Aprendizaje Automático y Minería de Datos

Facultad de Ciencias

Proyecto Final

Autores:
Nicholas J. Loughran Cree
Eduardo del Rio Ruiz

Junio 2021

Stroke Prediction

Proyecto Final

1. Introducción

En este proyecto buscamos aplicar los algoritmos estudiados durante el curso de Aprendizaje Automático y Minería de Datos, sobre una base de datos con el objetivo de obtener una predicción fiable. Todas los documentos de este proyecto se encuentran en la carptea OneDrive asociada [Cree and del Río Ruiz, 2021]. En este caso en específico se ha tomado una base de datos sobre personas que han y no han sufrido ataques cerebro-vasculares (también conocidos como derrames cerebrales o apoplejía)[fedesoriano, 2021], luego trataremos de reconocer si los atributos de los que disponemos son relevantes a la hora de que te de un ataque o no, y mediante estos tratar de predecir si a una persona le va a dar un ataque o no, con el menor margen de error posible.

Cabe a destacar que estamos tratando un caso delicado. Dado que estos algoritmos pueden fallar, pueden dar falsos negativos o falsos positivos. Hemos considerado que aunque un falso positivo es algo grave, ya que estarías diagnosticando que le va a dar un ataque a alguien que no, sería bastante peor diagnosticarle que no le va a dar un ataque, a alguien a quien si le va a dar. En otras palabras, daremos mayor importancia a los falsos negativos que a los falsos positivos.

2. Descripción de los datos

La base de datos utilizada ha descargada de la página Kaggle.com la cual facilita bases de datos para casos de estudio como este. Nosotros hemos elegido la base de datos "Stroke Prediction dataset, 2021". De acuerdo a la Organización de la Salud Mundial los ataques al corazón son la 2a mayor causa de muerte en todo el planeta, responsable de aproximadamente el 11 % de las muertes totales.

La base de datos contiene atributos sobre 5110 personas. A continuación se listan los atributos explicados con brevedad con sus respectivas estadísticas:

1. **ID**: identificador único.
2. **Gender**: Mujer (2994, 58.6 %), hombre (2115, 41.4 %) u otros (1).
3. **Age**: edad del paciente en años. El rango se encuentra entre los 0.08 años y los 82 años.
4. **Hypertension**: el paciente tiene hipertensión (498, 9.7 %) o no lo tiene (4612, 90.3 %).
5. **Heart_disease**: el paciente tiene problemas de corazón (276, 5.4 %) o no los tiene (4834, 94.6 %).
6. **Ever_married**: el paciente ha estado casado (3353, 65.6 %) o no lo está (1757, 34.4 %).
7. **Work_type**: tipo de trabajo. Empresa privada (2925, 57.2 %), autónomo (819, 16 %), menor de edad (687, 13.4 %), funcionario(657, 12.9 %) o desempleado(22, 4 %).
8. **Residence_type**: tipo de residencia. Urbana (2596, 50.8 %) o rural (2514, 49.2 %).
9. **Avg_glucose_level**: nivel medio de glucosa en sangre. El rango de datos varía entre 55.12 y 217.74. Para alguna de las personas, este atributo no estaba disponible

(N/A) por lo que se ha sustituido con la media del resto de valores para no afectar demasiado los cálculos.

10. **BMI**: índice de masa corporal. El rango de datos varía entre 10.3 y 97.6.

11. **Smoking_status**: indica si el paciente fumaba (885, 17.3 %), fuma (789, 15.4 %), nunca ha fumado (1892, 37 %) o si se desconoce (1544, 30.2 %).

12. **Stroke**: el paciente ha sufrido un ataque cerebro-vascular (249, 4.9 %) o no lo ha sufrido (4861, 95.1 %). Atributo objetivo de la predicción.

3. Modelos Utilizados

En primer lugar se han refinado los datos para que pudiesen ser más legibles para los modelos utilizados: algunos atributos consistían en números que estaban guardados como cadenas de texto que fueron pasados al tipo número y otros atributos binarios que necesitaban ser guardados con cadenas de texto para ser leídos bien por los correspondientes nodos.

Para la ejecución de los modelos se dividieron los datos del dataset en particiones de un 80 % de entrenamiento y un 20 % de test. Estos porcentajes obtenían los mejores resultados tras ser ejecutados por cada modelo.

