Árboles de Modelos GAMLSS para la predicción de tiempo de estancia hospitalaria

Juan Camilo España Lopera

Universidad de Antioquia
Facultad de Ingeniería
Departamento de Ingeniería Industrial
Medellín, Colombia
2018

Árboles de Modelos GAMLSS para la predicción de tiempo de estancia hospitalaria

Juan Camilo España Lopera

Trabajo de investigación presentado como requisito parcial para optar al título de:
Magíster en Ingeniería

Directora:
Ph. D. Olga Cecilia Úsuga Manco
Codirector:
Ph. D. Juan Sebastián Jaén Posada

Línea de Investigación:
Investigación de Operaciones y Estadística
Grupo de Investigación:
Innovación y Gestión de Cadenas de Abastecimiento - INCAS

Universidad de Antioquia
Facultad de Ingeniería
Medellín, Colombia
2018

Agradecimientos

Mi más sincero agradecimiento a los profesores Olga Cecilia Úsuga Manco y Juan Sebastián Jaén Posada, directora y codirector, por la paciencia en todo el desarrollo de la investigación, por apoyar con todo su conocimiento, por aportar sus perspectivas diferentes en el trabajo, por las exigencias que permitieron mejorar la investigación, por ayudarme a consolidar el amor por el conocimiento y la investigación y principalmente por su calidad humana y profesional.

A los profesores Juan Guillermo Villegas y Pablo Andrés Maya por compartir más que su conocimiento el amor por lo que hacen, amor que aumentó mi curiosidad y mis ganas de aprender cada día más. Gracias por disfrutar lo que hacen y transmitir esa emoción en sus estudiantes, mi más profunda admiración y agradecimiento. También agradecer porque en su rol de coordinadores del grupo de investigación encontré siempre una voz de aliento, un apoyo permanente y una cercanía que me ayudaron a sentirme siempre respaldado.

Al grupo de investigación INCAS y a sus integrantes por los consejos y por el esfuerzo que siempre ponen para hacer las cosas con excelencia, por la calidad humana que tienen los miembros y por el ambiente siempre respetuoso, cercano y profesional.

A mis amigos y familia por el apoyo, principalmente a mi madre, por la paciencia que tuvieron en todos los momentos en que no estuve disponible para compartir con ellos y apoyarlos en los momentos difíciles.

A la Universidad de Antioquia por cambiar mi vida y abrirme las puertas de un futuro mejor, por mostrarme un universo de conocimientos y de culturas que me hicieron cambiar mi forma de ver el mundo. Mi agradecimiento infinito y eterno para mi alma mater que me llena de orgullo y amor.

Resumen

El tiempo de estancia hospitalario conocido en la literatura como length of stay (LOS) es una variable fundamental en la administración hospitalaria debido a su impacto en la eficiencia de los recursos. Es por esto que su modelado y predicción han sido temas de interés de investigadores durante las últimas décadas. Adicionalmente, las características de su distribución, sesgada y de valores positivos, se reflejan en diferentes fenómenos, razón por la cual los trabajos desarrollados sobre la predicción de esta variable pueden ser replicable en diversos problemas. Aunque se han utilizado diferentes técnicas para la predicción de LOS, sigue existiendo una falencia en la exploración de técnicas que tengan en cuenta las características anteriormente mencionadas. Dado este escenario, en el presente trabajo se desarrolló la metodología de Arboles de Modelos GAMLSS (AMG), en la cual se utilizan los modelos aditivos generalizados de localización, forma y escala (GAMLSS) y se incorpora un procedimiento de selección de modelos dentro de varios candidatos que se generan. La metodología se aplica en tres casos de estudio con diferentes características, comparándola, con otras técnicas que fueron seleccionadas de acuerdo a la revisión de la literatura, posteriormente se realiza una validación de la metodología con base en la simulación de dos escenarios planteados.

Dentro de los principales logros de la investigación se encuentran: el diseño e implementación en software de una metodología que facilita la generación y selección de modelos de predicción de variables continuas, la inclusión del criterio de información de Akaike (AIC) como métrica para comparar y seleccionar modelos, lo que permite dar un enfoque basado en la bondad de ajuste y no solo en la disminución de residuales como la mayoría de las técnicas actuales; otro logro es la mejora encontrada en algunas de las métricas con la que se miden los modelos generados, con respecto a otras técnicas con las que se comparó. Finalmente, se destaca la creación de una estructura de modelos nuevos que se basa en un árbol de decisión y un modelo GAMLSS en cada nodo, que permite modelar fenómenos en los que se puede encontrar subgrupos de observaciones provenientes de distribuciones diferentes.

Tabla de Contenido

Lista de Tablas

Lista de figuras

Lista de abreviaturas

LOS: Length of saty

AIC: Akaike information criterion

BIC: Bayesian information criterion

SBIC: Schwarz's bayesian information criterion

RVC: Residuals variance criterion

MAE: Mean absolute error

RMSE: Root mean square error

MSE: Mean square error

GAMLSS: Generalized additive models for location, scale and shape

GLM: Generalized linear models

CART: Classification and regression tree

SVM: Support vector machine

CHAID: Chi-square automatic interaction detection

DRG: Diagnosis related group

Glosario

Técnica: Es un conjunto de supuestos y características que definen cierta estructura que puede tomar un modelo para representar la relación entre una variable respuesta y sus covariables.

Metodología: Es un conjunto de procedimientos que permite encontrar generar y seleccionar un modelo adecuado para representar ciertas relaciones entre variables para un caso específico.

Modelo: Es una representación parcial de la estructura de un problema. En el caso de modelos de predicción se busca representar la relación entre las covariables para explicar una variable respuesta.

Cubista: Es una técnica en la que la estructura de los modelos está definida por un árbol de decisión bajo el algoritmo M5 y en cada nodo una regresión lineal múltiple.

Desempeño: El desempeño de un modelo es la capacidad de éste para representar la realidad adecuadamente, generalmente se definen indicadores para medir la idoneidad del modelo, algunos se enfocan en la disminución de residuales, mientras que otros se enfocan en la bondad de ajuste y penalización por complejidad.

1. Introducción

Los hospitales se ven obligados continuamente a diseñar estrategias que les permitan, en primer lugar, ser más eficientes con el uso de sus recursos, ya que los gastos han ido incrementando debido al envejecimiento de la población, uso de tecnología médica costosa y a la naturaleza de las enfermedades actuales (Weir, D'Entremont, Stalker, Kurji, & Robinson, 2009); en segundo lugar, ser más efectivos con los servicios brindados, debido al aumento en la presión por parte de entidades regulatorias y usuarios cada vez más exigentes (H. C. Liu, 2013); en tercer lugar, mejorar la experiencia del cliente dado que es una ventaja competitiva y factor de éxito crítico (Vaish, Vaish, Vaishya, & Bhawal, 2016). Estas estrategias frecuentemente se diseñan sin tener información que las soporte, están basadas en la intuición de los directivos, disminuyendo la probabilidad de que tengan éxito (Weir et al., 2009). Es por esto por lo que se hace indispensable desarrollar estrategias soportadas por la información disponible en los hospitales.

La eficiencia en el uso de los recursos es uno de los objetivos más importantes en la literatura de administración hospitalaria (Al Taleb, Abul Hasanat, & Khan, 2017; Barnes, Hamrock, Toerper, Siddiqui, & Levin, 2016; Nouaouri, Samet, & Allaoui, 2015; Turgeman, May, & Sciulli, 2017). Para alcanzar este objetivo los hospitales utilizan diversas estrategias, entre ellas programar adecuadamente las cirugías y otros servicios prorrogables que generarán hospitalización, de tal manera, que los recursos no estén subutilizados pero que tampoco sean insuficientes para atender la demanda (Azari, 2015; Barnes et al., 2016; Gustafson, 1968). Este problema ha sido abordado por diferentes estudios por medio de la predicción del LOS (Al Taleb et al., 2017; Barnes et al., 2016; Nouaouri et al., 2015; Turgeman et al., 2017). Otra estrategia que se plantea continuamente dentro de los hospitales es la disminución del LOS debido al impacto que tiene en los costos de atención de los pacientes (Awad, Bader–El–Den, & McNicholas, 2017; V. Liu, Kipnis, Gould, & Escobar, 2010; Widyastuti, Stenseth, Wahba, Pleym, & Videm, 2012). Para lograr esta disminución es fundamental identificar las causas y por otro lado las características de las personas, que influyen tanto positiva como negativamente en el LOS, y de esta manera facilitar las acciones encaminadas a disminuirlo (V. Liu et al., 2010; Turgeman et al., 2017; Widyastuti et al., 2012).

Desde otra perspectiva, para lograr el objetivo de mejorar la efectividad en los servicios médicos prestados, las entidades implementan la estrategia de monitorear preventivamente que la capacidad de personal, de espacio e instalaciones, en sus unidad de emergencias no se desborde, ya que esto podría ocasionar un aumento de la probabilidad de que la condición de un paciente se agrave (Azari, 2015; Golmohammadi, 2016). El LOS también ha sido utilizado para mejorar la efectividad de los servicios médicos al utilizarlo para la generación de alertas sobre posibles complicaciones o cambios en las condiciones de paciente, que permiten enfocarse en la atención prioritaria a algunos pacientes (Awad et al., 2017; Azari, Janeja, & Mohseni, 2012), ya que cuando el tiempo real de un paciente ha superado el tiempo estimado con una diferencia significativa, se debe a que el paciente sufrió alguna complicación o existen ineficiencias en la atención prestada (Barnes et al., 2016).

El tercer objetivo que tienen las instituciones de salud para lograr el éxito, es la mejora en la satisfacción de sus usuarios (Vaish et al., 2016). Algunos autores coinciden en que el tiempo de espera juega un rol fundamental en dicha satisfacción, no solo por la cantidad de tiempo de espera sino por el hecho de ser informado de dicho tiempo (Carter & Potts, 2014). Se afirma que: "La percepción del tiempo de espera y la información sobre este tiempo influyó en la satisfacción de los usuarios" (Fontova-Almato, Juvinya-Canal, & Suner-Soler, 2015). Otros estudios también han encontrado evidencia de los efectos negativos que tiene la incertidumbre en el contexto de pacientes hospitalizados, ya que es un factor que está correlacionado positivamente con la ansiedad (Pahlevan Sharif, 2017). Esto evidencia que es fundamental tener una predicción acertada del LOS para poder informar a los pacientes y lograr una mejora en su satisfacción con la atención prestada.

En los tres objetivos mencionados anteriormente y las estrategias para lograrlo, se evidencia la importancia de la predicción del LOS y por qué esta variable ha sido el interés de muchos estudios en literatura relacionada con la administración en salud (Azari et al., 2012; Barnes et al., 2016; Cai et al., 2016; Nouaouri et al., 2015; Perez, Chan, & Dennis, 2006; Pilotto et al., 2016; Widyastuti et al., 2012). El LOS es medido restando la fecha de egreso de la fecha registrada de ingreso al hospital, midiendo un episodio individual de hospitalización (Turgeman et al., 2017). La mayor parte de los estudios se enfocan en predecir el tiempo en el momento del ingreso, con la información disponible en ese punto (Azari et al., 2012; Perez et al., 2006; Turgeman et al., 2017), sin embargo también existen estudios que han estudiado el problema de predecir el LOS, no solo en el inicio de la estancia, sino actualizando esta predicción a medida que se registran eventos que permitan ajustarla (Barnes et al., 2016; Cai et al., 2016).

El problema de predecir el LOS no ha sido investigado únicamente por su impacto en la administración hospitalaria, sino también por el interés que despierta su modelamiento y predicción en el contexto académico, debido a las características y la complejidad que estas conllevan (Turgeman et al., 2017). Una de las características que hace difícil el modelamiento de esta variable es que su distribución es altamente asimétrica, por lo cual limita el uso de algunos modelos a causa de sus supuestos, por ejemplo modelos de regresión lineal qué han sido foco de críticas de algunos investigadores (Carter & Potts, 2014; Faddy, Graves, & Pettitt, 2009). Otra característica es que existen muchas variables predictoras que pueden influir en el LOS, razón por la cual, las técnicas utilizadas deben tener componentes claros de selección de variables. Estas razones han conseguido que la literatura relacionada con la predicción de LOS se enfoque en la exploración y comparación de técnicas (Tanuja, Acharya, & Shailesh, 2011; Turgeman et al., 2017; Verburg, De Keizer, De Jonge, & Peek, 2014). Se identifican dos oportunidades en esta área de estudio, la primera es la no existencia de acuerdos en la literatura relacionados con la idoneidad de alguna de las técnicas exploradas, que hacen que el diseño de nuevas técnicas, sea un área de vigente interés. La segunda oportunidad se presenta con relación a la selección de los modelos que una técnica puede generar, ya que este no ha sido un tema que se haya abordado con profundidad.

Otro aspecto importante en la predicción del LOS es la dificultad en la generalización o extrapolación de los modelos, debido a que, según las características del país, ciudad o región, éste puede cambiar. En primer lugar, por la información disponible, ya que cada hospital puede llevar en sus registros información diferentes, de acuerdo con sus

necesidades y regulaciones. En segundo lugar, por las características de la población, ya que la influencia de una variable puede ser diferente en poblaciones distintas. En tercer lugar, por la existencia de regulaciones, que pueden influir en LOS y que son difíciles de detectar y controlar en este tipo de estudios, como por ejemplo regulaciones que le exijan a un hospital cumplir con unos parámetros mínimos de atención o cumplir ciertas condiciones para autorizar el egreso del paciente. Dado esto algunos autores mencionan la necesidad de que las investigaciones hagan un énfasis en la metodología con la que se obtiene el modelo para que sea replicable en diferentes contextos (Awad et al., 2017; Perez C., Marquez G., Acosta M., & Mezura M., 2017).

Dentro de las técnicas que se han usado para modelar el LOS se resalta la técnica cubista. Esta técnica se compone de un árbol de decisión y en cada nodo final se ajusta una regresión lineal (Kuhn & Johnson, 2013). La técnica cubista presenta características interesantes en el contexto de predicción del LOS, ya que ha mostrado un mejor MAE comparada con otras técnicas (Turgeman et al., 2017). Además sintetiza aprendizajes de investigaciones anteriores al utilizar dos de las técnicas más comunes en la literatura, regresión lineal y árboles de decisión. Sin embargo, es una técnica nueva que tiene muchas variantes por estudiar.

En la presente investigación se reconoce la necesidad de explorar algunas alternativas a las técnicas, que como la cubista hayan tenido buenos resultados, utilizando el conocimiento que ellas recogen, y buscando mejorar su desempeño, además de ampliar el espectro de problemas para los que se puede aplicar.

En lo que resta del documento se presenta, en la sección 2 los objetivos de la investigación, en la sección 3 se presenta la revisión de la literatura que permiten tener un marco de referencia y conocer las tendencias en las investigaciones relacionadas con predicción de LOS en cuanto a técnicas y selección de modelos. En la sección 4 se presenta la metodología desarrollada y su implementación en software. En la sección 5 se presentan los análisis y discusión de los casos de estudio para la aplicación de la metodología desarrollada, en la que se compara con otras técnicas existentes. En la sección 6 se valida la metodología con dos escenarios de simulación planteados. En la sección 7 se presentan las conclusiones, en la sección 8 los trabajos futuros, en la sección 9 las referencias utilizadas y finalmente en la sección 10 se presentan los Anexos.

2. Objetivos

2.1 Objetivo general

Diseñar una metodología que permita generar y seleccionar un modelo de predicción del LOS, que recoja aprendizajes de estudios previos y que tenga un desempeño competitivo con otras técnicas, desde el punto de vista de bondad de ajuste y de residuales del modelo seleccionado.

2.2 Objetivos específicos

1. Identificar en la literatura las técnicas usadas en la predicción del tiempo de estancia hospitalaria sus ventajas y desventajas.
2. Identificar las características más comunes de los problemas de predicción de tiempo de estancia hospitalaria.
3. Comparar e identificar las oportunidades en las técnicas identificadas en la literatura para predecir el tiempo de estancia hospitalaria.
4. Proponer y diseñar una técnica para predecir el tiempo de estancia hospitalaria que tenga un procedimiento propuesto para la selección del modelo.
5. Validar la técnica y procedimiento propuesto a partir de la de diferentes casos de estudio.

3 Revisión de la literatura

En esta sección se presenta un marco de referencia teórico de la problemática abordada y se analizan los resultados de diferentes investigaciones en el área. Para esto, la sección se dividirá en 3 partes, la primera parte es sobre la importancia del tiempo de estancia hospitalaria (LOS), sus características y las aplicaciones más importantes. En la segunda parte se presentan las técnicas utilizadas y su evolución histórica, finalmente en la tercera parte se presenta una revisión sobre la selección de modelos y los diferentes criterios que se han utilizado.

3.1 Tiempo de estancia hospitalaria, características y aplicaciones:

El tiempo de estancia hospitalaria (LOS) es un indicador que se mide para un paciente en un solo episodio de hospitalización. Los días se cuentan desde el registro de ingreso hasta el egreso del paciente; este último puede presentarse por el alta del paciente o la muerte (Jiménez, López, Dominguez, & Fariñas, 1999; V. Liu et al., 2010; Marazzi, Paccaud, Ruffieux, & Beg, 1998). En las diferentes investigaciones se han encontrado coincidencias en relación con ciertas características que describen esta variable. La primera es que solo toma valores positivos, en algunas investigaciones se toma como variable continua, al tener en cuenta la hora de ingreso y de egreso (Faddy et al., 2009; V. Liu et al., 2010), en otros casos se toma como un conteo omitiendo la hora (Carter & Potts, 2014; Verburg et al., 2014). La segunda característica es que es una variable en la que hay presencia de valores extremos correspondientes a casos con estancias muy largas, aunque el promedio se mantenga en un rango relativamente pequeño. La tercera característica es el sesgo a la derecha que se debe a la presencia de valores extremos y a la concentración de la variable en un rango pequeño, lo cual genera una distribución leptocúrtica. Algunos autores han analizado el ajuste del LOS a diferentes distribuciones, dentro de las que se encuentran la distribución exponencial (Awad et al., 2017), Log-normal, Gamma, Weibull (V. Liu et al., 2010; Marazzi et al., 1998) y adicionalmente se ha estudiado el ajuste a distribuciones discretas como la Binomial negativa y la Poisson (Carter & Potts, 2014; Verburg et al., 2014). Por otro lado se ha criticado el ajuste de modelos de regresión lineal, debido a que la estancia no se ajusta a una distribución normal (Carter & Potts, 2014; Faddy et al., 2009; Turgeman, May, Ketterer, Sciulli, & Vargas, 2015).

El LOS es un variable de mucha importancia porque se ha encontrado que existe relación entre ésta y otras variables claves en la administración hospitalaria, entre ellas está el consumo de recursos (Burns & Wholey, 1991; Marazzi et al., 1998), el costo de atención de los pacientes (Jiménez et al., 1999; Marazzi et al., 1998) o la mortalidad en un hospital (Marazzi et al., 1998; Widyastuti et al., 2012). Dada esta asociación, se ha utilizado el LOS como sustituta de estas variables y se ha usado para la toma de decisiones. Una de las estrategias más populares en los hospitales es la identificación de grupos de pacientes con estancia similares para diseñar esquemas de atención diferenciados y eficientes, también se utilizan los mencionados grupos para analizar indicadores de manera desagregada (Fetter, Shin, Freeman, Averill, & Thompson, 1980; Marazzi et al., 1998).

Estos grupos son conocidos como grupos relacionados de diagnóstico (DRG's), cuyo objetivo es definir unos tipos de casos que se espera que reciban los mismos servicios y tengan los mismos resultados (Fetter et al., 1980).

Existen dos enfoques con los que se ha estudiado el LOS, el primero busca entender la asociación entre el LOS y otras variables, el segundo se enfoca en modelar y predecir el LOS con base en unas variables predictoras. Los primeros son de tipo causal y fueron el enfoque inicial en el estudio de esta variable. Algunos trabajos iniciales se presentan en los años 70 donde se documentaron variaciones en el LOS causadas por diferencias geográficas (Burns & Wholey, 1991), posteriores trabajos como el de V. Liu et al., (2010) también han dirigido sus esfuerzos en identificar variables que permitan explicar adecuadamente el LOS. Dentro de las variables que se han encontrado como significativas están la severidad y factores relacionados con las instalaciones del hospital y de la comunidad que rodea al paciente, las cuales juegan un rol fundamental en las variaciones de esta variable, aunque las relaciones causales no hayan sido comprendidas en su totalidad (Burns & Wholey, 1991). También, son incluidas con frecuencia la comorbilidad y el diagnóstico o enfermedad principal del paciente, como factores críticos para explicar el LOS (Azari et al., 2012; V. Liu et al., 2010). A continuación, se profundizará en tres de estas variables, ya que son de alta complejidad y adicionalmente permiten una comprensión más amplia del fenómeno.

