2021-1

Modelo predictivo de la prognosis a los 6 meses de pacientes de accidentes cerebrovasculares usando técnicas de machine learning

TESIS DE PREGRADO

AUTOR: ANDRÉS FELIPE LUNA

ASESOR: CARLOS FELIPE ARBOLEDA

Contenido

1 Introducción

Un accidente cerebrovascular ocurre cuando el suministro de sangre al cerebro se reduce, o también, se interrumpe. Esto hace que el tejido cerebral deje de recibir oxígeno y nutrientes. Las partes afectadas en el cerebro comienzan a morir en cuestión de minutos. Hay tres categorías que conforman la totalidad de los accidentes cerebrovasculares. El primero se llama accidente cerebrovascular isquémico que se da cuando los vasos sanguíneos del cerebro se bloquean o se estrechan. El segundo se llama un accidente cerebrovascular hemorrágico que se produce cuando un vaso sanguíneo se rompe o cuando tiene una filtración. El tercer y último es el accidente isquémico transitorio y ocurre cuando hay un bloqueo de sangre en el cerebro por no más de 10 minutos (Mayo Clinic, 2020).

De acuerdo con el World Stroke Organización, anualmente hay 13.7 millones casos nuevos de accidentes cerebrovasculares (ACV). Globalmente, el 25% de las personas mayores a 25 años van a sufrir por lo menos un ACV en su vida. En el 2016, 9.5 millones de personas sufrieron de ACV isquémico, y más de 2.7 millones de personas murieron a causa de derrame isquémico. En el 2016 se registraron 4.1 millones de ACV's hemorrágicos a nivel mundial, de estos 2.8 millones murieron. En el 2016 había aproximadamente 80 millones de personas vivas que habían sufrido un derrame cerebral en algún momento de su viva. Note que a pesar de que la incidencia de los accidentes cerebrovasculares hemorrágicos es menor, la mortalidad es mayor (World Stroke Organization, 2019).

Solo en Estados Unidos hay 7 millones de personas vivas que tuvieron un derrame cerebral en algún punto de su vida. Cada año más de 795.000 personas tienen un ACV (Virani, y otros, 2020). De estos, el 87% son de tipo isquémico, 10% de tipo hemorrágico y 3% son ataques isquémicos transitorios. Al día, en Estados Unidos 19 personas mueren por un accidente cerebrovascular. Adicionalmente, los derrames cerebrales son la quinta causa de muerte en Estados Unidos luego de enfermedades del corazón, cáncer, enfermedades respiratorias crónicas y accidentes (Virani, y otros, 2020).Los ACVs son la principal razón para la discapacidad de largo plazo en Estados Unidos (Virani, y otros, 2020).

En Colombia, los accidentes cerebrovasculares constituyen la segunda causa de mortalidad y se estima que al año se presentan 45.000 nuevos casos (Portafolio, 2018). Además, según la Asociación Colombiana de Neurología los accidentes cerebrovasculares son el principal motivo de discapacidad (El Hospital, 2018).

Con el avance de la tecnología, machine learning se ha ido asentando en diversos campos, incluido el de la salud. Existen muchas aplicaciones de machine learning dentro de esta disciplina. Una de ellas es predecir la prognosis de un caso según el estado del paciente y la enfermedad. Entender y ser capaz de evaluar con precisión una prognosis tiene relevancia documentada para los pacientes, sus familias y el sistema de salud en su conjunto. La importancia de predecir las posibilidades de supervivencia es relevante para asistir tratamientos personalizados para los pacientes, así como para ayudar a mejorar el desempeño a nivel institucional a la hora del manejo de recursos y la atención a pacientes. Este es el caso actualmente con el área de cáncer, donde el aprendizaje estadístico se ha convertido en una herramienta importante para tener prognosis precisas (Gupta, y otros, 2014.).

Hay una necesidad creciente en la salud de encontrar y construir modelos predictivos que permitan saber con un margen de tiempo razonable si el tratamiento brindado mejora la prognosis del

paciente o si es necesario cambiar el plan de tratamiento. En este trabajo se pretende aplicar la tecnología de machine learning para el caso de los accidentes cerebro vasculares y al tratamiento de estos. De esta manera, se busca que machine learning sea un instrumento adicional para los médicos y profesionales de la salud.

Este trabajo se centra en construir un modelo que sea capaz de predecir la prognosis a los seis meses de una persona que tuvo un ACV isquémico. Los modelos construyen a partir de técnicas de machine learning.

2 Revisión de la literatura

Con el avance de la tecnología en el campo médico y el avance de las técnicas de statistical learning se ha desarrollado un interés para predecir la prognosis de los pacientes después de un accidente cerebrovascular. Heo et al. (Heo, y otros, 2019)usaron redes neuronales, random forest y regresión logística para predecir la prognosis de un paciente a los 3 meses el evento inicial. Lo que encontraron estos autores es que se puede predecir la prognosis a los 3 meses de un paciente de un ACV isquémico, con un AUC de 0.888 usando redes neuronales. Los autores del paper *Machine learning to predict mortality after rehabilitation among patients with severe stroke* querían predecir la mortalidad a los 3 años luego de salir de la rehabilitación con un algoritmo basado en arboles de decisión. El mejor modelo fue random forest con la implementación del minority oversampling technique, el cual logró un AUC de 0.928 (Scrutinio, y otros, 2020). Yu et al. (Yu, Park, Lee, Pyo, & Lee, 2020) usaron técnicas de machine learning al considerar un árbol de decisión. El objetivo era clasificar la severidad del accidente cerebrovascular. El árbol se construye con 13 de 18 variables propuestas. Los datos usados son los del National Institutes of Health Stroke Scale entre las personas mayores de 65 años. Con esta técnica se logró tener un accuracy del 91.11%. Xie et al. (Xie, Jiang, Gong, & Li, 2018) querían predecir el nivel de discapacidad, discapacidad o no, a los 90 días con la información recogida en las primeras 24 horas de admisión al hospital. Los predictores incluían la información de escáneres, demografía e información clínica. En este trabajo se usó Extreme Boosting y Gradient Boosting como los algoritmos de predicción. El algoritmo que mejor se desempeño fue Gradient Boosting con una AUC de 0.748.

De los trabajos revisados, todos se detienen en la construcción de modelo y las métricas de desempeño de sus modelos. Estos no interpretan la información que les transmita los resultados para poder generar un plan de acción basado en los resultados del modelo de predicción.

3 Objetivos

El objetivo principal de este trabajo es construir un modelo basado en árboles de decisión que tenga una buena capacidad predictiva de la prognosis a los 6 meses, de pacientes que han sufrido accidentes cerebrovasculares isquémicos. Para ello, se debe hacer un adecuado preprocesamiento de los datos y se deben calibrar todos los modelos candidatos.

Sin embargo, no solo se busca que el modelo tenga un buen desempeño, sino que además arroje resultados interpretables. Lo anterior pues los profesionales de la salud podrían usar este modelo para tomar mejores decisiones clínicas con el fin de mejorar la prognosis en 6 meses de los pacientes de ACVs.

4 Datos

La base de datos utilizada es la de The International Stroke Trial (IST) (Sandercock, Niewada, & Czlonkowska, 2011) fase 2 que se descargó gratuitamente en la página de La Universidad de

Edimburgo. IST fue un estudio conducido entre 1991 y 1996, incluyendo la fase de prueba que se realizó entre 1991 y 1993. En este se tomó a pacientes de accidentes cerebrovasculares isquémicos agudos. En ese sentido, no hay una persona en este estudio que haya tenido un ACV hemorrágico. El objetivo de IST era establecer cómo reaccionan los pacientes con la administración de aspirina y/o heparina. El estudio se hizo con una partición totalmente aleatoria para saber quién recibe aspirina, heparina, las dos o ninguna. El IST tiene todas las variables de la información base de los pacientes y el resto de las variables, que son de seguimiento, tienen una completitud de más de 99%.

Este estudio fue totalmente aleatorizado y anonimizado. El 100% de los pacientes tienen la información base y una muy buena porción tiene de datos de seguimiento a los 14 días y 6 meses. El uso de esta base de datos no representa un riesgo para la confidencialidad de los pacientes.

La base de datos contiene 19.435 pacientes de 467 hospitales en 36 países. El dataset incluye las siguientes variables de información base: edad, genero, si la persona tiene fibrilación auricular, si la persona tomo aspirina antes de la llegada al hospital, la presión arterial sistólica, el nivel de conciencia, el déficit neuronal y tiempo trascurrido entre el derrame y la atención médica. También, hay variables del tratamiento, que se le dio al paciente en el hospital, como: número días con aspirina, número de días con heparina, administración de esteroides y administración de glicerina. Finalmente, hay variables indicadoras de una situación médica que sucedió en el hospital como otro ACV o una embolia pulmonar.

5 Metodología

En primer lugar, se seleccionaron las variables que se van a usar en el modelo predictivo, es decir, todas las variables de información base, todas las variables de seguimiento a los 14 días y la variable que se va a predecir, a saber, si la persona está muerta a los 6 meses del accidente. Algunas variables de las seleccionadas se eliminaron manualmente porque no aportaban información a si la persona hubiese muerto a los 6 meses, por ejemplo: fecha del evento o comentarios adicionales. Estas variables se excluyeron del análisis completamente.

En segundo lugar, se hizo la exploración de los datos. En esta parte se hicieron distintas gráficas para ganar mayor entendimiento de los datos. De esta manera, se generó una mayor claridad frente a la relación entre las variables y especialmente entre las variables independientes con la variable dependiente.

En tercer lugar, se hizo el preprocesamiento de los datos. En este proceso se limpia la base de datos, de manera que se pueda utilizar para entrenar los algoritmos. Principalmente, se hizo un tratamiento de missing data, así como de juntar categorías en ciertas variables categóricas. Además, se verificó si había variables con cero o muy poca varianza y también se verificó si había variables explicativas muy correlacionadas entre sí. Finalmente, se hizo una selección de variables con Boruta antes de correr los modelos.

En cuarto lugar, se entrenaron distintos algoritmos de machine learning. Es fundamental decir que este los algoritmos que se van a usar son aprendizaje supervisado de clasificación porque la variable target es binaria. Se va a evaluar el desempeño de los algoritmos y se va a escoger el modelo que tengo el AUC más alto.

En quinto lugar, se va a hacer un análisis de los resultados arrojados por el modelo, en cuanto a la importancia de cada una de las variables explicativas y se va a interpretar dichos resultados.

Finalmente, se van a presentar las conclusiones de este trabajo.

