Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Desarrollo de un método de diagnóstico de migrañas a partir de un cuestionario realizado al paciente

Trabajo de especialización
Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Buenos Aires, 2020

DESARROLLO DE UN MÉTODO DE DIAGNÓSTICO DE MIGRAÑAS A PARTIR DE UN CUESTIONARIO REALIZADO AL PACIENTE

En el presente trabajo se desarrollaron tres clasificadores para identificar pacientes con migraña. Se utilizan como variables de entrada las respuestas a una encuesta de salud de los Estados Unidos, estudiando un total de 3056 sujetos. Con el modelo de regresión logística se obtuvieron los valores de 0.699 para ROC-AUC, sensibilidad de 0.6712 y especificidad de 0.7281. Con el método *Random Forests* se obtuvo un ROC-AUC de 0.7221, sensibilidad de 0.7192 y especificidad de 0.7250. Finalmente, para *XGBoost* se obtuvieron los valores de 0.7255 para ROC-AUC, 0.6917 de sensibilidad y 0.7593 especificidad. Las tres técnicas aquí implementadas han logrado un desempeño similar y variables importantes en común para la clasificación. Es posible a partir de estos resultados realizar una selección de preguntas específicas para el diagnóstico de migrañas, pudiendo funcionar estos clasificadores como asistencia a los profesionales de salud para disminuir los tiempos de consulta.

Palabras claves: migraña, clasificación, regresión logística, Random Forest, XGBoost.

MIGRAINE DIAGNOSIS METHOD BASED ON PATIENT QUESTIONNAIRE

In this work three classification methods were developed to predict patients who suffers migraines. Data from a health interview survey in the United States is used here, studying a total of 3052 subjects. With the logistic regression model a ROC-AUC score of 0.6997 was obtained, a sensitivity of 0.6712 and specificity of 0.7281. *Random Forests* showed a ROC-AUC score of 0.7221, 0.7192 sensitivity and 0.7250 specificity. Finally, for *XGBoost* a 0.7255 ROC-AUC score was obtained, 0.6917 sensitivty and 0.7593 specificity. All the methods showed a similar performance and important features -for the classification- in common. From these results, it is possible to create a short and specific questionnaire for migraine diagnosis, which could asist healths professional to reduce medical visits duration.

Keywords: migraine, classification, logistic regression, Random Forest, XGBoost.

Índice general

ESTRUCTURA DEL TRABAJO

El presente trabajo se encuentra separado en capítulos. En el Capítulo 1 se introduce el tema en estudio, su contexto actual y motivación. Se detallan allí las hipótesis y objetivos de este trabajo y se describe el *dataset* aquí utilizado.

El capítulo 2 describe el procesamiento aplicado a los datos para su la limpieza y acondicionamiento. El mismo consiste en la selección de variables según lo detallado en la introducción, imputación de datos faltantes y eliminación de datos atípicos univariados.

El capítulo 3 corresponde al análisis exploratorio de datos, allí se muestra una descripción detallada del *dataset* a través de diversas visualizaciones.

En el capítulo 4 se desarrollan tres métodos de clasficación. En primer lugar regresión logística con regularización *Lasso*, el cual es utilizado como *benchmark* para técnicas más complejas. En segundo lugar se implemente la técnica de *Random Forests* y por último *XGBoost*. Se describen aquí los resultados obtenidos con cada uno de ellos.

Por último, en el capítulos 5 se discuten los resultados obtenidos y la posible utilidad de la metodología aquí propuesta como herramienta médica de diagnóstico.

Aquellos interesados en reproducir el proceso aquí presente encontrarán de utilidad el capítulo 6.

1. INTRODUCCIÓN

1.1. Sobre el tema en estudio

La migraña (o hemicránea o jaqueca) es una enfermedad que tiene como síntoma principal el dolor de cabeza, pulsátil, unilateral u opresivo, acompañado de náuseas o vómitos, sensibilidad a la luz o los sonidos, usualmente muy intenso e incapacitante para quien lo sufre. El mecanismo exacto de esta enfermedad es aún desconocido, pero tanto factores genéticos como ambientales influyen en el padecimiento de dicha condición. Se cree que la activación del sistema trigénimo vascular (neuronas en el nervio trigenimal, que irriga vasos sanguíneos al cerebro) y la despolarización cortical propagada podrían jugar un rol importante [1, 2]. Dentro de la clínica médica, las migrañas se dividen en dos subtipos: migraña con aura y migraña sin aura. Esta enfermedad afecta al 11 % [3, 4] de la población adulta y tiene tres veces más incidencia en mujeres que en hombres, siendo su mayor prevalencia en la etnia caucásica [5, 4]. En los Estados Unidos, dentro de las consultas médicas generales, alrededor del 4 % corresponden a consultas por migraña, mientras que dentro de las consultas con especialista en neurología alrededor del 20 % es por esta enfermedad [6].

El proceso de diagnóstico de esta patología está fuertemente focalizada en entrevistar al paciente. El mismo se realiza en base a las respuestas del sujeto que las padece; se procede, posteriormente, a realizar exámenes físico para encontrar otras condiciones o problemas que puedan estar aumentando las cefaleas: inspección estructural de cuello y cabeza, exámenes neurológico, estudios cardiovasculares, entre otros [7].

Considerando la alta incidencia de esta enfermedad y que su procedimiento principal de diagnóstico es a partir de preguntas realizadas al paciente, sería de particular interés automatizar este proceso. No reemplazaría la entrevista médico-paciente, pero podría funcionar a modo de pre-consulta para acortar el cuestionario y así la duración de las visitas médicas. También podría funcionar a modo de auto-consulta, para que el paciente sea consciente de un posible diagnóstico y así pueda brindarle información clara y detallada al especialista durante su visita.

1.2. Hipótesis y objetivos

Frente a esta información se plantea la hipótesis: *es posible realizar un primer diagnóstico de migraña en base a un cuestionario*. Para verificar o refutar esta hipótesis se aplicarán técnicas de clasificación que predigan si el paciente ha sufrido o no migrañas en los último 3 meses.

1.3. Datasets de trabajo

En el presente trabajo se utiliza como fuente de datos la encuesta nacional de salud (*Health Interview Survey*, NHIS[1]) del Centro de control y prevención de enfermedades (*Centers for disease control and prevention*, CDC[2]). El objetivo principal de esta encuesta es monitorear la salud de la población de los Estados Unidos a través de la recopilación y el análisis de datos sobre una amplia gama de temas de salud. Así, se logra vincular la salud de la población con diferentes aspectos demográficos y socioeconómicos. La NHIS es una encuesta que consiste en entrevistar a distintos hogares modelo; estos hogares son entrevistados cada año. La selección de ellos no es al azar, sino que se diseña para ser representativo de diversos grupos étnicos y niveles socieconómico. El plan de muestreo se rediseña después de cada censo decenal (el plan actual fue implementado en 2016).

1.4. Población de estudio

Se utiliza aquí el *dataset* llamado *sample adult* (adulto modelo), que corresponde a los sujetos entrevistados mayores de 18 años durante el año 2018. La misma cuenta con 25417 registros y 742 variables, de las cuales se conservan y transforman aquellas indicadoras de:

- características del paciente (sexo, edad, peso, estatura, etnia)
- enfermedades crónicas (cardíacas, respiratorias, renales, hepáticas, de columna)
- antecedentes médicos (infartos, derrames cerebrales, varicela)
- hábitos (frecuencia e intensidad de ejercicio, cantidad de cigarros que fuma el paciente, horas de sueño, horas de trabajo frente a una computadora)
- medicamentos que toma al momento de la encuesta

En cuanto a la población de estudio, el presente *dataset* contiene la información relevada de adultos mayores de 18 años. Se consideran sólo los pacientes que:

- No padezcan enfermedades mentales graves que provoquen la pérdida del contacto con la realidad, tales como esquizofrenia
- No padezcan sindromes que deterioren las capacidades psíquicas, tales como la demencia senil
- No padezcan cáncer al momento de la encuesta
- No padezcan limitaciones físicas
- No estén gestando al momento de la entrevista

[1] `www.cdc.gov/nchs/nhis/`
[2] `https://www.cdc.gov/`

De esta manera, se trabaja únicamente con pacientes que no se encuentren gestando y que gocen de buena salud: esto se define, a nivel objetivo, como la constatación de la ausencia de enfermedades o de factores dañinos en el sujeto en cuestión [8].

Así, como primer paso se selecciona la población de estudio y las variables de interés según lo enunciado. Adicionalmente, se creó un *dataset* que contiene metadata de las variables a utilizar:

- `code`: código original de identificación de cada pregunta.
- `question`: pregunta de la encuesta.
- `levels`: niveles de las respuestas, únicamente para variables categóricas.
- `tipo`: tipo de variable, *id*, *discreta* o *contínua*.
- `category`: categoría a la que corresponde cada pregunta; *id*, *entorno* (variables relacionadas al entorno del paciente), *sujeto* (características físicas del paciente), *antecedentes* (enfermedades que ha padecido y valores anómalos en análisis clínicos), *enfermedades crónicas* y *hábitos*.
- `temp`: temporalidad de la pregunta; *presente* (valores como altura y peso), *constante* (preguntas sobre si ha padecido cierta enfermedad alguna vez), *12 meses* (enfermedades o dolencias que ha padecido en el último año) y *3 meses* (enfermedades o dolencias que ha paedecido en los últimos 3 meses).
- `object`: el tipo de objeto; *int*, *bool*, *float* o *category*.

Este *dataset* es utilizado para agrupar las variables y visualizarlas por categoría.

