Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE CIENCIAS EXACTAS Y NATURALES
FACULTAD DE INGENIERÍA

Aprendizaje Automático
Trabajo Práctico N°1

GRUPO N° 10

Amena, Santiago
Opazo, F. Ayelén
Rivero González, Leandro

EQUIPO DOCENTE

Cotik, Viviana
Henrión, Guillermo
Bujía, Gastón
Pepino, Leonardo

Modelo de predicción de accidente cerebrovascular

Amena, Santiago; Opazo, F. Ayelén; Rivero Gonzalez, Leandro

Mayo 2021

1. Resumen

El presente trabajo tuvo por objetivo predecir qué pacientes son más propensos a sufrir un ataque cerebrovascular (*stroke*). El mismo se desarrolló a partir de un modelo de árbol de decisión entrenado en base a atributos de personas que sufrieron o no este tipo de lesión. Las cuatro variables más asociadas al evento fueron *age*, *heart_disease*, *avg_glucose_level* e *hypertension*. Para evaluar el modelo se utilizó la métrica $f\beta - score$, con $\beta = 2$. El modelo con mejor performance hallado particionó por *age* en las 2 primeras capas del árbol, siguiendo por *avg_glucose_level*, *bmi* y *smoking_status*. Si bien los resultados del f2score son relativamente bajos (alrededor de 0.4), es posible introducir mejoras si se utiliza un ensamble de árboles en lugar de un único árbol de decisión.

2. Introducción

Bajo el objetivo de predecir qué paciente va a sufrir un accidente cerebrovascular, se describen a continuación los tipos de variables que componen el dataset y las metodologías utilizadas para su preprocesamiento. A su vez, se muestra el análisis de correlaciones entre los atributos y los criterios utilizados para la selección de las variables predictoras del evento. Se presentan, además, las estrategias para el balanceo de los datos y la medida de performance utilizada para la implementación y evaluación del algoritmo en distintos conjuntos de entrenamiento y validación. Se presenta el árbol de decisión con mejor performance encontrado y su correspondiente análisis y se prueban distintas profundidades de los árboles, con y sin poda, evaluando en el conjunto de evaluación aquel que maximiza la performance en el conjunto de validación. Finalmente, se realiza el reentrenamiento del árbol sin poda y se compara la performance en el conjunto de prueba con los modelos anteriores.

3. Datos

El dataset utilizado está compuesto de 5.110 observaciones de características de personas que han padecido o no han padecido un accidente cerebrovascular. El mismo está compuesto por 11 atributos, de los cuales 3 son numéricos (*age, avg glucose level* y *bmi*) y el resto categóricos: *gender, work type, residence type, smoking status, ever married, hypertension, heart disease* y, finalmente como variable target, si ha padecido o no un ACV (*stroke*). De estas últimas, las cuatro primeras son nominales [1], mientras que las 4 últimas son binarias. En la tabla 1 se observan las primeras filas del dataframe generado.

El dataset posee valores faltantes en la variable *bmi* (índice de masa corporal), mientras que *smoking status* (si la persona fumó alguna vez) posee valores categorizados como *unknow*. En el primer caso hay un 4 % de datos faltantes y en el segundo un 30 %. A su vez, los datos se encuentran desbalanceados, siendo las personas con *stroke* solo un 4 % del total de la muestra.

[1] Si bien *smoking status* podría pensarse como una variable ordinal, uno de sus valores es *unknow*, por lo cual no es posible asegurar un orden inherente en el atributo

4. Metodología

4.1. Imputación de valores faltantes

Como se explicó anteriormente, la variable *bmi* contiene valores faltantes, por lo cual se optó por realizar una imputación de los mismos. Para esto se analizaron las correlaciones con cada una de las variables restantes, con el fin de identificar los posibles predictores. Además, para el caso de las variables numéricas se analizaron scatterplots para identificar posibles relaciones no lineales. En el caso de *age*, se observó una relación positiva pero decreciente con *bmi*, con lo cual se creó la variable *age* al cuadrado para tener en cuenta esta no-linealidad en el modelo. Finalmente se probaron dos modelos diferentes (regresión lineal y K-nearest neighbors) con diferentes parámetros. Los resultados para ambos modelos se muestran en la figura 1. Se puede observar que el modelo lineal tiene una mejor performance. Sin embargo, debido a que KNN a partir de 25 vecinos tiene una performance similar, y considerando que el modelo lineal introduciría colinealidades en las variables afectando al árbol de decisión, se optó por utilizar KNN.