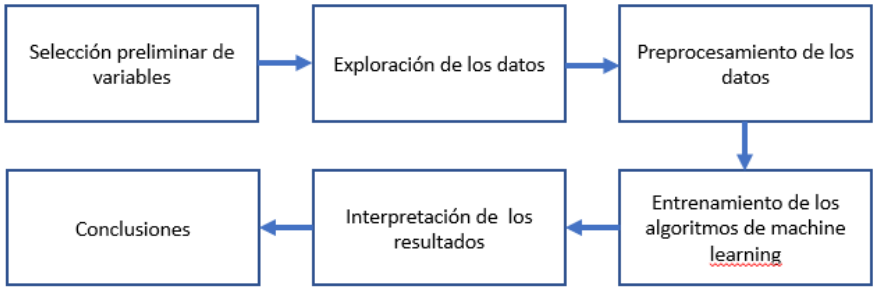

Ilustración 1. Metodología del presente trabajo

6 Selección preliminar de variables

Se quiere explicar si la persona muere o no a los 6 meses, usando solo la información que se tenga hasta los 14 días posteriores después del evento inicial. Se eliminaron las variables que no tenían sentido práctico para la predicción. Entre estas están todas las variables que son comentarios de los doctores, todas las fechas de un evento (por ejemplo, embolia pulmonar), porque existe otra variable que dice si un individuo sufrió de dicho evento. Al final la base de datos tenía 19.435 individuos y 60 variables, incluyendo la variable target.

En la base de datos modificada hay 156 personas de las cuales no se tiene información de si murió o no a los 6 meses. Estas personas se excluyeron del análisis.

7 Análisis exploratorio

En primer lugar, es necesario explorar la composición de la variable que se quiere predecir, es decir, ver si está balanceada o no. En la ilustración 2 se observa que el 77% de los individuos no murieron a los seis meses y el 23% sí. Esto es, 14.910 personas estaban vivas a los seis meses y 4.369 estaban muertas.

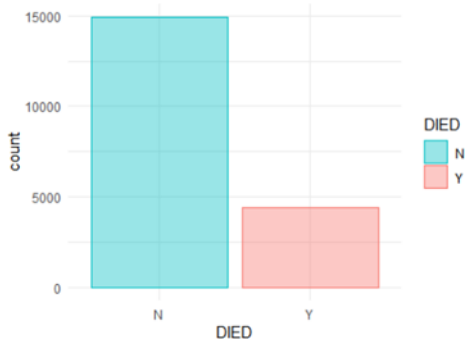

Ilustración 2. Niveles de la variable target

Solo hay 4 variables numéricas y el resto son factores. Las variables numéricas son AGE, ONDRUG, RDELAY y RSBP. La variable ONDRUG es tiempo que el paciente estuvo tomando drogas (aspirina o heparina). La variable RDELAY es el tiempo trascurrido entre el ACV y la atención médica. Finalmente, la variable RSBP es la presión arterial sistólica del paciente.

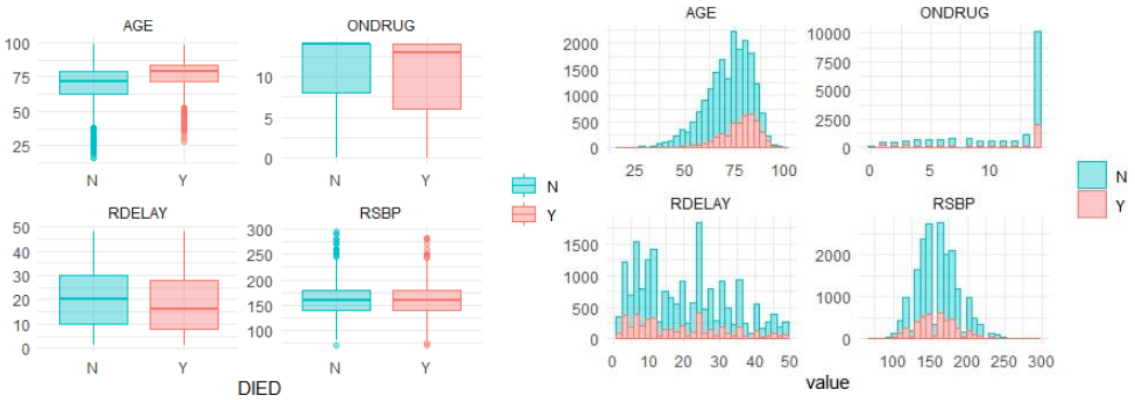

Ilustración 3. Histogramas y boxplots de las variables numéricas

Las personas que tiene derrames generalmente son viejos o no son jóvenes (Virani, y otros, 2020). Esto se puede apreciar en la ilustración 3, pues la mayoría de estos pacientes tiene más de 50 años y es expresamente raro un caso de una persona joven con un ACV. Frente a las variables ONDRUG, se puede apreciar claramente que interesa la mayoría de los sujetos que estuvieron con medicación 14 días. Ahora, la mediana de demora en horas desde el evento hasta que comienza el tratamiento es de 20 horas y un máximo de 49. Desde el punto de vista médico tiene sentido que la mediana de la presión sanguínea sea tan elevada, porque el cuerpo trata de compensar la falta de oxígeno en el cerebro bombeando más sangre (Madell, 2020).

Note que los pacientes que mueren a los 6 meses son de mayor edad que los pacientes que no mueren (ilustración 3). La mediana de la edad de los pacientes que mueren es de 79, mientras la media de los pacientes que viven es de 72. Entonces, la edad parece ser un factor significativo para predecir la muerte después de un derrame. Sin embargo, con la variable RDELAY pasa algo contraintuitivo, porque en esta muestra los pacientes que mueren tienen una demora menor y las personas que sobrevivieron tuvieron una demora mayor (ilustración 3). Es reconocido a nivel médico que es mejor tratar a una persona con ACV rápido, porque se minimizan los posibles daños al cerebro, y, por consiguiente, hay una mejor posibilidad de que la persona sobreviva.

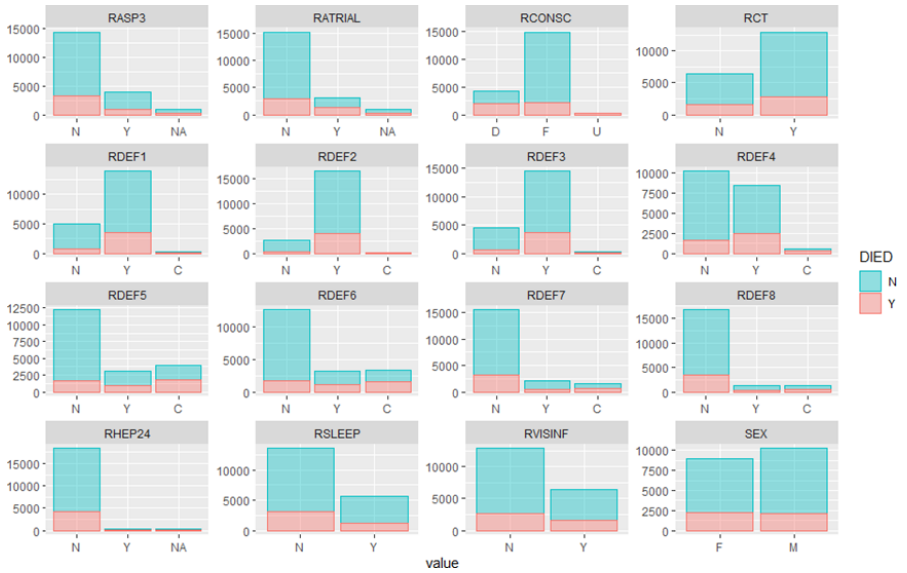

Ilustración 4. Barplots de variables categóricas

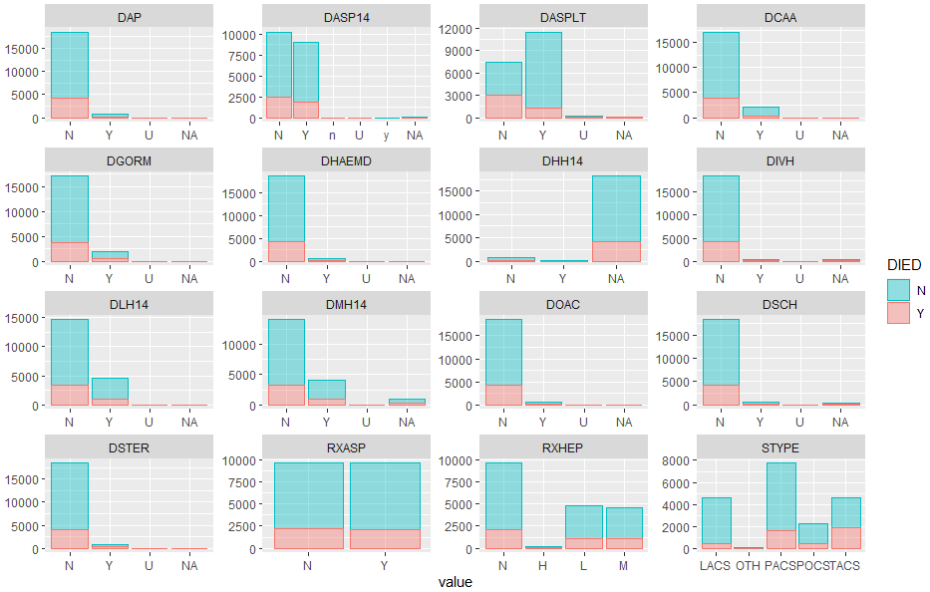

Ilustración 5. Barplots de variables categóricas

De las gráficas de las variables categóricas (ilustración 4, 5, 6, 7) hay dos grandes conclusiones: primero, muchas variables tienen NAs, y deben ser corregidos; segundo, hay variables que tienen niveles que tienen muy pocos individuos, sobre todo en las variables de seguimiento. Hay 30 variables que tienen NAs. Pero es preocupante ver la cantidad de NAs en la variable DPLACE y DHH14. En cuanto a segundo punto, note como en las variables después de RDEF8 hay niveles que se llaman U (Unkown). Como este nivel generalmente muy pocos individuos, vale la pena tratar estos datos.

