Comportamiento de los servicios de internación de los Hospitales de la Provincia de Buenos Aires

Ciencia de Datos – UTN FRBA

Jonathan Order
146.548-0
UTN Facultad Regional Buenos Aires

Nicolás Order
152.254-1
UTN Facultad Regional Buenos Aires

Agustina Descalzo
152.287-5
UTN Facultad Regional Buenos Aires

RESUMEN

El presente trabajo tiene como objetivo realizar un análisis sobre el rendimiento de servicios de internación de los centros de salud públicos existentes en la Provincia de Buenos Aires, para así luego poder estimar mediante técnicas de regresión el porcentaje de ocupación futuro de los mismos, siendo Random Forest el más adecuado de los modelos probados. Esto permitirá luego tener información para una más acertada toma de decisiones posterior.

1. INTRODUCCIÓN

La provincia de Buenos Aires tiene 135 municipios divididos en 12 Regiones Sanitarias que cuentan con 77 hospitales públicos provinciales, 272 hospitales municipales, 5 unidades de pronta atención y 1795 centros de atención primaria para la salud, según el Censo del Sistema Único de Registro, realizado en el año 2010.

2. DATASET

El Gobierno de la Provincia de Buenos Aires posee un portal de datos abiertos de distintas áreas. Se utilizó el dataset “Rendimiento de Hospitales”, que muestra distintas métricas relacionadas con la performance de las áreas de consultorios externos, así como de las salas de internación.

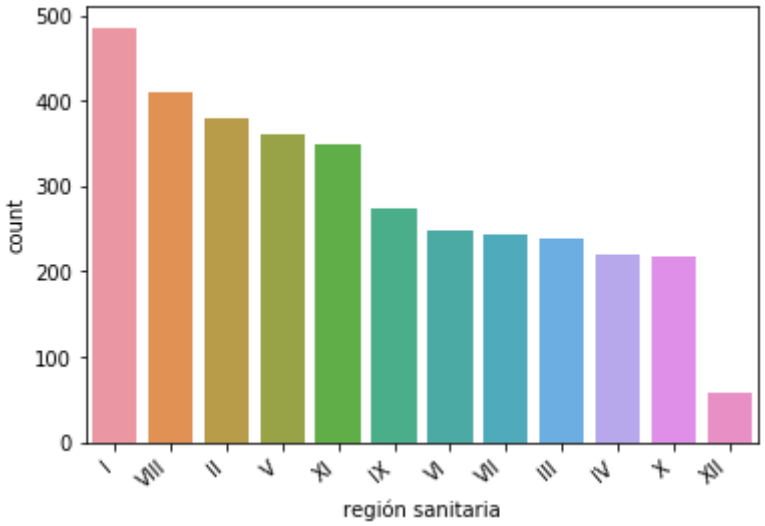
Las features con las que cuenta el dataset se explican según:

Egresos	Salida del paciente del establecimiento, sea cual sea su causa
Pacientes Día	Permanencia de un paciente internado, en días
Promedio días de estadía	Días de estadía del período, divido total de egresos del período
Porcentaje de ocupación	Total de pacientes por día dividido las camas disponibles por cantidad de días del período
Giro de camas	Total de egresos del período, dividido promedio de camas disponibles
Días de estadía	Total de días que el paciente permaneció internado
Tasa de mortalidad	Número de defunciones en el período, dividido número de egresos