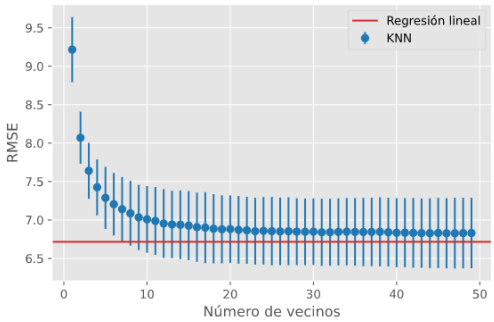

Figura 1: Error medio y desvío de *bmi* para diferentes números de vecinos y error medio de la regresión lineal.

4.2. Entrenamiento de árbol de decisión

Para el entrenamiento del árbol se realizó, en primer lugar, una exploración de las variables independientes y de sus correlaciones con *stroke*. Dichas correlaciones se muestran en la tabla 2. Dado que la correlación lineal suele no ser una buena medida de asociación entre variables categóricas, se realizaron tablas de contingencia con sus respectivos p-valores de tests de independencia; los cuales se muestran de la tabla 3 a la 9.

De acuerdo con el análisis de correlación y el de independencia, las cuatro variables más asociadas al evento de lesión son *age*, *heart_disease*, *avg_glucose_level* e *hypertension*. En la figura 5 se observan, para las variables contínuas, sus funciones de densidad según si tuvieron o no *stroke*; mientras que en la figura 6 se aprecian los porcentajes de ACV para las poblaciones con una enfermedad de base (izq: *heart_disease*, der: *hypertension*) vs. aquellas que no la tienen.

Luego de analizar cada variable por separado, se dividió al dataset en conjuntos de entrenamiento y test, y se entrenó un árbol de decisión base sobre el conjunto de entrenamiento. Partiendo de este modelo, se utilizó un algoritmo de *grid search* con cross-validation de 10 folds para buscar el mejor criterio de decisión y la profundidad máxima. Para validarlo se utilizaron dos enfoques: por un lado, se armaron 50 conjuntos de entrenamiento y validación aleatorios con proporción 80/20; por otro, se utilizó *crossvalidation* de 50 folds.

La métrica utilizada para la validación fue $f\beta - score$ con β=2, dado que por las características del problema se requería darle mayor peso al porcentaje de lesiones que fueron correctamente identificadas que al porcentaje de diagnósticos acertados. Los resultados se muestran en los boxplots de la figura 2. En ella se muestra la mayor varianza del segundo método respecto al primero. Esto se corresponde con que este método tiene un análisis más amplio de los resultados, ya que que todos los datos son utilizados para testear en algún momento. Debido a esto es que se utiliza el método de k-fold para evaluar los modelos.

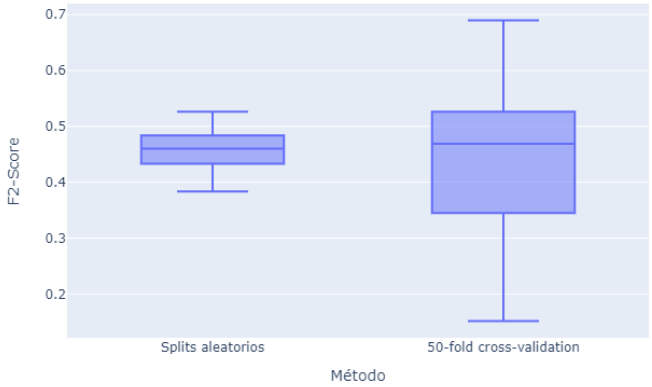

Figura 2: *Boxplot* de resultados de validación para la métrica de $f2 - score$, a la izquierda se observa la misma para 50 conjuntos de entrenamiento tomados al azar, a la derecha usando la técnica de $50 fold$.