2. PREPROCESAMIENTO

2.1. Transformación de variables

Dado que se trata de un cuestionario extenso, muchas preguntas son dependientes de la respuesta de otra: por ejemplo, la pregunta *¿Cuántos cigarrillos fuma a la semana?* sólo se realiza si el sujeto ha respondido afirmativamente a la pregunta *¿Fuma?*. Las variables de este tipo fueron transformadas de manera tal que abarquen a la población completa y no sólo a un grupo condicional a una respuesta anterior. Así, la pregunta *¿Cuántos cigarrillos fuma a la semana?* se respondería para toda la población asignando como respuesta *0 cigarrillos a la semana* a aquellos sujetos no fumadores. Otras transformaciones relevante al *dataset* original es el reemplazo de las respuestas *Desconocido* y *se rehusa a responder* por valores nulos -de esta manera es posible, si se desea, imputar los datos faltantes por diversas técnicas- y las respuestas a preguntas binarias (No/Si) por los valores 0 y 1.

Tras esta selección y transformaciones se obtiene un *dataset* de 47 variables y 13479 registros. El listado completo de variables se detalla en el *Apéndice*.

2.2. Imputación de datos faltantes

En la Tabla 2.1 se muestra el porcentaje de valores faltantes para cada variable. Dado que se busca trabajar con modelos como la regresión logística que no admite faltantes en sus valores de entrada, se decidió imputar todos los campos vacíos. Para ello se utilizó la técnica *Multivariate imputation by chained equations* (MICE); esta técnica asume que los valores faltantes son *Missing at Random* (MAR) [9]. Se entiende por MAR a un valor cuya probabilidad de ausencia es función de los parámetros observados pero no del valor faltante en sí mismo [10]. Así, por ejemplo, si los hombres son menos propensos a responder la pregunta *¿Padece depresión?*, dicha pregunta es MAR, dado que no depende de que el paciente padezca la enfermedad o no, sino de su género.

Una vez imputados los datos faltantes se realizó una correción sobre la variable `BMI` (índice de masa corporal). La misma se define como:

$$BMI = \frac{Peso(kg)}{Altura(m)^2} \tag{2.1}$$

Así, los faltantes en `BMI` son consecuencia de los datos ausentes en peso y altura; completos esos dos valores, el índice de masa corporal se re-calculó para todos los pacientes. Una vez implementada la técnica de imputación MICE se obtuvo un *dataset* completo en todas sus variables.

2.3. Tratamiento de datos atípicos univariados

Para la detección visual de *outliers* univariados, se realizó un *boxplot* para cada una de las variables contínuas presentes en el *dataset* (Figura 2.1) y se analizaron los datos alejados de forma independiente. Se observa que las variables relacionadas al consumo

Código	Pregunta	Faltantes (%)
id	identificación única	0
REGION	Región de USA en la que habita el individuo	0
SEX	Sexo	0
AHEIGHT	Altura (cm)	5.7
AWEIGHTP	Peso (Kg)	7.29
BMI	Índice de masa corporal	7.49
AGE_P	Edad	0
RACERPI2	Etnia	0
R_MARITL	Estado civil	0.22
DOINGLWA	¿Trabaja actualmente?	0.06
WRKLYR4	¿Trabajó en los últimos 12 meses?	0.15
HYPYR1	¿Ha sufrido hipertemsión en los últimos 12 meses?	0.15
HYPMED2	¿Toma actualmente medicamento para la presión (recetada por un médico)?	0
CHLYR	¿Ha tenido colesterol alto en los últimos 12 meses?	0.39
CHLMDNW2	¿Toma actualmente medicamento para reducir el colesterol (recetado por un médico)?	0
CHDEV	¿Alguna vez fue diagnosticado con algún tipo de enfermedad coronaria?	0.07
ANGEV	¿Algna vez fue diagnosticado con angina pectoris?	0.08
MIEV	¿Alguna vez tuvo un infarto?	0.05
HRTEV	¿Alguna vez fue diagnosticado con una enfermedad del corazón?	0.04
STREV	¿Alguna vez sufrió un derrame cerebral?	0.01
EPHEV	¿Alguna vez fue diagnosticado con efisema?	0.01
COPDEV	¿Alguna vez fue diagnosticado con EPOC?	0.01
ASP	Toma aspirina actualmente?	0.03
AASSTILL	¿Tiene asma?	0.06
ULCYR	¿Ha tenido una úlcera en los últimos 12 meses?	0.02
DIBEV1	¿Ha sido diagnosticado alguna vez con diabetes o prediabetes?	0.02
DIBPILL1	¿Toma actualmente medicamento para la diabetes?	0.01
INSLN1	¿Toma actualmente insulina?	0
AHAYFYR	¿Ha tenido rinitis alérgica en los últimos 12 meses?	0.06
SINYR	¿Ha sido diagnosticado con sinusitis en los últimos 12 meses?	0.03
CBRCHYR	¿Ha sido diagnosticado con bronquitis crónicaen los últimos 12 meses?	0.01
KIDWKYR	¿Ha sido diagnosticado con algún tipo de falla renal en los últimos 12 meses?	0
LIVYR	¿Ha sido diagnosticado con algún tipo de falla hepática en los últimos 12 meses?	0.03
ARTH1	¿Ha sido diagnosticado alguna vez con alguna forma de artritis, artritis reumatoide, gota, lupus o fibromialgia?	0.04
PAINECK	¿Ha sufrido dolor de cuello durante más de 24 horas seguidas en los últimos 3 meses?	0.01
PAINLB	¿Ha sufrido dolor de cintura durante más de 24 horas seguidas en los últimos 3 meses?	0.01
PAINFACE	¿Ha sufrido dolor de en la cara o la mandíbula durante más de 24 horas seguidas en los últimos 3 meses?	0
FLA1AR	¿Posee alguna limitación física?	0.14
CIGSDAY	Números de cigarrillo que fuma al día (todo tipo de cigarros)	0
VIG	¿Cuántos minutos a la semana realiza de ejercicio vigoroso?	1.69
MOD	¿Cuántos minutos a la semana realiza de ejercicio moderado?	2.92
ALC	En promedio ¿Cuántos vasos de alcohol bebe a la semana?	0
APOX	¿Ha tenido varicela?	4.71
AHEP	¿Ha tenido hepatitis?	1.45
LIVEV	¿Ha sido diagnosticado alguna vez con una enfermedad hepática crónica?	1.39
ASICPUSE	¿Con qué frecuencia utiliza una computadora?	1.49
ASISLEEP	En promedio y en un periodo de 24 horas ¿Cuántas horas duerme?	2.4
target	¿Ha tenido migrañas o cefaleas severas en los últimos 3 meses?	0

Tab. 2.1: Porcentaje de valores faltantes en cada variable.

de cigarrillos, alcohol y actividad física presentan desviación hacia los valores menores; la mayoría de la población presenta hábitos de consumo de alcohol y cigarrillos bajo-moderado, lo mismo con el hábito de ejercicio. Sí llaman la atención los datos tan altos en actividad física. Se detallan a continuación los valores descartados y su motivo:

- `ALC - consumo de alcohol semanal`: Se observa un único caso extremadamente alejado del resto. Dado que es un único sujeto y su valor duplica a su inmediato inferior, se considera un error de imputación y se lo descarta.

- `VIG - actividad vigorosa semanal`: Dado que una semana tiene en total 10080 minutos, se considera imposible físicamente realizar actividad física durante la mitad de ella o más (la mitad implicaría entrenar 12 horas por día). Así, se descartan aquellos valores mayores a 5040 minutos semanales.

- `MOD - actividad moderada semanal`: Análogamente al caso anterior, se descartan los valores mayores a 5040 minutos semanales.

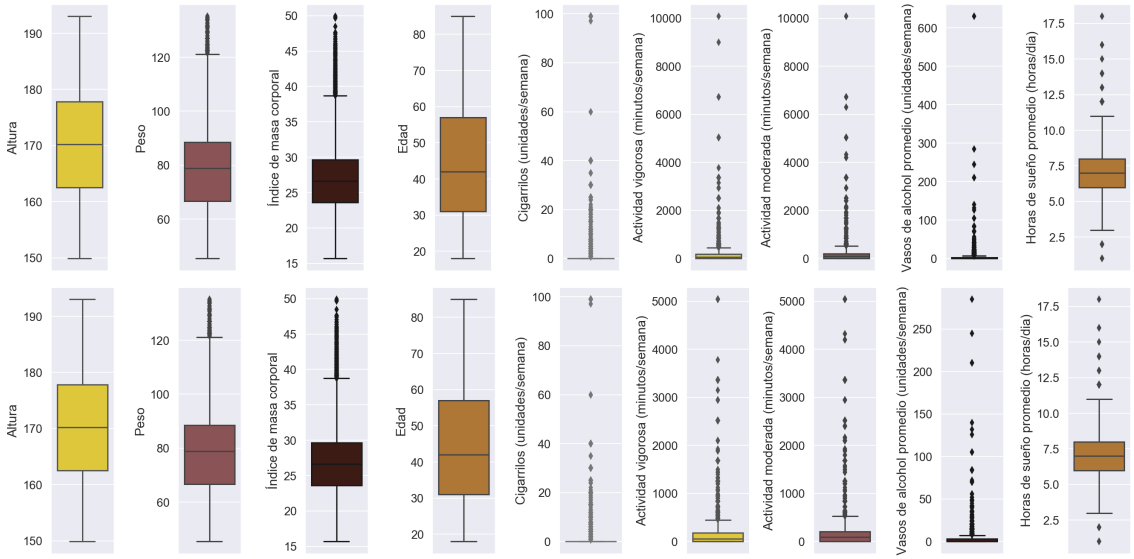

Fig. 2.1: Arriba: distribución de las variables contínuas en el *dataset* filtrado e imputado. Abajo: distribución de las variables contínuas tras remover datos atípicos univariados.

