Escuela Politécnica Superior de Linares

Grado en Ingeniería Telemática

Universidad de Jaén

Escuela Politécnica Superior de Linares

Trabajo Fin de Grado

DESARROLLO Y EVALUACIÓN DE UN MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA ESTUDIAR Y PREDECIR EL COMPORTAMIENTO DEL ABSENTISMO EN PRESTACIONES SANITARIAS.

Alumno: Pedro D. Romero Aceituno

Tutor: Carmen Martínez Cruz

Depto.: Informática (Lenguajes y Sistemas Informáticos)

Febrero, 2019

Trabajo de Fin de Grado en Ingeniería Telemática

DESARROLLO Y EVALUACIÓN DE UN MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA ESTUDIAR Y PREDECIR EL COMPORTAMIENTO DEL ABSENTISMO EN PRESTACIONES SANITARIAS.

Alumno: **Pedro D. Romero Aceituno**

Tutor: **Carmen Martínez Cruz**

VºBº a la defensa del TFG

Depto.: Informática (Lenguajes y Sistemas Informáticos)

Índice

1. RESUMEN.

El absentismo de pacientes en las prestaciones sanitarias es un problema importante en el sector sanitario a nivel internacional. El hecho de que los pacientes no atiendan a sus citas supone sustanciales pérdidas económicas, baja productividad en los proveedores de salud y problemas de acceso. También se relaciona con complicaciones en la salud de los pacientes y falta de adherencia en los tratamientos. Es por ello, que el desarrollo de nuevas herramientas e intervenciones que mejoren las actuales ratios de asistencia a las citas y reduzcan sus efectos negativos se antojan fundamentales para la mejora de una sanidad cada vez más costosa y con mayores retos.

En los últimos años se ha producido un aumento exponencial de los datos de salud disponibles motivado por la digitalización de historias clínicas, sistemas electrónicos de facturación, bases de datos genéticas o sistemas personales de monitorización de la salud. De igual forma se han desarrollado nuevas herramientas de análisis y aprendizaje automático que permiten obtener conocimiento e incluso hacer predicciones sobre los datos. En el presente trabajo se pretende abordar el problema del absentismo desarrollando una solución que usa tecnologías de aprendizaje automático o machine learning para mejorar las ratios de pacientes que atienden a sus citas y mitigar sus efectos negativos.

Uno de los mayores retos actuales que supone el uso de estas nuevas tecnologías es su incorporación a la práctica clínica real. Por tanto, el proyecto pretende desarrollar y evaluar una solución para el absentismo en un marco clínico real, el conjunto de clínicas Mission Neighborhood Health Center, organización de salud de atención primaria ubicada en San Francisco (EEUU). El objetivo principal será el desarrollo y evaluación de un modelo predictivo con el que conocer las probabilidades de atención de los pacientes a sus citas. Posteriormente, estas predicciones serán utilizadas para soportar un sistema de overbooking inteligente con el que generar recomendaciones de sobre citación y optimizar la planificación de citas de los centros.

2. INTRODUCCIÓN.

En este primer capítulo se hace una introducción del trabajo a alto nivel. En el siguiente apartado se realiza una descripción del proyecto, explicando las fases, introduciendo la tecnología a utilizar y arquitectura. En el segundo, se describe el contenido de la memoria y su estructura.

2.1. Descripción del proyecto

La no asistencia de pacientes a sus citas supone un grave problema en los sistemas sanitarios, impactando en los recursos económicos y en la salud de la población. Aplicaciones tecnológicas que hagan más efectivas las intervenciones actuales para reducir el absentismo y su impacto, supondrían un ahorro de costes, mejoras en el acceso y en la continuidad de cuidados de los pacientes. Lo expuesto supone la motivación principal para realizar el trabajo, en el cual se pretende desarrollar un modelo predictivo con el que conocer mejor los factores relacionados con el absentismo y predecir las probabilidades de asistencia de los pacientes a sus citas. Una vez construido el modelo se pretenden usar estas predicciones para alimentar un sistema de overbooking inteligente con el que mejorar las ratios de asistencia, productividad y uso de recursos en el entorno clínico real. Todo el proyecto se desarrolla en el marco de Mission Neighborhood Health Center, un conjunto de clínicas situadas en San Francisco, California.

En el trabajo se distinguen tres fases bien diferenciadas:

En primer lugar, se ha realizado una revisión sistemática de la literatura de referencia con el objetivo de conocer mejor la naturaleza del problema. Se han estudiado los factores que pueden explicar la no asistencia de pacientes a sus citas. Se han evaluado las intervenciones aplicadas en la actualidad para paliar el problema. Por último, se han revisado los trabajos anteriores de tecnologías analíticas y predicción aplicadas al problema.