La primera variable es la comorbilidad que se define como la aparición de dos o más condiciones médicas en un mismo paciente, este término es usado cuando se enfoca en cómo otra condición influencia el manejo de una condición de referencia. La aparición de varias condiciones en un paciente es una prioridad internacional de la salud, debido a que cada vez más pacientes viven con este fenómeno (Lawson et al., 2017). En algunos estudios sobre mortalidad o LOS se excluyen pacientes con comorbilidad debido a que ésta puede ser una variable de confusión que aumente la incertidumbre de los análisis, aunque por otro lado esto crea limitaciones para la generalización de los resultados (Charlson & Horwitz, 1984). La comorbilidad es una variable difícil de medir debido a que pueden existir una cantidad de combinaciones de enfermedades muy amplia, por lo cual se suelen utilizar índices que permitan analizar esta variable de una manera más fácil. Uno de estos índices es el comorbidity point score (COPS), el cual captura la aparición de enfermedades en los últimos 12 meses y con base en éstas se genera un score de 0 a 701 (V. Liu et al., 2010). Este índice ha mostrado ser una variable importante para la predicción de LOS (V. Liu et al., 2010).

La segunda variable de suma importancia es el diagnóstico o enfermedad principal del paciente. El manejo de esta variable en este tipo de análisis es complejo debido a la gran variedad de enfermedades identificadas, esto ha llevado a investigadores y diferentes instituciones a definir agrupaciones de enfermedades como el DRG, el cual se creó para agrupar pacientes que requieren un LOS similar (Burns & Wholey, 1991; Marazzi et al., 1998). El grupo relacionado de diagnóstico es ampliamente utilizado en la literatura para segmentar la población en las estimaciones de LOS y en muchos casos incluso para trabajar con el promedio de LOS de estos grupos como predicción del mismo. También es utilizado para tener estándares con los cuales los hospitales puedan compararse y medir su eficiencia (Burns & Wholey, 1991). Existen otras clasificaciones de enfermedades como Charlson index (Azari et al., 2012; Charlson, Pompei, Ales, & MacKenzie, 1987)

que cuenta el número de enfermedades por paciente y las clasifica de acuerdo a la severidad, aunque este estudio se enfoca en clasificar las enfermedades para medir la tasa de mortalidad, también es utilizado en estudios para analizar el LOS (Azari et al., 2012). Otra agrupación es la clasificación internacional de enfermedades versión 10 (CIE-10), realizada por la Organización mundial para la salud (OMS-Organización Mundial de la Salud, 1990). Esta clasificación es ampliamente usada, existen 14400 códigos de diagnósticos que se dividen en varias categorías (Barnes et al., 2016).

La tercera variable que es importante explicar en el estudio de LOS es la severidad, esta es una medida que permite estimar una probabilidad de muerte de un paciente de acuerdo a diferentes modelos. Algunos de los modelos utilizados para medir la severidad son el charlson index que clasifica las enfermedades y también indica la severidad (Charlson et al., 1987). Otro sistema que mide esta variable es el modelo acute physiology and chronic health evaluation (APACHE III), el cual es un ranking de clasificación de enfermedades que se aplica dentro de las 24 horas siguientes a la admisión de un paciente a cuidados intensivos. Este modelo APACHIE III utiliza un score de 0 a 299 con base en algunas variables demográficas y otras sintomáticas como presión sanguínea, temperatura corporal, ritmo cardiaco, entre otras (Woods, MacKirdy, Livingston, Norrie, & Howie, 2000). También se ha utilizado el intensive care national audit and research centre (ICNARC) score que mide el riesgo para unidades de cuidados intensivos en Inglaterra (Harrison, Parry, Carpenter, Short, & Rowan, 2007). Finalmente se tiene el mortality probability model (MPM) que mide la probabilidad de muerte de pacientes en cuidados intensivos (Lemeshow et al., 1993).

El segundo enfoque, más común en la literatura reciente, es la predicción y modelamiento del LOS (Al Taleb et al., 2017; Azari et al., 2012; Faddy et al., 2009; Tanuja et al., 2011). Estas investigaciones son de tipo correlacional y pueden realizarse con dos objetivos, el primero es el modelamiento de la variable LOS como respuesta explicada por un conjunto de variables explicativas, en este caso se estudia el tipo de relación que tiene cada una de las covariables con las variable respuesta y el grado de asociación entre las mismas (Burns & Wholey, 1991; Marazzi et al., 1998). Cuando se realizan estudios con este objetivo se presentan muchas similitudes con el enfoque anterior, ya que al final lo que se quiere es entender las variables que están impactando el crecimiento o decrecimiento del LOS, principalmente para identificar variables explicativas que pueden ser intervenidas para disminuir el LOS, un ejemplo de este tipo de variables es la cantidad de pacientes por cada enfermera disponible, ya que son variables que pueden ser modificadas por decisiones en la administración hospitalaria. Sin embargo, muchos modelos predictivos tienen el objetivo principal de predecir acertadamente el LOS, sacrificando incluso la capacidad de entender las relaciones entre variables explicativas y variables respuestas para tener mayor precisión en la predicción, esto se presenta más en las técnicas computacionales que se conocen como técnicas de tipo caja negra (Dayhoff & DeLeo, 2001; Mobley, Leasure, & Davidson, 1995).

La predicción del LOS se ha realizado para diferentes tipos de población, existen casos en los que se hace una predicción para un hospital y todos sus pacientes (P. Liu et al., 2006), otros casos se han enfocado en predecir LOS para unidades particulares como cuidados intensivos (Perez et al., 2006), algunos estudian el problema para un grupo de pacientes con enfermedades comunes, como por ejemplo, accidentes cerebrovasculares

(Al Taleb et al., 2017), o para grupos poblacionales comunes, como por ejemplo, personas de tercera edad (Pilotto et al., 2016). El estudio para poblaciones específicas se presenta principalmente debido a que se ha evidenciado que no ha sido posible encontrar un modelo que prediga adecuadamente y de manera general la estancia hospitalaria de todos los pacientes (Widyastuti et al., 2012). Este tema se ha abordado en algunos estudios, dividiendo la población y definiendo un modelo por cada grupo de la población, por ejemplo, por grupo relacionado de diagnóstico (Marazzi et al., 1998) o creando agrupaciones basados en generación de cluster (Azari et al., 2012; Charlson et al., 1987). Estas aproximaciones han permitido avanzar en estrategias que permitan tratar los problemas que surgen con el modelado de esta variable.

3.2 Técnicas para predicción del LOS y su evolución histórica

En esta sección se definen los conceptos modelo, técnica y metodología en el contexto de la predicción de variables, también se define la relación entre las medidas de desempeño de estos tres conceptos, posteriormente se plantea una descripción de las técnicas utilizadas en la predicción del LOS, adicionalmente se propone una clasificación de las mismas, y finalmente se estudia su evolución en el tiempo.

Los conceptos modelo, técnica, y metodología, en el contexto de predicción, son fundamentales en el presente trabajo, por lo cual, se hace indispensable definirlos dentro del marco de la investigación. Una técnica se entiende como un conjunto de procedimientos y supuestos que permiten proponer y estimar un modelo que cumpla una estructura determinada (Flach, 2012). Por otro lado, un modelo es el resultado de una técnica, el modelo es una representación parcial o simplificada de un problema que, en el caso de predicción, es una combinación de variables explicativas que definen el comportamiento de una variable respuesta (Flach, 2012).

Las metodologías son una conjunto de procedimientos que permiten generar varios modelos pertenecientes a una o varias técnicas y seleccionar un modelo final, a diferencia de las técnicas, las metodologías cubren algunos procedimientos que no están dentro de los que son abarcados por la técnica, por ejemplo, en el caso de modelos lineales generalizados, la selección de variables, de la distribución, de la función de enlace, la definición del criterio de para medir el desempeño de los modelos y la selección del modelo final con base en candidatos, no están incluidas dentro de los procedimientos que enmarcan la técnica de modelos lineales generalizados. En muchos casos estas actividades son desarrolladas por los investigadores basado en su experiencia, sin embargo, en la literatura también se han desarrollado metodologías que abarcan este tipo de actividades (Ojeda & Rocco, 2014). Es importante resaltar que para la medición de un modelo, técnica o metodología se hace a través de los indicadores que se obtienen con los modelos que éstas metodologías o técnicas seleccionan como el modelo final ya que no existen métricas propias para medir el desempeño de una técnica o metodología (Al Taleb et al., 2017; Ojeda & Rocco, 2014; Turgeman et al., 2017)

Para predecir el LOS se han utilizado técnicas de diferentes campos de estudio que han generado varias perspectivas para abordar el mismo problema. Para tener una estructura

que permita entender estas perspectivas y la evolución de esta área de estudio se han propuesto clasificaciones de técnicas, una de ellas es la propuesta por Awad et al., (2017) la cual se presenta en la Figura 1. En esta clasificación se identifican cuatro grupos de técnicas que, los autores consideran, resumen adecuadamente las diferentes aproximaciones que se encuentran en la literatura.

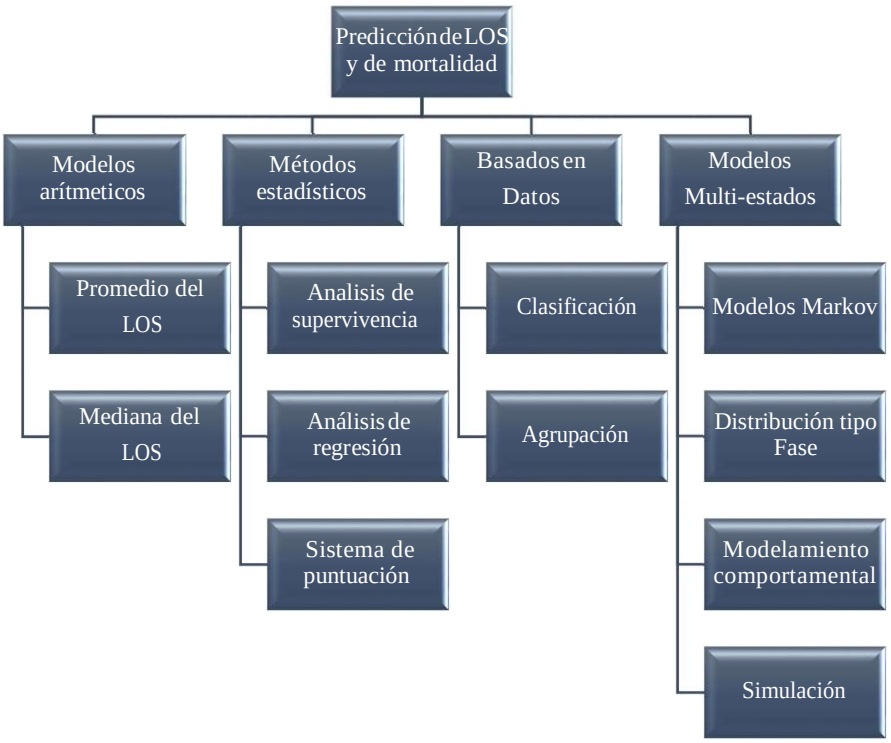


Figura 1 Clasificación de técnicas de predicción del LOS. Adaptado de Awad et al., (2017)

A pesar de que se reconoce la complejidad del LOS, algunos modelos aritméticos siguen siendo usados como una predicción de esta variable. Los más comunes son el cálculo de medidas de tendencia central como la media y la mediana, medidas que pueden generar errores debido a la naturaleza de la variable que generalmente no es simétrica. Adicionalmente estos cálculos univariados desconoce el efecto de otras variables sobre el LOS (Marshall, Vasilakis, & El-Darzi, 2005). Los métodos estadísticos se usan para suplir la necesidad que dejan los modelos aritméticos, usando variables explicativas y distribuciones de datos de diferentes tipos. Los principales métodos utilizados son análisis de regresión y análisis de supervivencia (Awad et al., 2017). En estudios recientes ha crecido la popularidad de los métodos basados en datos, los cuales pertenecen a técnicas de minería de datos. Dentro de este grupo, se encuentran métodos de clasificación y agrupación. Finalmente está el grupo de modelos multi estados basados en cadenas de markov, en los que generalmente se asumen transiciones de los pacientes por diferentes estados, estos sirven para analizar el flujo del paciente en los departamentos de un hospital (Awad et al., 2017).

Con base en esta clasificación y de acuerdo con la revisión de literatura que se realizó, se propone una clasificación propia para agrupar las técnicas de predicción de LOS. En la propuesta se incluyen dos nuevas categorías: los métodos empíricos y las técnicas

combinadas, los primeros son una práctica ampliamente utilizada en los hospitales (Gustafson, 1968), dentro de este grupo se encuentran las estimaciones de expertos, que pueden ser especialistas, médicos generales o un grupo interdisciplinario, que de acuerdo con el diagnóstico inicial, definen el número de días que estará el paciente. Las segundas son mezclas de técnicas existentes con las que se busca mejorar los resultados, aprovechando las bondades de dos técnicas diferentes. Finalmente, se presenta otro cambio en el grupo de técnicas basadas en datos que será llamado en la propuesta como técnicas computacionales con el fin de tener un concepto más amplio, dentro de ellas se encuentran investigaciones de machine learning, minería de datos y otras ramas de las ciencias de la computación que aparecen con frecuencia en la literatura. En la Figura 2 se presenta la clasificación propuesta con los grupos de técnicas encontradas en la revisión. Adicionalmente se presenta el porcentaje de artículos en los que se utiliza cada técnica, con respecto al número total de artículos de la revisión de la literatura en el que se encontraron aplicaciones de técnicas, esto con el objetivo de analizar la frecuencia de uso de las mismas.

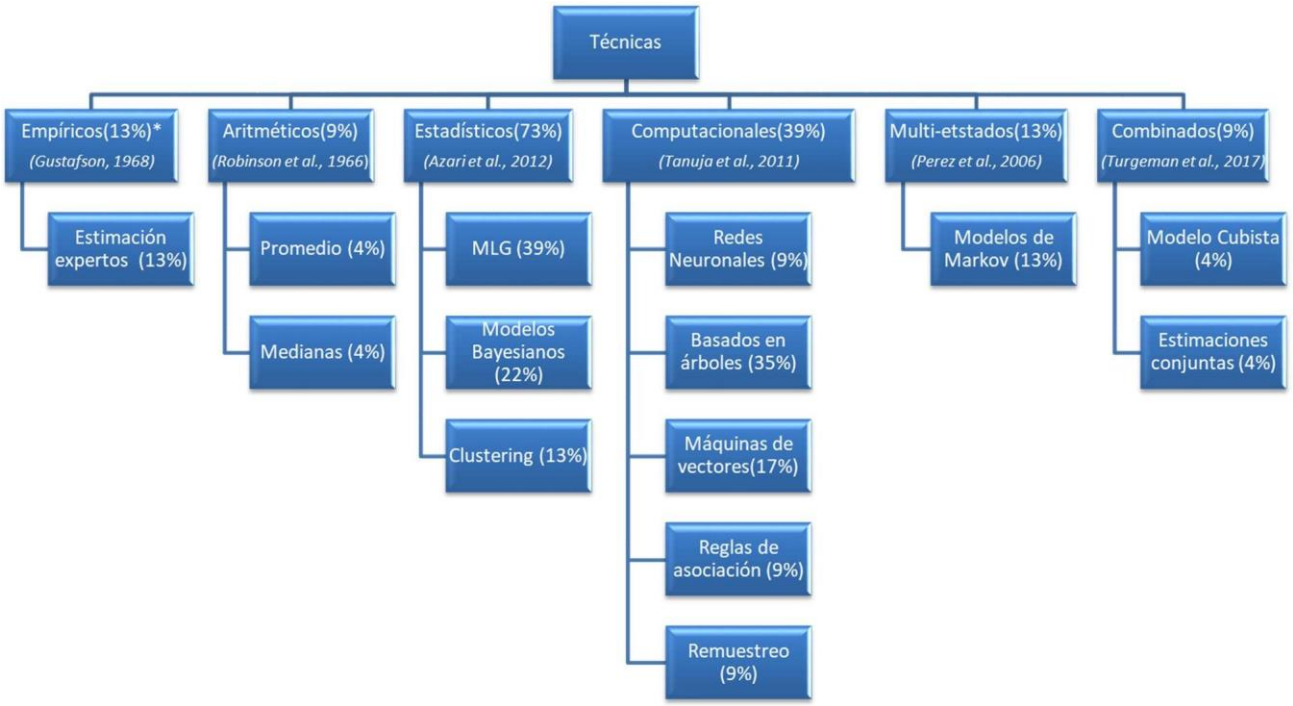


Figura 2 Clasificación de técnicas de predicción del LOS propuesta

Para complementar la clasificación presentada en la Figura 2, se presenta en la Tabla 1 el listado completo de las técnicas utilizadas, su clasificación y algunas referencias en las que se utilizan estas técnicas en el contexto de predicción del LOS. En esta tabla se muestra una lista de 30 técnicas que se dividen en los 6 grupos definidos: técnicas empíricas presentes en el 13% de los artículos, técnicas aritméticas presentes en el 9% de los artículos, técnicas estadísticas presentes en el 74% de los artículos, técnicas computacionales presentes en el 39% de los artículos, técnicas multi-estados presentes en el 13% de los artículos y técnicas combinadas presentes en el 4% de los artículos.

Tabla 1 Lista de técnicas de predicción LOS

Clasificación	Técnica	Referencia
Empíricos	Panel de expertos	(Barnes et al., 2016; Gustafson, 1968; Robinson)
Aritmético	Media	(Gustafson, 1968)
	Mediana	(Gustafson, 1968)
Estadístico	Análisis Discriminante	(Altman, Angle, Brown, & Sletten, 1972)
	Clasificador ingenuo de Bayes	(Azari et al., 2012; P. Liu et al., 2006)
	Redes Bayesianas	(Al Taleb et al., 2017; Azari et al., 2012)
	K-medias	(Azari et al., 2012)
	K-Vecinos cercanos	(Houthooft et al., 2015; Tanuja et al., 2011)
	Estimador de Odds	(Gustafson, 1968)
	Regresión de riesgos proporcionales de cox	(Verburg et al., 2014; Widyastuti et al., 2012)
	Regresión Binomial Negativa	(Carter & Potts, 2014; Verburg et al., 2014)
	Regresión de Poisson	(Carter & Potts, 2014; Verburg et al., 2014)
	Regresión Gamma	(V. Liu et al., 2010)
	Regresión lineal	(Jiménez et al., 1999)
	Regresión logística	(Azari, 2015; Barnes et al., 2016)
Computacionales	Boosting discreto adaptativo	(Houthooft et al., 2015)
	Bagging	(Azari et al., 2012)
	Boosting	(Azari et al., 2012)
	Árbol C4.5	(Pendharkar & Khurana, 2014; Perez et al., 2006)
	Árbol CART	(Houthooft et al., 2015)
	Árbol CHAID	(Pendharkar & Khurana, 2014)
	Bosques aleatorios	(Azari et al., 2012)
	Redes Neuronales	(Tanuja et al., 2011)
	Reglas de asociación	(Azari et al., 2012; Turgeman et al., 2017)
	Máquina de vector de relevancia	(Houthooft et al., 2015)
	Máquina de soporte vectorial	(Azari et al., 2012; Turgeman et al., 2017)
Multi-estados	Distribución tipo Fase	(Faddy et al., 2009)
	Cadenas de markov	(Faddy et al., 2009; Perez et al., 2006)
Combinados	Estimación conjunta	(Pendharkar & Khurana, 2014)
	Modelo cubista	(Turgeman et al., 2017)

Adicionalmente a la identificación de estos grupos de técnicas, también pueden diferenciarse las técnicas en el tipo de variable respuestas que utilizan, algunas tienen en cuenta variables respuestas categóricas y son conocidas como técnicas de clasificación, en este tipo de métodos se agrupan los días de estancia en rangos y se predice el rango en el que queda cada paciente (Mobley et al., 1995; Tanuja et al., 2011). Por otro lado, están las técnicas de regresión que estiman el LOS para variables respuestas continuas y la predicción, a diferencia de los métodos de clasificación, no es de pertenencia a un grupo sino que se predice el valor exacto que va tomar la observación de acuerdo a las covariables (Marazzi et al., 1998; Turgeman et al., 2017). A continuación se presentarán algunas aplicaciones que se han encontrado de los diferentes tipos de técnicas mencionados.

