"Aplicación de modelos de Clasificación por peso sobre los nacidos vivos en Argentina"

Gomez Lucas, Sabini Manuel, Tudanca Valentín

lucas.sgomez@live.com; manusabi13@gmail.com; vatudanca@gmail.com

1. Introducción

El presente trabajo busca realizar un análisis sobre los datos relevados por la Dirección de Estadística e Información en Salud (DEIS), más precisamente sobre el relevamiento de lo nacidos vivos registrados en el año 2018 en todo el territorio Argentino.
Se realiza un análisis exploratorio de datos buscando información relevante del dataset en cuestión y luego se ponen en práctica tres modelos de clasificación para predecir una característica del peso del recién nacido.
Todos los desarrollos fueron realizados en Python, a través de Jupyter Notebook.

2. Dataset

Para el análisis, se trabajará con el Data Set de "Nacidos vivos registrados en 2018", el cual tiene la siguiente información:

- Jurisdicción de residencia
- Tipo de Parto
- Sexo
- Edad de la madre
- Semanas de gestación
- Instrucción de la madre
- Intervalo de pesos al nacer (en gramos)
- Cantidad de nacimientos

Con un total de 30.099 muestras en las que se indican la cantidad de nacimientos agrupados por características, en distintos rangos.

3. Preprocesamiento

Se comenzó visualizando la cantidad y los tipos datos con los que se trabajaría. A fin de depurar la base de datos se verificó si contaba con valores nulos, pudiendo observar que si bien no existían campos vacíos había datos del tipo "Sin especificar". Para conocer que tanto repercutía en el análisis se Obtuvo la siguiente tabla:

Features	% de valores sin especificar
Jurisdiccion	2,15%
Tipo de parto	0,66%
Sexo	2,19%
Edad de la madre	2,51%
Semanas de gestacion	6,01%
Instrucción de la madre	8,86%
Intervalos de peso al nacer	3,80%

Cada muestra corresponde a un grupo de nacimientos con las características dadas. El total de nacimientos con dichas características se indica en "Cantidad de Nacimientos".

Con el fin de que cada muestra indique un recién nacido se agregaron tantas samples como nacimientos. Pasando de un Dataset de 30.999 muestras a uno de 699.486.

4. Métodos

❑ *Support Vector Machine (SVM)*

Este modelo consta de un clasificador lineal, que busca el hiperplano separador que maximiza el margen entre clases, siendo cada muestra mal clasificada penalizada por una función de costo C.

❑ *Logistic Regression*

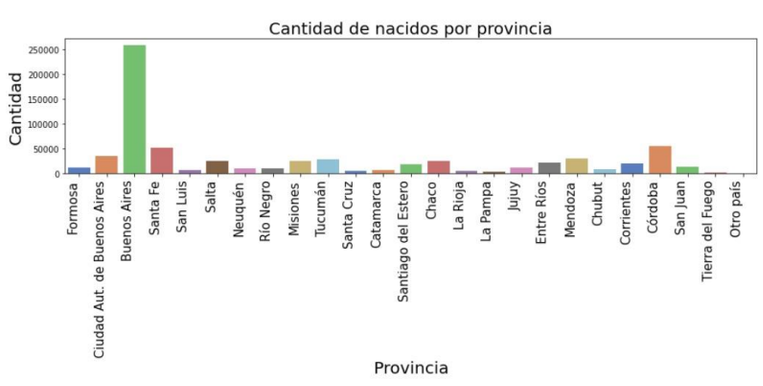
Es una regresión lineal precedida de una función de activación sigmoide. A cada muestra clasificada le asigna una probabilidad de pertenecer a cada clase existente en el problema. Si esta probabilidad es mayor a un cierto valor, entonces pertenece a una clase y viceversa.

❑ *KNN Classification*

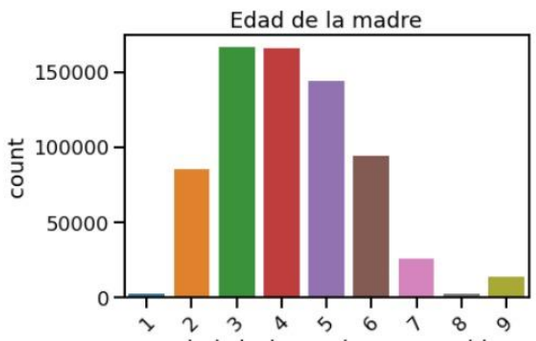
Este modelo es un clasificador lineal en el que la decisión se toma en función de las etiquetas de los "k" puntos vecinos. Se calculan las distancias de cada nuevo punto al resto, se encuentra los "k" vecinos más cercanos y en función de ello asigna la nueva etiqueta.