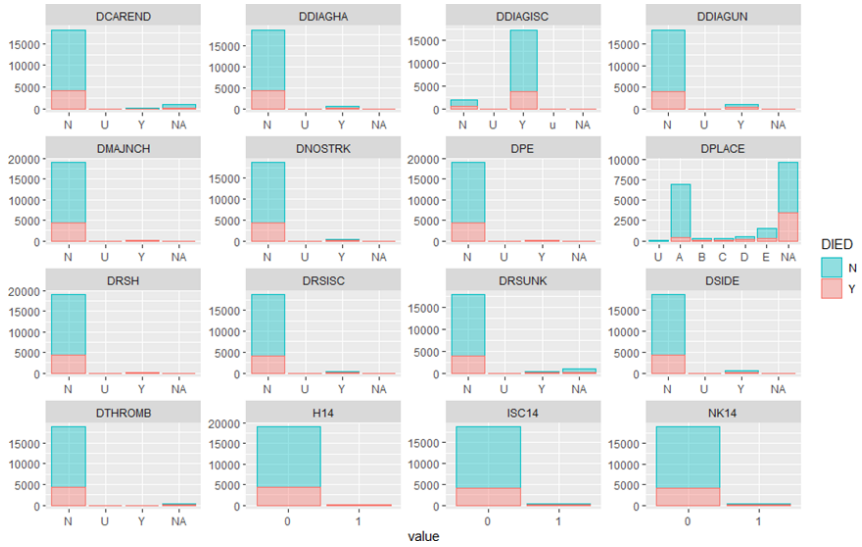

Ilustración 6. Barplots de variables categóricas

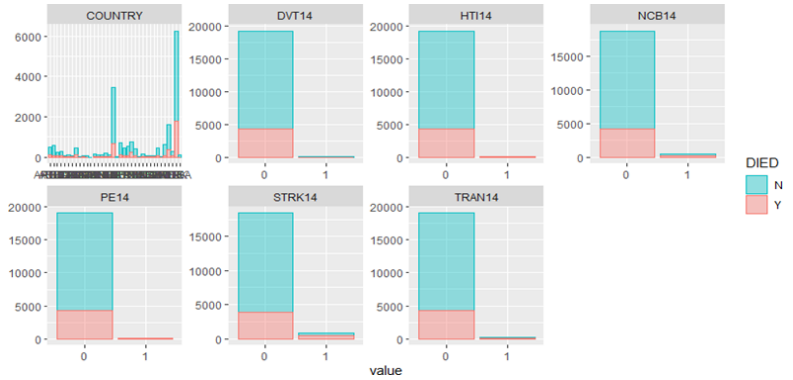

Ilustración 7. Barplots de variables categóricas

Hay una tendencia clara con la variable ONDRUG. En general, los pacientes toman aspirina o heparina más días después del derrame, tienen mayores probabilidades de estar vivos a los 6 meses. Ahora bien, la tasa de mortalidad si el paciente llega totalmente alerta (F) al hospital, es mayor, que si llega inconsciente(U) o mareado(D) (ilustración 8). Si, por el contrario, el paciente llega inconsciente al hospital la probabilidad de estar vivo a los 6 meses es muy baja, pero si le suministran drogas (aspirina o heparina) por varios días, la tasa de mortalidad se reduce, sustancialmente (ilustración 8).

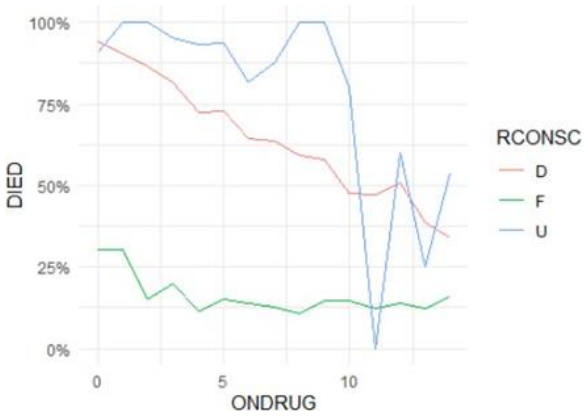

Ilustración 8. Relación entre la mortalidad, el nivel de conciencia y la calidad de días tomando drogas en el hospital

Si el paciente llega al centro médico con una presión sanguínea muy alta, de más de 230 mmHg, entonces la probabilidad de estar vivo a los seis meses es mínima, como se observa en la ilustración 9. Igualmente, es claro en la ilustración 9, que, si el paciente no llega con un campo visual reducido, representado por la variable RDEF5, entonces la probabilidad de estar vivo a los seis meses incrementa. Ahora, es lógico que los pacientes que están inconscientes y que no se les puede practicar la prueba de visión, tengan una tasa de mortalidad más alta, pues su situación médica es más grave.

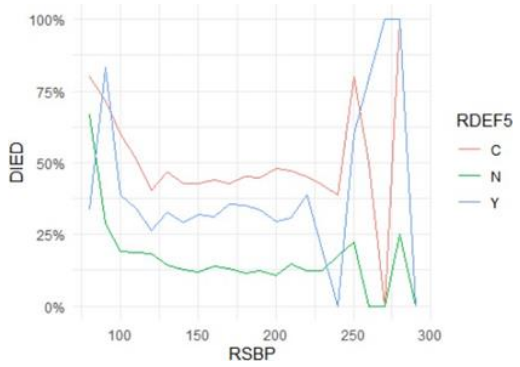

Ilustración 9. Relación ente mortalidad, la presión sanguínea y un campo visual reducido

En la ilustración 10 se ve la relación que existe entre la edad, el tipo de derrame y la tasa de mortalidad. Las personas que sufrieron un derrame tipo TACS, el cual afecta la parte anterior y parte media del cerebro en las arterias cerebrales, tienen menos probabilidades de estar vivos a los seis meses. El tipo de derrame que menos afecta la probabilidad de estar vivo a los 6 meses en adultos mayores, es LACS (ilustración 10) que se da por una obstrucción en venas secundarias en el cerebro. Note que la tasa de mortalidad para el derrame tipo PACS y POCS es similar a lo largo de los años. Además, es evidente, como con la edad la proporción de pacientes muertos a los 6 meses incrementa.

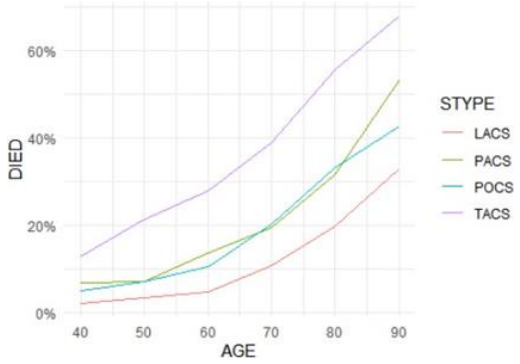

Ilustración 10. Relación ente mortalidad, la edad y el tipo de derrame

La ilustración 11 muestra los porcentajes de muertes en cada categoría, y las variables representan pruebas físicas que se le hacen al paciente cuando llega al hospital. RDEF1 es si la cara del paciente está caída de un lado, RDEF2 es si el paciente tiene un déficit de movimiento en el brazo y RDEF3 es si el paciente tiene un déficit de movimiento en la pierna. Note como en las tres variables la tasa de muerte a los 6 meses es mayor en el nivel C, que representa que los médicos no pueden hacer la prueba. Generalmente, esto se da cuando los pacientes llegan inconscientes al centro médico. Note, además, como la tasa de mortalidad es mayor, si los pacientes tienen la cara caída, con déficit del brazo y déficit de la pierna, frente a los que no tienen esas situaciones.

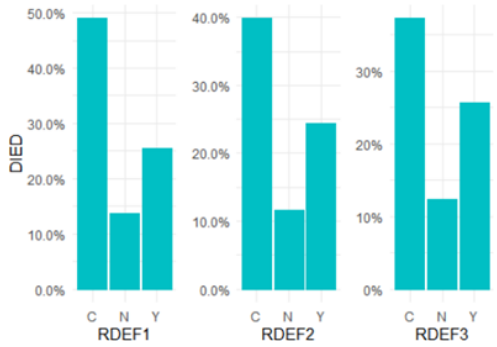

Ilustración 11. Relación entre mortalidad y las evaluaciones físicas realizadas

Según Spaccavento et al, el tiempo para recibir la atención médica después de ACV es factor crucial para reducir los daños en el cerebro, y mejorar las perspectivas de la recuperación (Spaccavento, Marimelli, Nardulli, & Macchitella, 2019). Este hecho, pareciera no aplicar a los datos, porque la mortalidad es más alta cuando los pacientes llegan rápido, como se ve la ilustración 12. Hay una tendencia clara de que la tasa de mortalidad a los 6 meses se reduce cuando la atención medica llega tarde. Adicionalmente, se observa cómo, en general, la tasa de mortalidad en las mujeres es más alta que la tasa de mortalidad de los hombres.

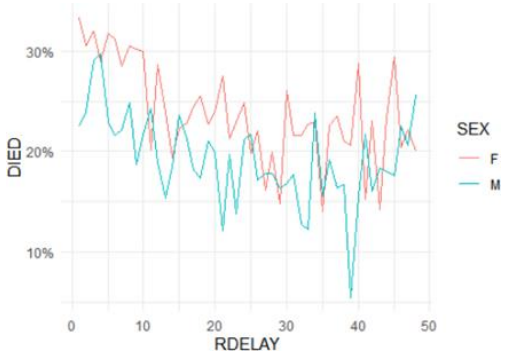

Ilustración 12. Relación entre mortalidad, el sexo y la demora en la atención medica

Las variables H14 y ISC14 son indicativas de si el paciente tuvo un derrame hemorrágico a los 14 días después del primer ACV y si el paciente tuvo un derrame isquémico en los 14 días siguientes, esto se ve en la ilustración 13. Si el paciente tuvo un ACV hemorrágico después, la tasa de mortalidad a los 6 meses se incrementa en un 30%. Ahora bien, si el paciente tuvo un ACV isquémico, entonces las probabilidades de que esté vivo a los seis meses se reducen drásticamente.

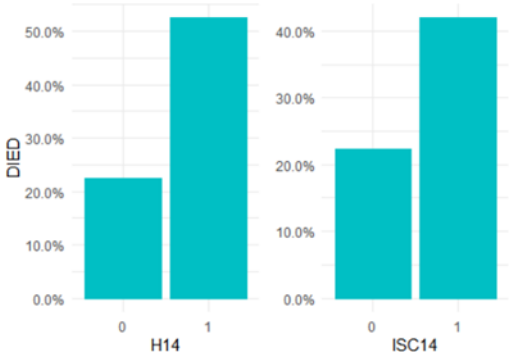

Ilustración 13. Relación entre mortalidad y derrames recurrentes en los primeros 14 días

8 Preprocesamiento de los datos

En primer lugar, hay muchas variables categóricas de este data set que tienen un nivel que se llama

"U", que es "unkown". Es decir, que no se conoce la información de esa variable para dicho individuo. La cantidad de individuos que está en el nivel 'U' es despreciable frente al tamaño del set de datos, como se ve en la ilustración 14. La variable que más observaciones tiene en el nivel 'U' es DCAA con 24 individuos. Por esta razón, se decidió tratarlos como datos faltantes.