En el caso de técnicas de clasificación se encuentran varias aplicaciones, Liu et al. (2006), por ejemplo, predice el LOS en un hospital geriátrico aplicando un clasificador Ingenuo de Bayes apoyado en un árbol de clasificación C4.5 para la selección de las variables más importantes, adicionalmente para mejorar el desempeño se utiliza una imputación Ingenua de Bayes para datos faltantes (P. Liu et al., 2006). En el trabajo de Tanuja et al. (2011) se comparan modelos de clasificación para la predicción de LOS en un hospital especializado, se utilizan cuatro modelos: redes neuronales, clasificador de Bayes ingenuo, modelo de k vecino más cercano y un árbol de decisión C4.5. Se encuentra que la red neuronal es la de mejor desempeño (Tanuja et al., 2011). En el trabajo de Al Taleb et al. (2017) en un hospital de Arabia Saudita en pacientes con problemas cerebrovasculares se utiliza un algoritmo J48 de árbol de clasificación que está basado en un árbol C4.5 y una Red Bayesiana, métodos que han sido utilizado en otras aplicaciones médicas (Kaur & Chhabra, 2014; Kumari, Vohra, & Arora, 2014) y la red Bayesiana tuvo un mejor desempeño (Al Taleb et al., 2017).

Para el caso de análisis de variables continuas , Marazzi et al. (1998) propone ajustar el LOS a tres distribuciones de probabilidad, Weibull, Gamma y Lognormal. En este estudio se ajustan estas distribuciones a 3260 muestras en las que se divide la población que inicialmente estaba formada por aproximadamente cinco millones de hospitalizaciones y se identifica que la distribución lognormal es la que se ajusta en la mayoría de las muestras. En este trabajo también se tuvo la limitación de no tener variables que explicaran el LOS, lo que reduce la capacidad de las instituciones de implementar estrategias para mejorar la eficiencia (Marazzi et al., 1998; Turgeman et al., 2017; Verburg et al., 2014). Otros trabajos abordan estas limitaciones con técnicas de regresión que tienen en cuenta las variables predictoras. Tal es el caso del trabajo de Carter & Potts (2014) quiénes, utilizando una regresión lineal, analizan la estancia hospitalaria de pacientes con operación total de rodilla en un hospital de servicios ortopédicos especializados. También se encuentra el trabajo propuesto por Verburg et al. (2014), en el que se realiza una comparación de ocho modelos de regresión diferentes, Regresión Lineal sin transformaciones, regresión lineal sin transformaciones eliminando LOS mayores a 30 días, regresión lineal con transformación logarítmica, modelo lineal generalizado con distribución Gaussiana y una función de enlace logarítmica, regresión de Poisson, regresión binomial negativa, regresión Gama con función de enlace logarítmica y regresión de riesgo proporcional de cox. En este estudio se encontró que el modelo de Cox y la regresión lineal con transformación logística tenían mejor desempeño para estancias inferiores a cuatro días, y en los otros casos el modelo lineal generalizado

con función de enlace logarítmica tuvo mejor desempeño, aunque la diferencia fue poca entre los diferentes modelos. También se concluyó que los modelos planteados no permiten predecir adecuadamente el LOS (Turgeman et al., 2017; Verburg et al., 2014).

Dentro de la familia de las regresiones también se han explorado métodos más sofisticados comúnmente utilizados en el campo de machine learning y minería de datos. Pendharkar & Khurana (2014) analizaron y compararon tres métodos de regresión con los datos de 88 hospitales de Pensilvania, arboles CART, arboles CHAID y regresión de vector de soporte, encontrando que aunque no habían diferencias significativas, el método CART es el más fácil de interpretar (Pendharkar & Khurana, 2014). En el trabajo de Houthooft et al. (2015) también se comparan varios métodos de regresión con los datos de un hospital universitario en su departamento de cuidados intensivos, los métodos utilizados son redes neuronales, regresión de vector de soporte, árbol CART, bosques aleatorios, regresión de vector de relevancia, encontrando que el método de mejor desempeño fue la regresión de vector de soporte.

Debido a las dificultades de encontrar un modelo que prediga adecuadamente el fenómeno de LOS, algunos autores han explorado métodos combinados. Por ejemplo Turgeman et al. (2017) aplica un modelo de árbol cubista, aunque el modelo no es nuevo ya que fue propuesto anteriormente (Quinlan, 1992), es menos común en la literatura de aplicación a la predicción del LOS. Este modelo permite estimar el valor de la estancia para cada paciente y no un rango de tiempo como los modelos de clasificación. Se aplicó en un hospital de adultos mayores con problemas cardiovasculares, este modelo obtuvo mejor desempeño que los otros cinco modelos con los que se comparó (árbol CART, árbol CHAID, máquina de vector de soporte, redes neuronales, modelo lineal generalizado con función de enlace Poisson).

En 1966 se presenta uno de los trabajos seminales en predicción de LOS (Robinson et al., 1966), el cual se realizó por parte de investigadores del campo de salud. En este estudio se realizaron análisis exploratorios para identificar estrategias para disminuir la incertidumbre con respecto al promedio de LOS que era la métrica más usada. Los métodos que plantea se basan en realizar una clasificación de pacientes por enfermedad o usando variables demográficas como el sexo y la edad utilizando reglas fijas. Posteriormente a esta clasificación se comparan la diferencias entre los grupos formados y verificando si existe una disminución en su variabilidad, también se realiza una predicción por parte de fisiatras y otra realizada por parte de las enfermeras (Robinson et al., 1966). Gustafson et al (1968) es uno de los primeros autores que utiliza modelos estadísticos para realizar la predicción de LOS, en este estudio compara 5 métodos dentro de los cuales están estimación por puntos basado en el criterio de expertos, modelos de regresión lineal múltiple, un método basado en histórico de medias, un método basado en odds de radio y finalmente un modelo subjetivo bayesiano (Gustafson, 1968). En el trabajo de Burns & Wholey (1991) también se analiza el LOS buscando las variables que mayor efecto tienen sobre el LOS sin plantear modelos estadísticos para predicción pero generando resultados importantes para modelos posteriores que deberán seleccionar las variables con mayor efecto en el LOS (Burns & Wholey, 1991).

En el año 1995 comienzan a aparecer aplicaciones computacionales que se estaban desarrollando en la época y empezaban a volverse más frecuentes, estas metodologías se

irían popularizando en las aplicaciones médicas, generando un campo de investigación relacionado con la computación, estadística y aplicaciones médicas (Mobley et al., 1995). A finales de los 90 también se aplica uno de los métodos más utilizados en el campo de la predicción del LOS que es el uso de las distribuciones tipo fase y modelos de markov (Jiménez et al., 1999; Marshall et al., 2005). Después del 2000 empiezan a aparecer modelos aplicados desde un campo de investigación de minería de datos (P. Liu et al., 2006; Silva, Cortez, Santos, Gomes, & Neves, 2008; Tanuja et al., 2011). Finalmente, en estudios más recientes, las mismas técnicas de minería de datos y algunas mejoras de éstas, dejan de reportarse en la literatura de minería de datos y se empiezan a asociar al campo de machine learning (Awad et al., 2017; Azari, 2015; Turgeman et al., 2017). En la Figura 3 se presenta un gráfico que muestra cómo ha cambiado, a través de las décadas, la participación de los grupos de técnicas que se identifican en el capítulo anterior. En ésta se observa que los métodos computacionales empezaron a aparecer en las últimas tres décadas, aun así, como resultado de la popularidad que han ganado diferentes corrientes de este enfoque, se ha vuelto el más usados en la actualidad. Por otro lado, el enfoque estadístico fue el que más se usó en la década de los 70 y 80, sin embargo, fue perdiendo participación frente a las técnicas computacionales, no obstante, en la actualidad sigue estando presente en muchas investigaciones.

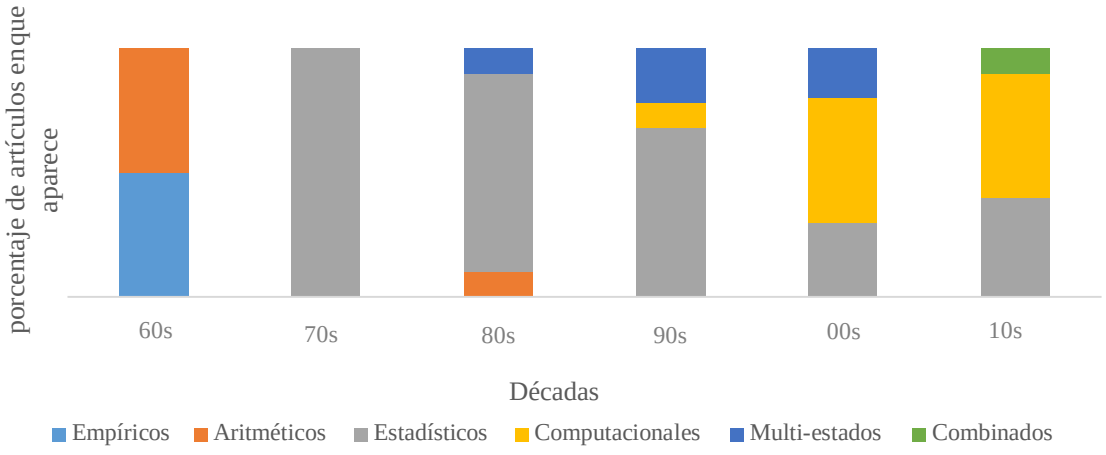


Figura 3 Evolución histórica de técnicas de predicción del LOS

Dentro de la amplia gama de técnicas encontradas es necesario resaltar la relevancia de la cubista, ya que tienen la bondad de aprovechar ventajas de diferentes técnicas. Ésta es una técnica basada en reglas que combina diferentes metodologías publicadas anteriormente y que ha ido evolucionando con el paso de los años (Kuhn & Johnson, 2013). La columna vertebral del modelo cubista se sustenta en la técnica árboles de modelos descrita por Quinlan (1992) en su algoritmo M5, donde se define un árbol de decisión y una regresión lineal en cada nodo, posteriormente éste sufre algunos cambios para dar origen a la técnica cubista. Esta técnica puede analizar gran cantidad de variables que es una ventaja de los árboles de decisión junto con la posibilidad de separar el

conjunto total de observaciones, en subgrupos con medias diferentes. Adicionalmente, puede aprovechar la precisión que permiten tener algunos modelos de regresión basados en distribuciones paramétricas.

Las características de la técnica cubista abren un campo de investigación interesante en el contexto de predicción del LOS. En primer lugar, existe evidencia de haber generado un modelo con menor MAE que otras cinco técnicas con las que se comparó. En segundo lugar, es una propuesta nueva en el contexto de predicción del LOS que brinda diferentes alternativas por explorar, por ejemplo, modificar la técnica de regresión lineal que se implementa en cada nodo del árbol, por una técnica más apropiada de acuerdo a la distribución sesgada del LOS (Marazzi et al., 1998). En tercer lugar, es un modelo que combina dos de las técnicas más usadas en predicción del LOS, árboles de clasificación y regresión lineal (Awad et al., 2017). En cuarto lugar, es una técnica que facilita la interpretación de resultados. Finalmente, esta técnica permite identificar subgrupos de la población con las características que definen estos subgrupos y que pueden ser importantes en la toma de decisiones de la administración hospitalaria (Turgeman et al., 2017).

3.3 Selección de modelos

Para un mismo problema pueden existir diferentes alternativas de modelo, es por esta razón que surge la necesidad de definir procedimientos y criterios que permita seleccionar un mejor modelo. Dichos criterios han sido sujetos de comparación sin que exista una postura unánime sobre la manera de seleccionar el modelo óptimo (Chai & Draxler, 2014; Chakrabarti & Ghosh, 2011; García Olaverri, 1996; Geweke & Meese, 1981; Guerra, Cabrera, & Fernández, 2003; Qi & Zhang, 2001). Parte de la controversia se debe a que los criterios no se han diseñado con el mismo fin, esto es un aspecto importante ya que el mejor modelo depende del uso que se le vaya a dar al modelo y el criterio de los investigadores (García Olaverri, 1996). En el trabajo de Geweke & Meese (1981) se afirma, por ejemplo, que criterios como el AIC están enfocados a minimizar el error de la predicción, y otros como el criterio de información bayesiano (BIC) están diseñados para trabajos en los que el objetivo es la estimación de parámetros.

Los criterios más usados como el AIC y el BIC tienen dos componentes, uno que mide la bondad de ajuste del modelo y otro que se utiliza como penalización por la inclusión de parámetros adicionales, este último componente es de gran importancia ya que generalmente los modelos con la inclusión de variables y por ende de parámetros, mejoran su bondad de ajuste, sin la penalización se seleccionaría siempre los modelos más amplios que generalmente redunda en un sobreajuste del mismo (Qi & Zhang, 2001). Por esta situación se crítica el uso de medidas de desempeño como criterio de selección de modelos, medidas como el error cuadrático medio (MSE) o el error absoluto medio (MAE). Sin embargo, en algunos trabajos es muy útil esta métrica ya que es fácil de interpretar y se hace en un contexto en el que se comparan modelos con una cantidad de parámetros similares, por lo cual la cantidad de parámetros no es un tema por el que haya que preocuparse (Qi & Zhang, 2001).

En la Tabla 2 se presentan las expresiones de algunos de los criterios de selección más utilizados en la literatura, Los criterios presentes son AIC propuesto por Akaike (1974), BIC propuesto por Sawa (1978), el criterio de información bayesiano de Schwarz (SBIC) propuesto por Schwarz (1978), el criterio de varianza de residuales (RVC) propuesto por Theil (1961) y también se presentan dos de las medidas sobre los residuales más usadas y que sirven, en muchas ocasiones, como criterio de selección de modelo, RSME y MAE. Dentro de las formulas de la Tabla 2 □□ □ □□, corresponden a los modelos con □ □ □' variables, T es el tamaño muestral, □□□, □□□, son los estimadores de la varianza del término del error en □□ □ □□, los cuales se definen con la expresión □□□= □□□/□.

Tabla 2 Criterios de selección de modelos

Criterio	Formula
AIC	$\square\square\,\square^{\square} + \frac{2\square}{\square}$
BIC	$\ln(\square^{\square}) + 2\square \; \frac{\square + 2}{\square - \square^{*}} \; \square^{\square} \; \square\square\square - \frac{2\square}{(\square - \square^{*})^{\square}} \; \square^{\square} \; \square\square$
SBIC	$\ln(\square^{\square}) + \square \cdot \frac{\square\square\square}{\square}$
RVC	$\square^{\square} \cdot \frac{\square}{\square - \square}$
MAE	$\square[\square\square - \square]$
RMSE	$\square\sqrt{\square^{\square}}$

Dentro de los criterios de selección mencionados, el AIC y BIC son los más populares porque han mostrado mejor capacidad para seleccionar el modelo adecuado (Qi & Zhang, 2001; Zhu, Li, & Liang, 2009). El indicador BIC inicialmente se creó para comparar dos modelos, aunque en la práctica se ha utilizada para comparar múltiples, debido a esta falencia se creó el criterio SBIC que sirve para comparar varios modelos y adicionalmente para aumentar la penalización causada por la inclusión de parámetros adicionales. Estos modelos tienden a seleccionar el modelo correcto a medida que el tamaño de la muestra aumenta (Zhu et al., 2009). En cuanto a las métricas MAE y RMSE, Willmott & Matsuura (2005) muestran evidencia de que el MAE puede tener mejor capacidad de identificar el modelos apropiado que el RMSE ya que este último es muy sensible a la presencia de valores extremos, por otro lado, el trabajo de Chai & Draxler (2014) contradice esta evidencia mostrando casos en los que el RMSE obtuvo mejores resultados en la identificación del mejor modelo.

Como se afirmó inicialmente, no existe evidencia contundente que pruebe que un solo criterio seleccione el mejor modelo para todos los casos, por el contrario, algunos autores proponen analizar varias métricas para tener varias perspectivas para la selección del modelo. Adicionalmente se deben tener claros los objetivos y contexto del problema ya que en el caso de modelos que pueden crecer mucho en el número de parámetros, es importante utilizar un criterio con penalización, mientras que para comparar modelos con un número de parámetros similar puede ser de gran utilidad analizar métricas como el MAE y RMSE. Dado lo anterior en el presente trabajo se considerarán varias métricas en la comparación de modelos, teniendo en cuenta algunos criterios con penalización, pero también utilizando métricas que permitan evaluar el modelo de forma más simple.

En las técnicas encontradas en la literatura existen muchos aprendizajes y algunas oportunidades. Dentro de las oportunidades se resalta la exploración de una gran variedad de técnicas, la combinación de las mismas para utilizar ventajas que cada una tiene y la mejora en cuanto a disminución de residuales del modelo generado por la técnica cubista, que se comparó con las técnicas más competitivas de la literatura (Turgeman et al., 2017). Por otro lado se identifican oportunidades en la técnica cubista en cuanto a poder incluir en cada nodo final modelos diferentes a la regresión lineal múltiple, también se identifican

oportunidades en cuanto a medir y comparar los modelos con criterios como el AIC y BIC que incluyen bondad de ajuste y penalización por complejidad, ya que muchos estudios únicamente utilizan métricas enfocadas en los residuales (Al Taleb et al., 2017; Barnes et al., 2016; Turgeman et al., 2017). También existe una oportunidad en el desarrollo de metodologías, que incluyan procedimientos que las técnicas no incluyen y que faciliten el proceso de selección de modelo final. Finalmente se identifica la oportunidad de que estas metodologías para desarrollo de modelos hagan parte de componentes de software que faciliten la generación de modelos de predicción de LOS.

4 Metodología: árboles de modelos generalizados aditivos de localización, escala y forma

En la revisión de la literatura se identificaron las características del LOS, las técnicas de predicción desde sus diferentes campos de estudio, la evolución de estas técnicas, algunos criterios para seleccionar modelos, y finalmente en la revisión de la literatura se identificaron algunos aprendizajes y oportunidades. Con base en estos aprendizajes y oportunidades se desarrolló la metodología Árboles de Modelos GAMLSS (AMG). Con esta metodología se aprovechan los aprendizajes y ventajas encontrados con la técnica cubista, al mezclar una técnica de árboles de decisión con una técnica de regresión en los nodos, se aprovechan las oportunidades encontradas al incluir otro tipo de regresiones con la utilización de la técnica de modelos aditivos generalizados de localización escala y forma (GAMLSS), en lugar de la regresión lineal múltiple. También se define una metodología que incluye procedimientos que generalmente no incluyen las técnicas como pruebas de hipótesis de distribuciones, selección de distribución, selección de variables y selección de modelo final. En la selección del modelo final se incluyen criterios basados en la bondad de ajuste con penalización por complejidad como AIC y BIC, pero también se dejan criterios basados únicamente en el comportamiento de los residuales como MAE y RMSE. Finalmente se implementa la metodología en un software estadístico con diferentes argumentos que permiten al usuario desarrollar un modelo utilizando los criterios que escoja y permitiéndole la posibilidad de incluir o no ciertos pasos de la metodología propuesta, con esta metodología se logran cubrir los principales retos encontrados en la literatura.

En la metodología AMG se aprovecha las ventajas que brinda la técnica de árboles de decisión, como la capacidad de analizar eficientemente una gran cantidad de variables e identificar aquellas que son de mayor importancia, facilitando de esta manera el procedimiento de selección de variables. También se incluyó la técnica GAMLSS en cada nodo final del árbol de decisión. Esta técnica tienen mayor flexibilidad, no solo con respecto a las regresiones lineales, sino también mayor flexibilidad con respecto a los modelos lineales generalizados y los modelos aditivos generalizados. Esta flexibilidad se refleja en el tipo de distribuciones de variables respuestas que puede modelar, ya que cuenta con una amplia gama de distribuciones. La flexibilidad también está en la posibilidad de modelar otros parámetros diferentes a la media de la distribución y también en la posibilidad de modelar componentes no lineales.