5. Resultados

Luego del proceso de validación se continuó el estudio con el mejor modelo encontrado. El gráfico del mismo se muestra en la figura 7. Se puede observar que la variable de mayor importancia es *age*. En primer lugar, el árbol distingue entre los mayores y menores de 53 años. Para el primer grupo, el porcentaje de pacientes que tuvieron un ACV es de 14 %, mientras que en el segundo es de 72 %. Luego, el algoritmo divide a cada grupo nuevamente de acuerdo con la edad, distinguiendo cuatro grupos. Finalmente, estos son divididos a partir del nivel de glucosa, índice de masa corporal y si fumó en algún momento.

A partir del modelo obtenido, se consideraron diferentes variaciones del parámetro de poda α, a través de la implementación de *grid search*. Se realizó un recorrido entre 0 y 0.05, con 100 pasos intermedios, y se obtuvieron los resultados que se muestran en la figura 3. En la misma se observa la performance de $f2-score$ para el conjunto de entrenamiento y validación para los diferentes alfas. El valor con el que se maximiza la performance en el conjunto de validación es 0.0145. Por otra parte, en la figura 4 se puede observar la profundidad del árbol en función del parámetro α. A medida que este se incrementa la profundidad es menor, evitando de esta manera el *overfitting*. Sin embargo, este modelo obtuvo una performance menor al árbol sin poda y con máxima profundidad de 3. En la tabla 11 se muestran los resultados de ambos modelos.

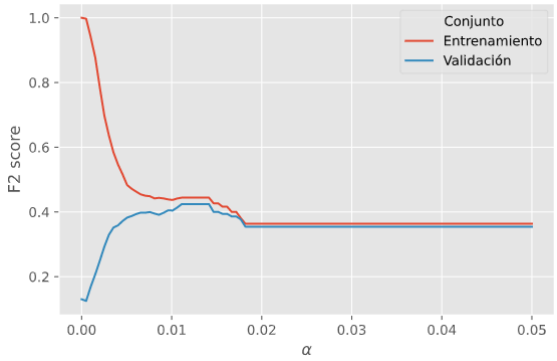

Figura 3: F2 score en función del α.

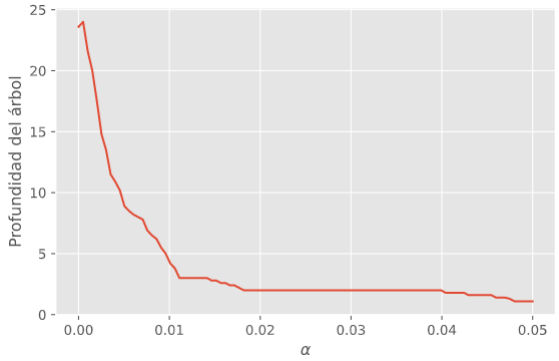

Figura 4: Profundidad del árbol en función del α.

Para el árbol sin poda pero con profundidad máxima, se realizó una tabla con los atributos más importantes, de acuerdo con la técnica de eliminación recursiva. El resultado se observa en la tabla 10. Los

atributos más importantes fueron *age*, *avg_glucose_level* y *smoking_status_unknown*. Finalmente, se reentrenó el árbol solo con los tres atributos más importantes como variables predictoras, y se obtuvieron los resultados mostrados en la tabla 11 en el conjunto de prueba. En la misma se observa que el modelo que mejor ajusta a los datos es el de máxima profundidad 3 y $\alpha = 0$.