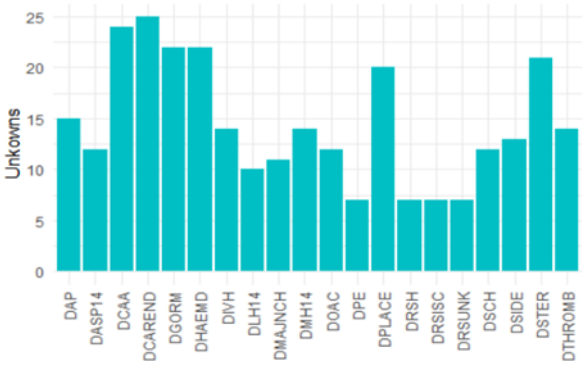

Ilustración 14. Número de unkowns por variable

Teniendo lo anterior es necesario tratar los missing data en toda la base de datos. Hay 17.994 observaciones que no tienen datos faltantes. La ilustración 15 muestra todas las variables que tienen por lo menos un dato faltante. La variable que más datos faltantes tiene en DPLACE, seguida por DCARED, DMH14, DRSUNK, RTRIAL y RASP3. Note que la variable DPLACE tiene casi el 50% de los datos faltantes, le falta 9.729 valores de 19.435. Por eso, se decido excluir dicha variable del análisis y del modelo.

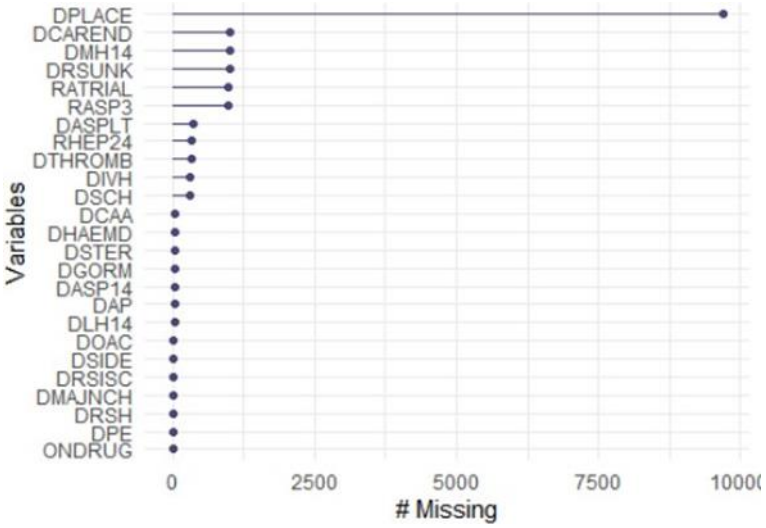

Ilustración 15. Número de datos faltantes por variable

Para la imputación de datos se va a asumir que los datos faltantes son MAR (Missing at Random) que significa el hecho de que sea faltante no fue totalmente aleatorio. Por eso, se puede predecir con otras variables explicativas. Entre las variables con missing data solo hay una que es numérica, ONDRUG, que solamente tiene un valor faltante. Por lo tanto, se decidió imputar este valor con la media de la columna, es decir, la media de esta variable. Para el resto de las variables se usó el paquete MICE para la imputación. Así, se resuelve el problema de missing data en este set de datos.

Adicionalmente, se observaron las correlaciones entre las variables explicativas. No obstante, hay unas variables que son categóricas y otras que son numéricas. Por lo tanto, se hicieron dos matrices de correlaciones: una para las variables numéricas y otra para las variables categóricas. Para las variables categóricas, se calculó la matriz de correlaciones policoricas. En la ilustración 16 se puede

ver que hay variables categóricas que están muy correlacionadas. Se eliminaron las variables que tiene correcciones absolutas mayores a 0.9. Así pues, se eliminaron las siguientes variables: HTI14, DVT14, DCAREND, STRK14, DRSH, TRAN14, NK14, PE14, DHH14, DRSISC y RCT.

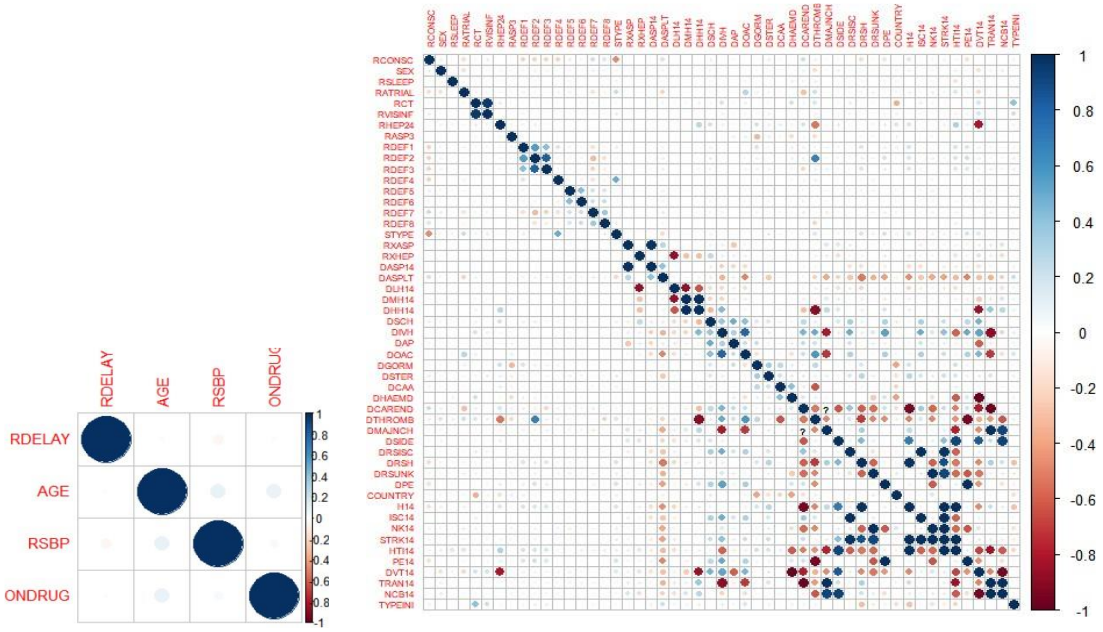

Ilustración 16. Matriz de correlaciones (variables numéricas y variables policoricas)

Las personas de este estudio son de países diferentes. La mayoría de los pacientes se encontraban en el Reino Unido e Italia. En estos países se concentra el 49% de los pacientes. La mayoría de los países restantes era países de Europa, pero igualmente había países del resto del mundo. En ese sentido, se creó la variable región del mundo la cual tiene 4 niveles: UK, Italia, resto de Europa y el resto del mundo.

Por otro lado, se juntaron 4 variables en una sola. Las que se juntaron fueron las variables que indicaban cuál había sido el diagnóstico final del evento inicial (derrame isquémico, derrame hemorrágico, derrame indeterminado o no fue un derrame). Las cuatro variables son binarias y se juntaron en una variable categórica, llamada TYPEINI.

El siguiente paso fue hacer la selección de variables a través del método de Boruta. Este algoritmo está construido alrededor de Random Forest y trata de capturar todas las variables que mejor ayudan a explicar la variable independiente. Después de correr Boruta se llegó a conclusión que se pueden eliminar las siguientes variables: RSLEEP, RHEP24, RASP3, DSCH, DAP, DCAA, DHAEMD y DTHROMB, como se puede ver en la ilustración 17. Hay 34 variables confirmadas. De estas, las de mayor importancia son: DASPLT que dice si el paciente salió del hospital con una prescripción a largo plazo de aspirina, ONDRUG que dice cuántos días el paciente tome medicamentos en el hospital, AGE dice la edad del paciente, RCONSC si el paciente llegó al hospital consciente o no.

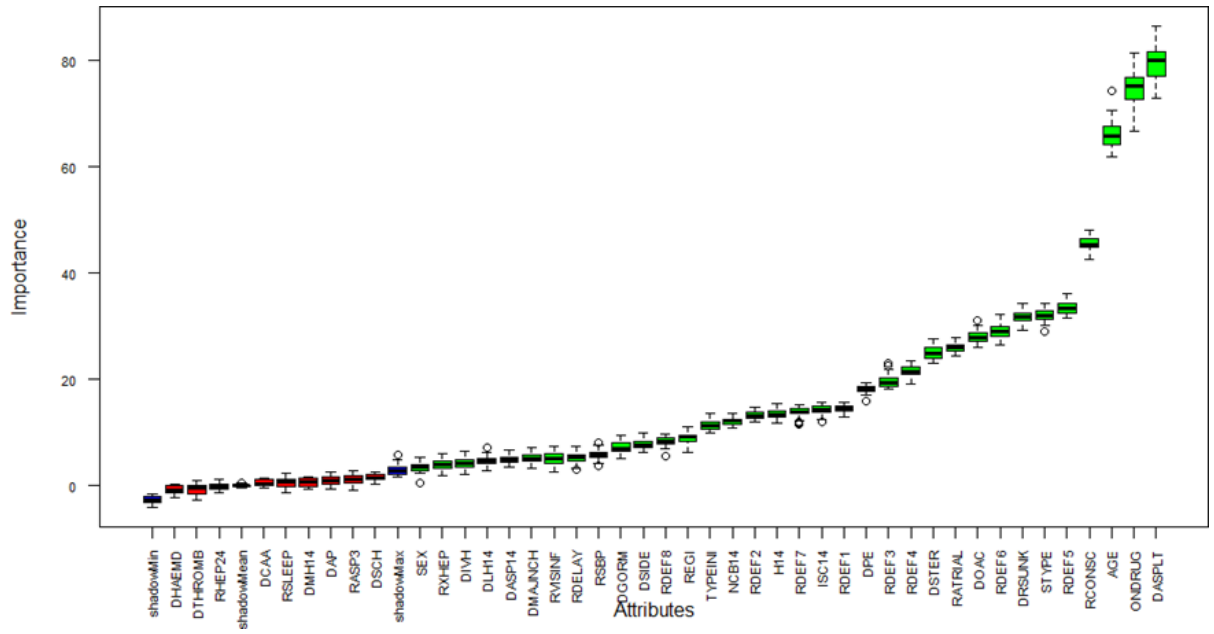

Ilustración 17. Variables aceptadas y rechazadas por Boruta

Finalmente, se hizo la división al set de entrenamiento y al set de prueba. El 75% de los datos van a ser parte del set de entrenamiento y el 25% van a ser parte de set de prueba. Es fundamental que se mantenga a proporción de la variable de respuesta. Por eso, se tomó una muestra totalmente aleatoria del dataset original. Vale la pena decir que a grandes rasgos se mantuvo a proporción de la variable de respuesta en los dos sets de datos.

9 Modelos de Statistial Learning

Por la naturaleza de los datos, se requiere usar algoritmos de clasificación, pues la variable de respuestas es una variable binaria. El objetivo de los algoritmos de clasificación es tener un modelo clasificador que tenga un buen desempeño. Dicho modelo se entrena con los datos de entrenamiento, y se evalúa el desempeño con los datos de prueba.

Los algoritmos que se van a usar son algoritmos basados en árboles de decisión para clasificación. En particular, se implementan Prunded Trees, Random Forest, Bagging, Gradient Boosting, Stochastic Boosting y Extreme Boosting.