Otra consideración importante, en el desarrollo de la metodología, es que en estas técnicas de regresión se usan todas las variables ingresadas para el modelo que se ajusta, luego se debe hacer una selección de las variables más relevantes para disminuir la complejidad del modelo y evitar el sobreajuste. Para esto se incluyó, posteriormente al ajuste del modelo GAMLSS, un componente de selección de variables usando el procedimiento stepwise ó paso a paso, que es el más común en la literatura. El procedimiento stepwise permite evaluar el modelo agregando o quitando una variable a la vez para seleccionar el modelo más balanceado entre complejidad y bondad de ajuste. Los dos últimos componentes mencionados (stepwise y ajuste de modelo GAMLSS) tienen un costo computacional alto que incrementa proporcionalmente al número de variables explicativas, es por esto que se adicionó un componente previo en el que se realiza una preselección de las variables que se utilizarán para ajustar el modelo GAMLSS. Este procedimiento se realiza con base en la información brindada por el árbol de decisión que genera un ranking de las variables más importantes para la predicción.

Antes de ajustar el modelo GAMLSS es necesario definir las distribuciones en la que se basará. Para esto se incluyeron dos componentes, uno en el que se evalúa la bondad de ajuste de las distribuciones definidas por el usuario y otro en el que se realiza una preselección de las distribuciones con mayor potencial. La bondad de ajuste sirve, además de permitir generar la información para la preselección de distribuciones, para entender y conocer el comportamiento de la variable respuesta, adicionalmente permite al usuario tener mayor información estadística para decidir sobre el ajuste del modelo, una falencia que tienen muchas de las técnicas encontradas en la literatura. Para aprovechar las ventajas que tienen las dos técnicas utilizadas en los componentes mencionados (Árboles de decisión, GAMLSS), se utiliza una idea extraída de los modelos cubistas, en los que se ajusta una regresión lineal (en este caso un modelo GAMLSS) en cada nodo de un árbol de decisión, en lugar de utilizar, como predicción, la media de las observaciones que pertenecen a un nodo, como es usual en los modelos de árboles de decisión. Finalmente, se incluye un componente de iteraciones boosting extraído también de los modelos cubistas, conocido como committees, este componente se utiliza para mejorar la predicción de los modelos ajustados y también para disminuir el sobreajuste.

4.1 Descripción de la metodología árboles de modelos generalizados aditivos de localización, escala y forma

En esta sección se presenta la metodología desarrollada en el trabajo, diseñada inicialmente para construcción de modelos de predicción de LOS, pero que puede ser empleada para predicción de diversas variables. La metodología se basa en el algoritmo M5 propuesto por Quinlan (1992) que posteriormente es mejorado en el marco de la creación de un software comercial por parte del mismo autor. La técnica fue nombrada cubista y su descripción en textos académicos aparece recientemente en el trabajo de Kuhn & Johnson (2013), luego de la publicación del código del algoritmo con una Licencia Pública General GPL.

Como se mencionó en la sección 3.2 las técnicas computacionales son las más comunes en la literatura reciente, estas técnicas generalmente no tienen supuestos de distribuciones de la variable respuesta como las técnicas estadísticas, esto hace que algunas técnicas desde el enfoque computacional se puedan aplicar y desde el estadístico no se aplican o se deban hacer ajustes como transformaciones de potencia. La técnica cubista proviene de un enfoque computacional, pero incorpora la regresión lineal que es una técnica estadística. En dicha regresión no se analiza ni se brinda información sobre los supuestos. En la metodología AMG se agregan características como las pruebas de hipótesis de bondad de ajuste, los criterios AIC y BIC y un amplio conjunto de distribuciones paramétricas que permiten a los investigadores tener más información para analizar y seleccionar el modelo final sin perder la flexibilidad que tiene la técnica cubista. Adicionalmente se exploran algunas variaciones en los procedimientos utilizados por dicha técnica para analizar su impacto sobre el desempeño del modelo desde el punto de vista del análisis de los residuales y de la bondad de ajuste del modelo. Son precisamente estas características adicionales las que diferencia la metodología AMG de otras como la técnica cubista o algoritmo M5, ya que estos últimos solo ajustan una distribución normal y en muchas ocasiones las variables respuestas, como es el caso del LOS, no tienen las características de dicha distribución, incluso con transformaciones.

En la Figura 4 se presenta un diagrama de flujo de la metodología AMG. En este diagrama se definen las actividades que se realizan en la metodología, las entradas manuales que requiere cada actividad, que en la implementación de software son argumentos definidos por el usuario. Se define adicionalmente el resultado de cada actividad y qué otras actividades utilizan cada uno de estos resultados. También se presentan los puntos de decisión del algoritmo que definen los bucles que debe realizar el algoritmo repitiendo una serie de actividades para diferentes nodos o committees. Luego de la presentación del diagrama de flujo, se describen en detalle cada uno de las actividades que la conforman. Estas actividades y las funciones que se utilizan para desarrollarlas están inmersas en el paquete `gmtree` que se explica en la sección 4.2. Esta función permite ejecutar automáticamente cada uno de los procedimientos diseñados en la metodología desarrollada.

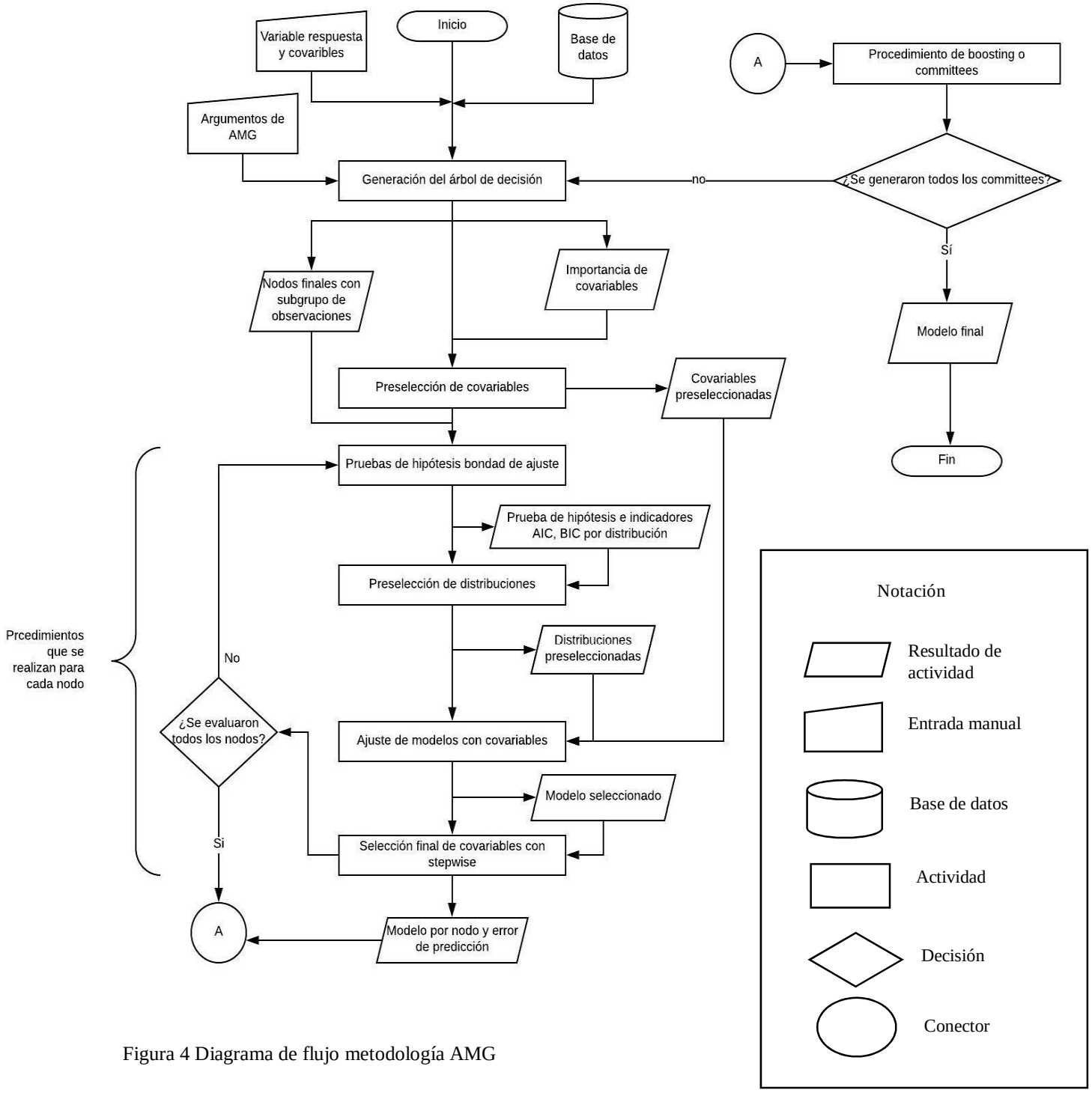


Figura 4 Diagrama de flujo metodología AMG

4.1.1 Generación de árbol de decisión

Para el componente de árboles de decisión existen varias alternativas que se podrían utilizar. En el trabajo se consideraron tres algoritmos que se encontraron en la literatura y tenían una implementación en el lenguaje R, árboles tipo CART (Classification and Regression Trees, Leo Breiman, 1996) utilizando la implementación de este algoritmo en el paquete `rpart` (T. Therneau & Atkinson, 2018). Árboles basados en el método CHAID utilizando la implementación de la función `ctree` (conditional trees) del paquete `partykit` (Hothorn, Hornik, & Zeileis, 2006; Hothorn & Zeileis, 2015). Finalmente, árboles basados en el algoritmo `M5` implementados en el paquete `RWeka` (Hornik, Buchta, & Zeileis, 2009; Witten & Frank, 2005). De estos tres algoritmos, el árbol tipo CART es el más común en árboles de regresión de la literatura de predicción del LOS, también es uno de los más populares en todos los campos debido a que fue uno de los precursores de esta técnica, otra característica que la hace interesante es que es una técnica diferente a la utilizada en la construcción de modelos cubistas, por lo que se estaría explorando una variante de dicha técnica.

Dado lo anterior, en la metodología AMG se utilizó un árbol tipo CART. Para construir este árbol se usa, dentro de la función `gamlss_tree`, la rutina Recursive Partitioning del paquete `rpart` (T. Therneau & Atkinson, 2018), implementada en el software R (R Core Team, 2018). Esta rutina incorpora las principales ideas del libro y software Classification and Regression Trees-CART (L Breiman, 1984). La construcción del árbol consiste en encontrar la variable que mejor divida la población en dos grupos, los datos son separados con dicha variable y luego este mismo proceso es aplicado a cada nuevo grupo, y así de manera recursiva hasta que se alcance un mínimo de cinco observaciones en un subgrupo o hasta que no se encuentre posibilidad de mejora. El modelo final puede ser representado como un árbol binario (T. M. Therneau & Atkinson, 2018). Para identificar la variable que mejor divida la población se evalúa, en el caso de árboles de regresión la expresión $\Delta_{\square\square} = \square\square_{\square} - (\square\square_{\square} * \square_{\square} + \square\square_{\square} * \square_{\square})$ con $\square_{\square} + \square_{\square} = 1$, donde $\square\square_{\square}$ representa la suma de cuadrado del nodo T, $\square\square_{\square}$ y $\square\square_{\square}$ representan la suma de cuadrado de los nodos hijos derecho e izquierdo respectivamente y $\square_{\square}$ $\square$ $\square_{\square}$ representan la proporción de observaciones en los nodos hijos derecho e izquierdo respectivamente. Esto es equivalente a maximizar el error entre grupos de un análisis de varianza. Después de construir el árbol de decisión la población inicial queda dividida en nodos finales, también conocidos como hojas, nodos que permiten obtener información interesante sobre el problema como algunas características que identifican observaciones con mayor dispersión, observaciones con valores más altos o más bajos en la variable respuesta.

4.1.2 Preselección de covariables

La selección de las variables explicativas o covariables es un tema de suma importancia en la literatura de modelos de predicción, en primer lugar porque son éstas las que determinan la capacidad de un modelo de explicar la variabilidad de la variable respuesta, en segundo lugar, porque definen la complejidad que puede alcanzar un modelo y junto con esta complejidad viene de la mano el tiempo de computo que requiere el ajuste de dicho modelo, finalmente porque el uso excesivo de variables, que no influyen en la variable respuesta, pueden generar modelos con relaciones espurias que redundan en un

sobreajuste del mismo. Por esta razón este tema se considera en muchas investigaciones y se trabaja de manera diferente según la técnica utilizada para modelar. Algunas técnicas como los árboles de decisión tienen procedimientos incorporados que permiten utilizar en los modelos finales, solo las variables relevantes y no todas las consideradas inicialmente. Por otro lado, en técnicas como GAMLSS, no existe un procedimiento incorporado, por lo cual se utilizan procedimientos adicionales como stepwise.

En el presente trabajo se realiza una preselección de las covariables que serán consideradas para ajustar los modelos en cada nodo final. Se utiliza el procedimiento que tienen incorporado los árboles de decisión generados en el paso anterior. Este paso se considera una preselección ya que la selección final de las covariables se define en la sección 4.1.6. El procedimiento que utilizan los árboles de decisión de la función `rpart` genera una importancia para cada variable. Esta importancia se calcula sumando las mejoras que aporta cada variable en cada una de las divisiones en las que es utilizada, dicha mejora se mide con base en la disminución del error de predicción (suma de cuadrados). El valor de importancia generado en este procedimiento es re-escalado para que la suma sea 100, posteriormente se eliminan de la lista de importancia de variables, las que tienen un valor menor a uno (Therneau & Atkinson, 2018). Las variables con importancia mayor a uno, según el árbol de decisión, son las preseleccionadas para ajustar los modelos en cada nodo final.

4.1.3 Pruebas de hipótesis bondad de ajuste

En AMG se pueden utilizar todas las distribuciones de probabilidad definidas en el paquete `gamlss` (qué se describirá con más detalle en la sección 3.1.5), dentro de las cuales se encuentran distribuciones para variables continuas, discretas o variables mixtas. Estas distribuciones pueden tener desde uno hasta cuatro parámetros y tienen la ventaja de poder modelar incluso variables que no siguen una distribución de la familia exponencial, como variables leptocúrticas, variables platicúrticas, variables con sesgo positivo o negativo y variables que representan conteos con alta dispersión, de hecho, las distribuciones de este paquete fueron pensadas como una flexibilización a los modelos lineales generalizados (Stasinopoulos & Rigby, 2018). Estas características, como se mencionó en la sección 2.1, son adecuadas para la distribución del LOS. En AMG para cada una de las distribuciones que el usuario quiere analizar, se realiza la siguiente prueba de hipótesis:

$$\square\square: \square \sim \square\,(\square, \square, \square, \square)$$

$$\square\square: \square \not\sim \square(\square, \square, \square, \square)$$

donde $\square$ pertenece a una de las distribuciones del paquete `gamlss` que como se mencionó pueden tener hasta 4 parámetros $\square, \square, \square, \square$. Para llevar a cabo estas hipótesis se utilizan dos pruebas estadísticas, Kolmogorov-Smirnov con la función `ks.test` y Anderson Darling con la función `ad.test`. Para evaluar las pruebas de hipótesis se utiliza una significancia definida por el usuario. Este paso también puede usarse como

restricción, definiendo en los argumentos de la función diseñada `gamlss_tree`, que solo se tengan en cuenta las distribuciones para las cuales la prueba de hipótesis es aceptada. De esta manera los modelos que se ajusten en cada nodo final solo tendrán en cuenta distribuciones que se ajustan a los datos analizados.

4.1.4 Preselección de distribuciones

Para seleccionar las distribuciones candidatas se utiliza la función `fitDist` (Rigby & Stasinopoulos, 2005). Con base en una variable y un conjunto de distribuciones definidas por el usuario, la función realiza, para cada distribución, la estimación de los parámetros sin utilizar covariables. Con estos ajustes se calculan indicadores para la selección de modelos como el criterio de información de akaike generalizado (GAIC) y el criterio de información bayesiano (BIC). En el caso de AMG se utiliza el GAIC para ordenar las distribuciones de menor a mayor, es decir, la de mejor ajuste de primera. De estas distribuciones se seleccionan las n distribuciones con mejor GAIC, donde n es definido por el usuario. Este paso ayuda a realizar una preselección de distribuciones con base en una idea del potencial que tiene la distribución para ajustarse a los datos analizados, esta preselección limita el número de modelos GAMLSS que se deben ajustar con covariables haciendo la ejecución de la función más eficiente.

4.1.5 Ajuste de modelos con covariables

En este paso se realiza, para cada una de las distribuciones preseleccionadas en los dos pasos anteriores, el ajuste de un modelo GAMLSS (Rigby & Stasinopoulos, 2005). Este tipo de modelos fue introducido con el fin de superar las limitaciones asociadas a los modelos lineales generalizados y a los modelos aditivos generalizados. En la ecuación (1) se presenta el modelo utilizado en AMG. Aunque los modelos GAMLSS permiten modelar diferentes parámetros de una distribución como μ, σ, ν y τ, en el desarrollo de este algoritmo se acotó al modelamiento de μ.

$$g_k(\theta) = X_k \beta_k + \sum_{j=1}^{J_k} Z_{jk} \gamma_{jk} \qquad (1)$$

Donde g_k es la función de enlace del parámetro θ, X_k es una matriz de n observaciones por J_k' variables explicativas, β_k es el vector de parámetros de longitud J_k'. Z_{jk} es una matriz $q \times q_{jk}$. Finalmente γ_{jk} es un vector de q_{jk} variables aleatorias, estas dos últimas componentes son el efecto aleatorio que puede estar conformado, en la forma general, por múltiples efectos aleatorios.

Estos modelos se ajustan para cada una de las distribuciones preseleccionadas y con el ajuste de estos modelos se obtiene el GAIC, el error absoluto medio MAE, el error cuadrático medio MSE y la raíz del error cuadrático medio RMSE. De acuerdo con la necesidad del usuario se definen los indicadores con los que se selecciona el modelo final.

4.1.6 Selección final de covariables con procedimiento paso a paso.

En el paso anterior se tiene un modelo con todas las covariables que se preseleccionaron en el paso definido en la sección 4.1.2, que garantiza que las covariables utilizadas son importantes para todo el conjunto de datos, sin embargo, no necesariamente son importantes para las observaciones que pertenecen a un nodo, es por esto que se adiciona este componente, que permite seleccionar, con base en las observaciones del nodo final, las variables que mayor importancia tienen. Se realiza un procedimiento paso a paso o stepwise utilizando la función **stepGAIC** con pasos hacia adelante y hacia atrás que permite obtener un modelo con parsimonia sin sacrificar de manera significativa el desempeño del mismo (Rigby & Stasinopoulos, 2005). Con este procedimiento se descartan algunas variables que no son relevantes para las observaciones del nodo que se está evaluando.

4.1.7 Procedimiento de boosting o committees

Este procedimiento es opcional, ya que el usuario define el número de iteraciones que se realizan, en caso de que el número de iteraciones sea una, este paso se omite. Cada iteración representa un nuevo árbol generado, pero antes de iniciar el nuevo árbol se modifica el valor de la variable respuesta de acuerdo con la expresión □□ = □ − □□□□ − □□, donde □□ representa el valor ajustado de la variable respuesta que se utilizará en la generación del árbol de decisión número □, □ es el valor real de la variable,
□□□□ es el valor predicho en el árbol □ − 1. Esta expresión tiene el objetivo de penalizar a las observaciones que son sobreestimadas o subestimadas, de tal manera que el próximo modelo genere una predicción menor o mayor respectivamente. Finalmente, cuando todos los árboles son generados la predicción final es el promedio simple del valor predicho de cada uno de ellos (Kuhn & Johnson, 2013). Este procedimiento es conocido como committees, nombre que se le asigna en la técnica cubista de donde es tomado.

