanto a la trasgresión administrativa desde dos puntos de vista totalmente diferentes:

> En el primero, intentando predecir todo aquello que esté relacionado con la violencia de género y en el segundo con todas aquellas infracciones que no tengan este carácter.

Palabras clave: Población, Galicia, violencia de género, no violencia de género, predecir

1. Introducción.

1.1. Objetivo.

Establecer unos cimientos fuertes y suficientemente genéricos para poder recopilar información de distintos tipos y fuentes adoptando la misma en un futuro cercano por la correspondiente Organización.

Esta recopilación de información, la mayor posible, servirá como base para la realización de un estudio poblacional por municipio, centrado en la Comunidad Autónoma de Galicia. Con este estudio se pretende poder establecer el pertinente conocimiento de cada municipio para poder tomar decisiones futuras en base a ello.

Este estudio finalmente intenta prever, en base al conocimiento de la Población, variables o características, la transgresión de leyes que se va a realizar desde dos puntos de vista, Clase, como son los siguientes:

- Infracciones cometidas que tengan relación con la violencia de género (VG).
- Infracciones cometidas que no guarden relación con la violencia de género (nVG).

1.2. Características o variables Poblacionales.

Esta información ha sido obtenida tanto de fuentes públicas [1][2][3] (online, mediante crawlers[1] generados, y off line) como de fuentes privadas.

El estudio de los datos está basado en el año 2016[2].

1.3. Medios.

Tanto la recolección de datos, como el proceso de tratamiento y estudio de los datos, mediante el lanzamiento de experimentos ha sido realizado con tecnología Phyton en diferentes cuadernos Jupyter Notebook.

No existen más requerimientos software ni hardware puesto que el proceso[3] es lanzado en un portátil de gama media.

2. Desarrollo.

Así pues se va a intentar predecir las clases definidas (normal y normalizada) mediante la generación de experimentos que se ejecutarán bajo una regresión Lineal, Lasso[4][6][7], y se analizará su conveniencia en función del porcentaje de ajuste que tenga el conjunto de datos escogido frente a los datos esperados, haciéndose esto mediante el método r2_score[11].

2.1. Ejecución del proceso.

La imagen de la **Figura 1** muestra el proceso de generación del estudio seguido de una manera más visual. No obstante a continuación se detallan las fases del mismo.

El hilo o proceso principal está guardado en el cuaderno *fMain*, desde donde se referencian todas las librerías[5] necesarias para su ejecución así como las distintas clases (.py) y demás cuadernos necesarios, para realizar las distintas tareas:

[1] Robots que descargan información de manera on-line y automática, hechos ad-hoc para este TFM.

[2] Debido a que las fuentes de datos que conforman la Clase corresponden al año 2016 todas las demás características han sido buscadas y seleccionadas en base a esto. No es muy abundante los datos referenciados a municipio existentes para este año. Con el paso de los años se está ganando en riqueza a nivel tanto de calidad de los mismos como de cantidad.

[3] Tiempo de ejecución estimado: 2 horas y 30 minutos.

[4] Internamente se comporta haciendo una selección de características dando más peso a unas sobre otras.

[5] Entre otras las de Pandas[9],Numpy[10] y Scikit-Learn.