Los árboles de clasificación son los árboles que se usan cuando la variable de respuesta es categórica. La idea fundamental de los árboles de clasificación es distribuir las observaciones, según unas condiciones que se evalúan en los nodos del árbol. En cada nodo se hace una bifurcación, los que cumplen con la condición y los que no. La división sucesiva del espacio hace que no haya regiones solapantes. Esta estructura de árbol mantiene hasta llegar a un nodo terminal. Cuando se está prediciendo un valor nuevo se recorre el árbol hasta alcanzar el nodo terminal. Para evaluar el corte de cada nodo se recurre al recursive binary splitting. El cual trata de encontrar el punto de corte u del predictor Xj, que produce la mayor pureza en el nodo. Ahora bien, los árboles de clasificación no suelen ser bueno predictores, porque los árboles suelen tener una varianza alta (Gareth, Witten, Hastie, & Tibshirani, 2013). Para reducir la varianza del modelo y tener una mejor predicción se hace el podado al árbol (pruning). En este caso, la prediccion se hará con el Pruned Tree.

Todos los métodos de aprendizaje estadístico tienen el problema del equilibrio entre sesgo y

varianza. El mejor modelo es aquel que logra dicho equilibrio, pues se reduce al máximo el error total del modelo. El sesgo es la diferencia promedio entre los valores predichos y los valores reales. La varianza es la variabilidad de las predicciones para un dato en específico. Un modelo con mucho sesgo le presta poca atención al set de entrenamiento y sobre simplifica el modelo. Por su parte, un modelo con varianza alta significa que el modelo le está prestando demasiada atención a set de entrenamiento y por lo tanto no puede generalizar bien para los datos de prueba. Las dos en exceso son malas, el mejor modelo es aquel que logre un equilibrio. Una manera de lograr esto es por medio del ensamblaje.

Los métodos de ensamble combinan la información de diferente árboles o modelos con el objetivo de lograr tener un equilibrio entre sesgo y varianza. Así, se logra que el nuevo modelo tenga una mayor capacidad predictiva, pues reduce la varianza o el sesgo. Los dos métodos más conocidos de ensamblaje es el Bagging, que incluye al Random Forest, y el Boosting. Para que los métodos de ensamblaje tengan mejores resultados es necesario que los modelos originales sean diversos entre sí.

El Bagging es una manera de ensamblaje en la cual se ajustan distintos modelos, cada uno con un subconjunto de los datos entrenamiento. Este subconjunto de datos se extrae con boostrap.Estas pseudo-muestras son del mismo tamaño que el set de entrenamiento, pero no son igualespor el hecho de que se extrajeron con un muestreo con repetición. Se ajusta un árbol a cada unade las muestras y luego se agregan en un promedio (Gareth, Witten, Hastie, & Tibshirani, 2013).Así, se logra que el árbol resultante tenga menos varianza y mejore la capacidad de predicción del modelo. En el proceso de calibración, el número de árboles no tiene un papel fundamental,porque por más de que se aumente este parámetro, no aumenta el riesgo de overfitting.

El Random Forest plantea una mejora al algoritmo de Bagging. Esta mejora consiste en decorrelacionar los árboles que se generan el proceso, es decir, hacer que los árboles que se plantean a lo largo de proceso no sean tan similares. Porque si así fuera, entonces el impacto sobre la varianza será mínimo, dado que la correlación entre los árboles es alta. Random Forest soluciona este problema haciendo que se seleccione una muestra m de los p predictores originales en cada división, donde $m < p$ (Gareth, Witten, Hastie, & Tibshirani, 2013). De esta manera, se consigue que la reducción en la varianza sea importante al hacer la agregación. El principal parámetro para calibrar en Random Forest es m, normalmente m es la raíz cuadrada el número de predictores originales. Al igual que en Bagging, calibrar el número de árboles no es crucial para tener un buen modelo, solo es necesario que haya un número lo suficientemente alto. Note que si $p = m$, entonces se está hablando de Bagging.

El Gradient Boosting un método de ensamblaje en el cual se busca un aprendizaje lento. La idea clave detrás del Gradient Boosting es ajustar secuencialmente el modelo inicial. Así, cada iteración usa la información del modelo de la iteración pasada para aprender de los errores, y extraer información que de los residuales (Hastie, Tibshirani, & Friedman, 2009). En ese sentido, se busca que el modelo vaya mejorando con cada iteración adicional. La iteración inicial se hace con un modelo de está sesgado, para que con cada iteración el modelo reduzca su sesgo. Esto hace que mejore la capacidad predictiva. Los principales parámetros que se emplean calibrar el modelo son: el número de árboles, la tasa de aprendizaje y la profundidad de los árboles. El modelo puede tener overfitting si el número de árboles es demasiado alto, para evitar esto se usar la tasa de aprendizaje. Entre menor sea la tasa de aprendizaje, se necesita más arboles tener buenos resultados.

Ahora bien, el Stochastic Gradient Boosting supone una mejora en cuanto a la capacidad predictiva del modelo. En esencia lo que hace el Stocastic Gradient Boosting es lo mismo que el Gradient Boosting solo que en cada iteración se selecciona una porción de los datos de entrenamiento. Esta selección es sin remplazo y de manera aleatoria (Hastie, Tibshirani, & Friedman, 2009). Esta submuestra luego es usada para entrenar el modelo. Los paramentos que se usan para calibrar estos modelos son los mismo que Gradient Bosting, más la proporción de los datos que se toman en cada iteración.

Finalmente, el siguiente modelo es Extreme Boosting, el cual permite añadir más incertidumbre, para que el ensamblaje tenga mejores resultados. Una de las posibles mejoras que se plantea con este modelo es saber cuántas variables se debe tener en cuenta en cada división (Morde, 2019). Lo que se espera es que los árboles de los cuales se aprenden estén decorrelacionados como en Random Forest. De manera que, modelo aprenda información completamente nueva en cada iteración. El parámetro adicional para calibrar este modelo es m.

10 Métricas

El accuracy es muy manipulable cuando las clases son desbalanceadas. Dado que el modelo puede predecir que todas las observaciones pertenecen a la clase mayoritaria, el accuracy podría seguir alto a pesar de que no se está prediciendo correctamente para la clase minoritaria. Por eso, es necesario usar una medida que sea más robusta. El AUC mide el área bajo la curva ROC. Un AUC de 1 significa que el modelo predice perfectamente la clase, por el contrario, un AUC de 0.5 dice que el modelo no tiene capacidad predictiva.

11 Calibración

Todos los modelos se calibraron con una grilla. Los modelos con pocos parámetros por calibrar, como Random Forest, se calibraron con una grilla normal. Así, se prueba todas las posibles combinaciones de los parámetros. Ahora bien, los modelos que tienen muchos parámetros por calibrar se calibraron con una grilla que llena el espacio (space-filling parameter grid). Esta trata un modelo en todas las áreas de espacio generado por los parámetros. La calibración se hizo mediante 10 fold cross validation.

Un árbol de decisión podado (pruned tree) es un modelo de árbol de decisión que esta calibrado. Normalmente, los árboles de decisión se calibran con dos parámetros: la profundidad y el costo de complejidad. No obstante, después de una grilla inicial, se observó que los modelos con un costo de complejidad muy cercanos a cero tenían mejor desempeño. Por eso, se decidió fijar ese valor a cero y calibrar el modelo solo con la profundidad. Note como en la ilustración 18, la profundidad que maximiza el AUC es 7.

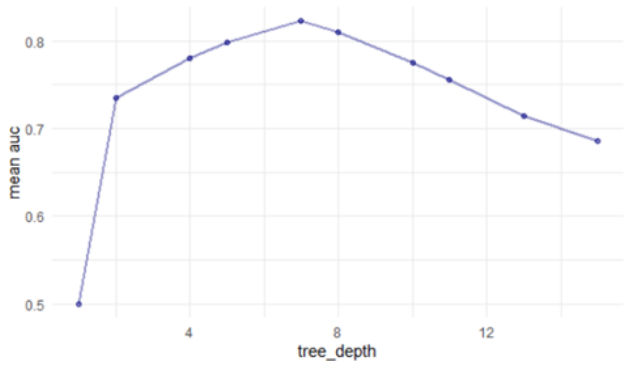

Ilustración 18. Calibración de pruned trees

El número de árboles en el algoritmo de Bagging no es un parámetro crucial para calibrar este modelo, pero es necesario que sea lo suficientemente alto. Se hizo la calibración para tener la seguridad de que el número de árboles escogidos era el número óptimo para tener el modelo que maximice el AUC. Es claro en la ilustración 19 que este parámetro no afecta mucho el AUC después cierto número de a de árboles. Por más de que se aumente el número de árboles excesivamente, los resultados no van a variar mucho. Con 1016 árboles se maximizó el AUC.

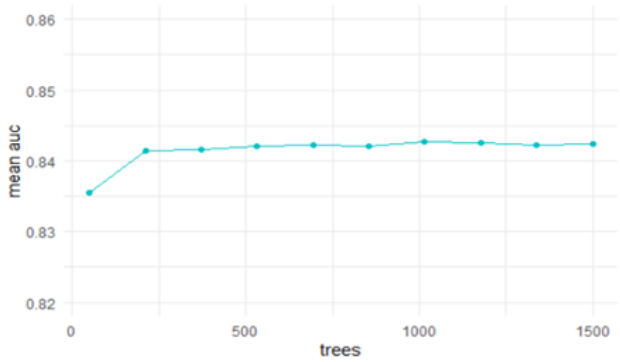

Ilustración 19. Calibración de Bagging

En Random Forest hay 2 parámetros para calibrar que son el número de árboles y el número de predictores que se pueden tratar en cada split de los árboles (m). A pesar de que solo se necesita que el número de árboles sea lo suficientemente grande, se tuvo en cuenta en la grilla para calibrar el modelo. En primera medida, en la grilla inicial se observa como a medida que el número de predictores candidatos en cada split se hace mayor el AUC en los datos de validación se hace menor. En ese sentido, se necesita que este número sea menor a 15 mientras que se necesita que el número de árboles sea mayor a 1.000. Se hizo una grilla más específica con la intención de tener mejores resultados. Es evidente en la ilustración 20, con lo resultados de la segunda grilla que el modelo final debe tener 6 candidatos para cada split y 1600 árboles. En la ilustración 20 se puede ver la representación gráfica de las dos grillas.