En el caso de la variable `CIGSDAY - cigarrillos por semana` no fue posible justificar los valores tan altos; no es improbable que se trate de sujetos que efectivamente fuman mucho semanalmente.

Tras la remoción de estos pocos *outliers* univariados el set de trabajo queda conformado por 13470 registros.

3. ANÁLISIS EXPLORATORIO Y DESCRIPCIÓN DEL DATASET

Según la *Migraine Research Foundation* aproximadamente el 12 % de la población -incluyendo niños- sufre de migrañas o cefaleas severas [4]; dentro de la población adulta, el 11 % sufre de esta condición [5]. En la Figura 3.1 se observa que el *dataset* de trabajo presenta una distribución de clase objetivo similar.

En la Figura 3.2 se muestra la correlación entre las variables contínuas presentes en el *dataset* de trabajo. Se observa que las únicas fuertemente correlacionadas son peso y altura con índice de masa corporal (como consecuencia de su definición, ecuación 2.1). Las variables restantes no presentan correlación mayor a 0.5.

Las variables categóricas han sido agrupadas en distintas categorías para su visualización en esta sección.

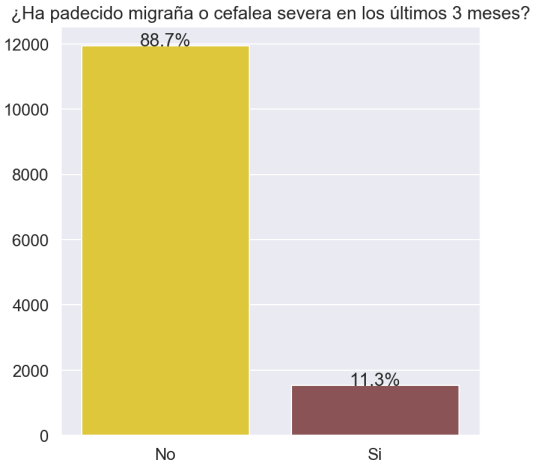

Fig. 3.1: Distribución de la clase objetivo en el set de datos en estudio.

Aquellas agrupadas como *entorno* corresponden a las preguntas relacionadas con el contexto habitual del paciente: su estado civil, región en la que reside y su condición laboral actual. En la Figura 3.3 se muestra la distribución de las variables descriptoras de dicha categoría; se observa que la mayor parte de la población de estudio corresponde a personas casadas y en actividad laboral.

Las variables dentro de la categoría *características* describen atributos físicos del paciente, tales como altura, peso, sexo, edad y etnia; se muestra su distribución en las Figura 3.4. La mayoría de los sujetos en estudio son de etnia caucásica mientras que la distribución de sexo se encuentra equilibrada entre hombres y mujeres. Las variables altura, peso e índice de masa corporal presentan una distribución similar entre las clases. En el caso de la variable edad, pareciera que aquellos pacientes que sufren migraña presentan una edad media menor que aquellos que no lo padecen. Esto en consistente con el hecho de que esta enfermedad tiene mayor incidencia en personas entre 18 y 44 años [4].

Las variables del grupo *hábitos* indican la frecuencia semanal de ejercicio, consumo de alcohol, cigarros, horas de sueño y frecuencia de uso de una computadora. Se aprecia en la Figura 3.5 que la mayoría de la población en estudio fuma y bebe poco, pero también realiza poca actividad física semanal. La mayoría utiliza una computadora todos los días y duerme entre 6 y 8 horas por día. Las distribuciones de ambas clases son similares.

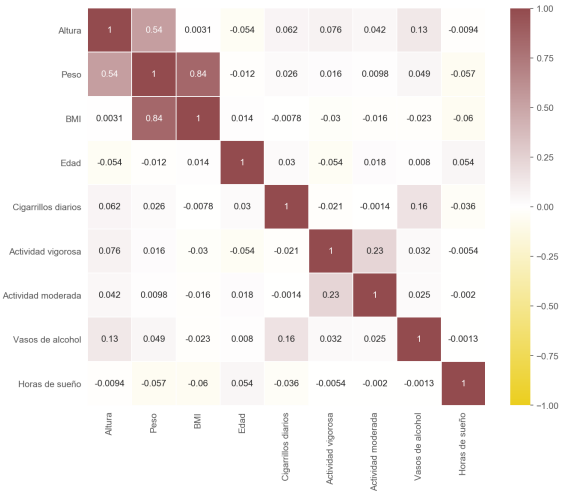

Fig. 3.2: Correlación entre variables contínuas.

Dada la baja proporción de pacientes con enfermedades crónicas, antecedentes médicos y consumo de medicamentos, estas categorías han sido estudiadas sólo para los pacientes que padecen migraña.

En la Figura 3.6 (arriba) se muestran las variables de la categoría *enfermedades crónicas* de aquellos pacientes que sufren migraña. Tal como su nombre lo indica, esta categoría describe el padecimiento de enfermedades como diabetes, asma, condiciones cardíacas y dolencias físicas como de espalda y cuello. Se observa en esta población que la dolencia crónica más presente con el dolor de cuello, cintura y en menor medida mandíbula, pudiendo estar asociadas a las cefaleas [11]. La rinitis alérgica afecta al 10 % de esta población, se trata de una enfermedad estacional capaz de provocar migrañas en quien la padece [12].

La categoría *antecedentes* contempla las enfermedades o episodios que el paciente ha sufrido alguna vez en el pasado: varicela, infarto, derrame cerebral, hepatitis. También contempla enfermedades que haya sufrido en el último año, como sinusitis o colesterol alto. En la Figura 3.6 (medio) se muestran los antecedentes de aquellos pacientes que han sufrido migraña. Se observa que la gran mayoría ha padecido varicela, lo cual no es posible afirmar que exista relación con la clase objetivo dado que es un enfermedad de muy alta incidencia en la población mundial adulta (por el rango etario aquí estudiado, la vacuna para esta enfermedad aún no existía durante la niñez de la mayoría de la población en estudio). Se observa, además, que más del 10 % de esta población han sufrido hipertensión, sinusitis y colesterol alto.

La categoría *medicamentos* indica las drogas que el paciente toma al momento de la encuesta. Se muestra en la Figura 3.6 (abajo) aquellas drogas tomadas (con o sin aviso de un profesional) por pacientes que sufren migrañas. En todos los casos la población que consume alguna de estas drogas es menor al 10 %, siendo hipertensión la de mayor presencia en este grupo.

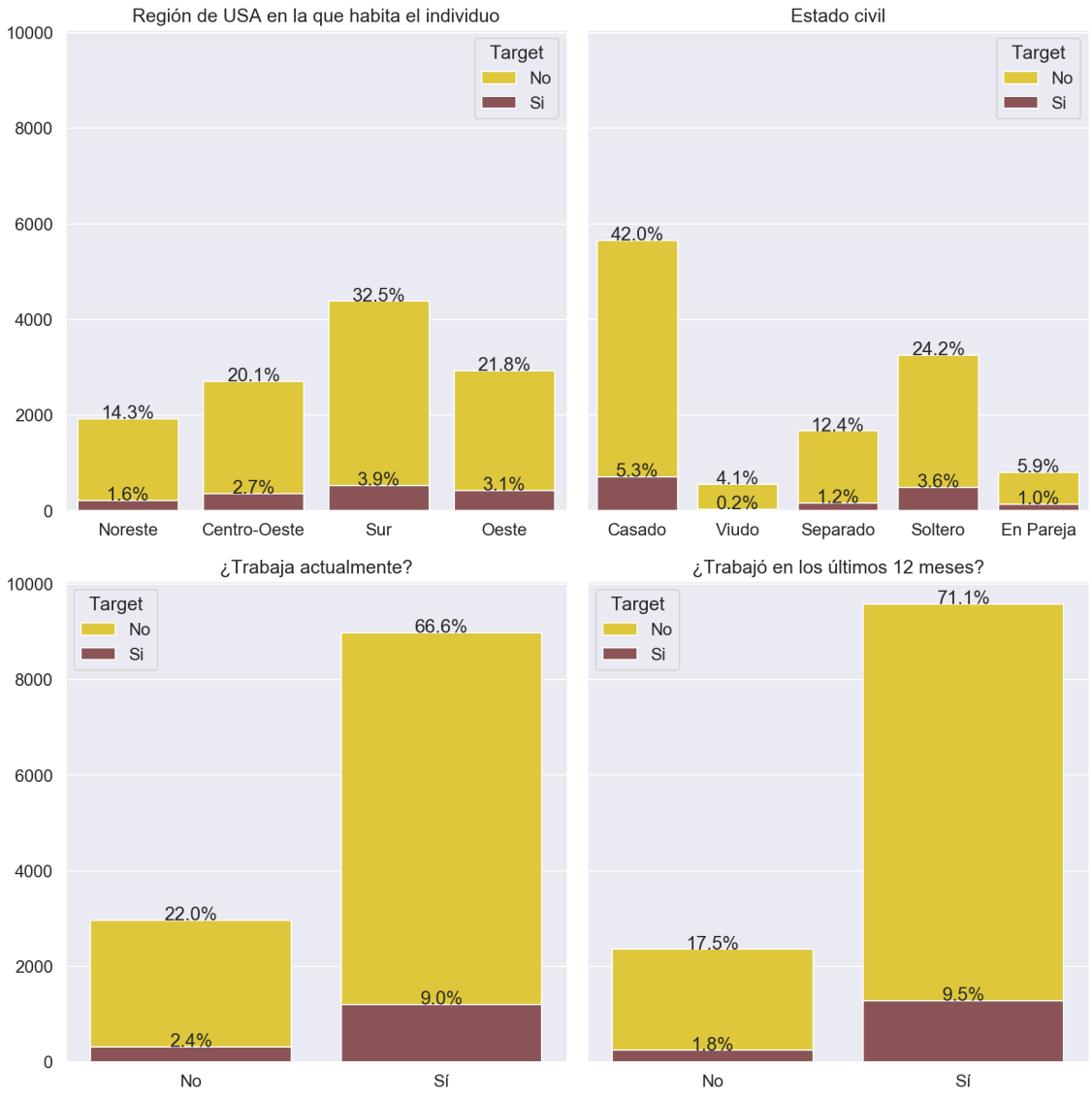

Fig. 3.3: Distribución de clase objetivo en las variables descriptoras del entorno del paciente.