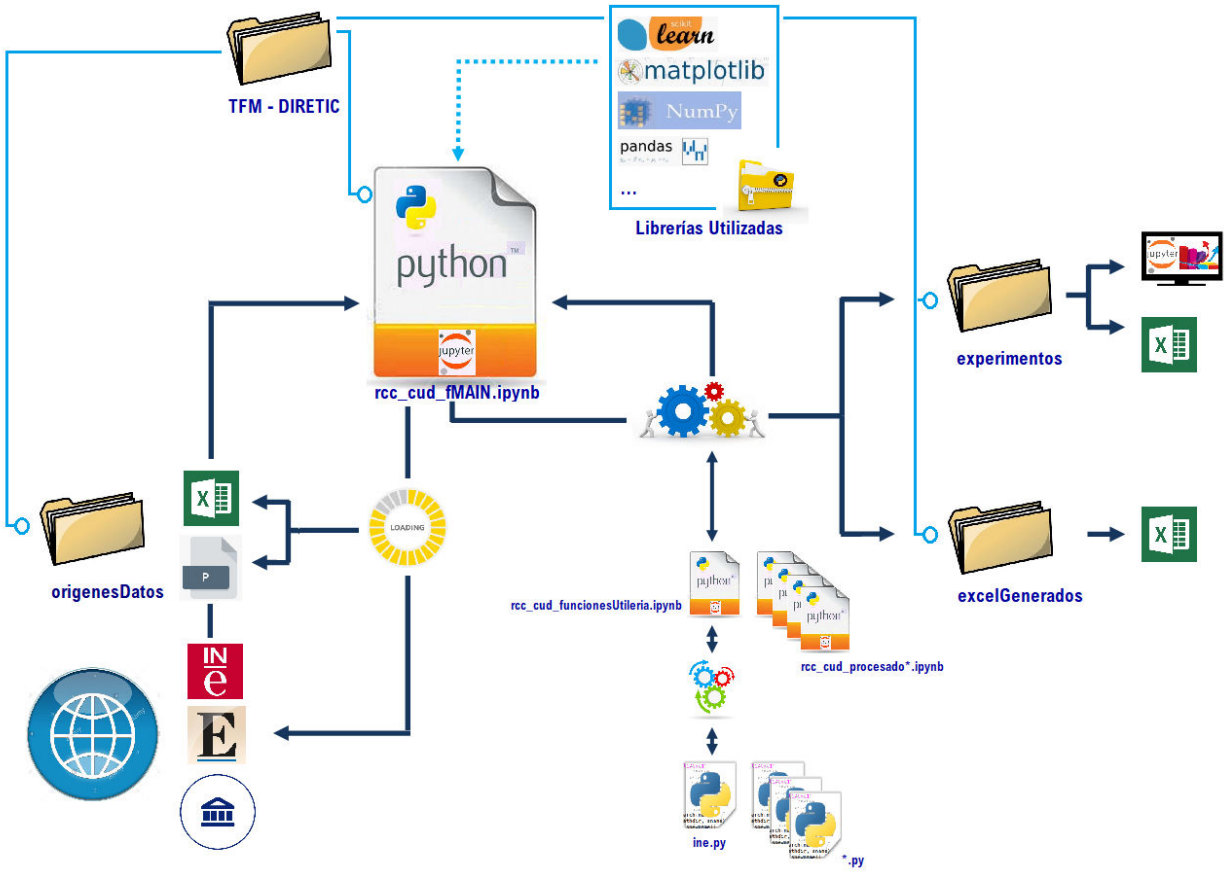

Figura 1. Ejecución proceso TFM – Estudio de Datos Poblacionales de Galicia

1) *Recolección datos*: En esta parte del proceso se recolecta todos los datos existentes de las distintas fuentes, tanto los previamente descargados como los que se hacen de manera online. Se trae información relativa a:
    a) *Datos estadísticos de la Población.*
    b) *Deuda del Municipio.*
    c) *Información relativa al número de empresas.*
    d) *Número de contratos establecidos.*
    e) *Referencia a la Actividad de la Población por sectores.*
    f) *Números relativos al Paro.*
    g) *Turismo.*
    h) *Información relativa a las armas y licencias.*
2) *Tratamiento de la Información*: Una vez descargados se hacen distintas tareas de tratamiento de datos para poderlos convertir en información que sea fácilmente integrada dentro del estudio.
    a) *Generación del Código INE para la posterior fusión.*
    b) *Obtención de datos cuantificables objetivos en mismas unidades*
    c) *Normalización, refiriendo este concepto a referir la variable medida en relación a la cantidad de población total del municipio.*
3) *Generación de los Experimentos*: Con la información dispuesta (variables y clase) se van conformando los distintos experimentos que van surgiendo de distintas combinaciones posibles entre las variables existentes a la vez que se va midiendo el resultado de cada una

de las predicciones lo que va determinando que conjuntos de valores son por los que se va apostando en las combinaciones. Se realizan un total de 75 experimentos por clase[6].

4) *Visualización de los Resultados*: Posteriormente se pasa a la realización de un análisis.

3. Resultados y discusión

Inicialmente se llega a un Accuracy máximo para *VG* de 0,*322759* y para *nVG* de 0,*427409987*.

Se procede a analizar los datos obtenidos. Tal y como se muestra en la **Figura 2**, existen valores que están bastante desproporcionados.