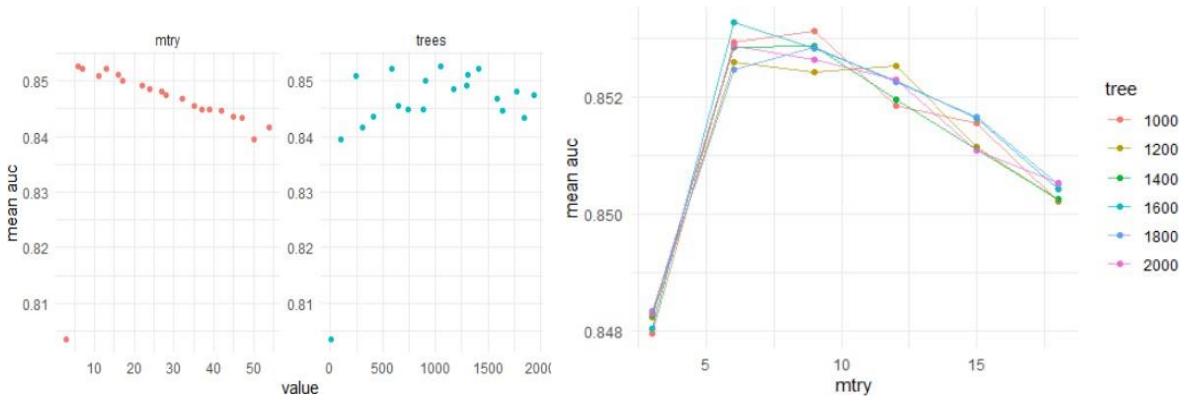

Ilustración 20. Calibración de Random Forest. Calibración preliminar (izq.). Calibración final (der.)

Con el algoritmo de Gradient Boosting hay 3 parámetros fundamentales que deben ser calibrados. El primero es número de árboles, el segundo es la profundidad de los árboles y por último la tasa de aprendizaje. Con la grilla inicial se ve una tendencia clara que la tasa de aprendizaje debería estar en el rango de (0, 0.025], pues se aprecia mejores resultados en esta región. En cuanto a profundidad de árbol que los modelos con una profundidad entre [4, 9] tiene mejor desempeño. Ahora bien, se observa en la ilustración 21 de la grilla inicial que no hay una tendencia clara con el

número de árboles, sin embargo, para hacer la búsqueda más focalizada se intentó un modelo con más de 500 árboles. Con la segunda grilla en la ilustración 21, es claro que los árboles no deben ser tan profundos, pues los mejores resultados fueron con una profundidad de 4 o 5. El mejor modelo fue con 500 árboles, con 4 nivales de profundidad y una tasa de aprendizaje de 0.0265.

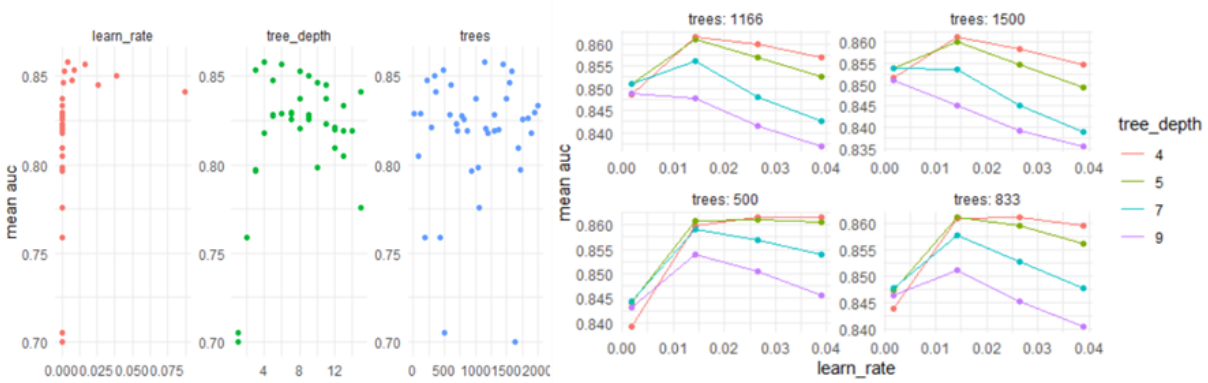

Ilustración 21. Calibración de Gradient Boosting. Calibración preliminar (izq.). Calibración final (der.)

A diferencia de los algoritmos anteriores se calibraron con una grilla normal, el Stochastic Boosting se calibró mediante una space-filling parameter grid. Esto para hacer la búsqueda el mejor modelo y el más eficiente. Los parámetros de este algoritmo son los mismos que los de Grandient Boosting, más sample size que es el porcentaje de los datos que se usa en cada iteración. Una particularidad de Stocastic Boosting y XGBoost es que se puede llegar al mismo desempeño con diferentes combinaciones de los parámetros. Por lo tanto, la combinación escogida es aquella que tuvo el mejor desempeño de los modelos ensayados. Este modelo tenía 651 árboles, 3 niveles de profundidad, una tasa de aprendizaje de 0.0242 y sample size de 0.260.

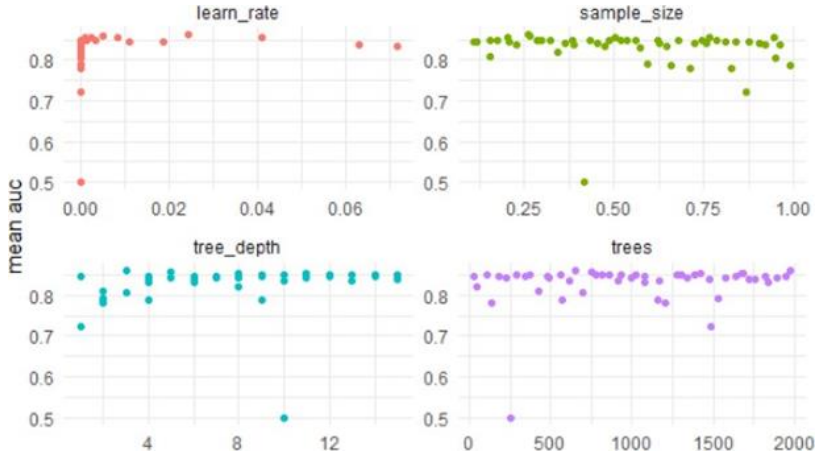

Ilustración 22. Calibración de Stochastic Gradient Boosting

El modelo de XGBoost tiene 5 parámetros por calibrar. De estos, el único nuevo frente a Stochastic Boosting es mtry que es el mismo que se calibró en Random Forest. Este parámetro dice cuál es el número de predictores candidatos en cada split. Note cómo los modelos que tienen una tasa de aprendizaje bajita y tiene una profundidad mayor a 4 y un mtry mayor a 20 tienen un AUC más alto. El modelo calibrado tiene un mtry de 28, 966 árboles, una profundidad de 5, una tasa de aprendizaje de 0.0101 y un sample size de 0.49.

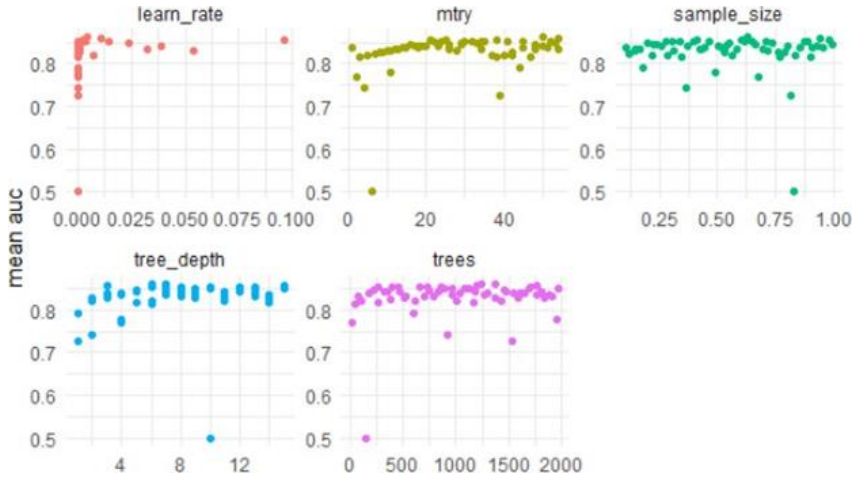

Ilustración 23. Calibración de XGBoost

12 Resultados

Luego de haber calibrado todos modelos se hace la comparación para saber cuál se desempeñó mejor. La medida de comparación se hace con el AUC de todos los algoritmos sobre el set de prueba. Gráficamente, es claro como en la curva ROC (ilustración 24) el modelo de Pruned Trees y Bagging fueron los que peor se desempeñaron de los 6 métodos, porque sus curvas no están tan elevadas como el resto. Con la curva ROC no es fácil identificar el mejor algoritmo del resto, pues su desempeño parece ser muy similar.

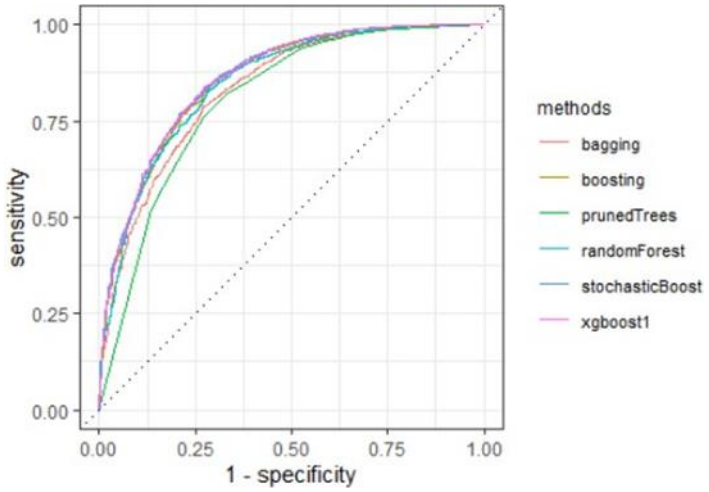

Ilustración 24. Resultados con la curva ROC

Ahora, es claro que el AUC de los modelos es similar, en el sentido de que no hay unas diferencias de más de 6 puntos porcentuales entre todos los métodos. Los mejores 2 mejores métodos, lo que tienen mejor AUC son el Schocastic Boosting y el XGBoost1. La diferencia entre estos dos es menor un punto porcentual. Pero, el algoritmo que tuvo un mejor desempeño fue Schocastic Boosting, con un AUC de 0.863.