6. Conclusiones

El presente trabajo tuvo como objetivo predecir qué paciente va a sufrir un accidente cerebrovascular a partir de un dataset de características de los pacientes. El análisis realizado encontró que la edad, el nivel de glucosa, el índice de masa corporal y si la persona fumó son predictores de lesión.

A lo largo del estudio se entrenaron diferentes modelos con distintos hiperparámetros, y se obtuvo el mejor modelo con el criterio de entropía y profundidad máxima de 3 nodos.

Si bien los resultados del $f2 - score$ son relativamente bajos, es posible introducir mejoras si se utilizara algún ensamble de árboles. Tal como se observa en la figura 2, el árbol de decisión es un algoritmo de alta varianza[2], con lo cual, utilizando una suficiente cantidad de árboles y promediando los resultados (*bagging*) o entrenando árboles en forma secuencial (*boosting*) es probable que se logre mejorar la performance del modelo.

7. Referencias

- Alpaydin, Ethem. *Introduction to Machine Learning* (2010). Massachusetts Institute of Technology.
- James, Witten, Hastie & Tibshirani. *An Introduction to Statistical Learning with Applications in R* (2015)
- Mitchell, Tom M. *Machine Learning* (1997).

8. Anexo

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Tabla 1: Dataset de variables relacionadas con la salud y en particular *stroke* (*target*).

Variable	Coef. de correlación
age	0.245
avg glucose level	0.132
bmi	0.042

Tabla 2: Coeficientes de correlación con *stroke*.

[2] James, Witten et. al., 2015

stroke	0	1
0	4632	229
1	202	47

significance=0.050, p=0.000

Tabla 3: Tabla de contingencia y test de independencia *heart disease*.

stroke	0	1
0	4429	432
1	183	66

significance=0.050, p=0.000

Tabla 4: Tabla de contingencia y test de independencia *hypertension*.

stroke	female	male
0	2853	2008
1	141	108

significance=0.050, p=1.000

Tabla 5: Tabla de contingencia y test de independencia *gender*.

stroke	unknown	formerly smoked	never smoked	smokes
0	1497	815	1802	747
1	47	70	90	42

significance=0.050, p=1.000

Tabla 6: Tabla de contingencia y test de independencia *smoking status*.

stroke	0	1
0	1728	3133
1	183	66

significance=0.050, p=0.000

Tabla 7: Tabla de contingencia y test de independencia *ever married*.

stroke	govt job	never worked	private	self employed	children
0	624	22	2776	754	685
1	33	0	149	65	2

significance=0.050, p=1.000

Tabla 8: Tabla de contingencia y test de independencia *work type*.

stroke	rural	urban
0	2400	2461
1	114	135

significance=0.050, p=0.000

Tabla 9: Tabla de contingencia y test de independencia *residence type*.

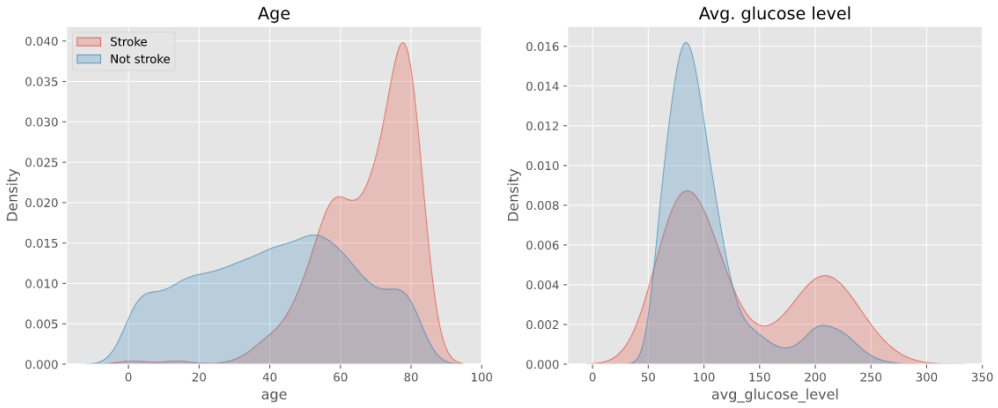

Figura 5: Kernel density plots para edad y nivel de glucosa promedio, según tuvieron o no ACV.