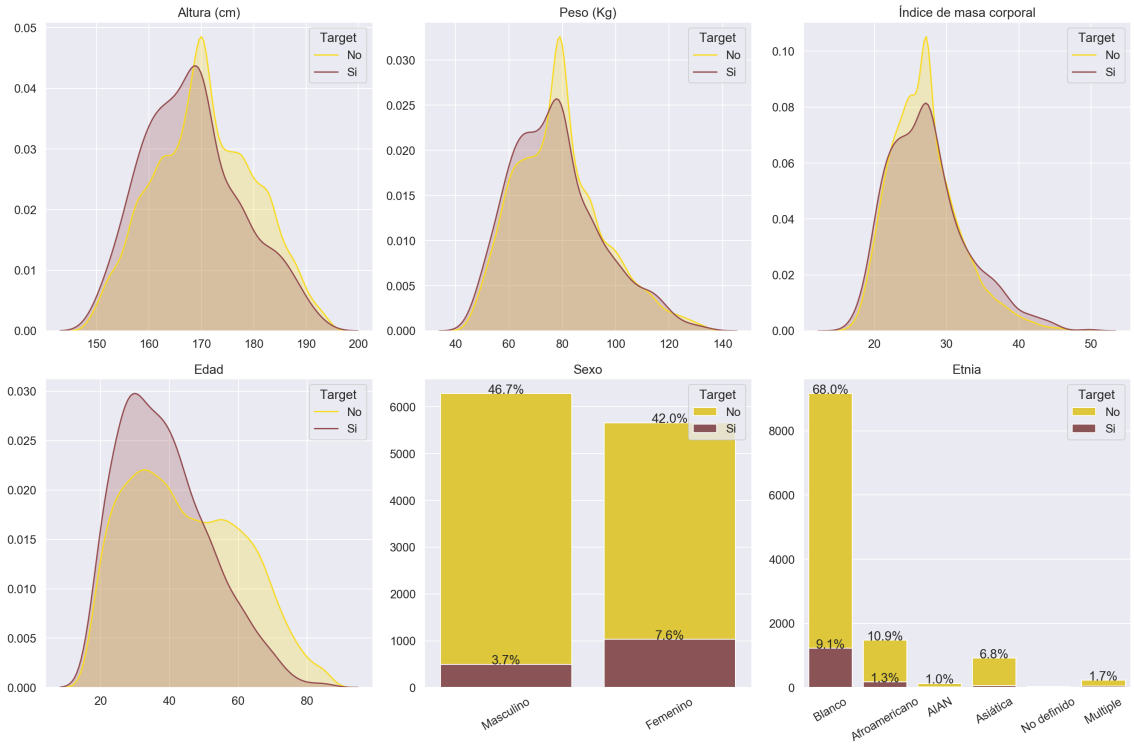

Fig. 3.4: Distribución de clase objetivo en las variables descriptoras de las características físicas del paciente.

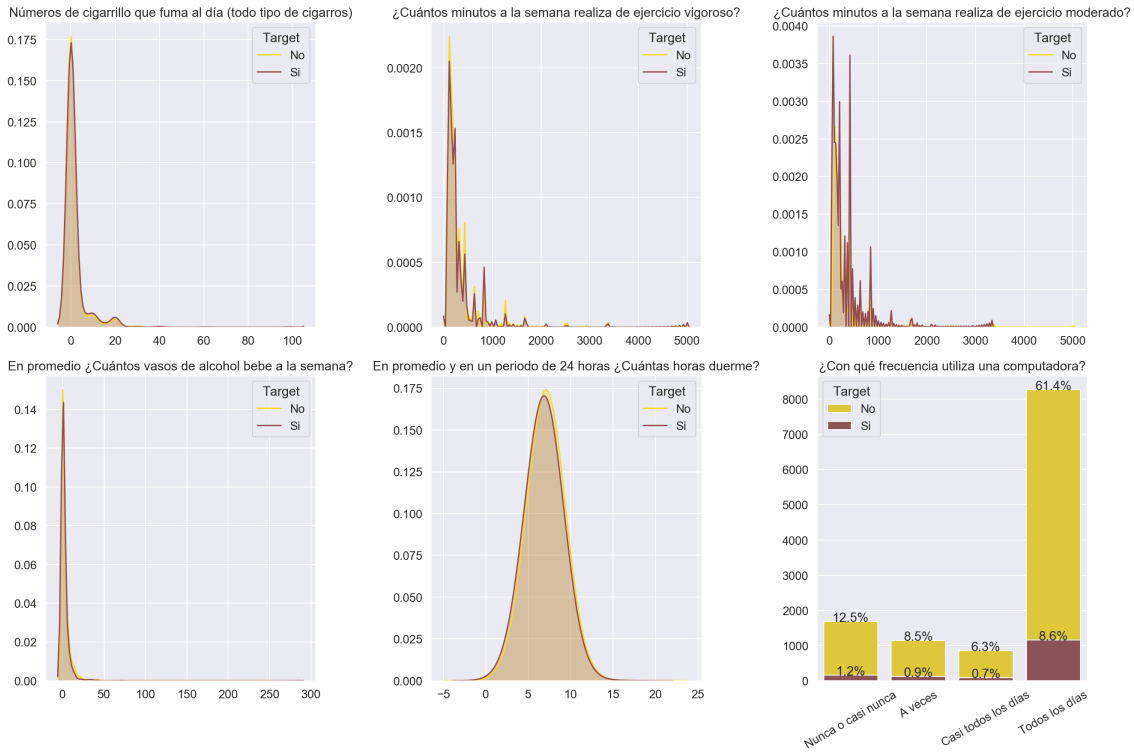

Fig. 3.5: Distribución de clase objetivo en las variables descriptoras de los hábitos del paciente.

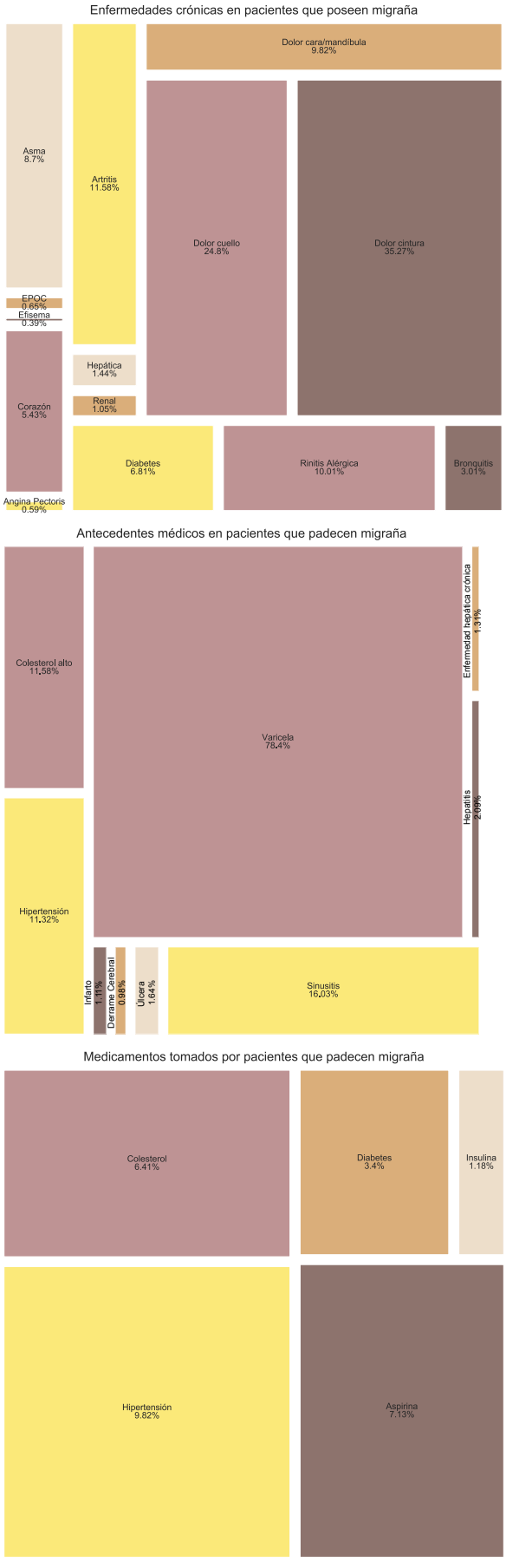

Fig. 3.6: Enfermedades crónicas (arriba), antecedentes médicos (medio) y medicamentos tomados (abajo) por pacientes que sufren migrañas.

4. CLASIFICACIÓN

Como se vio en el capítulo 3, la clase objetivo se encuentra fuertemente desbalanceada. Para equilibrar esta distribución se aplica la técnica de *undersampling*, la cual consiste en descartar casos de la clase mayoritaria (0: el paciente no posee migraña). En la Tabla 4.1 se detalla la proporción de cada clase en los *datasets* original y la muestra.

	Original		Balanceado	
0: No padece migraña	11942	88.7 %	1528	50 %
1: Padece migraña	1528	11.3 %	1528	50 %
Total	13470	100 %	3056	100 %

Tab. 4.1: Cantidad de pacientes en cada clase del *dataset* original y en el *dataset* balanceado.