4.2 Implementación de la metodología AMG en R

Paquete gmtree

Para la metodología descrita en la sección 4.1 se desarrolló un paquete en el lenguaje de programación R que comprende un grupo de funciones que ejecuta los pasos definidos sobre un conjunto de datos. Este paquete puede consultarse en el siguiente repositorio https://github.com/juancamiloespana/gmtree. El objetivo de este paquete es generar múltiples modelos AMG y seleccionar el de mejor desempeño de acuerdo al indicador definido por el usuario (AIC, BIC, MAE, RMSE). La función `gamlss_tree` ajusta modelos AMG compuestos de un árbol de decisión tipo CART y un modelo de regresión GAMLSS en cada hoja o nodo final creados por el árbol de decisión. La función `pred_gamlss` se utiliza para realizar predicciones de observaciones nuevas con base en el modelo ajustado en la función `gamlss_tree.` La función `test_dist` se utiliza para analizar el ajuste de las distribuciones del paquete `gamlss` que defina el usuario, utilizando dos pruebas de bondad de ajuste Anderson Darling y Smirnov Kolmogorov. La función `split_sample` se utiliza para dividir un conjunto de datos en dos subconjuntos, un conjunto para entrenamiento de modelos y otro para evaluación. El usuario define el porcentaje de datos que va a tener el conjunto de entrenamiento, y el porcentaje de datos restantes corresponderá al grupo de evaluación. La descripción detallada de la función se presenta en el Anexo 1.

5 Análisis y discusión de casos de estudio para validación

En esta sección se analizan tres casos de estudio de predicción del LOS que se utilizaron para validar la metodología AMG y para compararla con otras técnicas de la literatura. En la primera parte se describen los casos de estudio, su procedencia y algunas características de los conjuntos de datos. También se realiza una comparación de los tres casos de estudio y se describen las razones por las que fueron seleccionados, basado en sus diferencias y cómo éstas pueden impactar el desempeño de la metodología desarrollada. En la segunda parte se presenta una breve descripción y comparación de la distribución de la variable respuesta en cada uno los casos de estudio, analizando diferentes medidas de resumen y un análisis gráfico. En la tercera parte se presentan las funciones y los argumentos utilizados para generar el modelo de la metodología AMG y los modelos de las otras técnicas con las que se compararon los resultados. Por último, se presenta una comparación del desempeño del modelo seleccionado por la metodología AMG con respecto a los modelos generados por las demás técnicas. Esta comparación se realiza utilizando tres métricas diferentes, AIC, MAE, RMSE para tener una visión más amplia de las fortalezas y debilidades de cada una de las técnicas.

5.1 Descripción de los casos de estudio para validación

Los tres casos de estudio que se tuvieron en cuenta fueron seleccionados debido a que tenían características diferentes que permitían analizar el comportamiento de la metodología desarrollada bajo condiciones diversas. En primer lugar, los datos provienen de países diferentes. En segundo lugar, la cantidad de observaciones en cada caso varía de manera importante, lo que puede impactar en el ajuste de los modelos, más aún, teniendo en cuenta que la población se divide con el componente de árboles de decisión y la cantidad de observaciones utilizadas para ajustar cada modelo se reduce. En tercer lugar, la naturaleza y cantidad de variables explicativas difiere para cada caso.

El primer caso de estudio, que llamaremos Medellín, proviene de los registros médicos de un hospital de tercer nivel de Medellín, Colombia. El tercer nivel de atención es el que resuelve los problemas que exigen mayor grado de especialización, de mayor complejidad y que requieren uso de alta tecnología (Vignolio, Vacarezza, Álvarez, & Sosa, 2011). La información corresponde a todos los pacientes hospitalizados que tuvieron egresos en el año 2015. La base de datos cuenta con 27.872 registros y 16 variables predictoras y 1 variable respuesta que es el LOS en días medida como variable continua, teniendo en cuenta las horas de ingreso y egreso. De las 16 variables predictoras 11 son categóricas y 5 son discretas. En el Anexo 2 se presenta el listado de variables, la descripción y tipo de variable.

El segundo caso de estudio, que llamaremos Microsoft, proviene de un proyecto de Microsoft para predicción LOS en el que se utiliza una base de datos para evaluar varios modelos (Microsoft, 2018). Los datos fueron dados por un tercero y se les realizó un remuestreo de las variables rcount y lengthofstay que son el número de readmisiones en los últimos 180 días y el LOS respectivamente. También se realizó un remuestreo de las variables continuas para hacer que sus distribuciones fueran normales y se realizaron

correcciones de problemas que se tenía con las fechas de altas y número de documentos. En el Anexo 3 se presenta el listado de variables, la descripción y el tipo de variable.

El tercer caso de estudio, que llamaremos Arizona, proviene de los registros de hospitalización del seguro médico nacional de Estados Unidos que se registran en una base de datos conocida como Medpar. Los datos que se utilizaron son una muestra aleatoria de pacientes hospitalizados en 1991 en el estado de Arizona con problemas cardiovasculares y que recibieron uno de dos procedimientos, CABG (Coronary Artery Bypass Graft) y PTCA (Percutaneous Transluminal Coronary Angioplasty) que se utilizan para ciertos problemas cardiovasculares. La base de datos cuenta con 3589 registros, 5 variables predictoras y una variable respuesta correspondiente al tiempo de estancia (Hilbe, 2016). En el Anexo 4 se presentan las variables, la descripción y el tipo de variable.

Un aspecto importante es que las variables que están disponibles en cada caso de estudio son muy diferentes, solo coinciden variables demográficas como la edad y el sexo. Esto refleja la realidad de la predicción del LOS, en donde cada hospital lleva registro de diferentes tipos de variables. En la sección 3.1 se mencionan tres variables importantes en este tipo de estudios, el diagnóstico, la severidad de la enfermedad y la comorbilidad. En los tres casos estudiados se tiene el diagnóstico, severidad no se tiene en ninguno y la comorbilidad solo se tiene en el caso de Medellín. Esto evidencia uno de los principales problemas en de predicción del LOS, la disponibilidad de las variables, que según la literatura, tienen relación fuerte con el LOS. Esta es una oportunidad que tienen los hospitales y que probablemente es de las que más puede impactar el desempeño de los modelos de predicción del LOS.

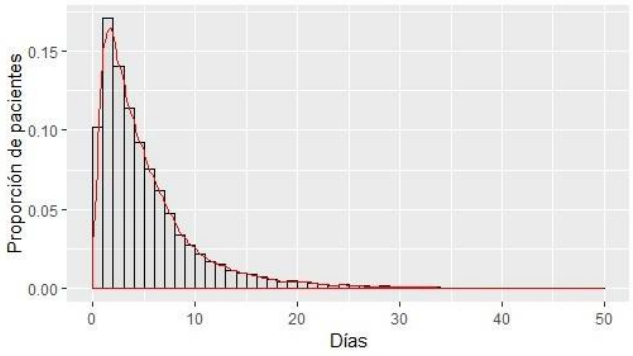
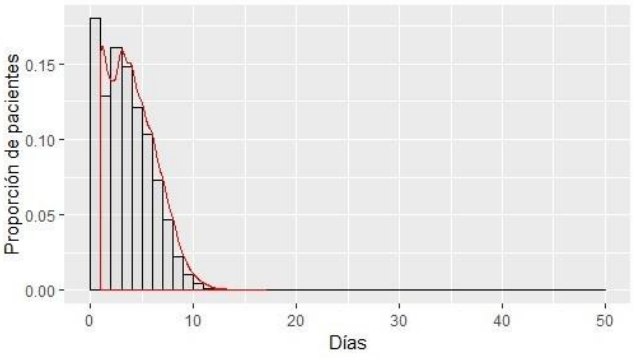
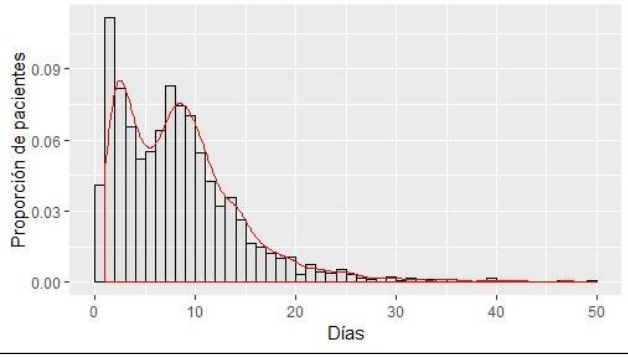
5.2 Análisis de distribución del LOS en los casos de estudio

En la Figura 5 se presenta, para cada uno de los casos de estudio, el histograma, la densidad y algunas medidas resumen del LOS. En los tres casos se observa, tanto en el histograma como en las medidas resumen (sesgo y curtósis), que el LOS es una variable leptocúrtica con sesgo positivo. Se observa también que el coeficiente de variación y rango son altos, con una concentración de las observaciones en valores cercanos a cero. La media se encuentra entre cuatro y nueve días y la mediana está entre cuatro y ocho días, en los tres casos la media fue mayor o igual a la mediana debido al sesgo de la distribución. Estas medidas coinciden con lo reportado en la literatura sobre las características de esta variable (Awad et al., 2017).

A pesar de las similitudes mencionadas, existen algunas diferencias en la distribución del LOS de cada uno de los casos de estudio. En primer lugar, en los datos del hospital de Medellín es donde se presenta la mayor variabilidad, como se evidencia en su rango que supera los 100 días y un coeficiente de variación cercano al 100%, seguido por los datos del caso de Arizona y en tercer lugar los datos de Microsoft, que tienen un rango equivalente al 15% del rango de los datos del hospital de Medellín y un coeficiente de variación cercano a la mitad. En segundo lugar, se observan diferencias importantes en las medidas de tendencia central, principalmente en los datos de Arizona que tienen una media y mediana del doble de los datos de Microsoft y tiene una media y desviación

mayor qué los datos de Medellín, a pesar de que este último tiene un rango aproximadamente un 30% mayor, lo que coincide con la evidencia gráfica que muestra que la cola de la distribución de los datos de Arizona es más pesada. Finalmente se resalta que los datos del LOS de Arizona tienen una distribución bimodal que no se observa en los otros casos, lo que puede sugerir que existen dos poblaciones dentro de las observaciones.

Figura 5 Histogramas y medidas resumen de LOS para casos de estudio

Caso	Histogramas	Medidas resumen
Medellín		Media: 6.01 días Coef. de variación: 98% Mediana: 4.00 días Rango: 110 días Curtósis: 28.14 Sesgo: 3.60
Microsoft		Media:4.00 días Coef. de variación: 60% Mediana: 4.00 días Rango: 16 días Curtósis: 2.97 Sesgo: 0.63
Arizona		Media: 8.83 días Coef. de variación: 78% Mediana: 8.00 días Rango: 82 días Curtósis: 16.17 Sesgo: 2.53

5.3 Ejecución de técnicas

Se comparó el desempeño del modelo seleccionado por la metodología AMG, utilizando la función `gamlss_tree`, con los resultados obtenidos con un árbol tipo CART

utilizando la función `rpart`, un modelo GAMLSS utilizando la función `gamlss_tree` con el argumento `arbol_activo` = F, que como se menciona en el Anexo 1 genera un modelo GAMLSS. Una Regresión lineal utilizando la función `lm` y un modelo cubista utilizando la función `cubist`. Para el AMG y GAMLSS se evalúan dos instancias, la primera es seleccionando el modelo de acuerdo con al criterio AIC, que se enfoca en seleccionar el modelo que maximice la función de verosimilitud y penalizando con el número de parámetros utilizados. La segunda instancia es ajustar el modelo que menor MAE genere. Con esto se obtuvo el resultado de 7 modelos diferentes. Para la medición de los indicadores definidos se divide el conjunto de datos, por medio de un muestreo aleatorio simple, en una muestra del 80% de las observaciones como conjunto de entrenamiento, con este se realiza la estimación de parámetros, y un 20% de las observaciones será apartada para ser utilizada como muestra de evaluación para medir los indicadores sobre ésta, estos porcentajes son los más comunes en la literatura (Pendharkar & Khurana, 2014; Tanuja et al., 2011; Turgeman et al., 2017).

Para usar las funciones mencionadas se deben definir unos argumentos que guían la ejecución de las mismas. A continuación se describen dichos argumentos y los valores definidos para la ejecución. La función `rpart` cuenta con los argumentos: `method`, este argumento define el método utilizado para seleccionar la mejor división posible de la población en cada nodo, para el caso de regresión el método utilizado es "ANOVA", esta selección la hace automáticamente la función escogiendo el argumento más adecuado según las características de la variable respuesta (T. M. Therneau & Atkinson, 2018).

El argumento `minbucket`, que corresponde al número mínimo de observaciones que puede tener un nodo final, se definirá de acuerdo al trabajo de Green (2010), en el que se propone que una regresión debe tener como mínimo un numero de observaciones mayor o igual a 104 + □, donde m es el número de predictores. Esta regla se utiliza para garantizar que las regresiones ajustadas en cada nodo tengan un número de observaciones adecuado para la estimación de parámetros. El argumento `minsplits`, que corresponde al número mínimo de observaciones que debe tener un nodo para intentar una división se define en relación al argumento anterior de la siguiente manera: `minbucket`*3 ya que existe una dependencia entre ambos argumentos. El argumento `maxdpeth`, es la profundidad máxima del árbol de decisión, su valor por defecto es 30, valor que se dejará ya que este argumento sirve como criterio de podado, el cual será controlado por los dos argumentos anteriores para asegurar el ajuste de los modelos en cada nodo. Finalmente, se tiene el argumento `cp`, que corresponde al parámetro de complejidad utilizado como criterio de parada. Para el caso del método anova, este argumento indica el valor mínimo que debe mejorar el □□ para aceptar una división, el valor que se utilizará en la ejecución es: 0.01 que es el valor por defecto, se deja el valor por defecto ya que como se mencionó anteriormente el podado será controlado por el número mínimo de observaciones en un nodo.

En la función `lm` se utiliza la estimación de mínimos cuadrados para generar los valores de los parámetros que forman el modelo y el único argumento de entrada son las covariables que se utilizarán. La definición de las variables para todas las técnicas se realizará de la misma manera, para asegurar igualdad de condiciones y se realiza con base en el procedimiento explicado en la sección 4.1.2 de la metodología AMG.

Los argumentos utilizados en la función cubist son los committeess que corresponden al número de iteraciones boosting que se realizan del algoritmo y se explican con más detalle en la sección 4.1.7, el valor que se utilizará es 1 para disminuir el tiempo computacional. El otro argumento de la función rules que corresponde al número máximo de reglas que pueden definirse para el modelo final, para calcular este valor se tomará el número mínimo de observaciones definido anteriormente 104 + □, y se dividirá el total de observaciones por este valor, esta cifra proporciona un número de reglas, que es equivalente al número de divisiones que tendrá la población, permitirá asegurar que el método cubista no tenga un número de observaciones en cada nodo alejado del valor que se definió para los árboles de decisión de la función rpart.

En la función gamlss incorporada dentro de la metodología AMG, que se utiliza para los modelos GAMLSS y AMG, se definen los siguientes argumentos: method que corresponde al algoritmo con el que se realiza la búsqueda de los coeficientes de los modelos GAMLSS, se utilizará el método RS()siglas de Rigby and Stasinopoulos creadores del algoritmo, se selecciona por ser mejor en cuanto a tiempo computacional. El argumento cyc, que es el número de ciclos que realiza el algoritmo RS(), se definirá en 50 para asegurar que no se aumente significativamente el tiempo computacional. El argumento family, que corresponde a la distribución que se ajusta, se tuvieron en cuenta 44 distribuciones que se presentan en el Anexo 5, las cuales tienen en cuenta todas las distribuciones reales positivas y las distribuciones de conteos.

En la función gamlss_tree se utilizan los argumentos de la función gamlss y adicionalmente: Steps, que corresponde a los pasos que se utilizan para realizar una selección de variables por medio de la metodología stepwise utilizando el indicador GAIC, se realizarán solo 2 pasos ya que su costo computacional es muy alto y adicionalmente porque ya existe una preselección de variables realizadas previamente por la función gamlss_tree. El argumento n_dist_mod, que corresponde al número de distribuciones para las que se ajusta el modelo con covariables se define en 44 para analizar todas las distribuciones ingresadas. El argumento var_sel, que corresponde al criterio que se usa para seleccionar el modelo final, como se corren dos instancias se definen dos argumentos "aicmodelo" y "MAE". committeess, que es el número de iteraciones boosting que se realiza del algoritmo, se utiliza el mismo que en el modelo cubista tomando el valor de 1. prueba_hip, que se utiliza para definir si se realizan pruebas de hipótesis para validar el ajuste de las distribuciones al conjunto de datos analizado, su valor por defecto es "TRUE" que indica que si se realizan las pruebas. signif, que es la significancia con la que se evalúan las pruebas de hipótesis de las distribuciones de los datos, será 0.005 que es el valor más común en la literatura. Finalmente está el argumento acepta_h, que se utiliza para decidir si solo se toman las distribuciones para las que se aprueba la hipótesis de bondad de ajuste. "TRUE" indica que solo se toman las distribuciones para las que se aprueba la hipótesis en otro caso se tienen en cuenta todas las distribuciones para realizar los ajustes de modelos con covariables, de definirá su valor en "FALSE" para analizar el indicador MAE en todas las distribuciones ajustadas.

5.4 Comparación de métricas

Como se mencionó en la sección anterior, para cada uno de los casos de estudio se utilizaron 5 técnicas diferentes (AMG, CUBISTA, CART, LM, GAMLSS), sin embargo dos de estas técnicas (AMG, GAMLSS) se evalúan con dos variantes que consisten en cambiar el criterio con el que se selecciona el modelo final, en una de las variantes se selecciona el modelo con mejor MAE y en la otra se selecciona el modelo con mejor AIC. De cada una de las técnicas resulta un modelo seleccionado, a este modelo seleccionado se le midieron tres indicadores AIC, MAE, RMSE y los resultados de estos modelos son con los que se comparan las técnicas. El primer indicador (MAE) está basado en la bondad de ajuste con penalización dada por el número de parámetros y los dos últimos (MAE, RMSE) como medida únicamente de los residuales. Estas tres perspectivas permiten tener una comparación más amplia sobre los criterios para los que funciona mejor cada una de las técnicas, ya que, como se mencionó anteriormente, según la necesidad del investigador, el criterio de selección puede cambiar. El indicador AIC no se calculó para los árboles CART y modelos cubistas ya que los modelos generados no permiten su cálculo. Para los modelos generados por la metodología AMG que tienen varios modelos paramétricos ajustados, uno por cada nodo, se sumaran los AIC del modelo de cada nodo. Esto implica una penalización mayor en estos modelos, porque, a pesar de usar las mismas variables, cada una de ellas ajusta un parámetro diferente en cada nodo. Las métricas presentadas corresponden a la medición en la muestra de evaluación.

En la Tabla 3 se muestran los resultados de los tres indicadores evaluados para cada uno de los 7 modelos seleccionados en el caso de estudio de Medellín. Como se observa, el MAE es menor que el RMSE en todos los casos, esto se debe a la presencia de valores extremos a la que es más sensible el RMSE. En el indicador MAE/MEDIA se observa que los modelos seleccionados tienen un porcentaje de error alto en todas las técnicas. Esto puede ser consecuencia de la baja correlación entre las variables predictoras y el LOS en este caso de estudio.

El modelo seleccionado con AMG-AIC es el que mejor indicador AIC tiene, sin embargo, ese mismo modelo, el de menor desempeño en MAE y RMSE. Por otra parte el modelo seleccionado con AMG-MAE es el segundo de mejor que mejor indicador MAE obtuvo y el tercero en RMSE, superado por el modelo seleccionado de regresión lineal y GAMLSS-MAE, esto último es lo esperado ya que en estos dos métodos la estimación de parámetros se realiza con el objetivo de minimizar la suma de cuadrados del error.

Tabla 3 Comparación de desempeño de técnicas caso de estudio Medellín

Técnicas	RMSE	AIC	MAE	MAE / MEDIA	MAE / MEDIANA
AMG-MAE	5,64	135197	3,51	59%	88%
AMG-AIC	7,42	115026	4,57	76%	114%
CUBISTA	5,67	NA	3,32	55%	83%
CART	5,68	NA	3,63	61%	91%
LM	5,47	140600	3,55	59%	89%
GAMLSS-AIC	7,26	140600	4,55	76%	114%
GAMLSS-MAE	5,48	115645	3,56	59%	89%

En la Tabla 4 se presentan los resultados de los indicadores de la ejecución de las 7 técnicas sobre el caso de estudio de Microsoft. En este caso el error MAE está entre 21 y 40% de la media, lo que muestra un porcentaje de erro bajo, que se debe a que para este conjunto de datos las variables predictoras si son significativas explicando el LOS. El modelo seleccionado con la técnica AMG-AIC obtuvo el mejor AIC de los 7 modelos seleccionados, no obstante, este modelo fue el que tuvo error más alto en MAE y RMSE. Por otro lado, el modelo seleccionado de la técnica AMG-MAE fue el segundo mejor modelo en el indicador de MAE después del modelo de la técnica cubista, por una diferencia muy pequeña y fue el de mejor desempeño en el indicador RMSE.