Para separar estos datos se utiliza la técnica de muestreo estratificado respecto al atributo "stroke". Se hace para que se asegure de tener un valor proporcional de datos con derrame cerebral en cada partición ya que solo corresponden a un 4.9 % de todos los datos y sin ello podrían ser incluidos todos en una partición y no en la otra, proporcionando datos sesgados.

3.1. Logistic Regression

El primer modelo utilizado para predecir ataques cardio-vasculares que se ha utilizado es regresión logística. El modelo se entrena con el 80 % de los datos utilizando el gradiente estocástico medio.

Antes de entrenarlos y de realizar las predicciones se normalizan a una desviación normal estándar con z-score los datos posibles (age, avg_glucose y bmi). En la regularización del método se aplica la distribución de Gauss sobre los coeficientes.

3.2. kNN

El siguiente modelo utilizado es el algoritmo k Nearest Neighbours (kNN). Este modelo solo acepta atributos numéricos por lo que los atributos binarios se convierten en números y los demás no numéricos son descartados antes de pasarse por el algoritmo. Los mejores resultados se obtuvieron con $k = 2$ y se determina el peso o la importancia de un vecino con la distancia entre ellos.

3.3. AdaBoost

Para el modelo utilizando AdaBoost se ha seleccionado RandomTree como el clasificador a utilizar porque producía mejores resultados que el clasificador SimpleCART que se había visto en clase.

3.4. Decision Tree

En el caso del algoritmo Decision Tree se ha decidido utilizar el ratio de ganancia como medida de calidad frente al índice de Gini, ya que se han obtenido mejores resultados.

3.5. Rule Association

En cuanto a las reglas de asociación se han tenido que reconvertir algunos atributos nominales a strings nuevamente para que sean legibles y se entiendan fuera del contexto del atributo (0s y 1s a Hypertension and No Hypertension, etc.), juntando después todas las columnas en un array. Hecho esto se extraen las reglas de asociación probando con distintos soportes (0.8 y el 0.5).

4. Resultados

La ejecución de algunos modelos no seleccionados para este proyecto no resultaban productivos. Modelos como SVM y Redes Neuronales requerían que los atributos leídos fueran de tipo numérico cuando la mayoría de los atributos trabajados en este dataset eran de tipo nominal. Por lo tanto solo realizaban las predicciones utilizando pocos atributos y no conseguían predecir bien los casos de ataque cerebro-vascular.

4.1. Logistic Regression

Se realiza la predicción sobre el 20 % de los datos, previamente normalizados de la misma manera que el entrenamiento. Utilizando el nodo Scorer se observa que se obtiene un accuracy del 95.108 %.

A primera vista, esto parece indicar que el modelo que se ha entrenado es de gran calidad. Sin embargo, al analizar la matriz de confusión se observa que no se acierta en ninguno de los casos que sí habían sufrido un ataque cerebro-vascular, andose todos los fallos como falsos negativos. Al observar la curva ROC, también se observa que para el caso de los positivos, el AUC es de 0.5, que indica que es un mal clasificador.

Concluyendo, la regresión logística no es buen modelo para este Dataset.

