Análisis Establecimientos Hospitalarios en la Provincia de Buenos Aires
Data Analysis & Machine Learning

Julián Boglio
Buying Assistant at ExxonMobil
Ing. Industrial UTN BA

Abstract - El objetivo de este proyecto es identificar las variables que mejor explican el comportamiento de los establecimientos hospitalarios pertenecientes a la Provincia de Buenos Aires y evaluar la factibilidad de un modelo que permita predecir la dependencia de un establecimiento según sus indicadores.

INTRODUCCIÓN

La Provincia de Buenos Aires (PBA) compone un distrito crítico debido a su cantidad de habitantes y su gran extensión geográfica.

Según el recuento de 2015, PBA tiene una población de 16.66 millones de personas.

La Provincia tiene 134 Municipios divididos en 12 Regiones Sanitarias, zonas en que se divide el territorio bonaerense tomando en cuenta su población y las instalaciones médicas y sanitarias disponibles.

(Figura 1)

Dada las problemáticas referidas a la disponibilidad de establecimientos sanitarios en PBA, el objetivo de este estudio es analizar si a partir de variables sencillamente medibles se puede realizar una predicción de las muertes por región sanitaria.

DATA

1. FUENTE

El set de datos utilizado proviene del Ministerio de Salud de la Provincia de Buenos Aires.

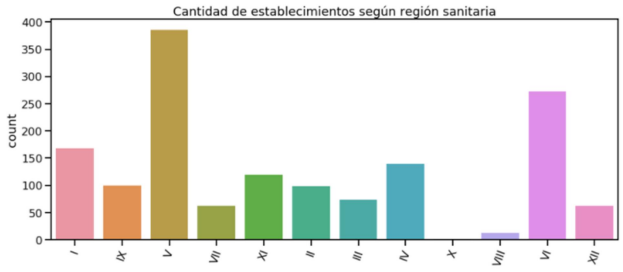
El mismo se compone de los diferentes indicadores de gestión hospitalaria desde 2005 a 2018 en los establecimientos relevados.

2. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Una de las problemáticas más importantes para trabajar con este dataset fue la limpieza de datos.

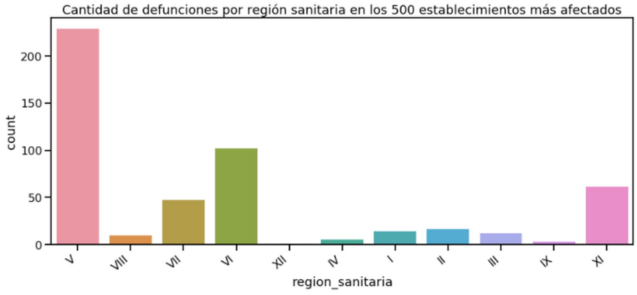
Sumado a eso, muchos establecimientos o incluso regiones sanitarias enteras no poseen información todos los años.

En primer lugar se verificó la cantidad de establecimientos según región sanitaria (Figura 2).



(Figura 2)

Luego hicimos el mismo procedimiento pero contabilizando los 500 establecimientos con mayor cantidad de muertes por año (Figura 3).



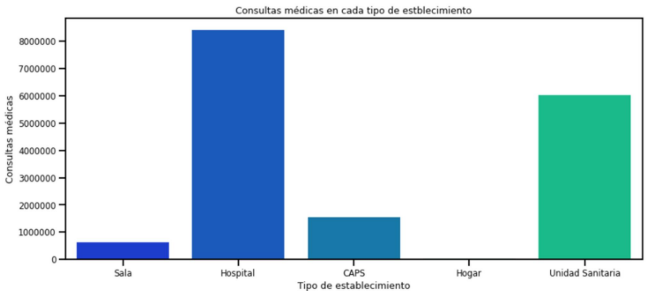
(Figura 3)

Luego logramos identificar que los establecimientos con tasa de mortalidad mayor al 50% son Hogares de Ancianos (Figura 4).

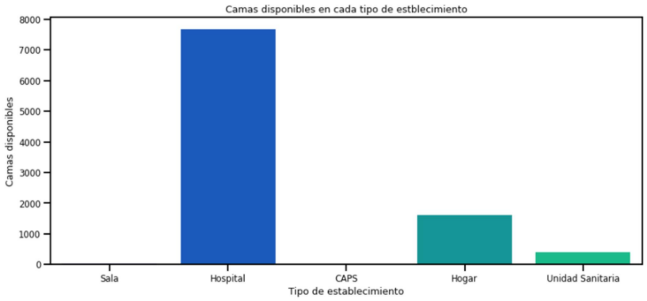
	establecimiento	region_sanitaria	defunciones	egresos
65	Hogar Municipal del Anciano	I	6.0	8.0
228	Hogar de Ancianos Dr. J. R. Lamas	II	4.0	5.0
245	Hogar de Ancianos Municipal Sto. D. de Guzmán	II	12.0	19.0
374	Hogar de Ancianos de Vedia	III	12.0	13.0
435	Hogar de Ancianos Carlos Stebb	IV	23.0	24.0
490	Hogar Geriátrico Nuestra Sra. de Luján	IV	12.0	15.0
723	Hogar de Ancianos Sarah Forrest de Cuelli	V	10.0	11.0
1204	Hogar de Ancianos Dr. Salvador Sallares	VI	10.0	13.0