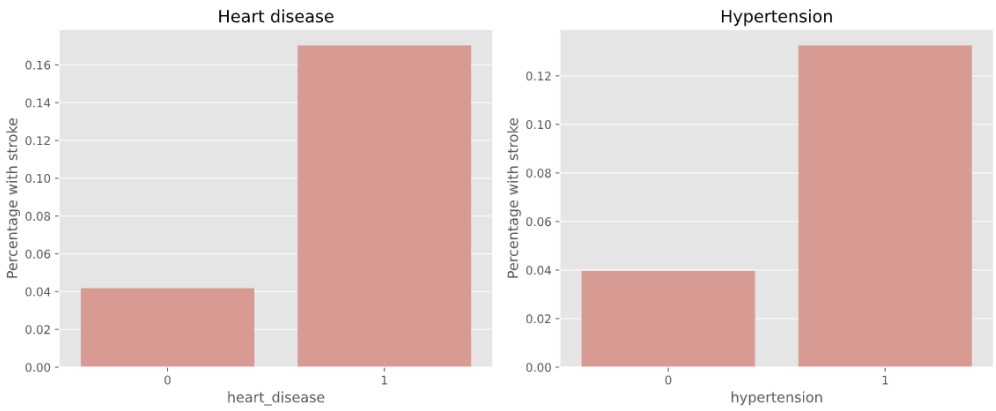

Figura 6: Porcentaje de *stroke* en relación a la cantidad de personas con una enfermedad de base y sin ella. A la izquierda la enfermedad de base es *heart_disease* y a la derecha es *hypertension*.

features	importance
age	1
avg_glucose_level	2
smoking_status_unknown	3
smoking_status_smokes	4
smoking_status_never_smoked	5
smoking_status_formerly_smoked	6
residence_type_urban	7
residence_type_rural	8
work_type_self_employed	9
work_type_private	10
gender_male	11
gender_female	12
bmi	13
ever_married	14
heart_disease	15
hypertension	16
work_type_never_worked	17
work_type_children	18
work_type_govt_job	19

Tabla 10: Importancia de los *features* en el árbol de decisión.

	precision	recall	f2-score
Con alfa	0,1	0,88	0,34
Sin alfa (max_depth=3)	0,14	0,74	0,40
3 mejores predictores	0,13	0,74	0,38

Tabla 11: Comparación de modelos

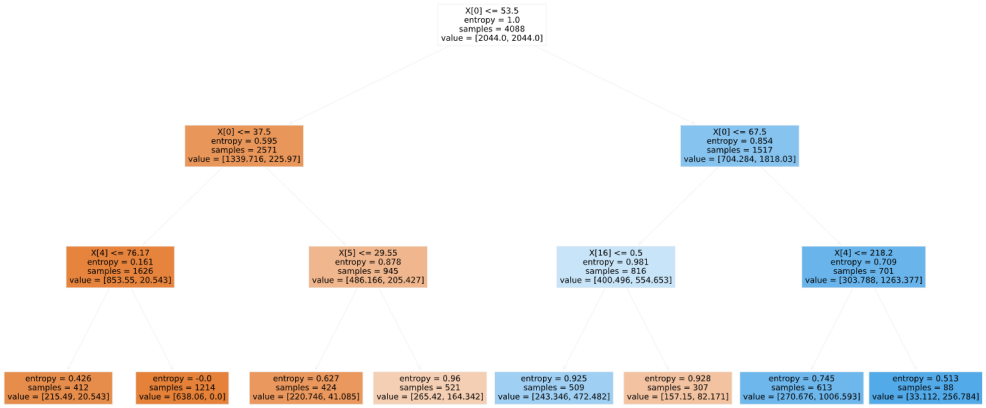

Figura 7: Gráfico de modelo de árbol de decisión optimizado.