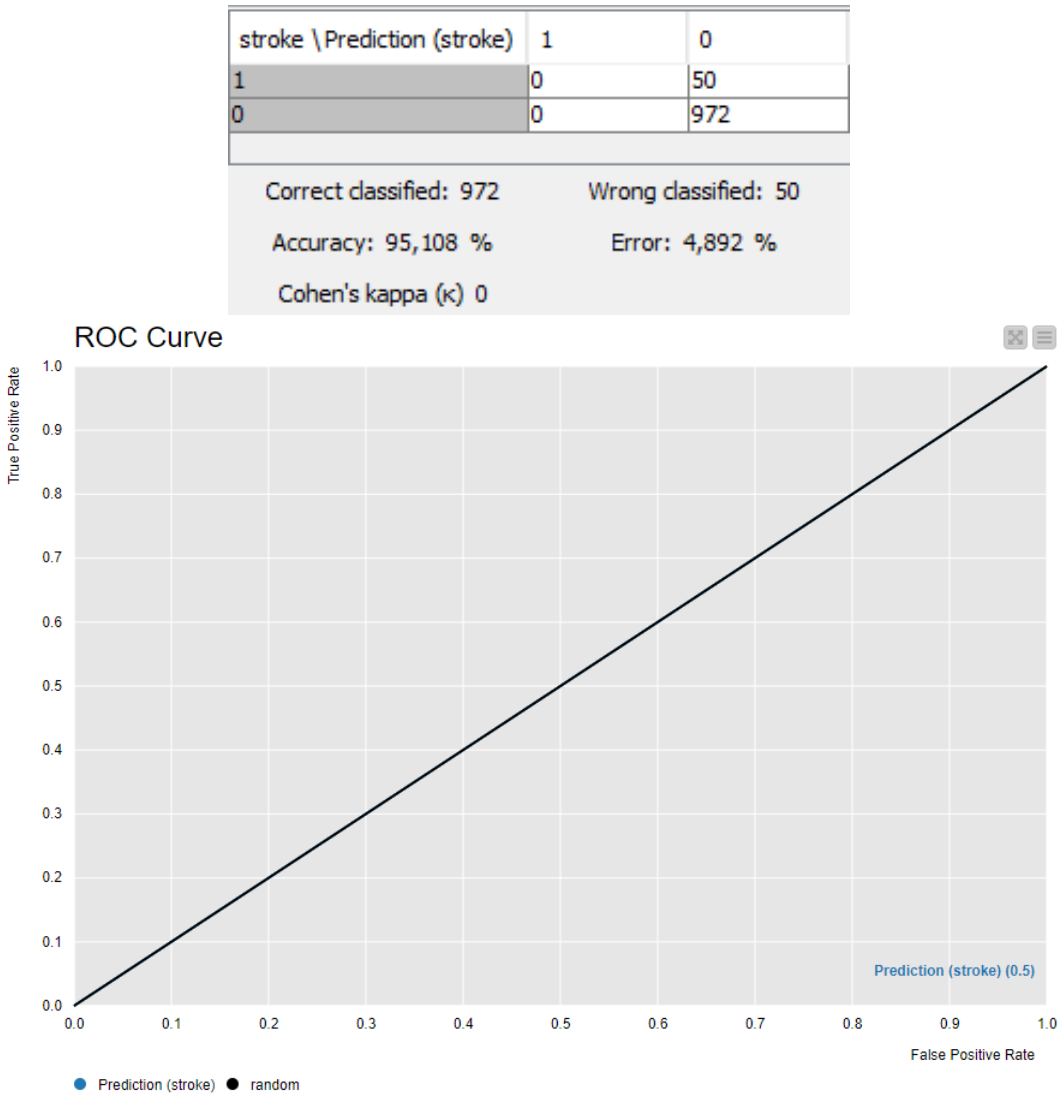

Figura 1: Resultados Regresión Logística

4.2. kNN

Se realiza la predicción sobre el 20 % de los datos. Para utilizar el kNN, todos los valores que tiene en cuenta deben de ser numéricos. Entonces, los atributos nominales binarios (heart disease, hypertension, residence type y ever maried) son cambiados de strings a números y el resto de atributos no numéricos son descartados. Utilizando el nodo Scorer se observa que se obtiene un accuracy del 92.368 %.

Este dato, aun siendo un poco peor que el anterior, sigue siendo bastante alto y otra vez parece indicar que el modelo que se ha entrenado es de gran calidad. Observando la matriz de confusión se detecta un incremento en el número de casos de verdaderos positivos, donde sí se ha detectado un derrame cerebral correctamente, pero este número sigue siendo bastante bajo. Solo acierta en un 20 % de las ocasiones. Al observar la curva ROC, el AUC es de 0.58, que indica que el clasificador sigue sin ser muy bueno.

El modelo de kNN no parece ser buen modelo para este Dataset.