Posteriormente, se ha creado un modelo predictivo para conocer las probabilidades de asistencia a la cita de nuevos pacientes y clasificarlos en “Atención” o “No Atención”. Esta fase, que es el núcleo principal del proyecto, consta a su vez de varias subfases donde se realiza la extracción, análisis y procesamiento de los datos, así como el entrenamiento y selección de los algoritmos de machine learning. La Figura 1 muestra de forma gráfica todo el proceso.

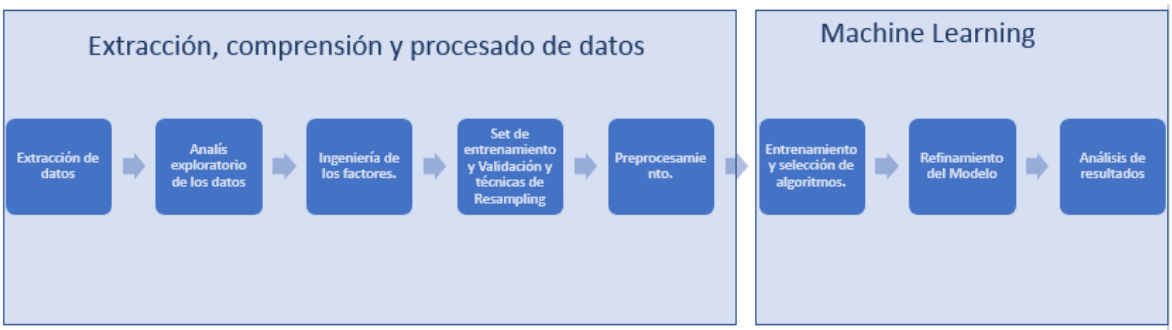


Figura 1.- Fases del proceso de construcción de modelos predictivos

En cuanto a los algoritmos predictivos utilizados en el proyecto, se ha apostado por la evaluación de un amplio espectro de los mismos con el fin de seleccionar con parámetros objetivos qué tipo de algoritmo se adapta mejor al conjunto de datos de nuestro problema. Por tanto, se han usado un gran número de algoritmos agrupados en tres categorías, lineales, no lineales y basados en árbol. Por cada tipología se han entrenado y evaluado los algoritmos, clasificándolos en función de su rendimiento y seleccionando los mejores en cada categoría para su implementación en la aplicación de overbooking. El lenguaje de programación elegido para dar soporte a todo el proceso anteriormente descrito es R, y su entorno de desarrollo RStudio.

Finalmente, se ha implementado una aplicación con el fin de evaluar la utilidad de los modelos predictivos en la práctica clínica real. La aplicación pretende realizar un sistema de recomendaciones para un overbooking inteligente en la programación de las clínicas. Para ello se leen datos del sistema de citación ubicados en la aplicación comercial Nextgen, a partir de estos se alimenta el sistema predictivo previamente creado que produce una serie de probabilidades y clasifica a las citas. Por último, se crean unas interfaces gráficas con el fin de evaluar la exactitud de las predicciones y de producir recomendaciones de overbooking fácilmente entendibles por los responsables de citación del centro. La Figura 2 ilustra a alto nivel la arquitectura del trabajo.

Puesto que los modelos predictivos se implementan usando el lenguaje R y la aplicación ha de ser distribuida, se ha decidido usar “Shiny” como framework para su desarrollo. Shiny es un paquete R que permite un entorno de trabajo para desarrollar web programadas en R, soportando a su vez HTML5/CSS3 y Javascript+ Node.js [47].