(Figura 4)

Vemos que las Unidades Sanitarias están casi tan afectadas como los Hospitales en cuanto a consultas médicas relevadas (Figura 5), aunque estos últimos tienen disponibilidad de atender pacientes en sus camas instaladas. (Figura 6)

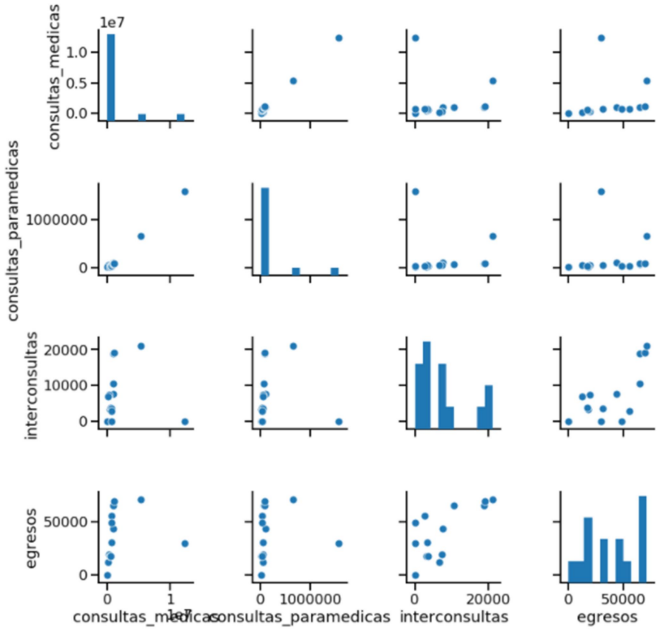


(Figura 5)

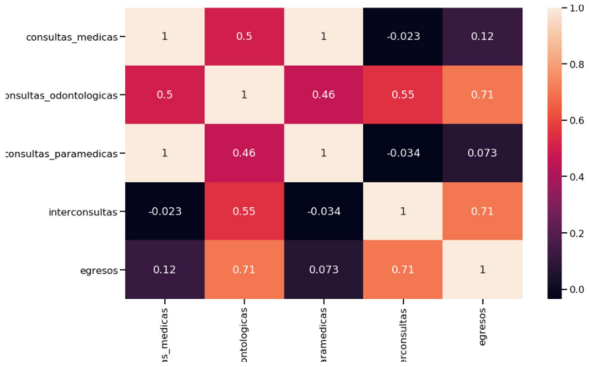


(Figura 6)

Graficamos la correlación perfecta entre consultas médicas y paramédicas, y además correlación positiva entre interconsultas y egresos. (Figuras 7 y 8).

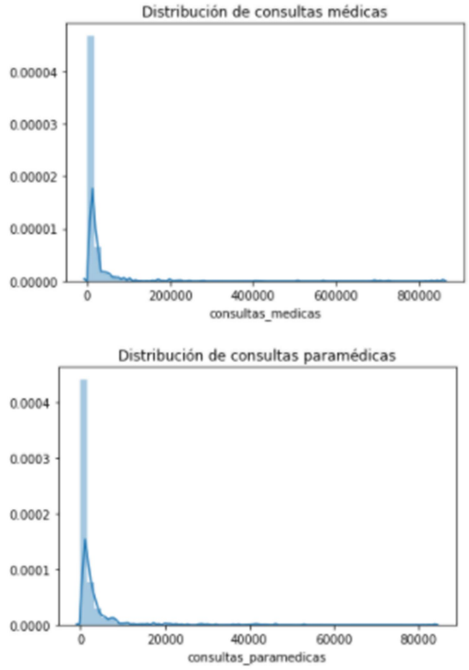


(Figura 7)



(Figura 8)

Para ello se decidió trabajar con todas las variables disponibles, y poder utilizarlas para identificar la dependencia de cada establecimiento. En la siguiente figura se aprecia la distribución de consultas médicas y paramédicas, las cuales muestran correlación perfecta.