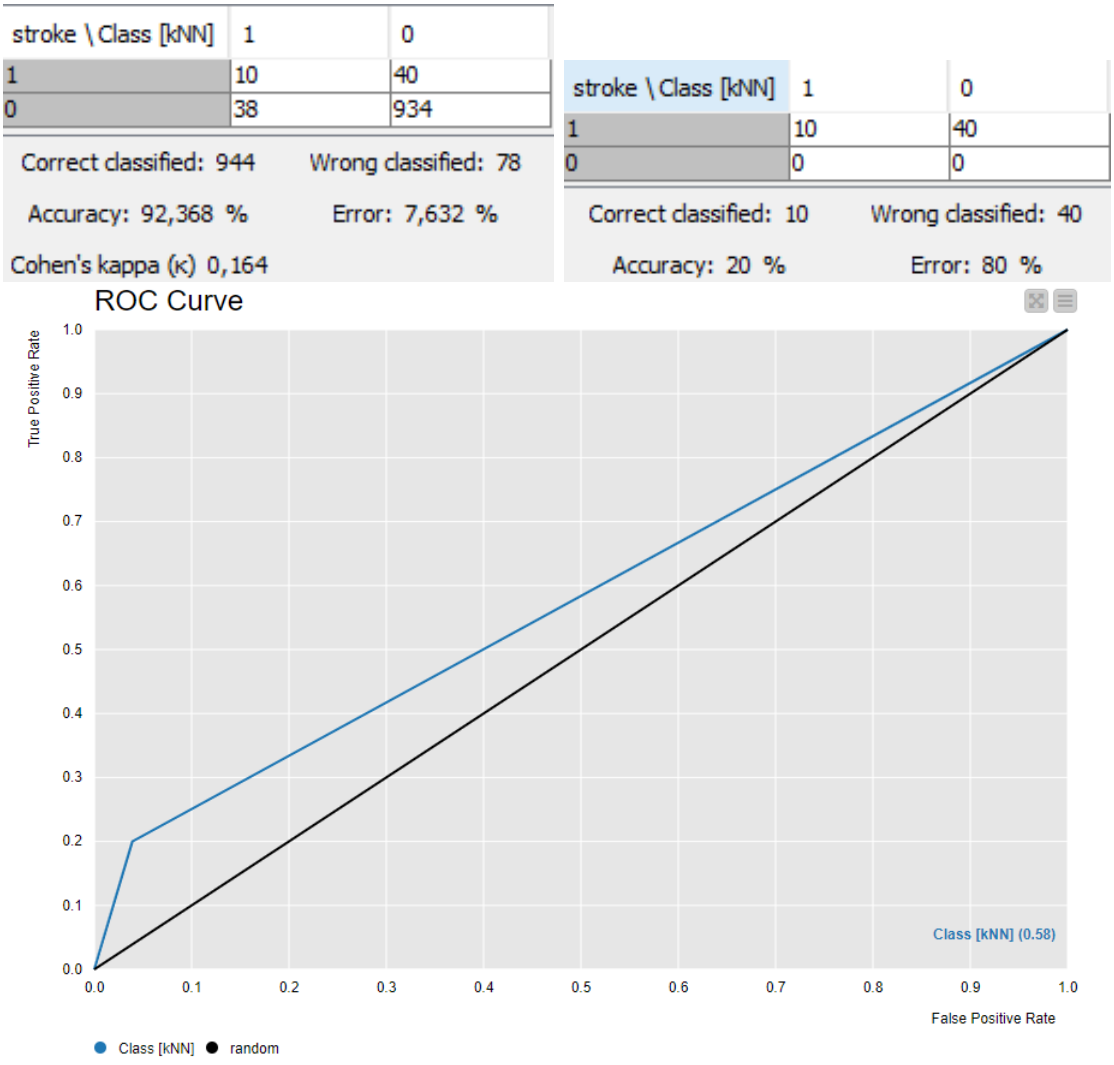

Figura 2: Resultados kNN

4.3. AdaBoost

Se realiza la predicción sobre el 20 % de los datos. Utilizando el nodo Scorer se observa que se obtiene un accuracy del 91.389 %.

Observando la matriz de confusión se detecta un ligero incremento en el número de casos de verdaderos positivos, aunque este número sigue siendo bastante bajo. Solo acierta en un 22 % de las ocasiones donde el paciente sí tiene apoplejía. También incrementa el número de falsos positivos y es bastante mayor que el número de verdaderos positivos, que significa que solo 11/60 predicciones de un ataque son fiables, indicando que no es muy buen clasificador.

Al observar la curva ROC, el AUC es de 0.585, que indica que el clasificador sigue sin ser muy bueno.

El modelo de AdaBoost no parece ser buen modelo para este Dataset.