5. Análisis Exploratorio de Datos

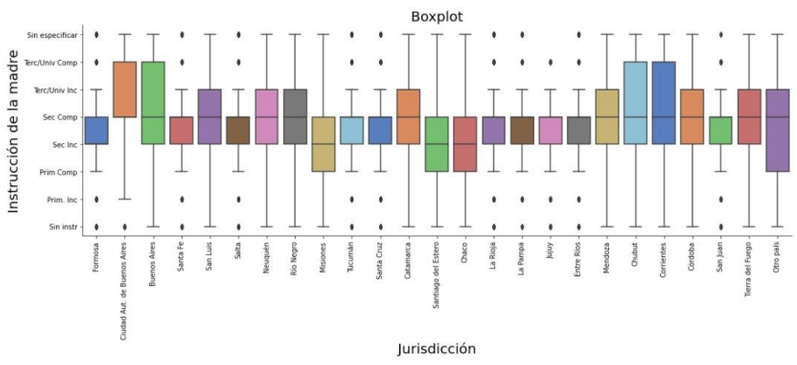
En primera instancia se visualizaron la cantidad de nacimientos por provincia, evidenciado que la gran mayoría de ellos se llevan a cabo en Buenos Aires Provincia:



Haciendo foco en las edades de las madres, podemos observar que la mayor cantidad de nacimientos se produce en el rango de entre 20 y 29 años, cayendo en el siguiente rango, entre 30 y 34 años.



Segmentando por la instrucción de la madre, se representó en un boxplot la tendencia en este sentido para cada una de las jurisdicciones, resultando destacable en este caso que si bien la gran mayoría de los datos se concentran en secundaría completa o incompleta, hay tres provincias en las que el 50% de las madres aun no completaron sus estudios secundarios. Se trata de Misiones, Chaco y Santiago del Estero.
A su vez Misiones presenta la mayor cantidad de madres que no han recibido ningún tipo de educación formal.

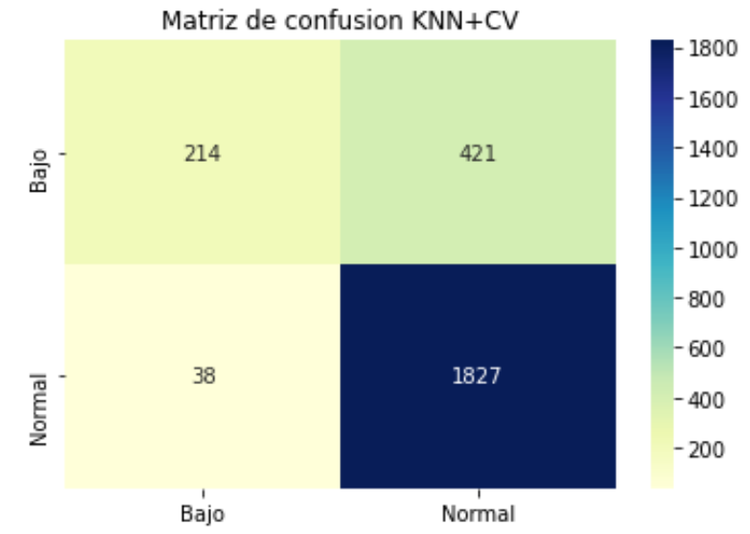


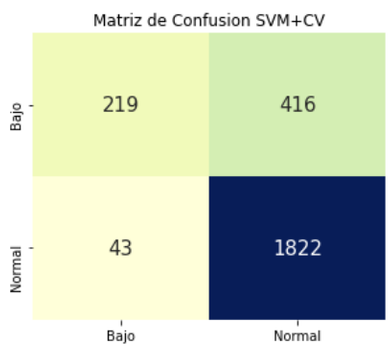
Se confirmó una correlación fuerte entre las semanas de gestación y los intervalos de pesos al nacer, lo que nos dio el puntapié inicial para iniciar la predicción.

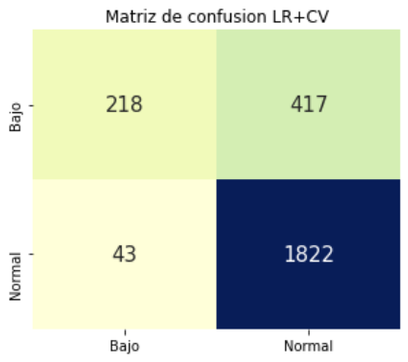
6. Resultados

La métrica utilizada para determinar el mejor modelo fue la precisión de la etiqueta "Peso Bajo", puesto resulta crítico que prediga mejor este valor que el de un recién nacido con peso normal.

Se entrenó con el 50% de la muestra. Si bien, por la forma de los datos los tres modelos resultaron similares a la hora de clasificar, el KNN para k=25, resultó ser levemente superior en la precisión sobre las predicciones de peso bajo, con un 85% siendo el accuracy de 82%.







7. Conclusiones

Podemos concluir que, con la base de datos de los nacidos vivos en Argentina durante 2018, y utilizando la información de las variables antes mencionadas es posible predecir si los recién nacidos tendrán un peso superior o inferior a los 3kg, con una precisión del 83%.