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

Gomez, Lucas Sebastián; Sabini, Manuel; Tudanca Valentín Alejandro

Universidad Tecnológica Nacional - FRBA

***Abstract* – En el siguiente trabajo se utilizan diferentes modelos de clasificación, aplicados sobre un dataset que contiene información sobre los nacidos vivos en Argentina durante 2018, con el fin de establecer si los neonatos tienen un peso superior o inferior a los 3kg.**

Keywords—SVM, Logistic Regression, KNN, Natalidad, Saluds)

I. 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 trata de datos públicos, relevados por diferentes organismos que el gobierno de la Nación pone a disposición de la comunidad.

La cantidad de nacimientos, las condiciones en que se llevan a cabo los mismos y las características tanto del neonato como de la madre, resultan estadísticas de interés que permiten evaluar políticas públicas con foco en la salud y en la sociedad. En el año 2018 hubo más de 700.000 nacimientos en Argentina [1].

A lo largo del presente informe 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.

II. DESCRIPCIÓN DEL 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.

Cada columna se encuentra duplicada, expresando el elemento por su nombre y por un "id" asignado por el creador del dataset.

III. ANÁLISIS EXPLORATORIO DE DATOS

A. Pre Procesamiento

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:

	Column_name	% valores Sin especificar
0	jurisdiccion_de_residencia_nombre	2.15
1	tipo_de_parto_nombre	0.66
2	sexo_nombre	2.19
3	edad_de_la_madre_grupos_nombre	2.51
4	semanas_de_gestacion_nombre	6.01
5	instruccion_de_la_madre_nombre	8.86
6	intervalos_de_peso_al_nacer_nombre	3.80

Se determinó eliminar los datos que no indican la jurisdicción de residencia y el tipo de parto.

Dado que cada línea representa una combinación de características y en la última columna se representa la cantidad de neonatos que cuentan con las mismas, se decidió descomponer el dataset de forma tal que cada sample contenga la información de un solo nacimiento. Para esto utilizamos la siguiente función:

```
[ ]  for j in range(samples):
         registro_a_repetir = nacidos_prep.iloc[j, :]
         n=nacidos_prep.iloc[j, n_col]-1
         nacidos_prep=nacidos_prep.append(([registro_a_repetir]*n), ignore_index=True)
```

De este modo se pasó de tener 30.099 muestras a tener 699.486 cada una de las cuales representa un recién nacido.

B. Estadísticas descriptivas

Apoyado en la creación de distintos tipos de gráficos se obtuvieron algunas características descriptivas de los datos tratados. 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:

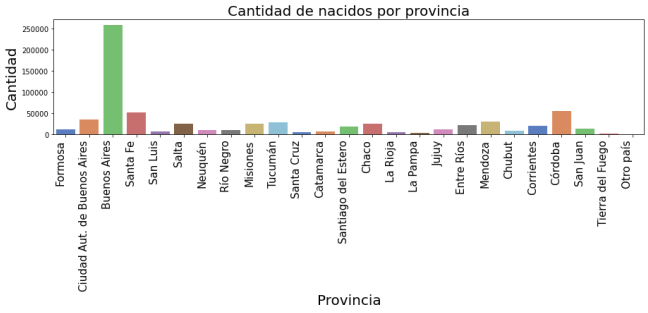

Figura 1. Cantidad de nacidos por Provincia

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.

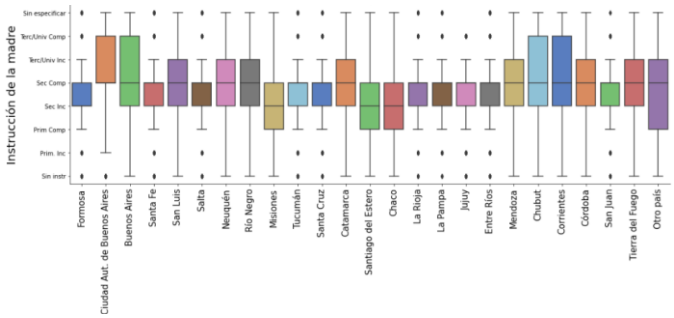

Figura 2. Cuartiles instrucción de la madre por provincia.

Por otro lado se pudo observar que la mayor cantidad de madres sin instrucción educativa se concentran en las provincias más pobres de nuestro país, compartiendo lugar con provincia de Buenos Aires:

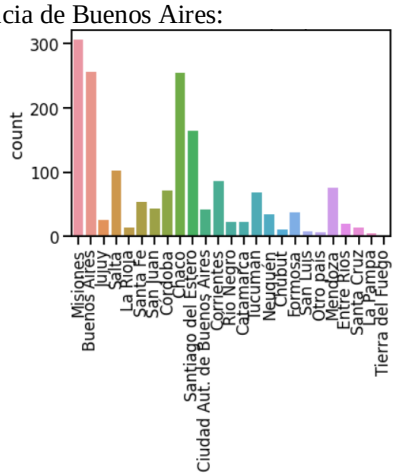

Figura 3. Madres sin Instrucción por provincia

Apoyándonos en el trazado de un heatmap se observó la correlación lineal entre las distintas features del dataset estudiado, confirmado la mayor correlación entre los rangos de las semanas de gestación y el peso de los neonatos. Esta información será de utilidad en adelante.