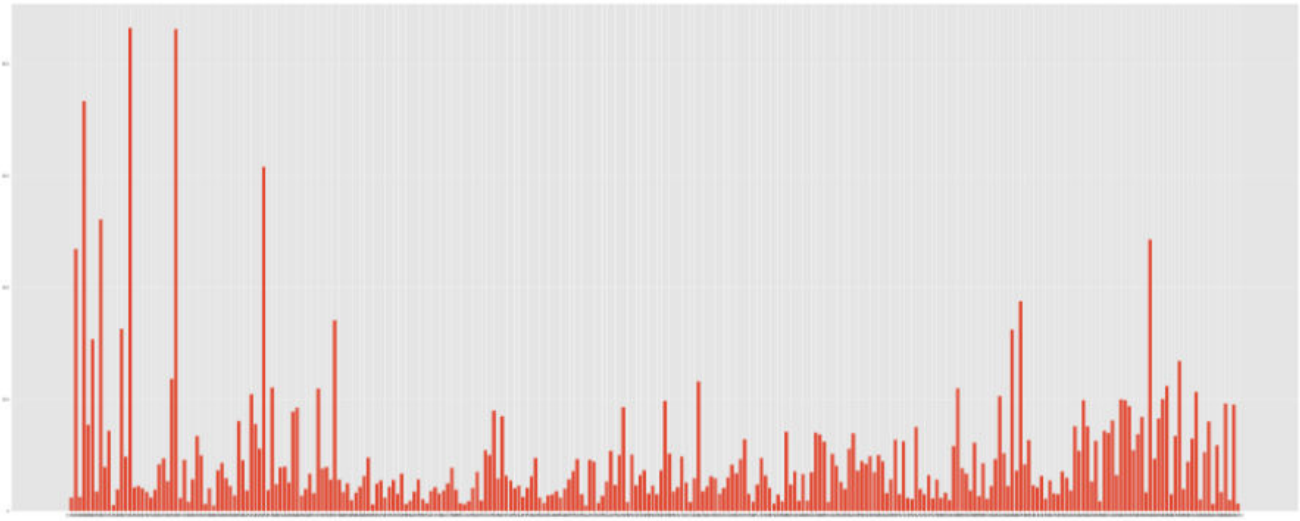

Figura 2. Gráfica de resultados pertenecientes a la diferencia entre el valor real y el predicho.

Tras analizar estos datos se llega a la conclusión de que algunos desfases se producen en municipios donde existe otra Fuerza y Cuerpo de Seguridad del Estado de la que no se poseen datos y los mismos no son los totales. Así pues se realizo una nueva generación de experimentos filtrando estos municipios. Los resultados obtenidos tras este experimento se muestran en la Tabla 1.

Nº	Características	Clase	Observaciones	Accuracy[7]
63	df_INE_2,df_Armas,df_numeroEstablecimientos	VG	% Armas	0,321656
101	df_experimento_63	VG	df_experimento_101	0,321656
201	**df_experimento_101**	**VG**	**experimento_201**	**0,371829**
63	df_INE_2,df_Armas,df_numeroEstablecimientos	NVG	% Armas	0,42741
101	df_experimento_63	NVG	df_experimento_101	0,42741
201	**df_experimento_101**	**NVG**	**experimento_201**	**0,439763**

Tabla 1. Tabla Resultados experimento filtrado FFCCSE.

Se sigue analizando los datos que tienen menos relación y se observa que esto tiene que ver con los municipios que poseen Policías Locales. Se analiza y se afina al hecho de que existe una relación con las que poseen en plantilla más de 15 vacantes. Así pues se procede a su filtrado y se vuelve a generar un nuevo experimento que cuyos resultados se pueden apreciar en la Tabla 2.

Nº	Características	Clase	Observaciones	Accuracy

[6] Existen 4 clases (VG, nVG, normalizado VG y normalizado nVG) que dan 300 experimentos..
[7] Acierto en la predicción. El valor debe estar entre 0 y 1 (siendo uno su valor máximo posible).

202	df_experimento_201	VG	experimento_202	**0,382096**
202	df_experimento_201	NVG	experimento_202	**0,504904**

Tabla 2. Tabla Resultados experimento filtrado Policía Local.

4. Conclusiones

A continuación, en la Tabla 3. se van a mostrar los experimentos que mejor puntuación han obtenido prediciendo las distintas clases[8].