Para la implementación de las distintas técnicas se separó el *dataset* en dos partes: desarrollo (*dev*) y testeo (*test*). La fracción de *test* se utiliza únicamente para la predicción del modelo final, sin intervenir durante el desarrollo del mismo. La fracción *dev* es separada en *train* y *val* (validación) para entrenar y optimizar los algoritmos. Para evitar el sobreajuste (*overfitting*) se aplica la técnica de validación cruzada (*cross validation*). Para ello, se generaron 5 conjuntos de datos: sobre el *dataset* balanceado se separan de forma sucesiva en entrenamiento (80 %) y validación (20 %). En la Figura 4.1 se detallan los procentaje de registros en cada uno.

4.1. Regresión logística (*benchmark*)

Para fijar un punto de referencia (*benchmark*) se aplica un Modelo Lineal Generalizado (GLM, por sus siglas en inglés *Generalized Linear Model*) utilizando como función de enlace (*link function*) una distribución binomial; esto es conocido como regresión logística. En secciones siguientes, se intenta mejorar el desempeño de esta regresión con técnicas más complejas.

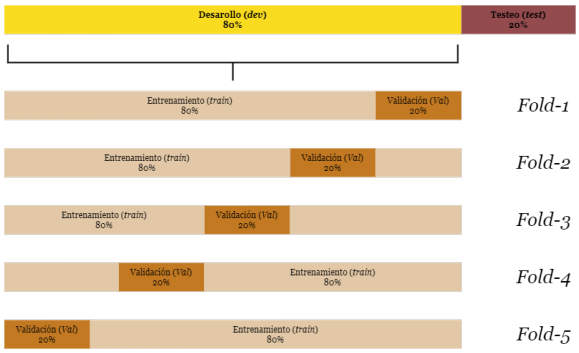

Fig. 4.1: Separación del *dataset* en desarrollo, entrenamiento, validación y testeo.

4.1.1. Selección de variables por regularización *Lasso*

Dada la gran cantidad de variables en el *dataset* (i.e., preguntas en la encuesta) es de particular interés conocer aquellas de mayor importancia a la hora de elaborar un diagnóstico. Para la selección de dichas variables se aplica una regresión logística con regularización *Lasso*: esta técnica permite realizar la selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico.

Dicha técnica busca minimizar la suma de mínimos cuadrados incorporando un parámetro de regularización, λ, el cual controla la penalidad: $\lambda = 0$ implica que ninguna variable sea eliminada, mientras $\lambda = \infty$ implicaría -en la teoría- que todas sean eliminadas.

Para encontrar el parámetro óptimo se utilizó la librería *GLMnet* [13], la cual es capaz de implementarla con validación cruzada (*k-Folds Cross Validation*). Se realizó, sobre el conjunto *dev*, con $k = 10$ utilizando como métrica a optimizar el área bajo la curva ROC (ROC-AUC por sus siglas en inglés *Receiver Operating Characteristic - Area Under the Curve*). El valor de λ obtenido fue de 0.009176; con él quedan seleccionadas las 21 variables de la Tabla 4.2.

Nombre	Pregunta	Categoría
HRTEV	¿Alguna vez fue diagnosticado con una enfermedad del corazón?	Antecedentes
STREV	¿Alguna vez sufrió un derrame cerebral?	Antecedentes
ULCYR	¿Ha tenido una úlcera en los últimos 12 meses?	Antecedentes
SINYR	¿Ha sido diagnosticado con sinusitis en los últimos 12 meses?	Antecedentes
APOX	¿Ha tenido varicela?	Antecedentes
AASSTILL	¿Tiene asma?	Enfermedades crónicas
AHAYFYR	¿Ha tenido rinitis alérgica en los últimos 12 meses?	Enfermedades crónicas
PAINECK	¿Ha sufrido dolor de cuello durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas
PAINLB	¿Ha sufrido dolor de cintura durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas
PAINFACE	¿Ha sufrido dolor de en la cara o la mandíbula durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas
CIGSDAY	Números de cigarrillo que fuma al día (todo tipo de cigarros)	Hábitos
MOD	¿Cuántos minutos a la semana realiza de ejercicio moderado?	Hábitos
ALC	En promedio ¿Cuántos vasos de alcohol bebe a la semana?	Hábitos
ASISLEEP	En promedio y en un periodo de 24 horas ¿Cuántas horas duerme?	Hábitos
HYPMED2	¿Toma actualmente medicamento para la presión (recetada por un médico)?	Medicamentos
AGE_P	Edad	Sujeto
FLA1AR	¿Posee alguna limitación física?	Sujeto
SEX_2	Sexo: femenino	Sujeto
RACERPI2_4	Etnia: Asiática	Sujeto
RACERPI2_6	Etnia: Múltiple	Sujeto
R_MARITL_2	Estado civil: Viudo	Sujeto

Tab. 4.2: Variables seleccionadas con la técnica de regularización *Lasso* para $\lambda = 0{,}009176$.

4.1.2. Combinación óptima de variables

Con las preguntas seleccionadas (Tabla 4.2) se realizaron modelos con diferentes combinaciones de variables sobre el conjunto *dev*. Dichas variables fueron seleccionadas según su categoría, modelando así regresiones que contemplen los temas de las preguntas. Por ejemplo, se modeló la variable *target* en función de las características del sujeto, en función de los hábitos, etc. El listado detallado de estos modelos se encuentra en la Tabla (Tabla 4.3).

Para seleccionar el mejor modelo de ellos se eligió aquel con mayor *deviance* explicada (aquel que la minimiza con respecto a la *deviance* nula). Este corresponde al modelo con

Expresión	Deviance explicada
target ∼ sujeto + hábitos + dolores + medicamentos + enfermedades crónicas + antecedentes	15.09 %
target ∼ sujeto + dolores	12.92 %
target ∼ sujeto + enfermedades crónicas + antecedentes	8.20 %
target ∼ sujeto + hábitos + dolores + medicamentos + enfermedades crónicas + antecedentes	7.14 %
target ∼ dolores	7.03 %
target ∼ sujeto + medicamento	6.14 %
target ∼ sujeto	6.09 %
target ∼ enfermedades crónicas + antecedentes	1.90 %
target ∼ hábitos	1.26 %
target ∼ medicamentos	0.70 %

Tab. 4.3: Combinaciones de variables utilizadas para generar los modelos. Se muestra la *deviance* explicada por cada uno.

todas las variables.

4.1.3. Búsqueda del punto de corte

Para seleccionar el punto de corte óptimo se estudian las métricas de desempeño en función del punto de corte para los 5 conjuntos de *train-val*. Como criterio para su elección, se busca el valor que maximice tanto la sensibilidad (también conocido como *True Positive Rate* o *recall*) como la especificidad (también conocido como *True Negative Rate*, TNR). Se muestra en la Tabla 4.4 el mejor punto de corte para cada conjunto de datos y sus métricas de *performance* (sensibilidad, especificidad y ROC-AUC).

Considerando las métricas mostradas en la tabla 4.4 se decide utilizar el punto de corte del conjunto número 4.

	Corte	ROC-AUC	Sensibilidad	Especificidad
1	0.4785	0.6708	0.6709	0.6706
2	0.5229	0.6728	0.6721	0.6735
3	0.4702	0.5133	0.5143	0.5123
4	0.5030	0.7198	0.7198	0.7198
5	0.4859	0.6462	0.6468	0.6456
Promedio	-	0.6446	0.6448	0.6444

Tab. 4.4: Medidas de desempeño obtenidas para cada uno de los 5 sets de entrenamiento.

4.1.4. Modelo final

Una vez seleccionadas las variables y el punto de corte, se entrena un modelo con el conjunto *dev* para luego predecir el conjunto *test*. Se muestran en la Tabla 4.5 los coeficientes obtenidos para cada variable. Se observa que de las 21 variables seleccionadas por el método de regularización alrededor de la mitad son significativas con nivel 0.05 (p-valor). Dentro de las características del sujeto el sexo (femenino), la edad y la etnia (asiática) están presentes, mientras que dentro de la categoría hábito son significativas las variables de consumo de alcohol, horas de sueño y consumo de cigarrillos. Dentro de las enfermedades crónicas las relacionadas a dolores de espalda son relevantes así como, dentro de la categoría antecedentes, padecer sinusitis; las preguntas restantes dentro de

Nombre	Pregunta	Categoría	Coeficiente	P-valor
Intercepto	-	-	0.9401	0.0029
SEX_2	Sexo: femenino	Sujeto	0.8272	0.0000
PAINECK	¿Ha sufrido dolor de cuello durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas	1.1837	0.0000
AGE_P	Edad	Sujeto	-0.0263	0.0000
PAINLB	¿Ha sufrido dolor de cintura durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas	0.7075	0.0000
SINYR	¿Ha sido diagnosticado con sinusitis en los últimos 12 meses?	Antecedentes	0.6878	0.0000
RACERPI2_4	Etnia: Asiática	Sujeto	-0.8643	0.0000
PAINFACE	¿Ha sufrido dolor de en la cara o la mandíbula durante más de 24 horas seguidas en los últimos 3 meses?	Enfermedades crónicas	0.7666	0.0010
ASISLEEP	En promedio y en un periodo de 24 horas ¿Cuántas horas duerme?	Hábitos	-0.1188	0.0021
ALC	En promedio ¿Cuántos vasos de alcohol bebe a la semana?	Hábitos	-0.0255	0.0048
CIGSDAY	Números de cigarrillo que fuma al día (todo tipo de cigarros)	Hábitos	0.0220	0.0139
HYPMED2	¿Toma actualmente medicamento para la presión (recetada por un médico)?	Medicamentos	-0.3339	0.0238
APOX	¿Ha tenido varicela?	Antecedentes	0.2148	0.0536
STREV	¿Alguna vez sufrió un derrame cerebral?	Antecedentes	0.8997	0.0661
HRTEV	¿Alguna vez fue diagnosticado con una enfermedad del corazón?	Antecedentes	0.4029	0.0795
R_MARITL_2	Estado civil: Viudo	Sujeto	-0.5002	0.0851
ULCYR	¿Ha tenido una úlcera en los últimos 12 meses?	Antecedentes	0.8422	0.0893
RACERPI2_6	Etnia: Múltiple	Sujeto	0.4076	0.1541
FLA1AR	¿Posee alguna limitación física?	Sujeto	0.6897	0.1847
MOD	¿Cuántos minutos a la semana realiza de ejercicio moderado?	Hábitos	-0.0002	0.1893
AHAYFYR	¿Ha tenido rinitis alérgica en los últimos 12 meses?	Enfermedades crónicas	0.2160	0.2193
AASSTILL	¿Tiene asma?	Enfermedades crónicas	0.1762	0.3393

Tab. 4.5

Corte	ROC-AUC	Sensibilidad	Especificidad
0.5030	0.6997	0.6712	0.7281

Tab. 4.6: Métricas de performance para el modelo de regresión logística.

esta categoría no son significativas.