Al igual que en el caso anterior, el indicador MAE también es menor que el RMSE para todos los modelos seleccionados, aunque para este caso, las diferencias entre estos dos indicadores son menores, debido a que en este conjunto de datos la asimetría es menor y en consecuencia el impacto que éste tiene sobre el RMSE también.

Tabla 4 Comparación de desempeño de técnicas caso de estudio de Microsoft

Técnicas	RMSE	AIC	MAE	MAE / MEDIA	MAE / MEDIANA
AMG-MAE	1,1	121722	0,85	21%	21%
AMG-AIC	2,09	94418	1,58	40%	40%
Cubista	1,2	NA	0,82	21%	21%
CART	1,24	NA	0,95	24%	24%
LM	1,18	126311	0,9	23%	23%
GAMLSS-AIC	1,83	111416	1,34	34%	34%
GAMLSS-MAE	1,18	126311	0,9	23%	23%

En la Tabla 5 se presentan los indicadores de los 7 modelos seleccionados para el caso de estudio de Arizona. Se obtuvieron resultados muy similares al caso de estudio anterior en cuanto a las mejores técnicas en cada métrica, en AIC el mejor desempeño lo obtuvo el modelo de AMG-AIC y al igual que en los casos anteriores este modelo fue el de mayor error en las dos métricas restantes. En el indicador MAE el mejor modelo fue seleccionado por la técnica cubista, al igual que en los dos casos anteriores, teniendo en segundo lugar al modelo de la técnica AMG-MAE, técnica que generó el modelo que

obtuvo el indicador RMSE. En este caso de estudio, al igual que en el primero, las diferencias entre las técnicas más competitivas de cada indicador son más estrechas que en el caso de estudio de Microsoft. Esto puede ser consecuencia de una menor capacidad predictora de las variables explicativas, que se refleja en el hecho de que en este conjunto de datos el error MAE está entre el 37% y 78% de la media, un poco mejor que en el caso de estudio de Medellón, pero con mayor porcentaje de error que el caso de estudio de Microsoft.

Tabla 5 Comparación de desempeño de técnicas caso de estudio Arizona

Técnicas	RMSE	AIC	MAE	MAE / MEDIA	MAE / MEDIANA
AMG-MAE	5,34	17182	3,41	39%	43%
AMG-AIC	9,27	15113	6,83	77%	85%
Cubista	5,71	NA	3,23	37%	40%
CART	5,35	NA	3,43	39%	43%
LM	5,57	17843	3,45	39%	43%
GAMLSS-AIC	9,43	15356	6,81	77%	85%
GAMLSS-MAE	5,57	17843	3,45	39%	43%

En los resultados presentados en esta sección se observa que la técnica AMG-AIC generó el modelo que obtuvo el mejor indicador AIC en los tres casos de estudio. Esto representa un resultado muy importante en el contexto de predicción del LOS, no solo por la mejora en el indicador sino también por la introducción de un indicador que no es común en la literatura reciente del área y que como se mencionó en la revisión de la literatura es uno de los criterios de selección de modelos más importantes por tener en cuenta la bondad de ajuste y por la penalización que se realiza por la complejidad del modelo, medida en el número de parámetros que se estiman.

En los indicadores basados en los residuales RMSE y MAE, el modelo de la técnica AMG-MAE tuvo resultados competitivos, ya que obtuvo el mejor RMSE en dos de los casos y en el tercer caso obtuvo el segundo mejor indicador. Este mismo modelo obtuvo el segundo mejor indicador MAE en los tres casos. El caso de estudio de Microsoft fue para el que se obtuvo mejores indicadores de los modelos generados por las técnicas AMG, lo que indica que en escenarios con mayor cantidad de datos y mejor calidad de variables es donde toman mayor ventaja sobre otras técnicas. Un aspecto importante de resaltar es la diferencia que se presenta en los modelos generados por las técnicas AMG-AIC y AMG-MAE, ya que el modelo con mejor AIC era el que peor indicador MAE y RMSE obtuvo y viceversa lo que muestra la importancia de definir el criterio con el que se escogen los modelos.

En el caso de estudio de Microsoft fue en el que mejores indicadores se obtuvieron. Esto resalta la importancia de tener variables explicativas que influyan en la variable respuesta. Adicionalmente en este caso es donde se tienen mayor cantidad de variables continuas, que son más fáciles de procesar por parte de las técnicas de modelado. En el caso de estudio de Arizona se tenían pocos registros y pocas variables, sin embargo, los indicadores de los modelos fueron mejores que en el primer caso, evidenciando que

existía información más relevante para la predicción. Otro tema que se observa en el análisis de los casos es que la diferencia entre la métrica MAE y RMSE es mayor mientras mayor sea la presencia de valores extremos, como se da en el caso de Medellín y Arizona, que tienen una mayor brecha entre las dos métricas, en contraste se tiene el caso de estudio de Microsoft que con menor dispersión presenta una menor brecha entre ambos indicadores.

El desarrollo e implementación de la metodología AMG permitió comprender que para el caso del LOS, y otros problemas con las mismas características, es importante tener presente que un conjunto de datos, por su naturaleza, puede provenir de poblaciones con distribuciones diferentes y que por lo tanto una técnica que permite identificar y segmentar estas poblaciones puede ayudar a modelar de manera más acertada el comportamiento de la variable respuesta. Adicionalmente se identificó una diferencia importante entre los modelos que son seleccionados basados en métricas de minimización de residuales como MAE y RMSE, con respecto a los modelos que son seleccionados con base en AIC, diferencia en cuanto a sus métricas de desempeño y en cuanto a la naturaleza de la distribución ajusta. En muchos problemas conocer la naturaleza de la distribución de los datos es fundamental por lo cual utilizar una métrica como el AIC será indispensable. Finalmente se resalta la importancia de tener covariables adecuadas que permitan tener una predicción ajustada, ya que sin estas covariables ninguna técnica arrojará un resultado bueno.

6 Estudio de simulación

El objetivo del estudio de simulación es verificar que los hallazgos que se obtienen con la utilización de la metodología AMG, en los casos reales, son estables y se repiten bajo ciertas condiciones y no son hallazgos de casos particulares. Además, el estudio permite explorar supuestos y condiciones para los cuales la metodología va a tener ventajas o desventajas. Los posibles escenarios que se pueden presentar en la predicción del LOS son innumerables, sin embargo, para definir un alcance razonable se plantearon dos escenarios, dejando la exploración de otros escenarios para investigaciones futuras.

En la literatura del LOS existe evidencia de que la distribución de esta variable puede ser diferente para ciertas características de la población analizada, el caso más común en las investigaciones es la identificación de distribuciones de diferentes enfermedades o tipos de ellas (Berki, Ashcraft, & Newbrander, 1984). Es para este tipo de casos que una metodología como AMG se vuelve pertinente, ya que el componente de árboles de decisión funciona adecuadamente en la identificación y separación de grupos de paciente. La identificación de estos grupos permite que se puedan ajustar modelos diferentes, en sus parámetros e incluso en las distribuciones.

En el trabajo de Berki, Ashcraft, & Newbrander (1984) se evidencia esta particularidad, incluso, se ajusta un modelo diferente para cada grupo de enfermedad. Algunas enfermedades tienen parámetros de distribución (media y desviación) diferentes, lo cual se esperaría sea una ventaja para la metodología AMG, ya que se adecúa a la estructura

de los modelos que genera. La generación de escenarios se basará en la estructura de la base de datos usada en el trabajo de Berki, Ashcraft, & Newbrander (1984).

En el primer escenario se consideran dos enfermedades con LOS promedio muy diferente, enfermedad de los ojos y Artritis. Se espera que el modelo generado sea capaz de separar la población que tiene una enfermedad de la población de la otra, este escenario se llamará Poblaciones diferentes. En el segundo escenario se plantea el caso opuesto, en este se consideran dos enfermedades con LOS promedio similar, Hernia y Próstata. Este escenario supone una dificultad para el modelo ya que probablemente no sea posible separar las poblaciones y la metodología AMG pierde una de las ventajas con respecto a las demás técnicas, este escenario se llamará Poblaciones similares.

6.1 Generación de escenarios

Para validar el adecuado funcionamiento de la metodología AMG se simularon dos escenarios con 100 repeticiones cada uno. Para la generación de las bases de datos se simularon las covariables identificadas en el trabajo de Berki, Ashcraft, & Newbrander (1984). En este trabajo se proporcionan para cada una de las enfermedades las medias de cada covariable o en el caso de las categóricas el porcentaje de clientes en cada categoría, así como la proporción de pacientes por cada grupo de enfermedades. Dada la poca información de las covariables, se asumió una distribución normal con la media dada por la investigación y una desviación del 10% de la media. En la Tabla 6 se presenta la participación de pacientes por enfermedad, la media y desviación del LOS por cada enfermedad y la lista de las 12 covariables utilizadas, sus respectivas medias o porcentajes de participación de las categorías.

Posterior a la simulación de las covariables se generaron las variables respuestas, para generar estas variables se utilizó la distribución exponencial ya que de acuerdo al trabajo de Awad et al. (2017) es la distribución que representa el comportamiento de esta variable. También se utilizó la distribución gamma debido a que es una distribución positiva y permite diferentes tipos de comportamiento incluyendo variables asimétricas y con diferentes tipos de curtósis. Con base en la desviación se calculó el parámetro de forma de la distribución gamma, asumiéndolo constante. Para la media se utilizó el modelo propuesto por Berki, Ashcraft, & Newbrander (1984), en el que la media dependía de las covariables de la Tabla 6, sin embargo, en este modelo no se garantizaba que los valores generados convergieran a la media del LOS dada en la Tabla 6, por esta razón se multiplicaron los coeficiente de la investigación por un valor igual a $\frac{\square}{\square_{\square}}$ donde $\square_{\square}$ es el promedio del valor generado por la ecuación propuesta en la investigación para calcular la media con base en las covariables. En la tabla 7 se presentan los coeficientes de cada una de las covariables y la distribución utilizada para cada una de las enfermedades que se utilizaron. A continuación, se presentan las ecuaciones utilizadas para generar los datos simulados para cada uno de los escenarios.

6.1.1 Escenario Poblaciones diferentes:

Para los datos simulados del escenario poblaciones diferentes se utilizaron las distribuciones planteadas en las ecuaciones 2 y 3, cada una corresponde a una de las enfermedades de este escenario.

$$\square_{\square\square} \sim \square\square\square\square\square\square'\square\square_{\square}, \square_{\square\square\square\square}\square\square_{\square\square\square\square}\square \quad (2)$$

$$\square_{\square\square} \sim \square\square\square\square\square\square\square\square\square\square\square\square\square'(\square_{\square}) \quad (3)$$

donde:

Gamma* = Parametrización de distribución gamma propuesta por (Rigby & Stasinopoulos, 2005)

Exponencial*=Parametrizaciónde distribuciónexponencialpropuestapor(Rigby&Stasinopoulos, 2005)

$\square_{\square\square\square\square}$ = □□□□□ □□ □□□ □□ □□□□□□□□□ □□□ □□□□□□□□□□ □□ □□□ □□□□, □□□ □□□□□ 6.

$\square_{\square\square\square}$ = □□□□□□□□□ó□ □□ □□□ □□ □□□□□□□□□ □□□ □□□□□□□□□□ □□ □□□ □□□□, □□□ □□□□□ 6.

$\square_{\square\square}$ = □□□ □ □□ □□ □□□□□□□□ □□□□□□□□□ □□□ □□□□□□□□□□ □□ □□□ □□□□ en el paciente i.

$\square_{\square\square}$ = □□□ □ □□ □□ □□□□□□□□ □□□□□□□□□ □□□ □□□□□□□□ en el paciente i.

□ = □□□□□□□□□□□□ □□□ □□□□□ □□ □□□□□□□□□, □ □{1 = □□□□□□□□□□□□ □□ □□□ □□□□, 2 = □□□□□□□□}

$$\square_{\square} = \square\square\square\ \square\square\square\ \square_{\square\square}\square_{\square\square}\square$$

□ = □□□□□□ □□ □□□□□□□□□□□□,
□ ∈ {1,2, … ,12} □□□□ □□□□□□ □□ □□□□□□□□□□ □□ □□□ □□□□□□ 6 □ 7 □□□□□□□□□ □ □□□□ □□□□□□□□□□□

$\square_{\square\square}$ = □□□□□□□□□□□□ □□ □□ □□□□□□□□□□□ □ □□□□ □□□□□□□□□□□□□ □, □□□ □□□□□ 7.

$\square_{\square\square}$ = □□□□□□□□□□□□ □ □□□□ □□ □□□□□□□□□□□□□ □, $\square_{\square\square} \sim \square(\square = \square_{\square\square}, \square = 0.1 * \square_{\square\square})$

$\square_{\square\square}$ = Media de la covariable j de la enfermedad e, Ver tabla 6.

6.1.2 Escenario Poblaciones similares:

Para los datos simulados del escenario poblaciones similares se utilizaron las distribuciones planteadas en las ecuaciones 4 y 5, cada una corresponde a una de las enfermedades de este escenario.

(4)

(5)

donde:

Gamma* = Parametrización de distribución gamma propuesta por (Rigby & Stasinopoulos, 2005)

Exponencial* = Parametrización de distribución exponencial propuesta por (Rigby & Stasinopoulos, 2005)

= 6.

= 6.

en el paciente i.

en el paciente i.

= {1 = 2 = }

=

=

= 6 7

= 7.

= ~ (= 0.1 *)

= Media de la covariable j de la enfermedad e, Ver tabla 6.

Tabla 6 Covariables utilizadas y sus respectivas medias o porcentajes

Diagnóstico principal – nombre variable	Enfermedad de los ojos	Hernia	Próstata	Artritis	Indice j
% de pacientes	44%	9%	5%	11%	NA
LOS Media	5	8	9	14	NA
LOS desv	4	10	7	12	NA
% F - genero	45%	13%	0%	63%	1
Edad – edad	51	4	65	60	2
Distancia domicilio-hospital (millas) - distdh	78	26	70	80	3
Método de pago 1 (%blue cross/blue shield) –mp1	38	48	32	35	4
Método de pago 2 (Medicaid/medicare work compensation) – mp2	38	16	60	53	5
Número de diagnósticos –numdiagn	3	2	3	2	6
Tiempo de servicios de radiología – tsrad	4	0	23	18	7
Tiempo de servicios de laboratorio – tslab	12	6	25	29	8
Intensidad en servicios de enfermería – isenf	118	121	115	116	9
Día de admisión (% adm juev-Sab) - admjs	31	27	18	18	10
Tipo de admisión (% de adm urgencia) - admurg	24	9	16	2	11
Tipo de admisión (% de adm emergencia) - admemer	46	23	15	0	12

Tabla 7 coeficientes de regresión de los modelos simulados y distribuciones

Variables	coeficientes Artritis	coeficientes Enferm. Ojos	coeficientes Hernia	coeficientes próstata	Indice j
Distribución	Exponencial	Gamma	Exponencial	Gamma	NA
genero	0.075	0.132	-0.047	0.000	1
edad	-0.006	-0.038	0.047	0.027	2
distdh	0.047	-0.038	-0.016	0.092	3
mp1	-0.031	0.094	-0.012	-0.269	4
mp2	0.013	0.019	-0.086	-0.264	5
numdiagn	0.072	-0.019	-0.078	0.129	6
tsrad	0.072	-0.377	-0.109	-0.081	7
tslab	0.003	-0.868	-0.428	0.119	8
isenf	0.082	0.189	0.093	-0.005	9
admjs	0.047	-0.075	-0.078	0.027	10
admurg	-0.006	-0.226	0.195	0.022	11
admemer	0.132	-0.321	-0.086	0.092	12

6.2 Análisis de resultados

Para analizar la predicción en cada uno de los escenarios planteados se utilizaron las siete técnicas mencionadas en los casos reales con el objetivo de ver si los resultados encontrados allí se reflejaban de manera constante en otros conjuntos de. En las Tablas 3 y 4 se presentan los intervalos de confianza al 95% de las métricas de desempeño para cada uno de los modelos generados, analizar los intervalos permite comprobar que las diferencias no se deben a un efecto aleatorio, sino que son resultado de diferencias estadísticas significativas.

Tabla 8 Resultados de las técnicas para escenario de Poblaciones diferentes

Intervalos de confianza 95% de confianza					
Técnicas	RMSE	AIC	MAE	MAE / MEDIA	MAE / MEDIANA
AMG-MAE	(7.00,7.12)	(21697, 21833)	(3.37,3.39)	(46.6%, 46.9%)	(64.8%, 65.2%)
AMG-AIC	(8.73,8.86)	(21171, 21249)	(4.35,4.46)	(60.2%, 61.7%)	(83.7%, 85.8%)
CUBISTA	(7,27, 7,29)	NA	(3.45,3.46)	(47.7%, 47.9%)	(66.3%, 66.5%)
CART	(7.09,7.10)	NA	(3.94, 3.96)	(54.5%, 53.9%)	(75.8%, 75%)
LM	(6.99,6.99)	(36816, 36985)	(3.66,3.67)	(50.6%, 50.8%)	(70.4%, 70.6%)
GAMLSS-AIC	(7.08, 7.30)	(22561, 22646)	(3.68,3.86)	(50.9%, 53.4%)	(70.8%, 74.2%)
GAMLSS-MAE	(7.02,7,03)	(25837, 27185)	(3.65,3.67)	(50.5%, 50.8%)	(70.2%, 70.6%)

Tabla 9 Resultados de las técnicas para escenario de Poblaciones similares

Intervalos de confianza 95% de confianza					
Técnicas	RMSE	AIC	MAE	MAE / MEDIA	MAE / MEDIANA
AMG-MAE	(7.89, 7.97)	(13945, 14040)	(5.3, 5.34)	(62.61%, 63.08%)	(82.48%, 83.09%)
AMG-AIC	(9.66, 10.14)	(13904, 13969)	(6.67, 7.09)	(78.77%, 83.65%)	(103.76%, 110.19%)
CUBISTA	(7.91, 7.93)	NA	(5.41, 5.42)	(63.9%, 63.97%)	(84.18%, 84.27%)
CART	(7.59, 7.6)	NA	(5.61, 5.62)	(66.26%, 66.39%)	(87.29%, 87.45%)
LM	(7.58, 7.58)	(19272, 19328)	(5.58, 5.59)	(65.91%, 66.05%)	(86.82%, 87%)
GAMLSS-AIC	(9.81, 10.31)	(13941, 14000)	(6.93, 7.36)	(81.79%, 86.95%)	(107.74%, 114.53%)
GAMLSS-MAE	(7.91, 7.93)	(13953, 14011)	(5.43, 5.43)	(64.06%, 64.11%)	(84.38%, 84.45%)

De los resultados obtenidos se observa que el modelo generado por la técnica AMG_AIC es el que menor AIC obtuvo en los dos escenarios, hallazgo que coincide con los resultados encontrados en los casos reales, donde esta técnica también fue la que obtuvo el menor AIC. Otro resultado que coincide con los hallazgos de los casos reales es el

desempeño de la técnica AMG_MAE que fue la que menor MAE obtuvo en ambos escenarios, en el caso del indicador RMSE la técnica no fue superior a las demás. Con lo anterior se puede concluir que es un resultado que se obtienen con la metodología desarrollada, en comparación con otras técnicas van a ser estables bajo diferentes escenarios.

Otro aspecto que se quería validar con las simulaciones, es que los resultados de los modelos obtenidos con la metodología AMG no fueran sensibles a cambios pequeños y no estructurales, como los cambios que se hicieron en las 100 instancias que se corrieron de cada escenario variando aleatoriamente los valores de las variables sin hacer cambios estructurales como las distribuciones, coeficientes de regresión, desviaciones o la cantidad de subgrupos que tenían distribuciones del LOS diferentes. Esto se comprobó con base en la longitud de los intervalos de confianza que se generaron en cada escenario, donde se observan intervalos pequeños que variaban máximo un 1% del promedio de los resultados obtenidos en las 100 repeticiones de cada escenario.