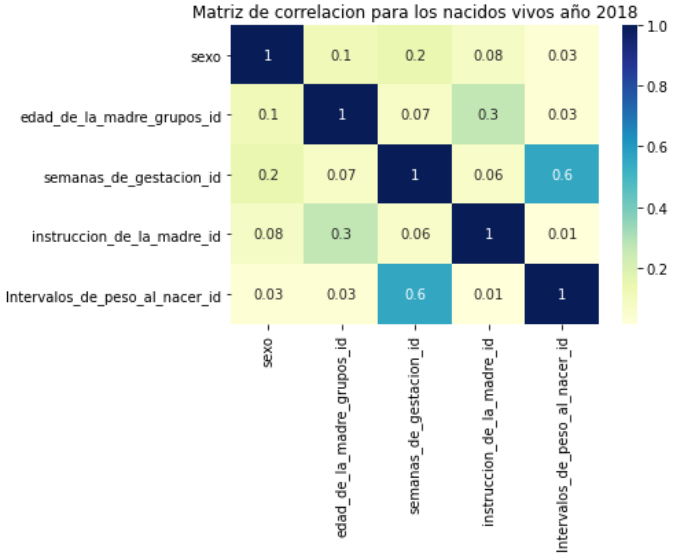

Figura 4. Correlación líneas entre pares de Features

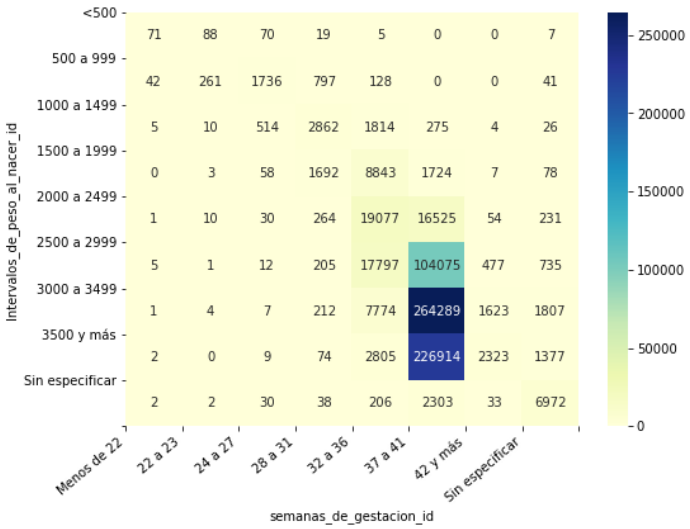

Figura 5. Correlación lineal entre intervalos de peso y semanas de gestación

IV. MACHINE LEARNING

A. *Objetivos*

Luego de realizado el EDA, se propuso la realización de Modelos de Machine Learning, particularmente de aprendizaje supervisado, con la información que cuenta el Dataset. Con el modelo SVM se obtuvieron los mejores resultados. Se buscó:

- Predecir si el recién nacido pesará más o menos de 3kg en función de:
  - Edad de la madre
  - Semana de gestación
  - Instrucción de la madre.
  - Sexo
  - Tipo de parto (simple o múltiple)

B. *Modelo SVM*

Es un conjunto de algoritmos de aprendizaje supervisado que se aplica para modelos de clasificación o regresión. Un modelo de SVM es un modelo que representa los puntos de muestra en el espacio separando las clases mediante un hiperplano definido como vector entre dos puntos [2][3]. Este método genera el hiperplano que maximiza la

distancia entre las muestras. Cuando nuevas muestras se introducen en el modelo el mismo predecirá la clasificación de la etiqueta según el lado que ocupen desde el punto de vista del hiperplano.

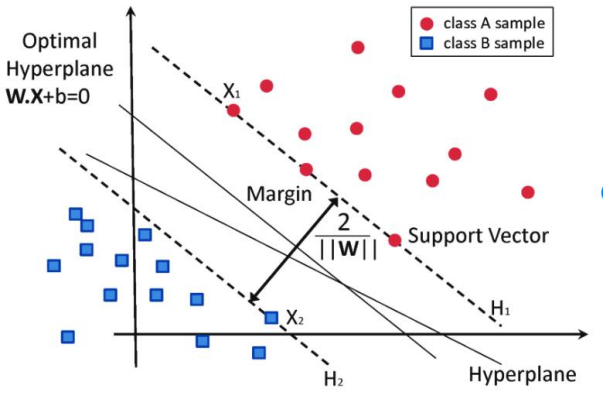

Figura 6. Support Vector Machine [4]

C. KNN Clasifier

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.