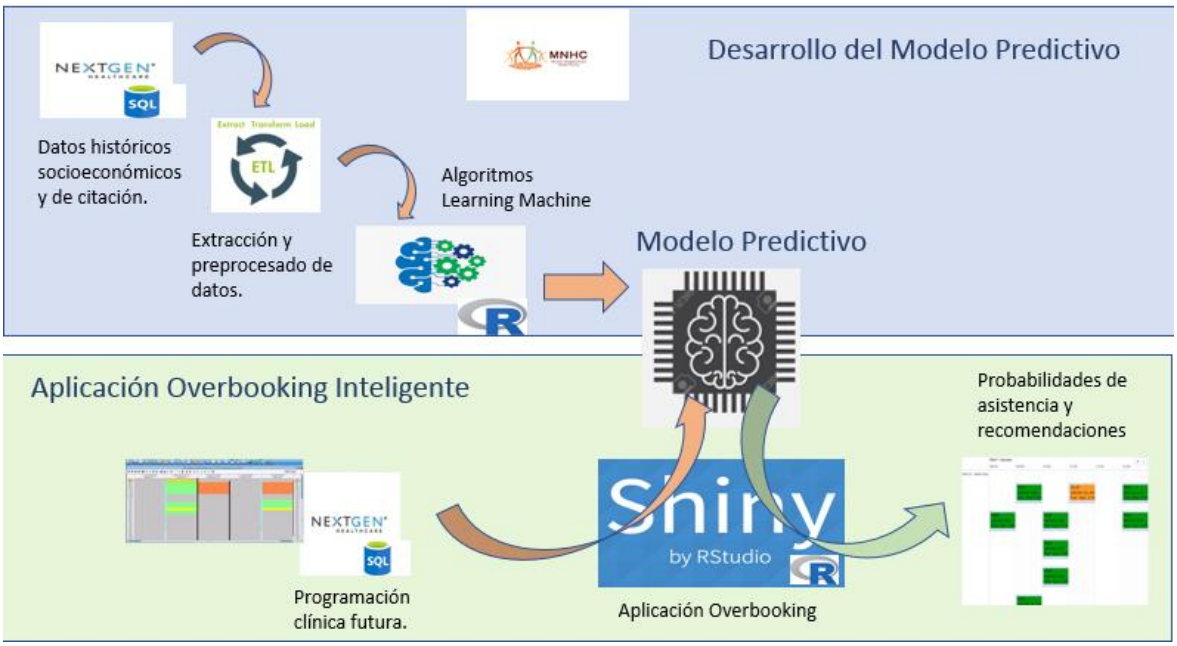


Figura 2.- Grafico resumen y arquitectura del proyecto.

2.2. Descripción de la memoria.

Una vez introducido el proyecto, vamos a describir cómo se estructura la memoria y qué se puede encontrar a lo largo de la misma.

Para comenzar, en el capítulo 3 se realiza un detallado estudio del problema a resolver, describiéndose el impacto, los factores asociado al absentismo y las intervenciones aplicadas actualmente y su efectividad.

En el capítulo 4 se describen los trabajos previos encontrados relacionados con el uso de tecnologías de machine learning para mejorar la asistencia de pacientes a sus citas.

En el capítulo 5 se definen los objetivos tantos principales como específicos relacionados con la elaboración del proyecto. Se incluyen también los objetivos docentes tenidos en cuenta.

Seguidamente en el capítulo 6, el más extenso de la memoria, se explican los materiales y métodos utilizados durante el desarrollo del proyecto. Este capítulo tiene tres partes: una primera donde se describe el estado del arte de las tecnologías utilizadas y la legislación vigente; la segunda donde se desarrolla todo el proceso de construcción del modelo predictivo; una tercera donde se diseña e implementa un prototipo para una aplicación de overbooking inteligente.

En la sección 6.1 se presentan el estado del arte, donde se describen las tecnologías utilizadas, el lenguaje de programación R y el entorno RStudio. También se desarrolla el estado actual de los algoritmos de Machine Learning, algunas técnicas de procesamiento de los datos y las métricas de evaluación de los resultados. En la sección 6.2 se describe a alto nivel la ley de protección de datos estadounidense HIPAA, que establece el marco legislativo de este proyecto.

La sección 6.3 es el núcleo del proyecto donde se desarrolla todo el ciclo de vida de la construcción de los sistemas predictivos creados. Desde la extracción del conjunto de datos, su análisis, la ingeniería de los factores predictivos, tratamiento de los datos, resampling, entrenamiento y selección de algoritmos, tunning de los parámetros y estudio de la importancia de los factores.

La última parte del capítulo 6, trata sobre el desarrollo de una aplicación que usa el modelo predictivo construido para realizar predicciones sobre datos nuevos e implementar un sistema de recomendación de overbooking basado en la probabilidad de asistencia.

Los resultados obtenidos serán expuestos y discutidos en el capítulo 7. Además de describir las métricas de rendimiento de los modelos y las predicciones sobre nuevos datos, se explicarán los mayores retos encontrados para el desarrollo del proyecto y se justificarán las decisiones tomadas en el proceso de desarrollo.

Por último, en el capítulo 8, se hará un resumen de las conclusiones del trabajo y líneas de trabajo futuras.

3. JUSTIFICACIÓN Y ESTUDIO DEL PROBLEMA.

En este capítulo se desarrolla una descripción del absentismo y sus consecuencias en el sector. Se investiga sobre distintos factores asociados a la no asistencia de los pacientes a sus citas, las intervenciones actuales y su efectividad para mitigar el problema.

3.1. Descripción del problema e impacto en sanidad.

Se define el absentismo de pacientes en la práctica clínica, conocido como no-show por su terminología anglosajona, como aquellos pacientes que no acuden a un evento previamente programado con un proveedor de salud, bien sea consulta, procedimiento diagnostico etc. El problema del absentismo se intensifica si el paciente no avisa al centro médico o proveedor de servicios con la antelación suficiente para completar el tiempo asignado con otro paciente.