Con relación a las ventajas de la metodología desarrollada se identificaron tres: la primera es la superioridad en los modelos generados medidos con el indicador AIC, adicionalmente, también mostro un desempeño superior en la métrica MAE. La segunda ventaja es el uso de distribuciones diferentes a la normal para ajustar los modelos, ventaja que se observa en que las técnicas basadas en GAMLSS superaron, con una diferencia importante, en el indicador AIC a la regresión lineal que utiliza únicamente la distribución normal, lo que evidencia que la distribución es un factor fundamental para la obtención de un buen AIC.

La tercera ventaja se observa en las diferencias observadas entre los dos escenarios planteados. En la métrica AIC la diferencias que se presentaron entre los modelos que ajustan una sola distribución (GAMLSS_MAE, GAMLSS_AIC, LM) y los modelos generados por la metodología AMG (AMG_MAE y AMG_AIC), son mayores en el escenario de poblaciones diferentes que en el escenario de poblaciones similares. Esto se debe a que la metodología AMG y GAMLSS se diferencian precisamente en que la primera separa la población antes de hacer al ajuste de la distribución, con el fin de generar modelos con distribuciones diferentes por cada subgrupo de la población encontrado. En el escenario de poblaciones similares, al hacerse difícil dicha separación, las dos técnicas terminan generando resultados similares porque la metodología AMG no encuentra una variable que le permita separar las poblaciones y ajustar modelos diferentes a cada una, que es precisamente su ventaja frente al modelo gamlss tradicional.

7 Conclusiones

En la presente investigación se evaluaron diferentes aspectos que juegan un papel fundamental en la construcción de modelos predictivos, fundamentalmente en el contexto de predicción de LOS, a continuación, se mencionan algunos de los más importantes que se abordaron con esta investigación.

En esta investigación se identificaron algunas oportunidades en el contexto de predicción del LOS que se abordaron con el desarrollo del trabajo. La primera es tener un estudio enfocado en la metodología que generaba un modelo de predicción más que en el modelo mismo y sus resultados, esta oportunidad se trabajó con el diseño e implementación en software de la metodología AMG, que facilitará la selección de modelos para predicción del LOS en un espectro amplio de posibles modelos. La segunda oportunidad está relacionada con el uso de métricas más enfocadas en la bondad de ajuste que en la minimización de residuales, como era el caso de la mayoría de artículos encontrados en la literatura. Para esto se utilizó, entre otras, la métrica AIC y esto llevo a la necesidad de incluir diferentes tipos de distribuciones, además de pruebas de hipótesis de ajustes de distribuciones que permitiera mejorar el ajuste de los modelos basados en AIC. Finalmente se identificó la oportunidad de mejorar el desempeño de los modelos de predicción del LOS, lo cual se logró generando el mejor modelo en todos los casos evaluados en la métrica AIC y el mejor modelo en la mayoría de casos evaluados con la métrica MAE.

En la literatura de predicción del LOS, principalmente desde los métodos computacionales que son los más comunes en la actualidad, no se aborda con profundidad el ajuste de los datos a las distribuciones paramétricas sobre las que se basan muchos modelos utilizados. Como consecuencia se tienden a ajustar regresiones que no cumplen los supuestos. Con la presente investigación se aborda esta problemática generando una metodología que utiliza métodos computacionales pero que permite el análisis de ciertas características importantes en los modelos de regresión como el ajuste de los datos a la distribución sobre la que se basa el modelo, generando dentro de la metodología AMG la versatilidad para que el usuario defina la rigurosidad con que se ajustará el modelo y sea consciente del incumplimiento de algunos supuestos para las conclusiones que se generen de estos.

Las construcciones de modelos predictivos con un desempeño adecuado dependen fundamentalmente de tener variables predictoras que expliquen gran parte de la variabilidad de la variable respuesta, no tenerlas implica errores más grandes y probablemente la imposibilidad de realizar una predicción acertada. Esto se evidencia dentro del primer caso de estudio, para el cual los modelos generaron predicciones con un margen de error muy amplio, y aunque existían diferencias en los modelos ajustados, no eran suficientes para lograr un desempeño apropiado. Este desempeño inferior al esperado se asocia a las variables explicativas, ya que las mismas técnicas en el caso de estudio dos mostraron un desempeño muy superior. Este aspecto se debe resolver desde las instituciones hospitalarias, garantizando que capturan la información adecuada.

En los modelos construidos con la metodología AMG se encontró que los que fueron seleccionado a través del AIC, que prioriza la bondad de ajuste, mostraban un mejor

comportamiento en este indicador con respecto a las otras técnicas, pero no fueron los que tuvieron mejor desempeño en los indicadores MAE y RMSE. Estos modelos, mostraban una tendencia a ajustarse mejor al modelo real de los datos, de acuerdo con la naturaleza del AIC que mide la bondad de ajuste. Adicionalmente la metodología desarrollada tenía la variante de seleccionar un modelo enfocado en los indicadores MAE y RMSE, que permitía tener predicciones muy competitivas en comparación con las técnicas que se comparó.

El desempeño del modelo generado con la técnica AMG-MAE superó los resultados que se obtienen en los modelos generados por las técnicas en las que se basa (árboles de decisión y modelos GAMLSS), analizándolos individualmente, y superando en casi todos los escenarios a la regresión lineal. Esto muestra que es una metodología competitiva que puede generar resultados de mucho interés en las investigaciones que se utilice. Adicionalmente, la metodología desarrollada tiene un potencial que no se alcanzó a explotar en el presente trabajo, potencial que se encuentra en el uso de los componentes aditivos de los modelos GAMLSS y en el ajuste de los parámetros de escala y forma de cada distribución. Explotar estos componentes adicionales puede mejorar considerablemente el desempeño de los modelos generados y permitir una flexibilidad mayor en cuanto al tipo de fenómenos que se modelan.

Los modelos generados con la técnica cubistas fueron los que mostraron mejor desempeño en el indicador MAE, aunque no fueron los de mejor indicador RMSE. Su desempeño superior con respecto a la metodología AMG puede ser causa de algunas características que diferencian ambos métodos. En primer lugar, en la técnica cubista se ajustan modelos lineales en nodos que no son finales, mientras que en la metodología AMG solo se hace en nodos finales. Los ajustes en nodos intermedios no se realizó debido al costo computacional que tenía para AMG particularmente, ya que no solo se ajusta un modelo en cada nodo, como el caso del cubista, sino que se ajustan, en cada nodo, tantos modelos como distribuciones de probabilidad se consideren. En segundo lugar, en la técnica cubista se ajusta la predicción que se realiza con base en observaciones cercanas, característica que no se implementó en la metodología AMG. Finalmente, otra diferencia está en el algoritmo utilizado para la generación del árbol de decisión que para AMG es un árbol tipo CART y para la técnica cubista el algoritmo es un M5.

La presente investigación se enfocó en generar las bases para un campo de investigación futuro en el que se puedan proponer metodologías que tengan en cuenta tanto el enfoque computacional como el enfoque estadístico en la construcción de modelos, para que no se conciban como campos de investigación independientes sino como complementarios, de esta manera podrá tener un conocimiento más amplio del fenómeno estudiado.

La metodología AMG ha mostrado ventajas en la generación de modelos con mejor AIC, principalmente en escenarios donde existen poblaciones con distribuciones claramente diferenciables en su media, y donde las distribuciones de las que provienen los datos no son simétricas. Adicionalmente se comprobó que estos resultados son estables en diferentes escenarios y no son sensibles a cambios no estructurales de los modelos que generan los datos. Finalmente es importante mencionar que la metodología AMG permite obtener modelos con muy buenos indicadores en AIC y en MAE, pero en el indicador RMSE la metodología tiene oportunidades de mejora.

8 Trabajos futuros

Es importante en investigaciones futuras explorar la validación de supuestos de los modelos generados y las consecuencias de su no cumplimiento. Algunos aspectos que se deben considerar son: la transformación de las variables respuestas antes de ajustar los modelos para disminuir el incumplimiento de supuestos y el análisis de penalización más altas en el AIC que generen un efecto más notorio en la selección del modelo.

Como se mencionó anteriormente los modelos GAMLSS que se utilizan en la metodología AMG tiene dos componentes que puede ser explorado en investigaciones futuras. El uso de los componentes aditivos y la modelación de los otros parámetros que componen las distribuciones pertenecientes a GAMLSS, también se pueden considerar componentes de la técnica cubista que no se utilizaron, como el ajuste de modelos lineales a todos los nodos del árbol generado y no solo a los nodos finales.

Un aspecto fundamental que se abordará en trabajos futuros es la búsqueda de algoritmos que permitan crear y seleccionar de manera eficientes los posibles modelos que se generen en el uso de la metodología AMG, esta es una problemática que se genera por la versatilidad de la metodología, ya que como se puede adaptar a problemas con estructuras muy diferentes, también se debe buscar la manera de que la metodología tenga la capacidad de encontrar de manera rápida un modelo adecuado para un conjunto de datos analizado.

9 Bibliografía

Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Al Taleb, A. R., Abul Hasanat, M. H., & Khan, M. B. Application of Data Mining Techniques to Predict Length of Stay of Stroke Patients, Informatics Health and technology International conference 1–5 (2017).

Altman, H., Angle, H. V, Brown, M. L., & Sletten, I. W. (1972). Prediction of Length of Hospital Stay. Comprehensive Psychiatry, 13(5), 471–480.

Awad, A., Bader–El–Den, M., & McNicholas, J. (2017). Patient length of stay and mortality prediction: A survey. Health Services Management Research, 30(2), 105–120. https://doi.org/10.1177/0951484817696212

Azari, A. (2015). Imbalanced Learning to Predict Long Stay Emergency Department Patients. In International Conference on Bioinformatics and Biomedicine (BIBM) Imbalanced (pp. 807–814).

Azari, A., Janeja, V. P., & Mohseni, A. (2012). Predicting Hospital Length of Stay (PHLOS): A Multi-tiered Data Mining Approach. In IEEE 12th International Conference on Data Mining Workshops (pp. 17–24). https://doi.org/10.1109/ICDMW.2012.69

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. Real-time prediction of inpatient length of stay for discharge prioritization, 23 Journal of the American Medical Informatics Association e2–e10 (2016). https://doi.org/10.1093/jamia/ocv106

Berki, S. E., Ashcraft, M. L. F., & Newbrander, W. C. (1984). Length of Stay variation within ICDA 8 Diagnosis related Group. Medical Care, 22(2), 126–142. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6422169

Breiman, L. (1984). Classification and Regression Trees. Routledge. New York. https://doi.org/10.1371/journal.pone.0015807

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(421), 123–140. https://doi.org/10.1007/BF00058655

Burns, L. R., & Wholey, D. R. (1991). The Effects of Patient , Hospital , and Physician Characteristics on Length of Stay and Mortality. Medical Care, 29(3), 251–271.

Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., & Gallego, B. (2016). Real-time prediction of mortality, readmission, and length of stay using electronic health record data. Journal of the American Medical Informatics Association, 23(3), 553–561. https://doi.org/10.1093/jamia/ocv110

Carter, E. M., & Potts, H. W. (2014). Predicting length of stay from an electronic patient record system: a primary total knee replacement example. BMC Medical Informatics and Decision Making, 14(1), 14–26. https://doi.org/10.1186/1472-6947-14-26

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute

error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and Recent Advances in Model Selection. In Handbook of the philosophy of science (Vol. 7, pp. 583–605). Elsevier B.V. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Charlson, M. E., & Horwitz, R. I. (1984). Applying results of randomised trials to clinical practice: impact of losses before randomisation. British Medical Journal (Clinical Research Ed.), 289(6454), 1281–1284. https://doi.org/10.1136/bmj.289.6454.1281

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases, 40(5), 373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks Opening the black box. Cancer, 91(S8), 1615–1635. https://doi.org/10.1016/S0967-067X(01)00020-4

Faddy, M., Graves, N., & Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. Value in Health, 12(2), 309–314. https://doi.org/10.1111/j.1524-4733.2008.00421.x

Fetter, R., Shin, Y., Freeman, J., Averill, R., & Thompson, J. (1980). Case mix definition by diagnosis-related groups. Medical Care, 18(2), 1–53. https://doi.org/10.1007/s13398-014-0173-7.2

Flach, P. (2012). Machine Learning: The Art and Science of Algorithms that Make Sense of Data (1st ed.). New York: Cambridge University Press. https://doi.org/10.1145/242224.242229

Fontova-Almato, A., Juvinya-Canal, D., & Suner-Soler, R. (2015). Influence of waiting time on patient and companion satisfaction. Revista de Calidad Asistencial, 30(1), 10–16. https://doi.org/10.1016/j.cali.2014.12.009

García Olaverri, C. (1996). ESTABILIDAD DE ALGUNOS CRITERIOS DE SELECCIÓN DE MODELOS. Qüestiió, 20(2), 147–166.

Geweke, J., & Meese, R. (1981). Estimating Regression Models of Finite but Unknown Order. International Economic Review, 22(1), 55–70. https://doi.org/10.2307/2526135

Golmohammadi, D. (2016). Predicting hospital admissions to reduce emergency department boarding. International Journal of Production Economics, 182(2016), 535–544. https://doi.org/10.1016/j.ijpe.2016.09.020

Green, S. B. (2010). How Many Subjects Does It Take To Do A Regression Analysis? Multivariate Behavioral Research, 26(3), 499–510. https://doi.org/10.1207/s15327906mbr2603

Guerra, C. W., Cabrera, A., & Fernández, L. (2003). Criterios para la selección de modelos estadísticos en la investigación científica. Revista Cubana de Ciencia Agrícola, 37(1), 3–10. Retrieved from http://www.redalyc.org/pdf/1930/193018072001.pdf

Gustafson, D. H. (1968). Length of stay: prediction and explanation. Health Services Research, 3(1), 12–34.

Harrison, D. a, Parry, G. J., Carpenter, J. R., Short, A., & Rowan, K. (2007). A new risk prediction model for critical care: The Intensive Care National Audit & Research Centre (ICNARC) Model. Critical Care Medicine, 35(4), 1091–1098. https://doi.org/10.1097/01.CCM.0000259468.24532.44

Hilbe, J. M. (2016). COUNT: Functions, Data and Code for Count Data. Cambridge University Press. Retrieved from https://cran.r-project.org/package=COUNT

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. Computational Statistics, 24(2), 225–232. https://doi.org/10.1007/s00180- 008-0119-7

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research. Retrieved from http://jmlr.org/papers/v16/hothorn15a.html

Houthooft, R., Ruyssinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., … De Turck, F. (2015). Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. Artificial Intelligence in Medicine, 63(3), 191–207. https://doi.org/10.1016/j.artmed.2014.12.009

Jiménez, R., López, L., Dominguez, D., & Fariñas, H. (1999). Difference between observed and predicted length of stay as an indicator of inpatient care inefficiency. International Journal for Quality in Health Care, 11(5), 375–384. https://doi.org/10.1093/intqhc/11.5.375

Kaur, G., & Chhabra, A. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. International Journal of Computer Applications, 98(22), 13–17. https://doi.org/10.5120/17314-7433

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (1st ed.). New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kumari, M., Vohra, R., & Arora, A. (2014). Prediction of Diabetes Using Bayesian Network. International Journal of Computer Science and Information Technologies, 5(4), 5174–5178. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.640.3573

Lawson, C., Pati, S., Green, J., Messina, G., Strömberg, A., Nante, N., … Kadam, U. T. (2017). Development of an international comorbidity education framework. Nurse Education Today, 55(2017), 82–89. https://doi.org/10.1016/j.nedt.2017.05.011

Lemeshow, S., Teres, D., Klar, J., Avrunin, J. S., Gehlbach, S. H., & Rapoport, J. (1993). Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA, 270(20), 2478–2486. https://doi.org/10.1001/jama.270.20.2478

Liu, H. C. (2013). A theoretical framework for holistic hospital management in the Japanese healthcare context. Health Policy, 113(1–2), 160–169.

https://doi.org/10.1016/j.healthpol.2013.08.009

Liu, P., Lei, L., Yin, J., Zhang, W., Naijun, W., & El-Darzi, E. (2006). Healthcare data mining: Prediction inpatient length of stay. In 3rd International IEEE conference Intelligent Systems (pp. 832–837). https://doi.org/10.1109/IS.2006.348528

Liu, V., Kipnis, P., Gould, M. K., & Escobar, G. J. (2010). Length of Stay Predictions Improvements Through the Use of Automated Laboratory and Comorbidity Variables, 48(8), 739–744. Retrieved from http://www.jstor.org/stable/25701529

Marazzi, A., Paccaud, F., Ruffieux, C., & Beg, C. (1998). Fitting the Distributions of Length of Stay by Parametric Models. Medical Care, 36(6), 915–927. Retrieved from http://www.jstor.org/stable/3767008

Marshall, A., Vasilakis, C., & El-Darzi, C. (2005). Length of stay-based patient flow models: Recent developments and future directions. Health Care Management Science, 8(3), 213–220. https://doi.org/10.1007/s10729-005-2012-z

Microsoft. (2018). Predicting Hospital Length of Stay. Retrieved July 5, 2018, from https://microsoft.github.io/r-server-hospital-length-of-stay/index.html

Mobley, B. A., Leasure, R., & Davidson, L. (1995). Artificial neural network predictions of lengths of stay on a post-coronary care unit. Heart and Lung - The Journal of Acute and Critical Care, 24(3), 251–256. https://doi.org/10.1016/S0147-9563(05)80045-7

Moores, B., Rahman, M. M., Settle, J. A. D., & Browning, F. S. C. (1975). On the predictability of the length of patient stay in a burns unit. Burns, 1(4), 291–296. https://doi.org/10.1016/0305-4179(75)90003-0

Nouaouri, I., Samet, A., & Allaoui, H. (2015). Evidential data mining for length of stay (LOS) prediction problem. In IEEE International Conference on Automation Science and Engineering (pp. 1415–1420). https://doi.org/10.1109/CoASE.2015.7294296

Ojeda, C. J., & Rocco, C. M. (2014). métodos multiobjetivo Metodología para selección de modelos de regresión lineal múltiple basada en métodos multiobjetivo Multiobjective Methods, (January).

OMS-Organización Mundial de la Salud. (1990). Clasificación internacional de Enfermedades. Retrieved from http://www.paho.org/hq/index.php?option=com_content&view=article&id=3561%3A2010-clasificacion-internacional-enfermedades-cie&catid=511%3Ahealth-information-analysis&Itemid=2560&lang=es

Pahlevan Sharif, S. (2017). Locus of control, quality of life, anxiety, and depression among Malaysian breast cancer patients: The mediating role of uncertainty. European Journal of Oncology Nursing, 27(2017), 28–35. https://doi.org/10.1016/j.ejon.2017.01.005

Pendharkar, P. C., & Khurana, H. (2014). Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. International Journal of Computer Science and Applications, 11(3), 45–56. Retrieved from http://www.tmrfindia.org/ijcsa/v11i33.pdf

Perez, A., Chan, W., & Dennis, R. J. (2006). Predicting the Length of Stay of Patients

Admitted for Intensive Care Using a First Step Analysis. Health Services & Outcomes Research Methodology, 6(2006), 127–138. https://doi.org/10.1007/s10742-006-0009-9

Perez C., N., Marquez G., A., Acosta M., H. G., & Mezura M., E. (2017). Full Model Selection issue in temporal data through evolutionary algorithms: A brief review. In IEEE Congress on Evolutionary Computation (pp. 2451–2457). https://doi.org/10.1109/CEC.2017.7969602

Pilotto, A., Sancarlo, D., Pellegrini, F., Rengo, F., Marchionni, N., Volpato, S., & Ferrucci, L. (2016). The Multidimensional Prognostic Index Predicts in-hospital length of stay in older patients: A multicentre prospective study. Age and Ageing, 45(1), 90–96. https://doi.org/10.1093/ageing/afv167

Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. European Journal of Operational Research, 132(3), 666–680. https://doi.org/10.1016/S0377-2217(00)00171-5

Quinlan, J. R. (1992). Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343–348). https://doi.org/10.1.1.34.885

R Core Team. (2018). R Project. R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Robinson, G. H., Davis, L. E., & Leifer, R. P. (1966). Prediction of Hospital Length of Stay. Health Services Research, 1(3), 287–300. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1067345/

Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. Econometrica, 46(6), 1273–1291. https://doi.org/10.2307/1913828

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. Retrieved from https://projecteuclid.org/euclid.aos/1176344136

Silva, Á., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. Artificial Intelligence in Medicine, 43(3), 179–193. https://doi.org/10.1016/j.artmed.2008.03.010

Stasinopoulos, M., & Rigby, R. (2018). gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. Retrieved from https://cran.r-project.org/package=gamlss.dist

Tanuja, S., Acharya, D. U., & Shailesh, K. R. (2011). Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay. Journal of Pharmaceutical and Biomedical Sciences, 7(7), 1–4. Retrieved from http://eprints.manipal.edu/757/1/Shailesh_K_R_et._al.%281%29.pdf

Theil, H. (1961). Economic forecast and policy. Amsterdam: North-Holland Pub. Co.