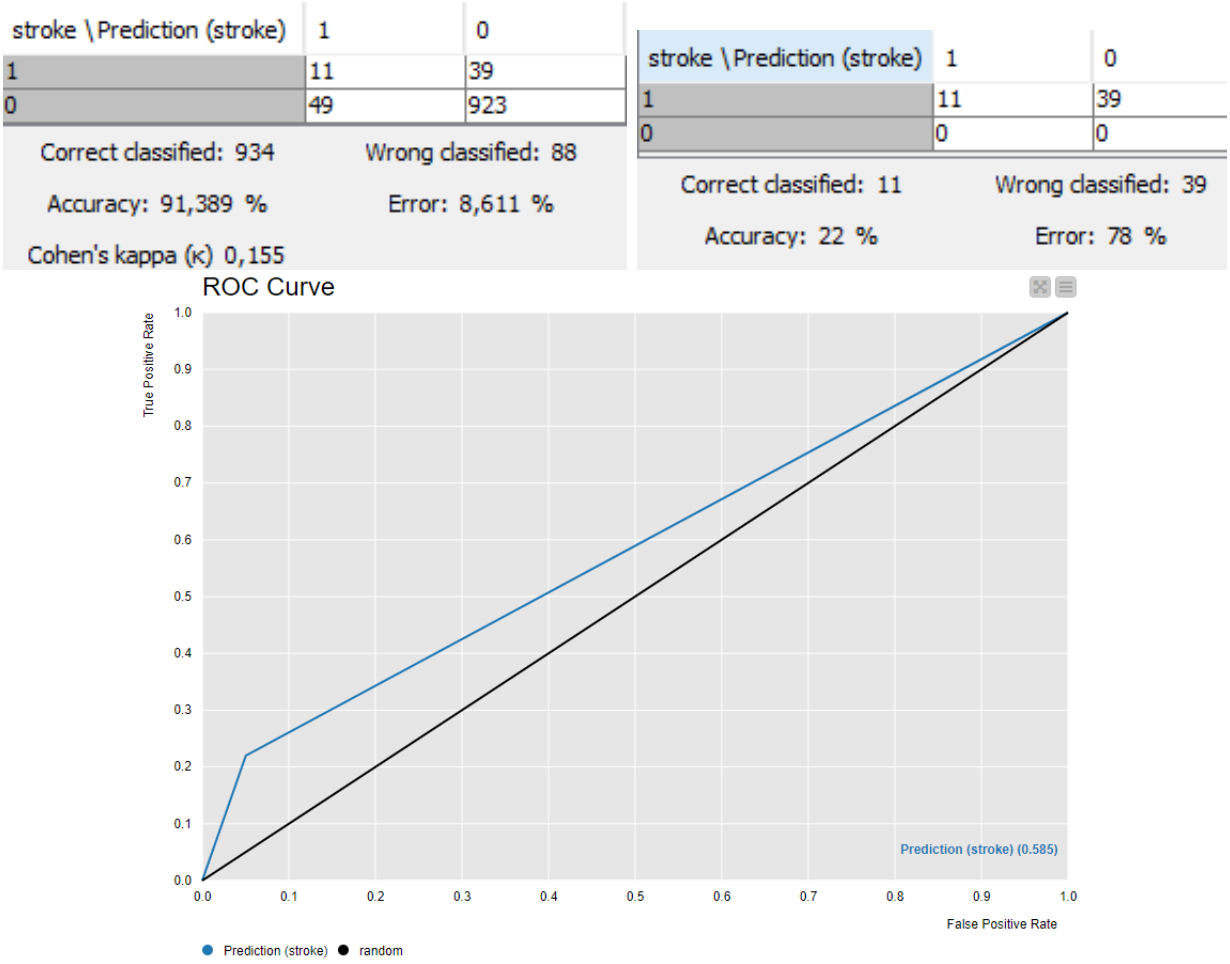

Figura 3: Resultados AdaBoost

4.4. Decision Tree

Se realiza la predicción sobre el 20 % de los datos. Utilizando el nodo Scorer se observa que se obtiene un accuracy del 91.292 %.