El primer y más obvio problema es el malgasto de tiempo y recursos. El coste del cuidado aumenta, así como la productividad de los proveedores de salud disminuye. Cuando la tasa de absentismo es alta, aumenta la dificultad en la planificación de recursos. En el sector privado la no asistencia de pacientes impacta directamente en los beneficios de la organización.

Existen algunos estudios que con mayor o menor rigurosidad hacen una estimación económica de las pérdidas originadas por el absentismo en Sanidad. Kheirkhah P et al. [1] estiman las pérdidas de media por paciente en $196 en un conjunto de clínicas de Veteranos en EEUU en 2008 siendo el coste anual de $14.58 millones. Un estudio similar realizado para unas clínicas de endoscopias en University of North Carolina [2] estimó las pérdidas diarias en una media de 725$ casi un 20% del total de las ganancias. En nuestro país, un estudio realizado en la agencia sanitaria costa del Sol [3] determinó que el coste económico fue superior a 3 millones de euros anual para una tasa de inasistencia del 13,8%. A nivel macro Beecham L [4] estimaba los costes derivados de la no asistencia de pacientes a sus citas en ($240 millones) en el NHS británico en el año 2000.

La no asistencia a las citas agudiza los problemas relativos al acceso a las prestaciones sanitarias. La pérdida de eficiencia producida por el absentismo aumenta las listas de espera, las esperas durante la prestación de los servicios médicos y la insatisfacción del paciente en general [5] [6] [7] [8].

Por último, el absentismo está relacionado con peores resultados en salud. Pacientes que no atienden sus citas presentan problemas en la continuidad de sus cuidados, poca adherencia a los tratamientos, escaso compromiso con los servicios médicos y pobres niveles de autocuidado. Existen numerosos estudios [9] [10] [11] [12] que relacionan el abandono de las citas con incrementos en la morbilidad de los pacientes y peor control de enfermedades crónicas y mentales.

En resumen, la no asistencia a las citas tiene un fuerte impacto en el sector sanitario, causando pérdidas económicas, empeorando la planificación y acceso a los servicios sanitarios y disminuyendo la calidad de los cuidados, así como los resultados de salud de los pacientes.

3.2. Factores asociados al absentismo.

Son diversos los estudios y metodologías recogidas en la bibliografía especializada destinados a conocer los factores asociados a la no asistencia de los pacientes a sus citas. Por un lado, están los estudios que realizan encuestas a los pacientes que no asisten a sus citas con el fin de conocer las causas de este comportamiento. Por otro, aún más comunes son los estudios retrospectivos que examinan conjuntos de datos extraídos de los sistemas de información de centros médicos y aplican técnicas estadísticas para determinar qué factores explican el absentismo.

Entre las causas auto declaradas por los pacientes se encuentran el no tener constancia de la cita, el olvido, problemas de comunicación con el centro, mejoría de los síntomas o problemas logísticos como atender a trabajo o tener cargas familiares [5] [11] [13].

Los factores más comúnmente identificados como causas asociadas a la no asistencia de citas son factores demográficos y socioeconómicos. Así, la edad, genero, raza, nivel de ingresos, desempleo o aseguramiento están relacionados con el absentismo [1][6][9]. La situación geográfica, ruralidad, distancia a la clínica, método y costes de transporte son otros factores a tener en cuenta [7][14]. La etnicidad, tratada por Nancarrow S et al. [10], también se asocia a un mayor riesgo de no asistencia en algunos países.

Existen factores atribuibles a los centros de salud y su operativa. El tiempo de espera a la cita es causa directa del aumento del riesgo de no-show [1] [8] [10]. El día de la semana o la época del año también explican fluctuaciones en las ratios de no asistencia. El tipo de

servicios ofrecido en el centro médico también es factor diferencial del riesgo de no-show [9].

El comportamiento previo del paciente y su historia de absentismos ha sido señalado como un factor clave en numerosos estudios realizados, véase Chang JT et al. [6] o Mani J et al. [7]. También son factores significativos si el paciente es nuevo o ya está establecido en el centro o el número de citas que tiene pendiente el paciente.

El estado de salud del paciente también determina el riesgo de no asistencia a la cita, siendo factores el diagnostico, tipo y/o dificultad del procedimiento. Pacientes con diagnósticos de salud mental tienen un riesgo mayor de no asistencia, así como aquellos con historial de abuso de consumo de sustancias [7] [11] [15]. La percepción de salud ha sido estudiada como factor de riesgo, aunque sin resultados significativos por Thongsai S et al.[16].