Nº	Características	Clase	Observaciones	Accuracy
202	df_experimento_201	VG	experimento_202	0,382096
201	df_experimento_101	VG	experimento_201	0,371829
21	df_INE_2,df_Armas	VG	Suma de todas las Armas	0,322759
22	df_INE_2,df_Armas	VG	Solo con Suma de todas las Armas	0,322759
51	df_INE_2,df_Armas,df_Licencias	VG	suma Armas y % Licencias	0,322759
202	df_experimento_201	NVG	experimento_202	0,504904
201	df_experimento_101	NVG	experimento_201	0,439763
16	df_INE_2,df_numeroHoteles	NVG	Ninguna	0,42741
63	df_INE_2,df_Armas,df_numeroEstablecimientos	NVG	% Armas	0,42741
102	df_experimento_16	NVG	df_experimento_16	0,42741

Tabla 3. Resultados con mejor Accuracy.

Se podría afirmar, dentro del prisma del estudio actual, y a tenor de los resultados obtenidos que:

1) Se puede observar que los mejores resultados obtenidos se producen cuando van asociados a lso conjuntos de datos relacionados tanto del Turismo como los relativos a la posesión de Armas y/o Licencias.
2) Incluso se puede contribuir, tras el estudio realizado dentro del mismo, a intentar romper con sesgos tradicionales de la Población o estereotipos puesto que:
    a) Se ha observado que aunque los porcentajes de población extranjera en Galicia, no son importantes, no existe una relación que determine la Clase.
    b) Además se ha comprobado que el número de infracciones no se relaciona con las variables económicas.

Referencias

[1] «Web del Instituto Nacional de Estadística, INE,» [En línea]. Available: http://www.ine.es. [Último acceso: 19 enero 2021].

[2] «Datos económicos de Expansión.com» [En línea]. Available: https://datosmacro.expansion.com/paro/espana/municipios/... [Último acceso: 21 enero 2021].

[8] Tras el primer set de experimentos se desecharon las clases normalizadas, tanto de VG como para nVG, por los malos resultados obtenidos.

[3] «Datos de deuda,» [En línea]. Available: https://www.hacienda.gob.es/. [Último acceso: 21 enero 2021].

[4] «Informe Telefónica sobre IA» [En línea]. Available: https://empresas.blogthinkbig.com/matematicas-del-machine-learning/. [Último acceso: 21 enero 2021].

[5] «Pirámide de Información» [En línea]. Available: https://soulimproveledge.com/piramide-del-conocimiento/#:~:text=B%C3%A1sicamente%2C%20la%20Pir%C3%A1mide%20del%20Conocimiento,alta%2C%20o%20de%20m%C3%A1s%20valor.&text=Un%20concepto%20algo%20m%C3%A1s%20dif%C3%ADcil,grado%20m%C3%A1s%20elevado%20del%20conocimiento%E2%80%9D. [Último acceso: 21 enero 2021].

[6] «Metodo Lasso» [En línea]. Available: https://www.cienciadedatos.net/documentos/31_seleccion_de_predictores_subset_selection_ridge_lasso_dimension_reduction. [Último acceso: 21 enero 2021].

[7] «Metodo Lasso» [En línea]. Available: https://es.wikipedia.org/wiki/LASSO_(estadística). [Último acceso: 19 enero 2021].

[8] «Dudas Phyton y su implementación» [En línea]. Available: https://es.stackoverflow.com/. [Último acceso: 21 enero 2021].

[9] «Documentación y dudas sobre Pandas y su implementación» [En línea]. Available: https://pandas.pydata.org. [Último acceso: 21 enero 2021].

[10] «Documentación y dudas Numpy y su implementación» [En línea]. Available: https://numpy.org/. [Último acceso: 21 enero 2021].

[11] «Librería y dudas sobre Modulo de IA y su implementación» [En línea]. Available: https://scikit-learn.org/stable/. [Último acceso: 21 enero 2021].

[12] «Informe Gartner» [En línea]. Available: https://www.gartner.com/en. [Último acceso: 21 enero 2021].

[13] «Jupyter Notebook» [En línea]. Available: https://jupyter.org/. [Último acceso: 21 enero 2021].

[14] «Python» [En línea]. Available: https://www.python.org/. [Último acceso: 21 enero 2021].

[15] «Cross Val Predict» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_predict.html. [Último acceso: 21 enero 2021].