3. Preprocesamiento de datos

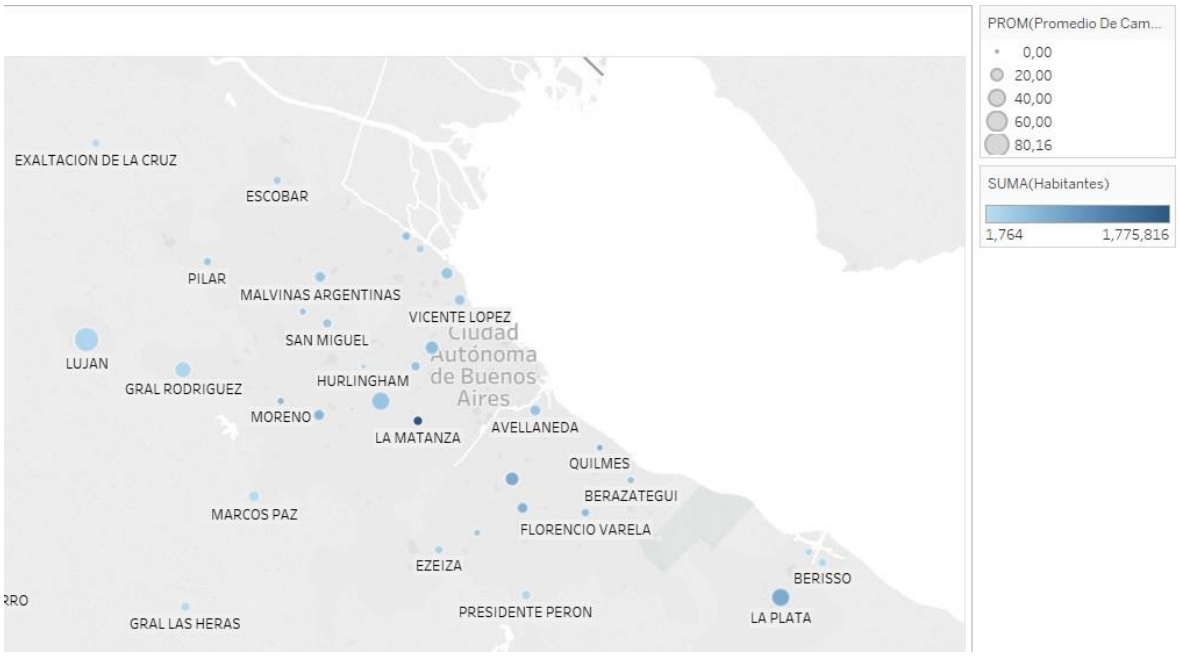
3.1. EDA

Se realizó una exploración de los datos para entender el comportamiento de las distintas features de nuestro set.



Cantidad de establecimientos según región sanitaria (CountPlot)

WordCloud de Partidos. Cuanto más grande el nombre, mayor cantidad de hospitales en el mismo.



Mapa de la Provincia y sus municipios. (Tableau)

En el último mapa se puede ver en el tamaño de cada círculo representado el promedio de camas disponibles en sus hospitales. Por otro lado, el color representa la cantidad de habitantes del municipio (dato obtenido de otro dataset de la provincia). Podemos ver cuáles son los puntos "calientes" en cuanto a necesidad de camas y disponibilidad de las mismas, cruzando el dato con la cantidad de habitantes del partido en cuestión.

Entre los puntos a destacar se encuentra La Matanza, de alta población (color) y baja cantidad de camas disponibles. Por el contrario, municipios como Luján, La Plata o Hurlingham cuentan con capacidad de camas holgada a pesar de ser populosos.

3.2. Preprocesamiento y Feature Engineering

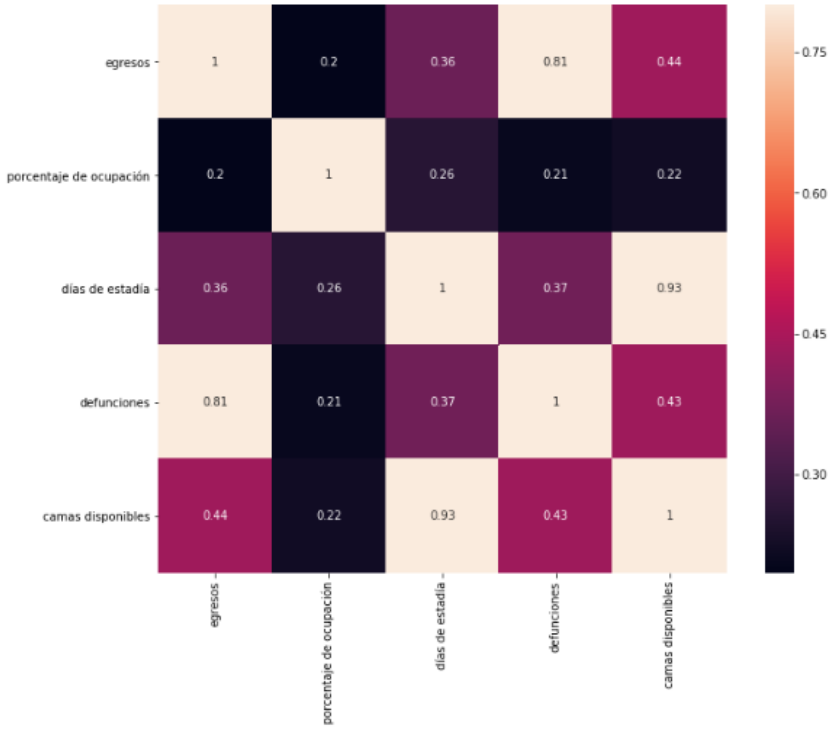
En una primera instancia, se eliminó el feature "Tasa de mortalidad hospitalaria" ya que esta se calculaba a partir de otras dos features presentes en el dataset. La feature "giro de camas", por su naturaleza de indicador calculado, fue reemplazada por una nueva denominada "camas disponibles". El objetivo era que las variables fueran independientes entre sí.