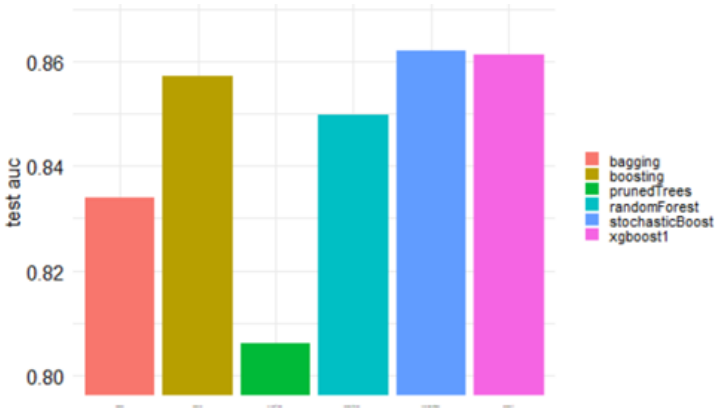

Ilustración 25. Resultados con AUC

Los mejores modeles fueron Stochastic Boosting y XGBoost y el peor fue Pruned Trees. Stochastic Boosting y XGBoost tienen una gran capacidad predictiva, pero no son fácilmente interpretables. Por su parte, Pruned Trees tiene una peor capacidad de predicción, pero es fácilmente interpretable. Es innegable que los modelos complejos son inferiores en términos de interpretabilidad, pero son muy superiores en términos de capacidad predictiva. Por lo tanto, es necesario hacer un análisis más a fondo sobre los resultados y que información se puede extraer de estos.

13 Interpretación de resultados

La etiología es una gran parte de los estudios de la medicina moderna. Investigar la causa de una condición médica o los factores de riesgo pueden ayudar a que las personas reconozcan más fácilmente la enfermedad y también para prevenirla. La importancia de las variables en un modelo como el de Stocastic Boosting, permite identificar los aspectos que permiten clasificar las observaciones en el modelo. Entonces, se usa la importancia global para tratar de analizar la importancia de las diferentes variables en este modelo en específico.

En primer lugar, se usa el método SHAP para analizar la importancia de cada una de las variables, al calcular el SHAP value de todas las variables y después organizarlas de mayor a menor. La gráfica contiene las 10 variables más importantes. Con la ilustración 26 se puede identificar que las variables más importantes son la edad de paciente, si le recetaron aspirina a largo plazo, la cantidad de días que le dieron aspirina o heparina al paciente mientras estaba en el hospital y si el paciente había llegado consciente al centro asistencial.

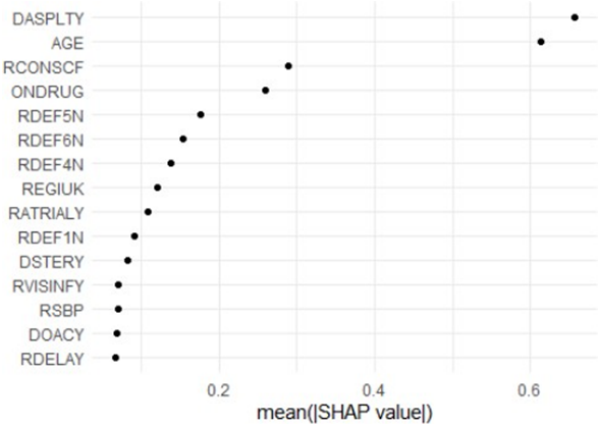

Ilustración 26. Importancia de las 10 variables más importantes

Sin embargo, la gráfica anterior no brinda información sobre qué tipos de valores de cada variable están más correlacionados con el resultado de la variable target. Dicho de otra manera, no da información sobre si, por ejemplo, una presión sanguínea alta está más correlacionada con desenlace positivo a los 6 meses o no. Dicha información se muestra en la Ilustración 27.

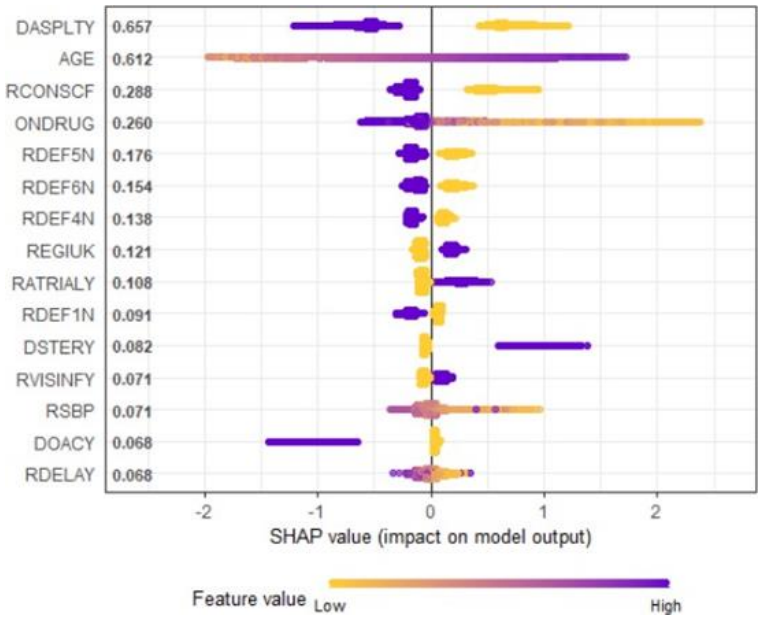

Ilustración 27. SHAP value de las variables más importantes

Note que, si el paciente le suministraron una prescripción de aspirina de largo plazo, es más probable que este vivo a los 6 meses. De la misma manera, es claro que entre mayor sea el paciente, menores chances tienen de estar vivo a los seis meses. Si el paciente llega a hospital totalmente consciente, hay mejores probabilidades de que este vivo a los 6 meses; además, entre mayor sea la cantidad de días de aspirina o heparina en la clínica, hay más probabilidades de estar vivo 6 meses después del accidente cerebrovascular original. Ahora, hay 3 factores de la evaluación inicial en la clínica que son muy importantes para saber si la prognosis de la persona va a ser favorable o no a los 6 meses: si la personas tiene pérdida de visión campo visual de uno de los ojos o de los dos (RDEF5), si la persona no tiene desorden visoespacial (RDEF6) y si no tiene disfasia, la capacidad de comunicarse verbalmente (RDEF4). Además, si el paciente tiene fibrilación auricular es más propenso a un resultado no favorables a los 6 meses. Con las variables relacionadas con el tratamiento, es claro como si el paciente recibió esteroides es más probable que muera a los 6 meses. Finalmente, la administraron otros anticoagulantes está más correlacionada con una prognosis fatídica. Pero es claro que, el estado con el cual llega el paciente al centro médico tiene un gran efecto sobre la prognosis del paciente a los 6 meses, al igual que el tratamiento con aspirina.

En el presente, la mayoría de los modelos de statistical learning que predicen enfermedades tiene un accuracy alto, pero tiene el problema de que no es interpretable sobre una observación o un solo paciente. De esta manera, se vuelve difícil hacer recomendaciones individuales para un caso en específico. Aunque el Stochastic Boosting y el XGBoost de este trabajo tienen una gran capacidad predicativa, no son tan buenos en términos de interpretabilidad.

Para interpretar la predicción de un solo paciente se usó LIME. Este paquete busca responder la pregunta de por qué cada predicción es como es. Así, se logra explicar la predicción de cada individuo. Los médicos o las personas interesadas pueden entender el modelo, para poder escoger

un plan de tratamiento que mejore la prognosis del paciente.

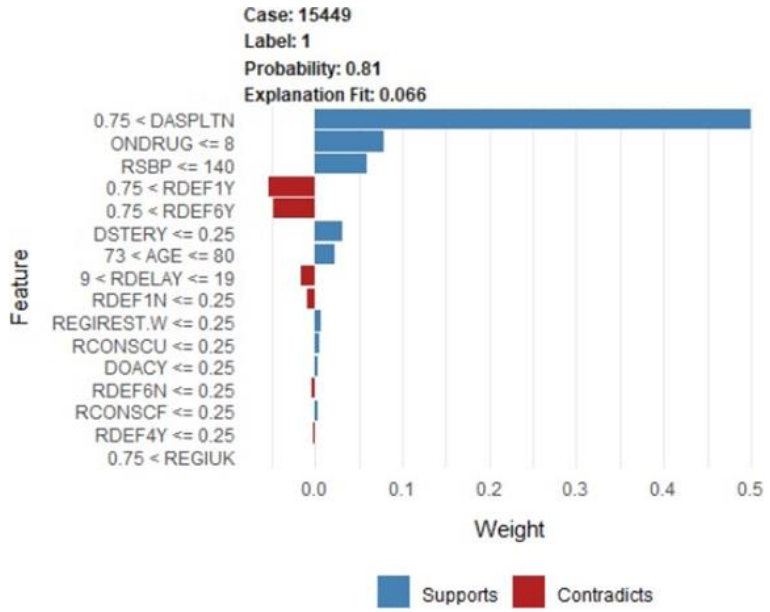

Ilustración 28. Interpretación individual para el paciente 15449

Recuerde que la variable que se está prediciendo es si la persona va a estar viva a los 6 meses o no. En la ilustración 28 se ve la explicación local del modelo para el paciente 15449. Note que el modelo predice que este paciente va a morir a los seis meses. Este resultado se da por tres predictores principales: DASPLT, ONDRUG y RSBP. A este paciente no se ha recetado o no se recetó una terapia de aspirina de largo plazo. Según el modelo, a las personas que se le receta aspirina a largo plazo tiene más chance de sobrevivir. Del mismo modo, al paciente 15449 no se le ha tratado al paciente con la droga más de 8 días, por eso que esta variable indica que este paciente puede morir. Finalmente, el tercer factor que contribuye a que esta sea la predicción del modelo es que este paciente tiene una presión sistólica de menos de 140 mmHg, pues es de 116 mmHg.

14 Discusión

En primer lugar, se van a discutir los resultados generales del modelo, es decir, la importancia de las variables. Luego, se va a discutir la importancia de los resultados individuales.

El accidente cerebrovascular isquémico ocurre cuando un coagulo tapa parcial o totalmente una arteria en el cerebro, interrumpiendo el flujo de oxígeno en el cerebro. La aspirina previene que las plaquetas en la sangre, producidas por el colesterol y otras sustancias grasas, se conviertan en coágulos (American Heart Association, 2020). Por eso, algunos médicos recetan aspirina a los pacientes de ACV's isquémicos, para prevenir el riesgo un segundo ACV isquémico (Hankey, 2016). En este sentido, es lógico que los pacientes que tengan una prescripción de aspirina de largo plazo tengan una mejor prognosis a los 6 meses, así como el tratamiento con aspirina dentro del hospital.

La edad es un factor clave para la mortalidad de pacientes de ACVs por múltiples razones. Muchos estudios han documentado este fenómeno, como en el paper de Yousufuddin & Young llamado *Aging and ischemic stroke* (Yousufuddin & Young, 2019). En general, las personas de más edad son más frágiles y por eso tienden a tener más complicaciones con un ACV y más complicaciones en la recuperación.