D. Regresión Logística

Se trata de una regresión lineal que es seguida de una función logística o sigmoide que es la que generará la salida binaria:

$$p(\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}, \qquad [5]$$

E. Predicción del peso de los bebes

En el desarrollo del estudio se utilizaron distintos modelos de aprendizaje supervisado. Tantos modelos de Support Vector Machine (SVM), K-Nearest Neigbors (KNN) y Logistic Regresion (LR), todos contenidos en la librería de Scikit-Learn.

Debido a la enorme cantidad de samples del dataset generado (+650.000) se decidió tomar una muestra aleatoria de 5000 Samples para entrenar los modelos, los tiempos de computo con el dataset original eran excesivamente grandes.

Se retiraron las features que indicaban los id's de las categorías quedando sólo aquellas que indicaban los nombres de las mismas. Para la feature que indica el peso del recién nacido, mediante un LabelEncoder, se asignó la categoría en la, que se encuentra el peso en Bajo o Normal, considerando como niños que tiene bajo peso aquellos que pesan menos de 3Kg al nacer. Luego de aplicada esta clasificación se encontró que el 26% de los niños que nacen tienen bajo peso. Para el resto de las features se generaron Dummies para comenzar a entrenar, dado que todas ellas eran variables categóricas.

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.

Para contrastar esto gráficamente se utilizó una matriz de confusión representada a través de un heatmap para cada uno de los modelos entrenados.

F. Resultados

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%.

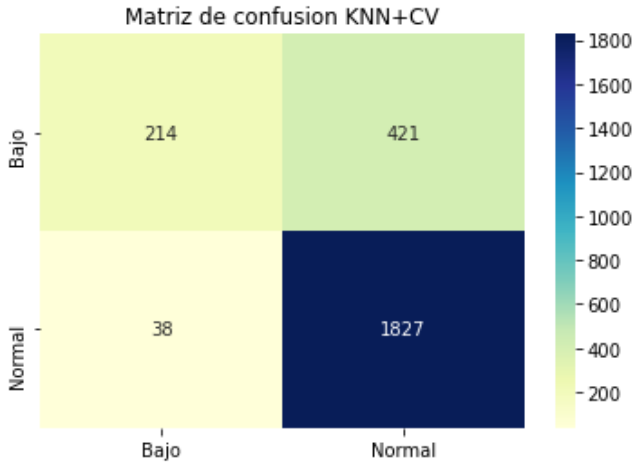

Figura 7. Matriz de Confusión - KNN

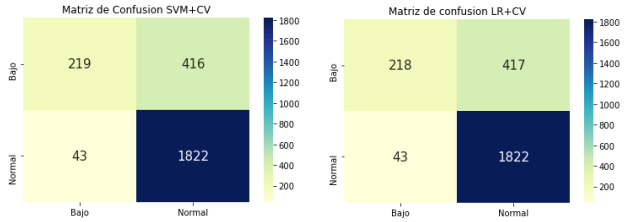

Figura 8. Matriz de Confusión – SMV | Logistic Regression

De la matriz de confusión obtenida el valor más crítico es el de los 38 neonatos que tienen bajo peso y son etiquetados como de peso normal. A continuación las métricas de clasificación para el mejor método:

	precision	recall	f1-score	support
0	0.85	0.34	0.48	635
1	0.81	0.98	0.89	1865
accuracy			0.82	2500

V. 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 85%.

Se estima que para mejorar el modelo se podría relevar información de las madres, tal como el estado socioeconómico y sus características físicas. Por otro lado, contando con los mismos datos, se estima que las predicciones mejorarían teniendo información que no sea agrupada por rangos.

Si bien escapa al alcance de este trabajo, a futuro resultaría beneficioso utilizar el mismo set de datos para trabajar con información relativa a la educación de la madre. Durante el

análisis exploratorio se observó una disparidad significativa en esta variable en las diferentes regiones del país.

REFERENCIAS

[1] Ministerio de Salud de la Nación, Secretaría de Coberturas y Recursos de Salud, Subsecretaría de Coberturas Públicas Sanitarias, Dirección Nacional de Información en Salud, Dirección de Estadísticas e Información en Salud, ISSN: 0301-4630 Boletín Número 157 Buenos Aires, Junio de 2018

[2] Fung, G. M., & Mangasarian, O. L. (2005). Multicategory proximal support vector machine classifiers. Machine learning, 59(1-2), 77-97.

[3] Albanese, D., Visintainer, R., Merler, S., Riccadonna, S., Jurman, G., & Furlanello, C. (2012). mlpy: Machine learning python. arXiv preprint arXiv:1202.6548.

[4] García-Gonzalo, Esperanza & Fernández-Muñiz, Zulima & Garcia Nieto, Paulino Jose & Sánchez, Antonio & Menéndez, Marta. (2016). Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers. Materials. 9. 531. 10.3390/ma9070531.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] Nong Ye (2003). *The handbook Data Mining*, Mahwa, New Jersey