Luego se eliminaron los valores nulos encontrados en las features a utilizar en el modelo predictivo: "egresos", "promedio de camas disponibles", "porcentaje de ocupación", "giro de camas", "días de estadía", "promedio días de estadías" y "defunciones" con la condición que se cumpliera la nulidad en todos los campos en simultáneo.

Posteriormente se realizaron pair plots para visualizar en un gráfico de dos dimensiones la naturaleza de los datos y tener una primera visualización de la posibilidad de la existencia de una correlación entre las distintas features, así como visualizar los outliers. Estos, al ser reconocidos, fueron eliminados mediante la utilización de cuantiles.

"Pacientes día" y "Días de estadía" eran campos con el mismo contenido por definición, y el PairPlot arrojó una alta correlación, por lo que fue eliminada la primera.

Eliminados los outliers, se analizó la correlación entre variables y se graficó mediante el siguiente heatmap.



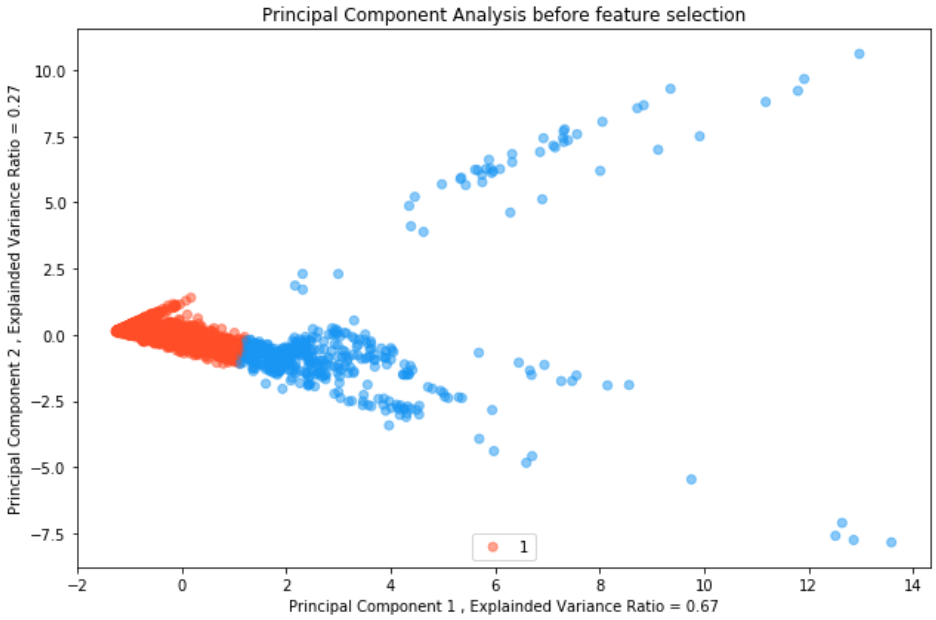
Del mismo pueden extraerse algunas conclusiones:

- Una fuerte correlación entre los días de estadía y las camas disponibles, e inversamente correlacionadas ambas dos con el porcentaje de ocupación (bastante lógico por la naturaleza de los campos en cuestión).

- Alta correlación entre egresos y defunciones, lo cual muestra que muchos de los egresos son por muerte de los pacientes y no por una recuperación.