(Figura 11)

Logramos identificar a la región sanitaria V como la más afectada por las muertes año a año (Figura 9).

		defunciones
anio	region_sanitaria	
2015	V	1530.0
2013	V	1502.0
2014	V	1477.0
2016	V	1452.0
2012	V	1439.0
2017	V	1412.0
2018	V	1311.0
2005	V	1289.0
2011	V	1127.0
2017	VIII	1124.0
2018	VIII	1077.0
2016	VIII	1053.0
2015	VIII	1014.0
2006	V	1000.0
2005	VIII	991.0

(Figura 9)

4. MODELOS DE CLASIFICACIÓN

Siguiendo con el pipeline del proyecto, se separó la variable a predecir del set de datos.

Luego se dividió el dataset en Train y Test y así entrenar a nuestros modelos con una parte de los datos y luego realizar una prueba con nuevos datos desconocidos para ellos. Luego pasamos a medir su performance a la hora de comparar la predicción con el valor real.

Se definió utilizar un modelo KNN. Este método de clasificación supervisada se basa en datos de entrenamiento para estimar la función de predicción por cada clase.

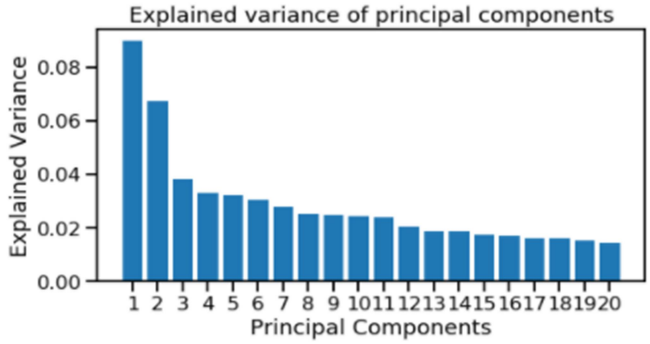
Luego se decidió contrastar la eficiencia de este modelo frente a una Support Vector Machines. Una SVM es un modelo que separa las clases en espacios lo más amplios posibles mediante un hiperplano definido como el vector entre los 2 puntos más cercanos de diferentes clases.

Los resultados fueron los siguientes:

MODELO	ACCURACY
KNN	99.45%
SVM	96.90%

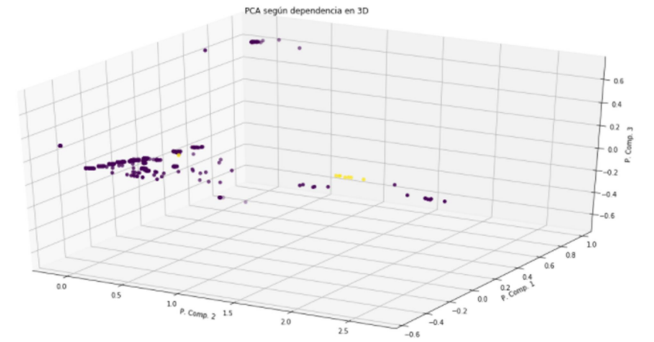
5. ANÁLISIS DE LOS COMPONENTES PRINCIPALES (PCA)

Se tomó la decisión de realizar un PCA. Esta herramienta reduce la dimensionalidad del set de datos, quedándose con las variables que más información aporten al modelo.

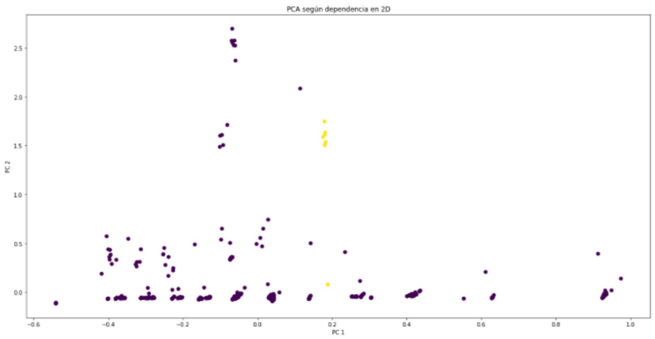


(Figura 11)

Se graficó en 3D (Figura 12) y en 2D (Figura 13), siendo la última la forma óptima de visualización.



(Figura 12)



(Figura 13)

REFERENCIAS

Información disponible en el portal de datos del Gobierno de la Provincia de Buenos Aires: https://catalogo.datos.gba.gob.ar/

RECONOCIMIENTOS

Agradezco profundamente a Martin Palazzo y su gran equipo por destinar muchas horas y brindarnos los contenidos para realizar este trabajo en una materia tan completa.

Destaco en todo el plantel una capacidad de inventiva y de generar importantes expectativas en el alumnado.

Ojalá esta experiencia haya sido una base sólida que permitan desarrollar estas habilidades en el futuro.

Por otra parte, agradezco a Ezequiel Vannucchi, mentor del proyecto, por la predisposición para orientarme y motivar la realización del proyecto.