Observando la matriz de confusión se detecta el mayor número de casos de verdaderos positivos (13) y solo acierta en un 26 % de las ocasiones donde el paciente tiene apoplejía. Al observar la curva ROC, el AUC es de 0.603, que aún no siendo muy alto, es mejor clasificador que los modelos anteriores.

El modelo de Árbol de decisión no parece ser buen modelo para este Dataset.

stroke \Prediction (stroke)	1	0
1	13	37
0	52	920

Correct classified: 933 Wrong classified: 89

Accuracy: 91,292 % Error: 8,708 %

Cohen's kappa (κ) 0,181

stroke \Prediction (stroke)	1	0
1	13	37
0	0	0

Correct classified: 13 Wrong classified: 37

Accuracy: 26 % Error: 74 %

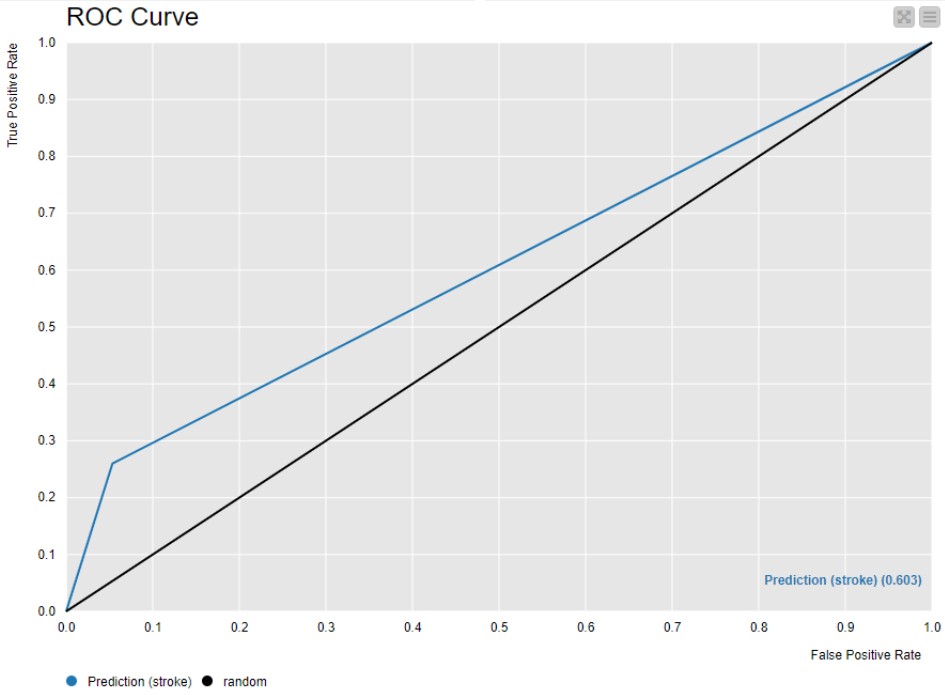

Figura 4: Resultados Árbol de decisión

4.5. Rule Association

Finalmente se quiere extraer las reglas de asociación del dataset. Como este trabajo quiere estudiar las predicciones de ataque cerebro vascular, se filtrarán las reglas para poder observar las que impliquen tener o no tener apoplejía. También se quieren reglas de alta confianza y soporte, así que se extraen las que superan el umbral de 0.8 en ambos casos.

Se extraen las siguientes reglas:

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule2	0.832	0.966	1.016	No Stroke	<---	[No Hyper Tension,No Heart Disease]
rule6	0.867	0.96	1.01	No Stroke	<---	[No Hyper Tension]
rule8	0.906	0.958	1.007	No Stroke	<---	[No Heart Disease]

Figura 5: Reglas de asociación con soporte y confianza $> 0{,}8$

Estas reglas parecen indicar que si no se sufre hipertensión o problemas de corazón, en el mayor de los casos no se sufrirá un ataque cerebro-vascular. Sin embargo, esto se podría deber a la poca frecuencia de datos positivos para todos estos atributos dentro del dataset. Por ello se decide ampliar el umbral de soporte (a 0.5) para observar que otras reglas podrían extraerse:

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule2	0.506	0.951	0.999	No Stroke	<---	[No Hyper Tension,Married,No Heart Disease]
rule4	0.512	0.962	1.012	No Stroke	<---	[No Hyper Tension,Female]
rule9	0.518	0.957	1.006	No Stroke	<---	[No Heart Disease,Private]
rule14	0.537	0.943	0.992	No Stroke	<---	[No Hyper Tension,Married]
rule16	0.54	0.958	1.007	No Stroke	<---	[Female,No Heart Disease]
rule18	0.543	0.949	0.998	No Stroke	<---	[Private]
rule19	0.558	0.953	1.002	No Stroke	<---	[Female]
rule23	0.573	0.942	0.99	No Stroke	<---	[Married,No Heart Disease]
rule25	0.613	0.934	0.982	No Stroke	<---	[Married]
rule28	0.832	0.966	1.016	No Stroke	<---	[No Hyper Tension,No Heart Disease]
rule32	0.867	0.96	1.01	No Stroke	<---	[No Hyper Tension]
rule34	0.906	0.958	1.007	No Stroke	<---	[No Heart Disease]

Figura 6: Reglas de asociación con soporte $> 0{,}5$ y confianza $> 0{,}8$

De estas, las datos más interesantes que se observan son ser Mujer, haber estado casado en algún momento y trabajar en empresa privada tienden a implicar que no se sufrirá un ataque cerebro-vascular, con un soporte entre 0.5 y 0.6. Esto se puede deber a la presencia alta de estos valores en sus respectivos atributos dentro de este dataset y la baja probabilidad general de sufrir un derrame cerebral.

5. Conclusiones

Los resultados obtenidos de los modelos implementados indican que los clasificadores son bastante pobres. En el mejor de los casos, solo se acertaba un 26 % de los casos que sí tenían derrame cerebral y de los que diagnosticaba el derrame, solo un 20 % eran verdaderos. En la tabla inferior, la columna de Positive Accuracy representa el porcentaje de casos positivos de derrame cerebral que fueron predichos correctamente y la columna de Diagnosis Accuracy indica el porcentaje de los predichos como casos positivos que acertaron.

Model	**True Positives**	**Positive Accuracy**	**Diagnosis Accuracy**	**AUC**
Logistic Regression	0	0 %	0 %	0.5
kNN	10	20 %	20.83 %	0.58
AdaBoost	11	22 %	18.33 %	0.585
Decision Tree	13	26 %	20 %	0.603

De este trabajo se pueden concluir varias cosas. Por un lado puede ser que la cantidad de casos de ataque cerebro-vascular en el dataset sea demasiado bajo (solo representaban un 4.9 % de los casos del dataset) y hagan falta más casos positivos para ser preciso en la predicción.

Por otro lado, los resultados pobres de los modelos también se puede deber a que los atributos que se trabajan en ellos no sean buenos indicadores de que se vaya a sufrir un derrame cerebral y no tengan una fuerte relación de causalidad entre ellos. En las reglas de asociación se observa que ser mujer implicaba no sufrir un ataque cerebro-vascular cuando dentro de los casos positivos formaban un 56 % de los casos(141 Female, 108 Male). Puede ser que no haya relación entre el género y la probabilidad de sufrir un derrame o puede que los resultados obtenidos se den porque había mayor cantidad de miembros de un género que del resto.

Formas en las que podría mejorarse este trabajo y variaciones sobre él:

- Incrementar casos positivos de stroke. Si se realizan los modelos con porcentaje de casos positivos en derrames cerebrales puede que se observen mejoras en las predicciones de los modelos.

- Igualar los porcentajes de los valores dentro de los atributos nominales. Si se igualan las cantidades de casos de Hombre y Mujer o del tipo de trabajo, se podría evitar algunos sesgos de precisión o de reglas de asociación y puede que los modelos obtengan mejores resultados.
- Trabajar con diferentes atributos. Si se realiza el mismo estudio sobe atributos que puedan tener más relación con derrames cerebrales como datos sobre el cerebro, genética, casos de derrames en la familia, niveles de estrés, etc. Se podrían obtener mejores resultados de predicción.

Referencias

N. L. Cree and E. del Río Ruiz. Proyecto final minería de datos, 2021. https://unicancloud-my.sharepoint.com/:f:/g/personal/nlc529_alumnos_unican_es/EmllwcXRAGxGo1pXX0SkcqABeCmbstMHXP80_0g_JAmKeg?e=gYNMe6.

fedesoriano. Stroke prediction dataset, 2021. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.