4. Clustering con PCA y K-Means

Se realizó un PCA para observar las componentes principales que se pueden obtener en nuestro set, para luego verificar la existencia de clusters en nuestro set de datos. Se puede observar que, con las primeras dos componentes, se abarca un 94% de la variación (67% con la primera y un 27% adicional con la segunda). Luego se aplicó K-Means para 2 clusters, ya que la curva de suma de distancias cuadráticas de las muestras a su centroide más cercano se aplana a partir del tercer cluster, quedando éste con muy pocas muestras.



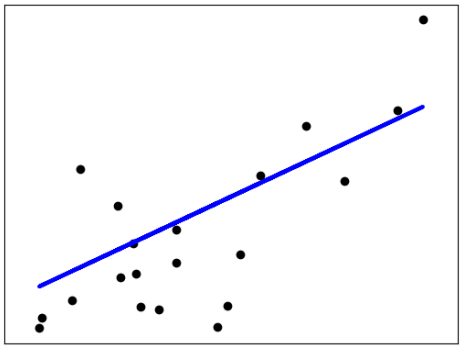
ScatterPlot de las dos principales componentes del PCA. Se pueden apreciar 2 clusters según K-Means.

5. Modelo de Regresión

Se dividio el set en 10% para test y 90% para train. Para cada uno de los modelos, se implementó primeramente Cross Validation y Grid Search, de manera de obtener mejores hiperparámetros para nuestros modelos predictivos.

Para poder estimar el porcentaje de ocupación de los servicios de internación de los hospitales de la provincia, se implementaron distintos modelos de regresión detallados a continuación, que arrojaron sus resultados.

5.1. Linear Regression



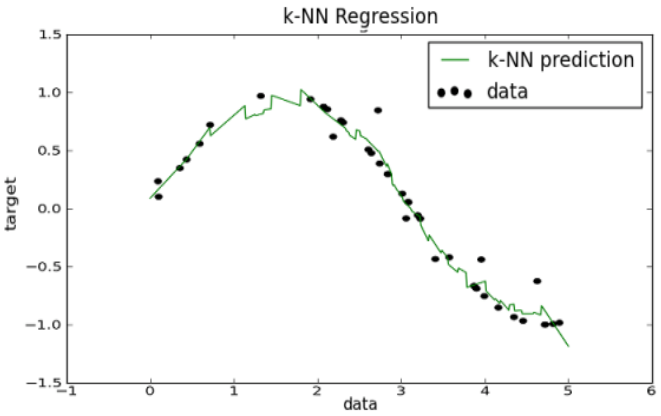
Este modelo intenta generar una función que minimice mejor la suma residual de cuadrados entre las respuestas observadas en el conjunto de datos, y las respuestas predichas por la aproximación lineal.

En una primera instancia se realizó una regresión simple, del tipo lineal, la cual arrojó un error MAE de 20.97 y una correlación de r2 = 12%

Grafico ilustrativo de como aproxima una recta a la serie de puntos de la cual surge.

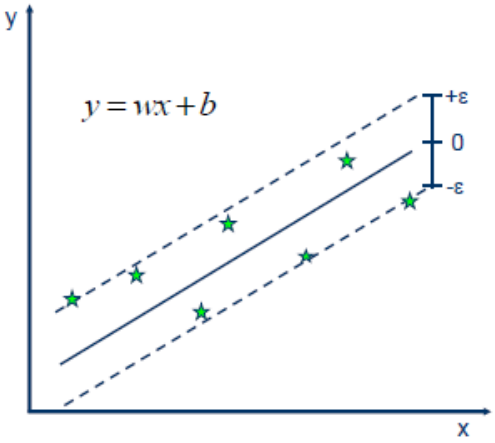
5.2. KNN Regression

Como segundo modelo de predicción, se probó realizar la regresión con KNN. Es un algoritmo simple que almacena todos los casos disponibles y predice el objetivo numérico basándose en una medida de similitud de los datos más cercanos (por ejemplo, funciones de distancia).



Se definieron 5 campos, con grid search en valores entre 1 y 5. Dando como mejor híper parámetro el valor 4. Esto arrojó un error MAE de 12,85 y una correlación r2 = 67%.