3.3. Intervenciones y su efectividad.

Para paliar el fuerte impacto económico y en la salud de los pacientes que representa el absentismo en las citas sanitarias se han desarrollado múltiples intervenciones con diferentes enfoques, efectividad y coste. Así las intervenciones pueden dividirse en aquellas que intentan modificar el comportamiento de los pacientes como educación y recordatorios por carta, llamadas telefónicas o sistemas automáticos por SMS. Mientras que otras intervenciones intentan mejorar la operativa de citación de los pacientes, como las estrategias de acceso avanzado facilitando la elección por parte del paciente del día y hora de la cita. Dentro de esta última clasificación se puede incluir la sobre citación (u overbooking en inglés) de pacientes como medida para paliar la infra utilización de recursos provocada por la no asistencia a las citas. En los siguientes párrafos se describirán las distintas intervenciones y su efectividad.

Existen distintos estudios que demuestran que el olvido o carencias en la comunicación centro médico – paciente es uno de los factores explicativos de la no asistencia a las citas. Por tanto, intervenciones relacionadas con el recordatorio del día y hora de la cita se han demostrado efectivas para reducir el número de no-shows. Uno de los primeros enfoques, que aún se sigue utilizando en numerosos sistemas de salud es el envío de cartas vía servicio postal. En un estudio publicado en el BMC Psychiatry [21] se demuestra la efectividad del envío de cartas para la reducción de la ratio de absentismo.

Según Shah SJ et al. [22] los recordatorios vía llamada telefónicas 7 y 2 días antes de la cita reducen de forma considerable el número de pacientes que no atienden su cita, así como mejoran el beneficio económico. En otros dos trabajos revisados [18] [23] se expresa que un sistema de recordatorios telefónico puede evitar el absentismo en las visitas médicas. En ambos estudios se concluye que la efectividad de estas intervenciones puede mejorarse si se realiza una estratificación de pacientes con mayor probabilidad de no-show. Por tanto, si se destinan los recursos de forma más intensiva en pacientes con determinadas características, mejora la efectividad y coste-efectividad de la intervención.

Diversos estudios [17] [24] [25] [26] [28] tratan la efectividad de los mensajes de texto automáticos para la reducción de la ratio de pacientes que no atienden a sus citas. La efectividad varía dependiendo de varios factores como el tipo de atención, el procedimiento, el número de mensajes [19], el contenido de los mensajes [27] o focalizar la intervención en función de la probabilidad de no asistir de los pacientes [20]. En cuanto a la efectividad comparada con otras estrategias como llamadas o cartas, existe evidencia, aunque limitada, de que los mensajes automáticos tienen la misma efectividad o mayor que el resto de intervenciones y un coste-efectividad menor que las llamadas telefónicas.

En algunas organizaciones sanitarias se están desarrollando estrategias de acceso avanzado que permiten al paciente seleccionar día y hora de la cita, mejoran el tiempo de espera (factor de impacto en el absentismo) y la comunicación con el paciente. Sin embargo, no hay evidencia de los efectos en la satisfacción del paciente y en los resultados clínicos.

Por último, otra intervención muy común para paliar el efecto del absentismo en el sector sanitario es el “overbooking”. Esta práctica consiste en citar más pacientes de la teórica capacidad diaria del centro, consiguiendo minimizar los efectos derivados de los no-show en uso de recursos, productividad y beneficio económico. Existe evidencia de que este tipo de intervenciones reduce perdidas e infrautilización de recursos, pero tiene asociado un coste en tiempo de espera de los pacientes y malestar de los proveedores clínicos al tener que hacer frente a la sobre utilización aquellos días en los que todos los pacientes asisten. Existen estrategias que combinan sistemas predictivos con el overbooking para realizar una sobre programación inteligente en función de las probabilidades de asistencia de los pacientes, de esta forma se consigue optimizar el uso de recursos minimizando tiempos de espera y el malestar de proveedores [29].

4. ANTECEDENTES.

Se ha realizado una selección de artículos publicados en revistas y bases de datos especializadas para analizar la existencia y los resultados de trabajos previos en los que se usa sistemas de machine learning para abordar el absentismo de pacientes. En resumen, podemos concluir que los trabajos encontrados son teóricos, reducidos a un solo centro o universidad, y con una implementación y evaluación en la práctica clínica reducida. No se ha encontrado ninguna aplicación comercial que aborde el problema estudiado. Existe alta variabilidad en las tecnologías empleadas y en los resultados, aunque la mayoría de los autores concluyen como positivo o potencialmente útil la aplicación de este tipo de tecnología para mejorar las intervenciones actuales. En el resto del capítulo se detallan los hallazgos de la revisión realizada y al final del mismo en la Tabla 1 es posible encontrar un resumen comparativo de los artículos.

Tras la revisión de los artículos se presenta un resumen descriptivo de los resultados obtenidos atendiendo a las diferentes variables estudiadas como el país, tipo de estudio, objetivo, factores asociados al absentismo, intervenciones, fuentes de datos, tecnología y resultados.