Algunos de los coeficientes obtenidos pueden asociarse con información conocida sobre la enfermedad. Por ejemplo la variable *dummy* sexo femenino presenta un valor positivo, indicando una mayor probabilidad de padecer migraña en mujeres. En el caso de la variable continua edad, su coeficiente es negativo; esto es consistente con lo enunciado previamente sobre la mayor incidencia de migraña en el rango etario 18-44 años. Cabe mencionar que dentro de las variables *dummy* correspondientes a etnia sólo se conserva la etnia asiática, siendo su coeficiente negativo. Así como la población caucásica es la que más padece de migrañas, estudios parecen indicar que dicha enfermedad tiene menor incidencia en la población de etnia asiática [14].

En la Tabla 4.6 se muestran las métricas de *performance* del modelo. Estos valores son utilizados como punto de referencia para las técnicas más complejas que se detallan en las secciones siguientes.

4.2. Random Forests

Random Forests [15] es un método de predicción conocido como *bagging*, en el que se combinan predictores débiles para resolver problemas complejos. Es un ensamble, en el que diversos árboles de decisión aprenden por separado y sus resultados se combinan en el paso final para devolver la predicción. La desventaja de los métodos de ensamble es que a medida que se complejizan tienden a perder interpretabilidad.

Con los mismos *datasets* detallados al comienzo del capítulo, se optimizan los parámetros que maximicen la métrica ROC-AUC y se realiza una predicción sobre el set de datos *test*. Se intenta superar el valor obtenido con la regresión logística $ROC - AUC = 0{,}6997$.

Parámetro	Límite inferior	Límite superior	Óptimo
n_estimators	5000	10000	6119
max_depth	100	5000	234
min_samples_split	2	500	245

Tab. 4.7: Rango de valores y óptimo obtenido para cada parámetro buscado de *Random Forests*.

4.2.1. Búsqueda de parámetros

El método *Random Forests* cuenta con una gran cantidad de parámetros que influyen en el desempeño del mismo. Una técnica ampliamente utilizada para este proceso es la Búsqueda Bayesiana; esta técnica logra encontrar parámetros óptimos de forma automática sin requerir largos tiempos de procesamiento como lo haría una técnica por fuerza bruta [16].

En el presente trabajo, dicha búsqueda es implementada para encontrar los valores óptimos de 3 parámetros importantes de *Random Forests*: cantidad de estimadores (*n_estimators*, la cantidad de árboles de decisión en el ensamble), profundidad (*max_depth*, la profundidad máxima a alcanzar por cada árbol) y (*min_samples_split*, la cantidad mínima de datos requeridas para crear una bifurcación en el nodo). Se realizan 100 iteraciones de la búsqueda con validación cruzada (*5-folds cross validation*) optimizando la métrica ROC-AUC. En la tabla 4.7 se muestra el rango de valores buscado y los óptimos obtenidos de cada parámetro.

4.2.2. Ensamble final

Una vez obtenidos los parámetros se entrenó el algoritmo con el *dataset* de desarrollo (*dev*) y se predijo el conjunto de *test*. Las métricas obtenidas se muestran en la tabla 4.8; se observa una mejoría en las métricas ROC-AUC y Sensibilidad con respecto a la regresión logística, pero igual desempeño con respecto a la especificidad.

En la Figura 4.2 se muestran las 10 variables más importantes para esta clasificación. Consistentemente con las variables relevantes en el modelo de regresión logística, entre las más influyentes se encuentran aquellas preguntas relacionadas a dolores de espalda, edad, sexo (femenino), etnia (asiática), horas de sueño y antecedentes de sinusitis. Aparecen en este clasificador también la altura y peso del paciente como relevantes, algo que no se observó en el modelo lineal generalizado.

ROC-AUC	Sensibilidad	Especificidad
0.7221	0.7192	0.7250

Tab. 4.8: Métricas de performance para *Random Forests*.

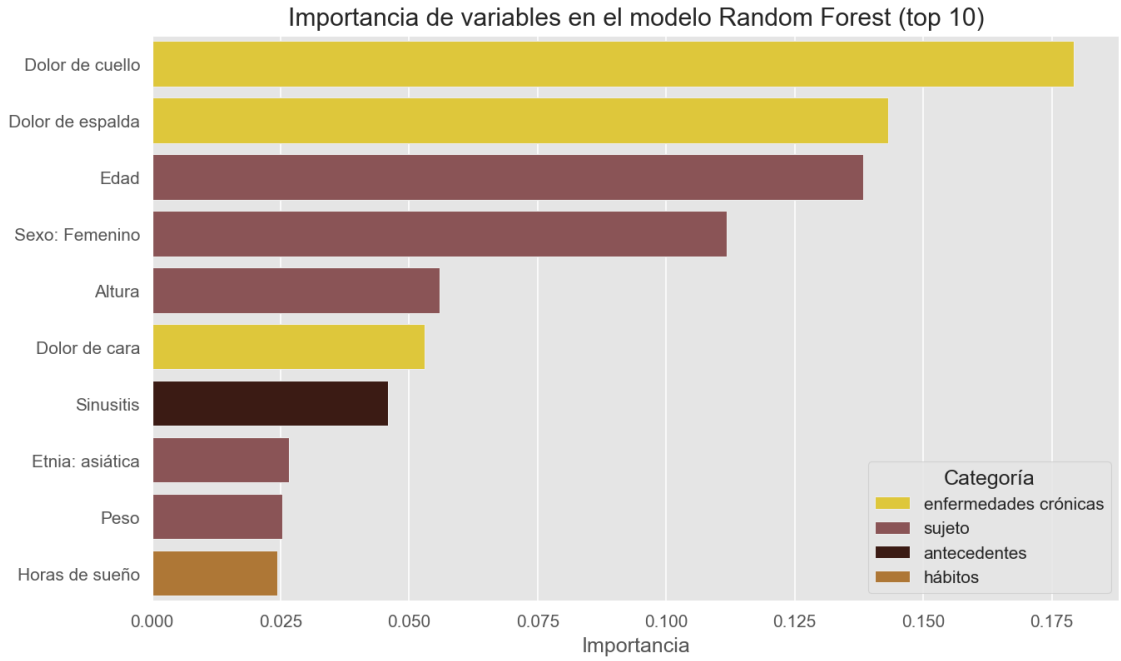

Fig. 4.2: Importancia de variables en *Random Forests*; se muestran las 10 más importantes.

4.3. XGBoost

El último método de predicción implementado es *XGBoost* [17], un ensamble del tipo *boosting*. Al igual que *Random Forests* esta técnica utiliza una gran cantidad de árboles de decisión, con la diferencia de que no aprenden de forma independiente, sino que los árboles siguientes asignan distintos pesos a los casos mal clasificados por el árbol anterior.

Análogamente a la sección anterior, se utilizan los *datasets* descriptos al comienzo del capítulo y se buscan los parámetros óptimos a través de una búsqueda Bayesiana.

4.3.1. Búsqueda de parámetros

XGBoost cuenta con muchos más parámetros que *Random Forests*; en este caso se decidió optimizar la cantidad de árboles en el ensamble (*n_estimators*), la profundidad de dichos árboles (*max_depth*), cantidad mínima de datos en un nodo (*min_child_weight*), tasa de aprendizaje (*learning_rate*), proporción de datos a utilizar en cada árbol (*subsample*), cantidad de variables a utilizar en cada árbol (*colsample_bytree*) y *gamma*, un parámetro de regularización. En la tabla 4.9 se muestra el rango de valores buscado y los óptimos obtenidos para cada uno.

4.3.2. Ensamble final

Con los parámetros obtenidos se entrenó el *dataset* de desarrollo (*dev*) para predecir el de testeo (*test*). En la tabla 4.10 se muestran las métricas de desempeño obtenidas,

Parámetro	Límite inferior	Límite superior	Óptimo
n_estimators	5000	10000	10000
max_depth	100	5000	5000
min_child_weight	2	500	6
learning_rate	0,0001	0,4	0,24
subsample	0,1	1	0,86
colsample_bytree	0,1	1	0,96
gamma	0	20	14,6

Tab. 4.9: Rango de valores y óptimo obtenido para cada parámetro buscado de *XGBoost*.

ROC-AUC	Sensibilidad	Especificidad
0.7255	0.6917	0.7593

Tab. 4.10: Métricas de performance para *XGBoost*.

observando un valor de ROC-AUC muy cercano al obtenido con *Random Forests*, una disminución en la sensiblidad con respecto al mismo (aunque levemente mayor que la sensibilidad obtenida con el modelo de regresión logística) y una especificidad mayor a las dos técnicas anteriores.