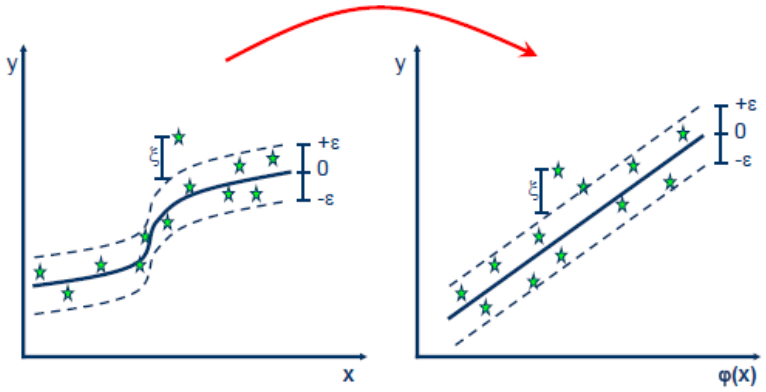
5.3. Support Vector Regression, con Kernel rbf



SVR utiliza los mismos principios que la SVM para la clasificación, con solo unas pocas diferencias menores. En primer lugar, dado que la salida es un número real, se vuelve muy difícil predecir la información disponible, que tiene infinitas posibilidades. En el caso de la regresión, se establece un margen de tolerancia (épsilon) en aproximación a la SVM. Sin embargo, la idea principal es siempre la misma: minimizar el error, individualizar el hiperplano que maximiza el margen, teniendo en cuenta que se tolera parte del error.

Grafico del hiperplano formado tomando como parámetro épsilon

Para este caso, fue aplicado un kernel Rbf. La aplicación de kernels permite transformar problemas no lineales, en lineales y a partir de ello estimar nuestra regresión.

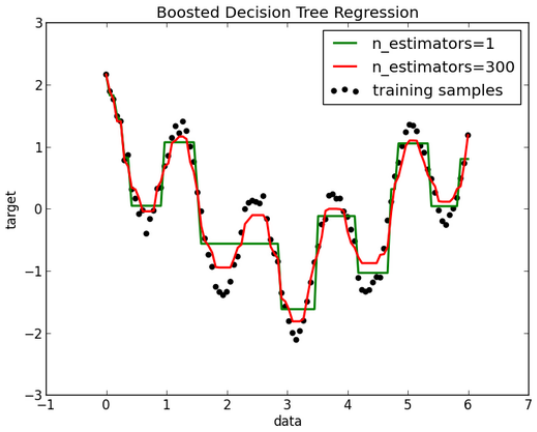


Además, como híper parámetros para este análisis, tomamos:

- C → valores entre 1 y 500. Obteniendo como mejor parámetro el número 285.
- Gamma → valores entre 1 y 50. Obteniendo como mejor parámetro el número 18.

Los resultados obtenidos fueron un MAE de 13.66 y un coeficiente de correlación r2 = 62%.

5.4. Random Forest



En un árbol de regresión, dado que la variable de destino es un número de valor real, ajustamos un modelo de regresión a la variable de destino utilizando cada una de las variables independientes. Luego, para cada variable independiente, los datos se dividen en varios puntos de división. Calculamos la suma del error cuadrado (SSE) en cada punto de división entre el valor predicho y los valores reales. La variable que da como resultado el SSE mínimo se selecciona para el nodo. Luego, este proceso continúa de forma recursiva hasta que se cubren todos los datos.

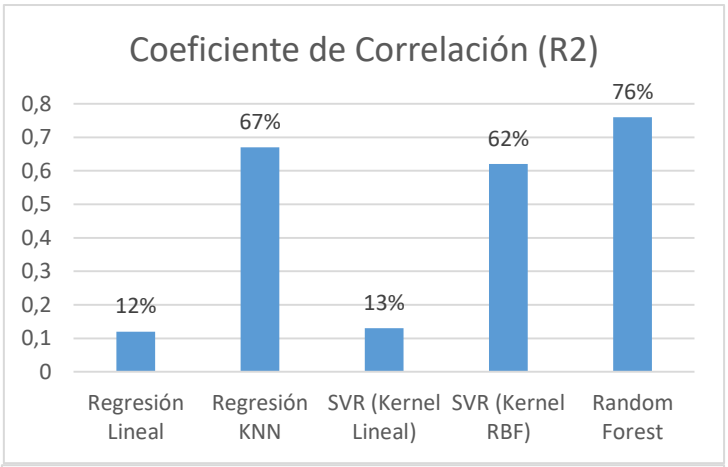
Y como resultado final de esta estimación, tuvimos como resultado un MAE de 10.97 y un r2 = 76%.