Según el origen de los estudios, Norte América es el continente más prolífico a la hora de emprender investigaciones e intervenciones relacionadas (65% de los artículos revisados), siendo Estados Unidos el país que más estudios produce. Este hecho está en consonancia con datos de otras industrias, donde el país americano es pionero siendo Google o Amazon ejemplos de empresas líderes que han alcanzado ventajas competitivas mediante el uso de estas tecnologías. Por detrás están Asia y Europa.

El diseño de los estudios es variado, predominando la construcción y validación de modelos estadísticos predictivos utilizando datos retrospectivos almacenados en las bases de datos clínicas y administrativas de los centros donde se estudia la asistencia de pacientes a sus citas. En dos de los estudios evaluados Cronin PR [30] y Percac-Lima S [31] validan las intervenciones asociadas mediante sendos ensayos clínicos aleatorios para confirmar el aumento en la efectividad de las mismas.

Atendiendo a los factores explicativos del absentismo, el denominador común en casi todos los trabajos es usar variables sociodemográficas asociadas al paciente como edad, sexo, raza, etnicidad, nivel de ingresos, grado de educación, ruralidad, distancia al centro, tipo de aseguramiento etc. Variables relacionadas con la cita como el día, hora, mes y sobre todo el tiempo de espera hasta el día de la cita se muestran con un alto valor explicativo en

los modelos. Otro de los factores más utilizado para explicar el absentismo en los estudios es el comportamiento previo de los pacientes y su historia previa de absentismo a las citas. Distintos autores desarrollan diferentes métricas para incluir este factor en los modelos, destaca el trabajo de Blumenthal DM et al. [34] donde se construye una variable basada en el análisis del lenguaje natural de historias clínicas electrónicas para modelar el comportamiento pasado de los pacientes, además se concluye que la inclusión de esta variable mejora en la precisión del modelo.

Algunos modelos incorporan datos clínicos para predecir la probabilidad de asistencia a la cita. En el último trabajo mencionado [34], también se incluye el historial previo de problemas psiquiátricos, factor relacionado con la no asistencia en la bibliografía. En el trabajo de Reid MW et al. [32] se incluyen el diagnóstico clínico e historial de abuso de sustancias. Kurasawa H et al. [37] incorpora en el modelo variables clínicas de control de la diabetes y medicación. La comorbilidad de los pacientes es introducida en el modelo propuesto por Daggy J et al. [38] a través del índice de Charlson. En el estudio de Woodward B et al. [33], la herramienta predictiva se basa en un modelo previo de predicción usando parámetros clínicos de VIH. Todas las variables comentadas han sido estadísticamente significativas para explicar la atención de pacientes a las citas en los diferentes modelos utilizados.

Las fuentes de datos más utilizadas en los estudios revisados son las historias clínicas electrónicas (HCE), así como las bases de datos administrativas, de registros y facturación. Se observa una tendencia a la confección de bases de datos clínicas para fines específicos y un patrón común es la combinación de varias bases de datos heterogéneas para extraer información y construir modelos más precisos.

Aunque aparecen algunos estudios con una población pequeña, la nota predominante es que las cohortes estudiadas superen los cientos de miles o millones de registros, refiriéndose en ocasiones a poblaciones enteras, siendo esta una de las ventajas en las estrategias de Big Data. Otra característica de los estudios es la alta dimensionalidad de las muestras, así por cada uno de los pacientes (o citas) se tienen en cuenta numerosas variables para el estudio.

En cuanto a las tecnologías más utilizadas para confeccionar los modelos predictivos, casi todos los estudios comienzan con técnicas de estadística clásica para describir la población, estudiar el peso y significancia de los distintos factores incluidos en los modelos. En cuando a las técnicas de predicción, los modelos de regresión son los más utilizados (90% de los estudios). En uno de los artículos Devasahay SR et al. [35] se realiza una comparativa

entre varios algoritmos de predicción: modelos de regresión, Support Vector Machine y árboles de decisión sin un resultado concluyente. Adel Alaeddini et al. [39] construyen un modelo hibrido con regresión logística e inferencia bayesiana concluyendo mejores resultados que usando ambas técnicas por separado. Finalmente se han encontrado otros dos trabajos [40] [41] que utilizan otros algoritmos predictivos como arboles de decisión o el gradient boosting algorithm (gbm).

El comportamiento de los modelos predictivos varía en función del estudio, aunque la mayoría de los autores declara la capacidad de los mismos para predecir de forma razonable la probabilidad de absentismo de los pacientes. Solo en un estudio [35] los autores concluyen que no es posible realizar predicciones solidas con el modelo construido. No existe un patrón homogéneo para describir el comportamiento de los modelos, siendo la métrica más común el AUC o ROC (área bajo la curva) variando entre el 0.60 y el 0.8 en función de los estudios. Otros parámetros de calidad del modelo como la precisión o sensibilidad son especificados en algunos trabajos con valores muy heterogéneos dependiendo de cada estudio en concreto.