En la figura 4.3 se muestran las 10 variables más importantes en la clasificación *XGBoost*, observándose consistencia con las dos técnicas anteriores. La diferencia con respecto a ellos es la aparición de otras variables de la categoría *hábitos*, siendo las horas de trabajo frente a una computadora y la actividad física moderada aquellas más importantes.

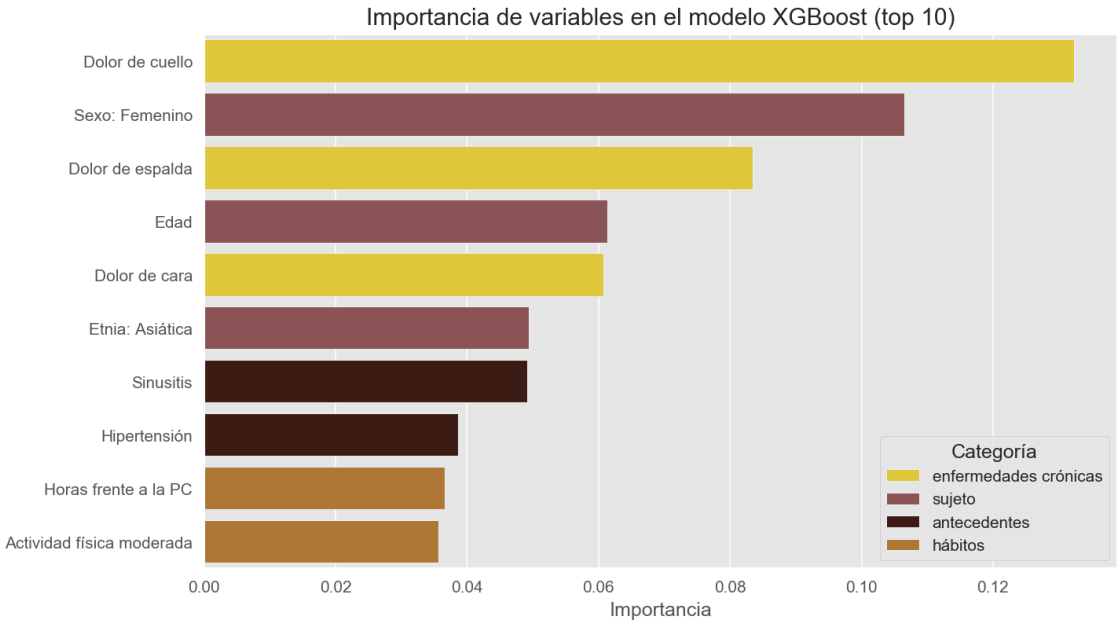

Fig. 4.3: Importancia de variables en *XGBoost*; se muestran las 10 más importantes.

4.4. Comparación

En la tabla 4.11 se resumen las métricas obtenidas para cada técnica implementada. Se observa en las métricas ROC-AUC y sensibilidad un leve incremento en los métodos de ensamble con respecto a la regresión logística. La especificidad incrementa con la técnica *XGBoost* con respecto al modelo lineal generalizado, pero disminuye levemente en el ensamble *Random Forests*. Esto deja en evidencia lo potente de un modelo simple como lo es la regresión logística el cual presenta una interpretación más directa de sus coeficientes, a diferencia de los ensambles en los que, si bien puede obtenerse la importancia de variables, no es posible conocer cómo influyen en la clasficiación.

Técnica	ROC-AUC	Sensibilidad	Especificidad
Regresión Logística	0,6997	0,6712	0,7281
Random Forests	0,7221	0,7192	0,7250
XGBoost	0,7255	0,6917	0,7593

Tab. 4.11: Comparación de métricas de las tres técnicas estudiadas.

En la tabla 4.12 se listan las variables más importantes para cada técnica; para los métodos de ensamble se listan las 10 más importantes mientras que para la regresión logística se listan aquellos variables estadísticamente significativas (nivel 0.05). Se observa que para los tres métodos las variables importantes de la categoría sujeto -aquellas variables relacionadas a características personales del paciente- incluyen el sexo (femenino), edad y etnia (asiática). En la categoría de enfermedades crónicas los tres clasificadores presentan las mismas tres variables de mayor importancia, aquellas relacionadas a dolores de espalda (cuello, espalda y cara/mandíbula). En cuanto a antecedentes, la variable común para todos es la pregunta que indica si el paciente ha sufrido sinusitis en los últimos 12 meses. Por último, en la categoría hábitos no se encontró coincidencia como en las categorías anteriores, influyendo en cada clasificador variables distintas.

Categoría	**Variable**	**Regresión Logística**	***Random Forests***	***XGBoost***
Sujeto	Sexo: femenino	x	x	x
	Edad	x	x	x
	Etnia: Asiática	x	x	x
	Altura		x	
	Peso		x	
Enfermedades crónicas	Dolor de cuello	x	x	x
	Dolor de espalfa	x	x	x
	Dolor de cara	x	x	x
Antecedentes	Sinusitis	x	x	x
	Hipertensión	x		x
Hábitos	Horas de sueño	x	x	
	Consumo de alcohol	x		
	Consumo de cigarrillos	x		
	Horas frente a la pc			x
	Actividad física moderada			x

Tab. 4.12: Variables importantes (top 10) de los métodos *Random Forests* y *XGBoost* y variables estadísticamente significativas (nivel 0.05) del modelo de regresión logística.

5. DISCUSIÓN Y CONCLUSIONES

Los tres clasificadores detallados en el capítulo 4 han logrado un desempeño similar; ninguno de los algoritmos de ensamble implementado han superado ampliamente al *benchmark* establecido con la regresión logística. Esto evidencia lo potente de una técnica simple como lo es un modelo lineal generalizado.

Dada la sencilla interpretabilidad de los modelos lineales y el desempeño similar a técnicas más complejas, la regresión logística podría ser de utilidad para elaborar un cuestionario corto y específico para el diagnóstico de migrañas. Dentro de los coeficientes obtenidos se observa que se refleja información conocida sobre esta condición: mayor incidencia en mujeres [4], menor incidencia en etnia asiática [14], menor incidencia en adultos mayores [4] y relación con dolores de espalda y cuello [11]. La pregunta con respecto a antecedentes temporalmente cercanos en sinusitis también tiene un coeficiente positivo; esto podría deberse a que los dolores de cabeza por sinusitis y las migrañas son frecuentemente confundidos [18]. Con respecto a los hábitos del paciente, se obtuvieron coeficientes negativos en la variable relacionada al sueño, indicando que los pacientes que duermen más horas tienen menos probabilidad de sufrir migrañas. La relación entre desórdenes de sueño y dolores de cabeza es conocida y estudiada, agravando la falta de sueño esta condición [19, 20]. En el caso del consumo de alcohol se obtuvo un coeficiente negativo, lo cual indicaría que un mayor consumo semanal de bebidas alcohólicas implicaría una menor probabilidad de sufrir migrañas. Esta relación no podría afirmarse, dado que el vínculo entre consumo de alcohol y cefaleas está fuertemente asociado al tipo de bebida alcohólica (vino blanco, tinto, espumantes, cerveza) [21] y la encuesta aquí utilizada como fuente de datos no contempla esa diferenciación. Esta consistencia entre los parámetros de regresión y datos conocidos sobre la enfermedad evidencian que este modelo es capaz de reflejar la realidad.

En los métodos de ensamble posteriormente aplicados se han obtenido como variables importantes preguntas en común con el modelo de regresión logística. Las preguntas relaciondas a características del sujeto, dolores crónicos y antecedentes de sinusitis han sido importantes para la clasificación. Sin embargo, en los ensambles de *Random Forests* y *XGBoost* no es posible interpretar cómo es la influencia de estas variables en la clase objetivo.

Como se mencionó anteriormente, las consultas médicas por migraña comprenden más del 4 % del total de consultas en los Estados Unidos. Dado que el diagnóstico de esta enfermedad se basa principalmente en una entrevista con el paciente, un clasificador como el desarrollado en el presente trabajo podría implementarse con el fin de reducir la cantidad de visitas o su duración. Con el avance de las comunicaciones hoy en día es posible realizar consultas médicas a distancia; una posible implementación podría ser a través de un sistema de pre-consulta en el que el paciente responda el cuestionario para luego pasar a la consulta con el profesional. En base a las respuestas del cuestionario y la salida del clasificador el profesional de la salud podrá decidir en qué temas ahondar y si son necesarios exámenes posteriores.

Queda para futuras investigaciones mejorar el desempeño de estos clasificadores. Dado que la incidencia de la enfermedad en la población mundial es de alrededor del 11 % el *dataset* final utilizado se reduce notablemente con respecto a su tamaño original. Podrían incorporarse resultados de encuestas de años anteriores, para así aumentar la cantidad de casos en estudio. Si los datos recopilados son suficientes, incluso podría entrenarse un algoritmo complejo de *deep learning*.

Es importante mencionar las limitaciones de las técnicas aquí implementados. La población en estudio -y utilizada para entrena los algoritmos- corresponde a población adulta de los Estados Unidos que gozan de buena salud. Si bien el *dataset* original presenta diversidad en cuanto a la etnia de los pacientes, la mayoría corresponde a etnia caucásica. Es importante tener presente esta información al momento de utilizar estos métodos predictivos, dado que un set de datos sesgado dará lugar a resultados sesgados. Se aconseja conocer los detalles sobre los datos de entrenamiento para no dar lugar a interpretaciones erróneas de la clasificación.