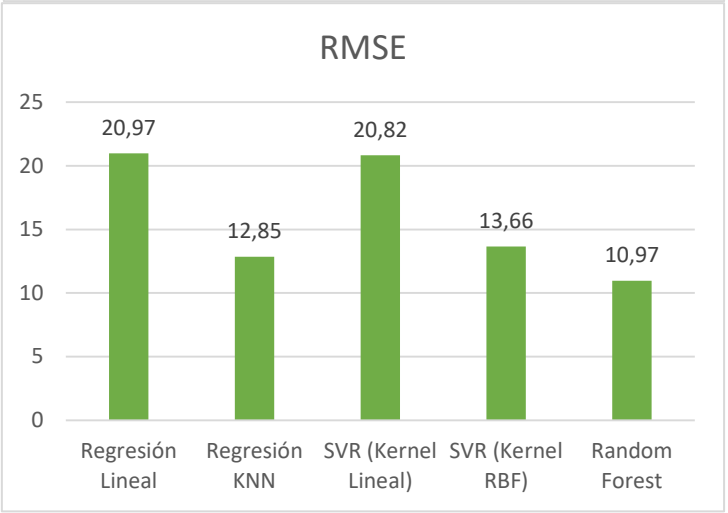
6. Conclusión

6.1 Resultados

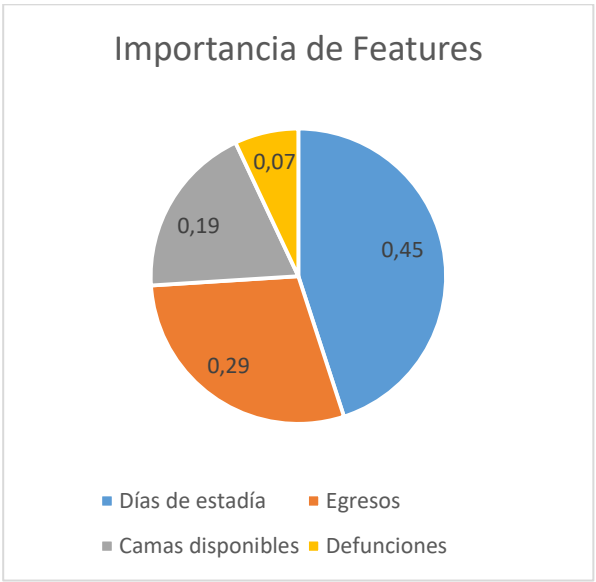
En el objetivo de poder predecir el porcentaje de ocupación de un determinado hospital en base a las features dadas, se puede ver que los modelos regresores lineales tuvieron mala performance, tanto en la baja correlación como el alto error. Los modelos no lineales tuvieron mejores resultados, destacándose Random Forest como el más adecuado para este caso.

A nivel beneficio de la predicción de la ocupación, es de mayor utilidad realizarlo en aquellos partidos con mayor densidad poblacional y menor disponibilidad de camas, como por ejemplo La Matanza, mencionada anteriormente. Esto permite una mejor distribución de los recursos, la posibilidad de planear ampliaciones en los hospitales, planes de derivación al partido más cercano con el menor porcentaje de ocupación predicho. A su vez, es una herramienta de gran utilidad en cuanto a planificación de infraestructura necesaria para soportar el sistema de salud en cada región.





En cuanto a Random Forest, la contribución de las features a la estimación del porcentaje de ocupación se ve en el gráfico de torta siguiente.



6.2 A mejorar a futuro

Los datos del dataset utilizado presentaban de manera “cruda” 25.000 registros, de los cuales sólo 3.000 estaban completos en sus datos. Creemos que con una mejor carga en el futuro las predicciones se tornarán más correctas.

Por otro lado, la utilización de datos de otros datasets podría colaborar para extender nuestro análisis a otros sectores del centro de salud más “críticos”, como los consultorios externos que presentan alta rotación de pacientes y necesidades diferentes que la internación.

REFERENCIAS

[1] Provincia de Buenos Aires – Rendimientos de establecimientos públicos de salud
http://catalogo.datos.gba.gob.ar/dataviews/247005/rendimientos-de-establecimientos-desalud/