El nivel de conciencia del paciente, con el cual llega al hospital es un valor clave sobre las probabilidades de seguir viviendo a los 6 meses. Otros estudios también han identificado a esta

variable como determinante sobre la prognosis de paciente, tal y como se ve el estudio *Early Predictors of Death and Disability After Acute Cerebral Ischemic Event* (Hénon, Godefroy, Leys, & Mounier-Vehier, 1995). En este trabajo se llega a la conclusión de que el estado de conciencia del paciente es un determinante de la prognosis de la persona a los 3 meses. Este resultado es consistente con los resultados del modelo de este trabajo, el cual dice que si la persona llega totalmente consciente hay mayores probabilidades de una prognosis favorable.

Cuando hay una lesión en el cerebro se manifiesta de diferentes formas, dependiente de la gravedad de la lesión y la ubicación de la misma. Los síntomas dan información del accidente cerebrovascular. Los más reconocidos por los médicos son dificultad para hablar o entender lo que otros están diciendo, parálisis en la cara, brazo o pierna, problemas para ver por uno o ambos ojos y alteraciones visoespaciales (Mayo Clinic, 2021). De estos, 2 son muy significativos para predicar la prognosis del paciente en este trabajo, problemas para ver por uno o ambos ojos y alteraciones visoespaciales. Note como la afectación de la parte del cerebro que controla la vista es un factor determinante para saber si la prognosis del paciente va a ser bueno o mala. Porque es importante saber si el paciente tiene problemas visoespaciales y problemas de un campo visual reducido. En el estudio de *Visual field defect after ischemic stroke-impact on mortality* (Sand, Naess, Thomassen, & Hoff, 2018) se encontró que las pernas con problemas de visión de tienen a tener una mayor tasa de mortalidad después de una ACV isquémico.

La fibrilación auricular es una condición médica que se caracteriza por unos ritmos cardiacos irregulares. Esta condición está asociada con la ocurrencia de eventos cerebrovasculares isquémicos (Wolf, Abbott, & Kannel, 1991), pues es más probable que se generen coágulos de sangre en el corazón. Por lo tanto, si el paciente tiene esta condición, entonces el riesgo de tiene un segundo ACV aumenta. En ese sentido, es lógico que el resultado de la importancia global de si el paciente tiene fibrilación auricular sea alta, y los que la tienen, tienen mayores posibilidades de tener una prognosis desfavorable.

El paquete Lime se ha usado para poder interpretar los resultados de modelos sumamente complejos como XGBoost. Vale la pena aclarar que Lime se usa para instancia individuales, es decir, para cada paciente. Lo valioso de poder interpretar los casos individuales es que se logra tener una herramienta para la toma de decisiones de los médicos. Si se usara el modelo solo como un agente de decisión, las consecuencias podrían ser peligrosas pues no se está tenido en cuenta la opinión de un médico experto. Por eso, es crucial que el modelo sea interpretable para los profesionales de la salud.

Como se había dicho antes, la ilustración 28 tiene tres predictores que hacen que el modelo clasifique a este paciente con una prognosis desfavorable. Hay múltiples trabajos en los cuales se llega a la conclusión que el tratamiento con aspirina previene un ACV recurrente, como *Risk of recurrent stroke and antiplatelet choice in breakthrough stroke while on aspirin* (Kim, Joon Kim, Park, & Jo, 2020). De esta manera los médicos podrían recetar una terapia con aspirina para que mejore la prognosis del paciente 15449. De la misma manera, paciente 15449 tiene una presión sistólica de 116, lo cual no es normal después de un ACV (Silvestrini & Provinciali, 2013), porque la reacción natural del cuerpo es bombear más sangre porque no está llegando a una parte del cerebro. Con esta información los médicos pueden hacer un plan de acción para que la prognosis sea más favorable para el paciente.

Con la interpretación individual de las predicciones se puede revolucionar la medicina, en el sentido de que se puede sacar un plan de acción, o, un plan de tratamiento para cada persona. Así,

a partir de un modelo estadístico de la prognosis se puede hacer un plan diferente de cómo se puede tratar una persona para que su desenlace sea lo mejor posible a los 6 meses.

Este trabajo tiene varias limitaciones. En primer lugar, los datos están muy completos, pero fueron tomados hace bastante tiempo. Con el avance de la tecnóloga en la medicina se tiene más datos y de mejor calidad para tener una clasificación más precisa. Además, durante este tiempo se han creado nuevos métodos de tratamiento para accidentes cerebrovasculares, que se pueden incluir en los predictores del modelo. En segundo lugar, hacen falta más variables de la historia medica de factores de riesgo para un ACV como la diabetes, el peso o si fuma o no. También, debido a que los datos provienen mayoritariamente de países de Europa, puede ser posible que los datos o los resultados no sean generalizables a otras regiones del mundo. Finalmente, la variable que se está intentando predecir esta desbalanceada esto afecta el desempeño del modelo.

15 Conclusiones

Este trabajo busca predecir la prognosis a los seis meses de los pacientes que tienen un accidente cerebrovascular isquémico. Para esto, se usaron los datos del *Internatinal Stroke Trial* fase de dos con su debido preprocesamiento. Luego, se construyeron distintos modelos de machine learning y se escogió el que tenía mayor capacidad predictiva. El mejor modelo fue Stochastic Boosting con un AUC de 0.863. Con base en los resultados de dicho modelo se hizo una interpretación global e individual de las predicciones del modelo. Hay 3 grandes conclusiones se este trabajo. Primero, el tratamiento con aspirina es fundamental para que el desenlace de un paciente a los 6 meses sea favorable. Segundo, el estado neurológico del paciente con el cual llega a hospital es importante para predecir la prognosis a los seis meses. Finalmente, se puede interpretar el modelo para cada predicción para montar un plan de acción para maximizar la favorabilidad de las prognosis.

16 Bibliografía

American Heart Association. (22 de Abril de 2020). *American Stroke Association*. Obtenido de Aspirin and Stroke: https://www.stroke.org/en/life-after-stroke/preventing-another-stroke/aspirin-and-stroke

El Hospital. (3 de Noviembre de 2018). Obtenido de BOGOTÁ ES LA CIUDAD COLOMBIANA CON MAYOR PREVALENCIA DE ATAQUE CEREBROVASCULAR: https://www.elhospital.com/temas/Bogota-es-la-ciudad-colombiana-con-mayor-prevalencia-de-ataque-cerebrovascular+127097#:~:text=Seg%C3%BAn%20la%20Asociaci%C3%B3n%20Colombiana%20de,emergencia%20para%20su%20pronta%20atenci%C3%B3n.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-Based Methods. En *An Introduction to Statistical Learning with Applications in R* (págs. Pg 303-335). New York: Springer.

Gupta, S., Tran, T., Luo, W., Phung, D., Broad, A., Campbell, D., . . . Khasraw, M. (2014.). Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Journals*.

Hankey, G. (2016). The benefits of aspirin in early secondary stroke prevention. *The Lancet*, 314-314.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and Additive Trees. En *The Elements of Statistical Learning* (págs. Pg 337-387). New york: Springer.

Hénon, Godefroy, Leys, & Mounier-Vehier. (1995). Early Predictors of Death and Disability After Acute Cerebral Ischemic Event. *AHJ Journals*.

Heo, J., Yoon, J., Park, H., Dae Kim, Y., Suk Nam, H., & Hoe Heo, J. (2019). Machine Learning–Based Model for Prediction of Outcomes in Acute Stroke. *Stroke AHA Journals*.

Kim, J.-T., Joon Kim, B., Park, J.-M., & Jo, S. (2020). Risk of recurrent stroke and antiplatelet choice in breakthrough stroke while on aspirin. *Nature*.

Madell, R. (06 de Febrero de 2020). *Healthline.* Obtenido de Blood Pressure Readings: https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained

Mayo Clinic. (20 de Noviembre de 2020). *Mayo Clinc*. Obtenido de Coma: https://www.mayoclinic.org/es-es/diseases-conditions/coma/symptoms-causes/syc-20371099

Mayo Clinic. (09 de Enero de 2021). *Mayo Clinic*. Obtenido de Accidentes cerebrovasculares: https://www.mayoclinic.org/es-es/diseases-conditions/stroke/symptoms-causes/syc-20350113

Morde, V. (07 de Abril de 2019). *Towards Data Science*. Obtenido de XGBoost Algorithm: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Parmar, P. (10 de Enero de 2018). *The Pharmaceutical Journal.* Obtenido de Stroke: classification and diagnosis: https://pharmaceutical-journal.com/article/ld/stroke-classification-and-diagnosis

Portafolio. (29 de Octubre de 2018). *Portafolio*. Obtenido de Derrame cerebral, la segunda causa de muerte en el país.

Potter, L. (22 de Marzo de 2021). *Geeky Medics.* Obtenido de Stroke Classification: https://geekymedics.com/stroke-classification/

Sand, K., Naess, H., Thomassen, L., & Hoff, J. (2018). Visual field defect after ischemic stroke-impact on mortality. *Acta Neurologica Scandinavica*.

Sandercock, P., Niewada, M., & Czlonkowska, A. (02 de Noviembre de 2011). *International Stroke Trial database (version 2),.* Obtenido de University of Edinburgh. Department of Clinical Neurosciences.: https://datashare.ed.ac.uk/handle/10283/124

Scrutinio, D., Ricciardi, C., Donis, L., Losavio, E., Battista, P., & Guid, P. (2020). Machine learning to predict mortality after rehabilitation among patients with severe stroke. *Nature*.

Silvestrini, M., & Provinciali, L. (2013). Elevated Blood Pressure in the Acute Phase of Stroke and the Role of Angiotensin Receptor Blockers. *International Journal of Hypertension*.

Spaccavento, S., Marimelli, C., Nardulli, R., & Macchitella, L. (2019). Attention Deficits in Stroke Patients: The Role of Lesion Characteristics, Time from Stroke, and Concomitant Neuropsychological Deficits. *Behavioural Neurology*.

Virani, S., Alonso, A., Bejamin, E., Bittencourt, M., Callaway, C., & Carlson, A. (2020). Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. *AHA/ASA Journals*, Chap 14.

Wolf, P., Abbott, R., & Kannel, W. (1991). Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke Journal of American Heart Association*.

World Stroke Organization. (15 de Enero de 2019). *World Stroke Organization*. Obtenido de Global Stroke Fact Sheet: https://www.world-stroke.org/assets/downloads/WSO_Global_Stroke_Fact_Sheet.pdf

Xie, Y., Jiang, B., Gong, E., & Li, Y. (2018). Use of Gradient Boosting Machine Learning to Predict Patient Outcome in Acute Ischemic Stroke on the Basis of Imaging, Demographic, and Clinical Information. *American Journal of Roentgenology*.

Yousufuddin, M., & Young, N. (2019). Aging and ischemic stroke. *Impact Journals Aging*.

Yu, J., Park, S., Lee, H., Pyo, C.-S., & Lee, Y. (2020). An Elderly Health Monitoring System Using Machine Learning and In-Depth Analysis Techniques on the NIH Stroke Scale. *Mathematics*.