En el presente trabajo fue posible implementar tres métodos de clasificación para predecir si el paciente padece o no migraña en base a las respuestas de un cuestionario. Para los tres métodos (regresión logística, *Random Forests* y *XGBoost*) se obtuvieron medidas de desempeño cercanos a 0.7, tanto para ROC-AUC, sensibilidad y especificidad. Los coeficientes del modelo lineal son consistentes con datos conocidos sobre la enfermedad. Es posible seguir trabajando sobre estoss algoritmos para mejorar su desempeño, pudiendo tener utilidad como herramienta de asistencia a profesionales de la salud para diagnosticar la enfermedad.

6. DISPONIBILIDAD DE LOS DATOS Y REPRODUCIBILIDAD DEL MÉTODO

Los datos utilizados en el presente trabajo son públicos y pueden descargarse de su fuente[1]. Los datos transformados y el código utilizado están disponibles en un repositorio público[2]. Es posible reproducir la metodología seguida aquí utilizando los *IPython Notebooks* allí presentes. El análisis aquí presentado fue realizado utilizando los lenguajes de programación *Python 3.7* y *R 3.6*.

[1] `www.cdc.gov/nchs/nhis/`

[2] `https://github.com/ponybiam/DMKD-UBA/tree/master/Trabajo-Especializacion`

Bibliografía

[1] Daniela Pietrobon. Migraine: new molecular mechanisms. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 11(4):373–86, aug 2005.

[2] Claudia Gasparini, Heidi Sutherland, and Lyn Griffiths. Studies on the Pathophysiology and Genetic Basis of Migraine. *Current Genomics*, 14(5):300–315, aug 2013.

[3] L. J. Stovner, K. Hagen, R. Jensen, Z. Katsarava, R. B. Lipton, A. I. Scher, T. J. Steiner, and J. A. Zwart. The global burden of headache: A documentation of headache prevalence and disability worldwide, mar 2007.

[4] Migraine Facts - Migraine Research Foundation. Consultado el 11/06/2020.

[5] T. J. Steiner, A. I. Scher, W. F. Stewart, K. Kolodner, J. Liberman, and R. B. Lipton. The prevalence and disability burden of adult migraine in England and their relationships to age, gender and ethnicity. *Cephalalgia*, 23(7):519–527, sep 2003.

[6] Stewart J. Tepper, Carl G.H. Dahlöf, Andrew Dowson, Lawrence Newman, Hank Mansbach, Martin Jones, Ba Pham, Chris Webster, and Reijo Salonen. Prevalence and diagnosis of migraine in patients consulting their physician with a complaint of headache: Data from the landmark study. *Headache*, 44(9):856–864, oct 2004.

[7] Mark W. Weatherall. The diagnosis and treatment of chronic migraine, 2015.

[8] salud — Definición — Diccionario de la lengua española — RAE - ASALE. Consultado el 24/05/2020.

[9] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, mar 2011.

[10] Daniel F. Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *American Statistician*, 50(3):207–213, 1996.

[11] Arani Vivekanantham, Claire Edwin, Tamar Pincus, Manjit Matharu, Helen Parsons, and Martin Underwood. The association between headache and low back pain: A systematic review, jul 2019.

[12] Min Ku, Bernard Silverman, Nausika Prifti, Wei Ying, Yudy Persaud, and Arlene Schneider. Prevalence of migraine headaches in patients with allergic rhinitis. *Annals of Allergy, Asthma and Immunology*, 97(2):226–230, aug 2006.

[13] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, feb 2010.

[14] Walter F. Stewart, Richard B. Lipton, and Joshua Liberman. Variation in migraine prevalence by race. *Neurology*, 47(1):52–59, jul 1996.

[15] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001.

[16] Jia Wu, Xiu Yun Chen, Hao Zhang, Li Dong Xiong, Hang Lei, and Si Hao Deng. Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, 17(1):26–40, mar 2019.

[17] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 785–794, New York, NY, USA, aug 2016. Association for Computing Machinery.

[18] Roger K. Cady and Curtis P. Schreiber. Sinus headache or migraine? Considerations in making a differential diagnosis. *Neurology*, 58(9 SUPPL. 6):S10–S14, may 2002.

[19] Jeanetta C. Rains and J. Steven Poceta. Sleep-Related Headaches, nov 2012.

[20] Leslie Kelman and Jeanetta C. Rains. Headache and sleep: Examination of sleep patterns and complaints in a large clinical sample of migraineurs. *Headache*, 45(7):904–910, jul 2005.

[21] Alessandro Panconesi. Alcohol and migraine: Trigger factor, consumption, mechanisms. A review, feb 2008.

APÉNDICE

Tab. 6.1: Diccionario de variables.

Código	Pregunta	Niveles
id	identificación única	-
REGION	Región de USA en la que habita el individuo	1 Noreste 2 Centro-oeste 3 Sur 4 Oeste
SEX	Sexo	1 Masculino 2 Femenino
AHEIGHT	Altura (cm)	-
AWEIGHTP	Peso (Kg)	-
BMI	Índice de masa corporal	-
AGE_P	Edad	00 menor de 1 año 85 mayor de 85 años
RACERPI2	Etnia	01 Blanco (incluye latinos) 02 Afroamericana 03 AIAN 04 Asiática 05 No definido 06 Multiple
R_MARITL	Estado civil	1 casado 2 viudo 3 separado 4 soltero 5 en pareja NaN desconocido
DOINGLWA	¿Trabaja actualmente?	0 No 1 Si NaN desconocido
WRKLYR4	¿Trabajó en los últimos 12 meses?	0 No 1 Si NaN desconocido
HYPYR1	¿Ha sufrido hipertemsión en los últimos 12 meses?	0 No 1 Si NaN desconocido
HYPMED2	¿Toma actualmente medicamento para la presión (recetada por un médico)?	0 No 1 Si NaN desconocido
CHLYR	¿Ha tenido colesterol alto en los últimos 12 meses?	0 No 1 Si NaN desconocido
CHLMDNW2	¿Toma actualmente medicamento para reducir el colesterol (recetado por un médico)?	0 No 1 Si NaN desconocido
CHDEV	¿Alguna vez fue diagnosticado con algún tipo de enfermedad coronaria?	0 No 1 Si NaN desconocido
ANGEV	¿Algna vez fue diagnosticado con angina pectoris?	0 No 1 Si NaN desconocido
MIEV	¿Alguna vvez tuvo un infarto?	0 No 1 Si NaN desconocido
HRTEV	¿Alguna vez fue diagnosticado con una enfermedad del corazón?	0 No 1 Si NaN desconocido
STREV	¿Alguna vez sufrió un derrame cerebral?	0 No 1 Si NaN desconocido
EPHEV	¿Alguna vez fue diagnosticado con efisema?	0 No 1 Si NaN desconocido
COPDEV	¿Alguna vez fue diagnosticado con EPOC?	0 No 1 Si NaN desconocido
ASP	Toma aspirina actualmente?	0 No 1 Si NaN desconocido
AASSTILL	¿Tiene asma?	0 No 1 Si NaN desconocido

Código	**Pregunta**	**Niveles**
ULCYR	¿Ha tenido una úlcera en los últimos 12 meses?	0 No 1 Si NaN desconocido
DIBEV1	¿Ha sido diagnosticado alguna vez con diabetes o prediabetes?	0 No 1 Si NaN desconocido
DIBPILL1	¿Toma actualmente medicamento para la diabetes?	0 No 1 Si NaN desconocido
INSLN1	¿Toma actualmente insulina?	0 No 1 Si NaN desconocido
AHAYFYR	¿Ha tenido rinitis alérgica en los últimos 12 meses?	0 No 1 Si NaN desconocido
SINYR	¿Ha sido diagnosticado con sinusitis en los últimos 12 meses?	0 No 1 Si NaN desconocido
CBRCHYR	¿Ha sido diagnosticado con bronquitis crónicaen los últimos 12 meses?	0 No 1 Si NaN desconocido
KIDWKYR	¿Ha sido diagnosticado con algún tipo de falla renal en los últimos 12 meses?	0 No 1 Si NaN desconocido
LIVYR	¿Ha sido diagnosticado con algún tipo de falla hepática en los últimos 12 meses?	0 No 1 Si NaN desconocido
ARTH1	¿Ha sido diagnosticado alguna vez con alguna forma de artritis, artritis reumatoide, gota, lupus o fibromialgia?	0 No 1 Si NaN desconocido
PAINECK	¿Ha sufrido dolor de cuello durante más de 24 horas seguidas en los últimos 3 meses?	0 No 1 Si NaN desconocido
PAINLB	¿Ha sufrido dolor de cintura durante más de 24 horas seguidas en los últimos 3 meses?	0 No 1 Si NaN desconocido
PAINFACE	¿Ha sufrido dolor de en la cara o la mandíbula durante más de 24 horas seguidas en los últimos 3 meses?	0 No 1 Si NaN desconocido
FLA1AR	¿Posee alguna limitación física?	0 No 1 Si NaN desconocido
CIGSDAY	Números de cigarrillo que fuma al día (todo tipo de cigarros)	-
VIG	¿Cuántos minutos a la semana realiza de ejercicio vigoroso?	-
MOD	¿Cuántos minutos a la semana realiza de ejercicio moderado?	-
ALC	En promedio ¿Cuántos vasos de alcohol bebe a la semana?	-
APOX	¿Ha tenido varicela?	0 No 1 Si NaN desconocido
AHEP	¿Ha tenido hepatitis?	0 No 1 Si NaN desconocido
LIVEV	¿Ha sido diagnosticado alguna vez con una enfermedad hepática crónica?	0 No 1 Si NaN desconocido
ASICPUSE	¿Con qué frecuencia utiliza una computadora?	0 Nunca o casi nunca 1 A veces 2 Casi todos los días 3 Todos los días NaN desconocido
ASISLEEP	En promedio y en un periodo de 24 horas ¿Cuántas horas duerme?	-
AMIGR	**¿Ha tenido migrañas o cefaleas severas en los últimos 3 meses?**	**0 No** **1 Si**