La mitad de los estudios evaluados usan los resultados de los sistemas predictivos para aumentar la efectividad de intervenciones relacionadas con el absentismo en la práctica clínica. Así, tanto Cronin PR [30], como Reid MW et al. [32] utilizan el riesgo de no asistencia a las citas para construir un sistema de overbooking, en ambos casos el estudio concluye una mejor utilización de los recursos sin penalizar los tiempos de espera de los pacientes o sobrecargar a los proveedores de salud. En el estudio de Percac-Lima S et al. [31] se usan las probabilidades del modelo para estratificar por riesgo de no asistencia a los pacientes dedicando así más recursos en un sistema de gestión de casos a estos pacientes con más riego. En un trabajo realizado en la Veterans Health Administration [36] se mejora una clásica intervención de recordatorios mediante llamadas telefónicas estratificando pacientes por riesgo de no asistencia a sus citas. Por último, Daggy J et al. [38] desarrollan un sistema de programación de citas en función de los resultados de un modelo predictivo.

Las limitaciones encontradas en los estudios responden de un lado a carencias en los datos por problemas en la heterogeneidad o fiabilidad de las fuentes, y la indisponibilidad de datos para incluir nuevos factores potencialmente explicativos del absentismo a los modelos. De otro lado, una limitación declarada en los estudios es su validación externa, bien sea por ser realizados en un tipo de población muy específica, o bien por el diseño en los modelos estadísticos que presentan deficiencias fuera de los datos usados para su construcción y test.

Atendiendo a los resultados y conclusiones declaradas por los autores en los diferentes estudios, en todos ellos (excepto uno [35]) se encuentra evidencia que los modelos construidos pueden realizar solidas predicciones de la probabilidad de no asistencia de los pacientes a sus citas médicas. En los trabajos donde existen intervenciones asociadas a los resultados de la predicción o estratificación de los pacientes por el riesgo de no asistencia, todas las intervenciones aumentan su efectividad y mejoran la ratio de absentismos o utilización de recursos en los centros donde se implementan. Los resultados sobre las intervenciones son evaluados mediante ensayos clínicos aleatorios [30] [31] o estudios observacionales retrospectivos [32] [36] [38].

La Tabla 1 a continuación resume los hallazgos encontrados en los trabajos previos revisados sobre tecnologías de machine learning en el campo del absentismo de prestaciones sanitarias.

Ref	Titulo	Organización	Factores	Intervención	Resultados del modelo	Tecnología
30	Success of automated algorithmic scheduling in an outpatient setting.	Department of Medical Dermatology at Massachusetts General Hospital (MGH)	· Socio-Demográficos. · Tipo de cita y tiempo de espera. · Comportamiento histórico del paciente	Sistema inteligente de overbooking basado en recomendaciones de un sistema predictivo	· AUC o c-statistic (0.71) usando 10.800 citas en el dataset de testing. · Validación externa a través de un ensayo clínico asociado a la intervención.	Modelo predictivo basado en regresión lineal.
31	Patient navigation based on predictive modeling decreases no-show rates in cancer care.	Massachusetts General Hospital (MGH)	· Edad, días de espera, tipo de aseguramiento y comportamiento histórico del paciente en la asistencia en los últimos 5 años.	Patient Navigation. Sistema basado en llamadas de un case manager para gestionar el acceso a los cuidados en pacientes con altas probabilidades de no atender a sus citas médicas.	AUC o c-statistic (0.80) * (Se ofrecen pocos detalles sobre el modelo.) Evaluación de los modelos con un estudio aleatorio	Modelo predictivo basado en hierarchical logistic regression.
32	Preventing patient absenteeism: validation of a predictive overbooking model.	VA Greater Los Angeles Healthcare System	· Datos sociodemográficos. · Historial de visitas. · Diagnósticos clínicos. · Historial de Abuso de sustancias.	Overbooking siguiendo las recomendaciones de un sistema predictivo	· 0.7496 ROC en el modelo prospectivo. · sensitivity, specificity, PPV, and NPV of 36%, 98%, 73%, and 86%	multivariable logistic regression model
33	Risk Prediction Tool for Medical Appointment Attendance Among HIV-Infected Persons with Unsuppressed Viremia.	Vanderbilt Comprehensive Care Clinic	· Datos sociodemográficos. ·Año de inicio en cuidados de VIH. · Riesgo HIV fallo de carga viral.	No hay una intervención asociada	No se detallan los datos del modelo.	Estadística clásica y Regresión lineal
34	Predicting Non-Adherence with Outpatient Colonoscopy Using a Novel Electronic Tool that Measures Prior Non-Adherence.	Massachusetts General Hospital	Género, historia de problemas psiquiátricos, tiempo de espera, nº de anteriores visitas, nivel de educación y NAR	No hay una intervención asociada	(AUC= 70.2%). (AUC = 64.3 %) en modelos sin el factor construido (NAR)	Regression model.
35	Predicting appointment misses in hospitals using data analytics.		Sociodemográficos. De acceso al centro. Tiempos de espera. Aseguramiento.	No hay una intervención asociada	Se reportan bajos resultados de precisión, sensibilidad y	Logistic Regression, Support Vector Machine y árboles de decision