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression

Trees. Retrieved from https://cran.r-project.org/package=rpart

Therneau, T. M., & Atkinson, E. J. (2018). An introduction to recursive partitioning using the rpart routines. https://doi.org/10.1017/CBO9781107415324.004

Turgeman, L., May, J. H., & Sciulli, R. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission, 78 Expert Systems with Applications 376–385 (2017). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2017.02.023

Turgeman, L., May, J., Ketterer, A., Sciulli, R., & Vargas, D. (2015). Identification of readmission risk factors by analyzing the hospital-related state transitions of congestive heart failure (CHF) patients. IIE Transactions on Healthcare Systems Engineering, 5(4), 255–267. https://doi.org/10.1080/19488300.2015.1095823

Vaish, A., Vaish, A., Vaishya, R., & Bhawal, S. (2016). Customer relationship management (CRM) towards service orientation in hospitals: A review. Apollo Medicine, 13(4), 1–5. https://doi.org/10.1016/j.apme.2016.11.002

Verburg, I. W. M., De Keizer, N. F., De Jonge, E., & Peek, N. (2014). Comparison of regression methods for modeling intensive care length of stay. PLoS ONE, 9(10), 1–11. https://doi.org/10.1371/journal.pone.0109684

Vignolio, J., Vacarezza, M., Álvarez, C., & Sosa, A. (2011). Niveles de atención, de prevención y atención primaria de la salud. Prensa Médica Latinoamericana, XXXIII(1), 11–14. Retrieved from http://www.scielo.edu.uy/pdf/ami/v33n1/v33n1a03.pdf

Weir, E., D'Entremont, N., Stalker, S., Kurji, K., & Robinson, V. (2009). Applying the balanced scorecard to local public health performance measurement: deliberations and decisions. BMC Public Health, 9(1), 127–133. https://doi.org/10.1186/1471-2458-9-127

Widyastuti, Y., Stenseth, R., Wahba, A., Pleym, H., & Videm, V. (2012). Length of intensive care unit stay following cardiac surgery: Is it impossible to find a universal prediction model? Interactive Cardiovascular and Thoracic Surgery, 15(5), 825–833. https://doi.org/10.1093/icvts/ivs302

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79–82. https://doi.org/10.3354/cr030079

Witten, I. H., & Frank, E. (2005). Data Mining : Practical Machine Learning Tools and Techniques (4th ed., Vol. 2). Cambridge: Morgan Kaufmann. https://doi.org/10.1186/1475-925X-5-51

Woods, A. W., MacKirdy, F. N., Livingston, B. M., Norrie, J., & Howie, J. C. (2000). Evaluation of predicted and actual length of stay in 22 Scottish intensive care units using the APACHE III system. Anaesthesia, 55(11), 1058–1065. https://doi.org/10.1046/j.1365-2044.2000.01552.x

Zhu, L., Li, L., & Liang, Z. (2009). Comparison of six statistical approaches in the selection of appropriate fish growth models. Chinese Journal of Oceanology and Limnology, 27(3), 457–467. https://doi.org/10.1007/s00343-009-9236-6

10 Anexos

Anexo 1

A continuación, se presenta la descripción, el uso, los argumentos, los valores y los ejemplos de cada una de las funciones implementadas en R.

---

gamlss_tree Ajusta Árbol de Modelos GAMLSS AMG

Descripción

gamlss_tree ajusta un modelo de Árbol de Modelos GAMLSS que se compone de un árbol de decisión tipo CART con la función rpart y un modelo GAMLSS con la función **gamlss** en cada nodo final del árbol.

Uso

```
gamlss_tree(form, datos, n_dist_mod=4, var_sel="aicmodelo",
            steps=2, porc_entre=0.8, committess=1,
           nom_dist=c("exGAUS","GIG","GG","BCCGo","BCPEo","G
                      A","GB2","BCTo","WEI3", "LOGNO", "EXP",
                      "PARETO2","IG","IGAMMA","NO"), cyc=50,
                      prueba_hip=TRUE,
                      acepta_h=FALSE,type="counts",
                      arbol_activo=TRUE)
```

Argumentos

form	Una fórmula compuesta de dos partes, una que indica la variable respuesta que se encuentra al lado izquierdo del operador ~, y la segunda parte las variables explicativas que se ubican al lado derecho del operador ~ y que se separan con el signo +.
datos	Un **data.frame** que contiene las variables que se refieren en el argumento **form**.
n_dist_mod	Es el número de distribuciones que se van a preseleccionar con base en el ajuste de la variable respuesta a las distribuciones seleccionadas por el usuario. Este argumento permite limitar el número de distribuciones con las cuales se ajustan los modelos GAMLSS con covariables y reducir el tiempo de cómputo del algoritmo.

var_sel	Es la variable que se utiliza para seleccionar el modelo, se puede utilizar los valores “aicmodelo” para seleccionar el modelo con mejor AIC o se utiliza el valor “MAE” para seleccionar el modelo con mejor indicador MAE.
steps	Es el número de pasos que se realizan en el procedimiento de stepwise.
porc_entre	Es el porcentaje de datos que se utilizarán como muestra de entrenamiento
comitteess	Es el número de iteraciones boosting que se realizan.
nom_dist	Es el vector que indica las distribuciones del paquete gamlss que se tendrán en cuenta durante la construcción de los modelos.
cyc	Es el número de iteraciones que se utiliza en la función gamlss para realizar el ajuste de los parámetros de los modelos en cada nodo.
prueba_hip	Es un argumento de tipo bandera. Cuando está en “TRUE” indica que se realizan las pruebas de hipótesis de bondad de ajuste de la variable respuesta con respecto a las distribuciones ingresadas por el usuario.
acepta_h	Es un argumento de tipo bandera, cuando está en “TRUE” se indica que solo se utilizarán las distribuciones qué aprueban la prueba de hipótesis de bondad de ajuste de la variable respuesta con una significancia del 0.05.
type	Es el tipo de distribuciones que se tendrán en cuenta durante la generación de modelos, sirve como complemento de las distribuciones definidas en el argumento nom_dist se utilizan los mismos tipos de la función fitDist del paquete gamlss.
arbol_activo	Es un argumento de tipo bandera que se utiliza para definir si se realiza el árbol de decisión inicial, cuando el argumento es diferente de “TRUE” la función ajusta modelos GAMLSS para las distribuciones definidas y selecciona el modelo que mejor resultados tenga de acuerdo con la métrica definida por el usuario.

Valores

La función retorna una lista con los siguientes valores, los cuales están definidos para cada uno de las iteraciones o comitteess que tenga el modelo:

arboles	Son los árboles de decisión ajustados con la función rpart durante la ejecución de la función.
modelosxnodo	Es una lista con los modelos GAMLSS que se ajustan por cada uno de los nodos finales del árbol de decisión.

valor_nodo	Es un vector con la predicción de cada nodo del árbol de decisión, que equivale a la media de la variable respuesta para el conjunto de observaciones que pertenecen a cada nodo.
dis_sel	Es un vector con la lista de distribuciones seleccionadas en cada uno de los modelos ajustados por nodo final.
nodos_train	Es una lista que guarda el conjunto de observaciones de entrenamiento, junto con sus variables explicativas, que pertenecen a cada uno de los nodos del árbol de decisión.
Form	Es la fórmula ingresada por el usuario para ajustar el árbol de modelos GAMLSS.
datos_test	Es el grupo de observaciones que fue separada por la función para ser utilizada en la evaluación de los modelos.
nodo_prueba	Es una lista que contiene una tabla por nodo con la información obtenida al utilizar la función probar_dist en los datos pertenecientes a cada nodo.
nodos_ajuste_cov	Es una lista qué contiene una tabla por nodo con la información de las métricas obtenidas para cada uno de los modelos ajustados con cada una de las distribuciones preseleccionadas.

Ejemplo

```
# Se carga la librería COUNT para utilizar el conjunto de datos

azpro

library(azpro)

data(azpro)

#Se ajusta el modelo GAMLSS

model_amg<-gamlss_tree(los~.,datos=azpro)

#Se muestran algunos de los valores arrojadas por la función

model_amg$arboles
model_amg$modelosxnodo
model_amg$nodos_ajuste_cov
```

test_dist	Probar ajuste de distribuciones del paquete **gamlss** a variable respuesta.

Descripción

test_dist analiza el ajuste de la variable respuesta a cada una de las distribuciones definidas por el usuario, calcula indicadores de bondad de ajuste AIC y BIC con la función **fitDist** y adicionalmente realiza dos pruebas de bondad de ajuste, Anderson Darling con la función **ad.test** y Smirnov Kolmogorov con la función **ks.test,** estas dos pruebas se evalúan con una significancia definida por el usuario generando una columna que indica si la prueba es aceptada o rechazada.

Uso

```
test_dist(y, type="realplus", extra="NO", signif=0.05)
```

Argumentos

y	Una fórmula compuesta de dos partes, una que indica la variable respuesta que se encuentra al lado izquierdo del operador ~, y la segunda parte las variables explicativas que se ubican al lado derecho del operador ~ y que se separan con el signo +.
type	Este argumento define un grupo de distribuciones par las cuales se evaluarán las pruebas y las métricas de bondad de ajuste. Es el mismo argumento de la función **fitDist** del paquete **gamlss.**
extra	Son las distribuciones adicionales a las que están contenidas en el argumento type.
signif	Es la significancia con la cual se evaluarán las pruebas de hipótesis realizadas.

Valores

La función retorna una **data.frame** que tiene una fila por cada una de las distribuciones y en cada columna tiene los siguientes valores:

Dist	Es el nombre de la distribución evaluada.
KS	Es el valor P resultante de la prueba Smirnov Kolmogorov.
AD	Es el valor P resultante de la prueba Anderson Darling.

SBC	Es el Criterio de Información Bayesiano.
AIC	Es el Criterio de Información de Akaike.
res_prueba_ks	Es el resultado de evaluar el valor P de la prueba de Smirnov Kolmogorv con la significancia definida por el usuario.
res_prueba_ad	Es el resultado de evaluar el valor P de la prueba de Anderson Darling con la significancia definida por el usuario.

Ejemplo

```
#Se carga la librería COUNT para utilizar el conjunto de datos #azpro

library(azpro)
data(azpro)

#Se evalúa la variable "los" que contiene el tiempo de #estancia hospitalaria del conjunto de datos azpro.

probar<- test_dist (y=azpro$los)
```

---

pred_gamlss_tree	Realizar predicciones utilizando un Árbol de Modelos GAMLSS AMG

Descripción

pred_gamlss_tree utiliza la lista que genera la función gamlss_tree, que contiene la información del modelo ajustado, para predecir observaciones de un nuevo conjunto de datos.

Uso

```
pred_gamlss_tree(newdata,objeto)
```

Argumentos

newdata	Es un data.frame que contiene las variables explicativas de un conjunto nuevo de observaciones, que debe contener las mismas variables utilizadas en el ajuste del modelo.

objeto	Es la lista que genera la función **gamlss_tree** y que contiene toda la información del modelo generado en ésta.

Valores

La función retorna una **data.frame** que contiene los datos ingresados en el argumento newdata y se adicionan las siguientes columnas:

pred_arbol	Es el valor predicho por el árbol de decisión de la función **rpart**.
pred_gam	Es el valor final predicho por la combinación del árbol de decisión y los modelos GAMLSS en cada nodo.

Ejemplo

```
#Se carga la librería COUNT para utilizar el conjunto de datos

azpro
library(azpro)
data(azpro)

#Se ajusta el modelo GAMLSS

model_amg<-gamlss_tree(los~.,datos=azpro)

#Se predicen los valores de los para un conjunto de datos
con #base en el modelo ajustado

pred<-pred_gamlss_tree(objeto=model_gam,newdata=azpro)
```

---

split_sample separa una **data.frame** en dos con un muestreo aleatorio

---

Descripción

split_sample Esta función toma una data.frame y genera una lista con dos data.frame utilizando muestreo aleatorio para dividir las observaciones, el porcentaje de datos que contiene cada uno de los conjuntos creados es definido por el usuario como argumento de entrada.

Uso

```
split_sample(datos,perc)
```

Argumentos

datos	Es un data.frame que se quiere dividir en una muestra de entrenamiento y una deevaluación o prueba.
perc	Es un valor entre 0 y 1 que indica la proporción de datos que se dejarán en la muestrea de entrenamiento.

Valores

La función retorna una lista con dos data.frame que contienen la misma estructura de los datos ingresados, los dos data.frame generados son:

train	Es un data.frame con una muestra aleatoria de los datos que se utilizarán para entrenamiento.
test	Es un data.frame con una muestra aleatoria de los datos que se utilizarán para evaluación o prueba.

Ejemplo

```
#Se carga la librería COUNT para utilizar el conjunto de
datos

azpro
library(azpro)
data(azpro)

#Se dividen los datos en dos conjuntos

separados<-split_sample(datos=azpro,perc=0.8)

#Se genera una lista con dos data.frame en el que el
#data.frame train contiene el 80% de las observaciones.

separados$train
separados$test
```

Anexo 2

Variables disponibles en el conjunto de datos del caso de estudio 1 correspondiente a datos reales de un hospital colombiano.

Variables	Descripción	Tipo de variable
cod_habilitacion	Departamento tratante del paciente	Categórico
mes	Mes de egreso del paciente	Categórico
capitulo	Capítulo del diagnóstico de acuerdo con ICD 10	Categórico
los_dias	Tiempo de estancia en días	continua
tipo_egreso	Razón por la que egresó el paciente	Categórico
Edad	Edad del paciente	Discreta
semana_epidemio	Semana del año en que ingresó el paciente	Discreta
genero	Género del paciente	Categórico
antibiotico	Si el paciente recibe antibióticos	Categórico
servicio_egreso_e	Servicio en el que se encontraba cuando egresó	Categórico
cod_aseg_grd_e	Código de asegurador	Categórico
comorbilidad	Número de enfermedades diagnosticadas	Discreta
dia_mes_ingreso	Día del mes que ingresó el paciente	Discreta
dia_semana_ingreso	Día de la semana que ingresó el paciente	Discreta
especialidad_e	Especialidad del diagnóstico principal	Categórico
dx_principal_e	Diagnóstico principal de ingreso ICD 10	Categórico
dx_medico_e	Diagnóstico principal según médico tratante ICD 10	Categórico

Anexo 3

Variables disponibles en el conjunto de datos del caso de estudio 2 correspondiente a datos reales modificados en el marco de un proyecto desarrollado por Microsoft.

Variables	Descripción	Tipo de variable
eid	Identificación de la admisión	Discreta
vdate	Fecha de visita	Categórica
rcount	Número de readmisiones en los últimos 180 días	Discreta
gender	Género del paciente	Categórica
dialysisrenalendstage	Indicador de enfermedades renales durante la estancia	Discreta
asthma	Indicador de asma durante la estancia	Discreta
irondef	Indicador de deficiencia durante la estancia	Discreta
pneum	Indicador de neumonía durante la estancia	Discreta
substancedependence	Indicador de dependencia de sustancia durante la estancia	Discreta
psychologicaldisordermajor	Indicador de trastorno psicológico mayor durante la estancia	Discreta
depress	Indicador de depresión durante la estancia	Discreta
psychother	Indicador de otros trastornos psicológicos durante la estancia	Discreta
fibrosisandother	Indicador de fibrosis durante la estancia	Discreta
malnutrition	Indicador de malnutrición durante la estancia	Discreta
hemo	Indicador de desorden sanguineo durante la estancia	Discreta
hematocritic	Promedio valor hematocrito durante la estancia	Continua
neutrophils	Promedio valor neutrófilos durante la estancia	Continua
sodium	Promedio valor sodio durante la estancia	Continua
glucose	Promedio valor de glucosa durante la estancia	Continua
bloodureanitro	Promedio de nitrógeno ureico en la sangre durante la estancia	Continua
creatinine	Promedio del valor de la creatinina durante la estancia	Continua
bmi	Promedio del indice de masa corporal durante la estancia	Continua
pulse	Promedio del pulso durante la estancia	Continua
respiration	Promedio de la respiración durante la estancia	Continua
secondarydiagnosisnonicd9	Indicador de un segundo diagnóstico durante la estancia	Discreta
discharged	Fecha de alta	Categórica
facid	Departamento hospitalario de la estancia	Categórica
lengthofstay	Tiempo de estancia	Discreta

Anexo 4

Variables disponibles en el conjunto de datos del caso de estudio 3 correspondiente a una muestra aleatoria de un conjunto datos reales del seguro médico del estado de Arizona en Estados Unidos.

Variables	Descripción	Tipo de variable
LOS	Tiempo de estancia en días	Discrteta
Procedure	Indicador de procedimiento 1= CABG, 0 = PTCA	Discrteta
Sex	Indicador de sexo 1= Hombres, 0 = Mujeres	Discrteta
Age75	Indicador de mayor de 75 años, 1>75 años, 0 < 75 Años	Discrteta
Admit	Indicador de tipo de admisión 1= Urgencia, 0 = Electiva	Discrteta
Hospital	Código del departamento hospitalario que trata al paciente	Categórica

Anexo 5

Lista de distribuciones utilizadas en la ejecución de la metodología AMG en los tres casos de estudio.

Distribución	Nombre Distribución	Tipo distribución
BNB	Beta Negative Binomial	Conteo
DEL	Delaport	Conteo
DPO	Double Poisson	Conteo
GEOM	Geometric	Conteo
GEOMo	Geometric Original	Conteo
GPO	Generalized Poisson	Conteo
LG	Logarithmic	Conteo
NBF	Negative Binomial Family	Conteo
NBI	Negative Binomial Tipo I	Conteo
NBII	Negative Binomial Tipo II	Conteo
PIG	Poisson Inverse Gaussian	Conteo
PO	Poisson	Conteo
SI	Sichel Original	Conteo
SICHEL	Sichel (mu as the mean)	Conteo
WARING	Waring	Conteo
YULE	Yule	Conteo
ZALG	Zero Adjusted Logarithmic	Conteo
ZANBI	Zero Adjusted Negative Binomial	Conteo
ZAP	Zero Adjusted Poisson	Conteo
ZAPIG	Zero Adjusted PIG	Conteo
ZAZIPF	Zero Adjusted Zipf	Conteo
ZIBNB	Zero Inflated Beta Negative Binomial	Conteo
ZINBF	Zero Inflated Negative Binomial Family	Conteo
ZINBI	Zero Inflated Negative Binomial	Conteo
ZIP	Zero Inflated Poisson	Conteo
ZIP2	Zero Inflated Poisson (mu as mean)	Conteo
ZIPF	Zipf	Conteo
ZIPIG	Zero Inflated PIG	Conteo
ZISICHEL	Zero Inflated Sichel	Conteo
BCCGo	Box-Cox Cole and Green	Real positiva
BCPEo	Box-Cox Power Exponential	Real positiva
BCTo	Box-Cox t	Real positiva
exGAUS	The ex-Gaussian	Real positiva
EXP	Exponential	Real positiva
GA	Gamma	Real positiva
GB2	Generalized Beta Type II	Real positiva
GG	Generalized Gamma	Real positiva
GIG	Generalized Inverse Gaussian	Real positiva
IG	Inverse Gaussian	Real positiva

IGAMMA	Inverse Gamma	Real positiva
LOGNO	Log Normal	Real positiva
NO	Normal	Real positiva
PARETO2	Pareto Type 2	Real positiva
WEI3	Weibull (mu as mean)	Real positiva