					especificidad en todos los modelos.	
36	Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior in the Veterans Health Administration.	Veterans Affairs health care facilities	Sociodemográficos Datos sobre la cita. Historial de asistencia.	Recordatorios mediante llamadas telefónicas 24, 48 y 72 horas para aquellos pacientes seleccionados como alto riesgo de no asistencia.	ROC (average) 0.712. Validación teórica sobre test dataset. Validación externa sobre una intervención.	Regresión logística, Redes Neuronales, Modelos de Árboles.
37	Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients With Diabetes.	University of Tokyo Hospital	Datos demográficos del paciente, día de la semana de la cita, comportamiento previo y variables clínica de control de la diabetes y medicación.	No existe una intervención asociada	AUC 0.958, precisión 0.757 y recall 0.659	Regresión logística.
38	Using no-show modeling to improve clinic performance.	Midwestern Veterans Affairs (VA) hospital	Datos sociodemográficos. Comorbilidad. Índice de Charlson. ratios de asistencia históricos. Detalles de la cita, día, tiempo de espera, hora, mes.	Sistemas de programación de citas en función de los resultados de un modelo predictivo.	ROC 0.82. Validación externa con un Montecarlo Model.	Regresion Model
39	A hybrid prediction model for no-shows and cancellations of outpatient appointments	Wayne State University	Se usan variables demográficas asociadas al paciente, así como comportamientos anteriores (no-show), día de la semana, clínica.	Se declaran usos futuros, pero no existe una intervención asociada al estudio.		Regresión logística. Bivariate Dirichlet distribution. Bayesian Inference.
40	Machine-Learning-Based No Show Prediction in Outpatient Visits	Madrid, the San Carlos Clinical Hospital	Demográficos. Detalles de la Cita. Asistencia previa.	No existe una intervención asociada	AUC 0.7404. Val externa solo en Test Data set.	Gradient Boosting algorithm
41	Predictive Analytics for Outpatient Appointments	Khoo Teck Puat Hospital, Singapore	Demográficos (Edad, Raza, Sexo) Día y hora de la cita, Departamento especialidad, Distancia a la clínica, no show historial de asistencia, SMS response, deuda del paciente, medico.	No existe una intervención asociada	No están descritos todos los parámetros. 60-65% Precisión. Eval externa solo con datos de test.	multivariate analysis, multiple logistic regression model, Recursive partitioning/ Decision Tree model.

Tabla 1.- Descripción de los estudios previos analizados.

5. OBJETIVOS.

Objetivo General: El objetivo principal del proyecto es evaluar si es posible intervenir eficazmente sobre los problemas derivados del absentismo de citas médicas usando tecnologías de aprendizaje automático o machine learning.

Objetivos Específicos: El objetivo principal se descompone a su vez en una serie de objetivos secundarios:

- Mejorar el conocimiento del absentismo, consecuencias, factores que lo producen e intervenciones eficaces aplicadas para paliar el problema.
- Determinar si es posible predecir la no asistencia a las citas.
- Evaluar y clasificar la eficacia de los distintos algoritmos predictivos.
- Estudiar y valorar la aplicación de los resultados teóricos a la práctica clínica real.
- Aplicar los resultados a un problema de negocio concreto. Grupo de clínicas MNHC.

Objetivo Docente.

Desde el punto de vista docente, el objetivo es aprender y presentar un conjunto de tecnologías muy novedosas de predicción basadas en el aprendizaje automático supervisado de datos y su aplicación en un problema de negocio concreto.

6. MATERIALES Y METODOS.

En el presente capítulo, se explican las tecnologías y herramientas utilizadas, así como el proceso de construcción de las aplicaciones del proyecto. El capítulo tiene tres partes diferenciadas: Una primera más teórica donde se describe el estado del arte de las tecnologías utilizadas (6.1) y la legislación vigente que constituye el marco del proyecto (6.2). La segunda parte, eminentemente práctica donde se desarrolla todo el proceso de construcción del modelo predictivo para el problema del absentismo en las clínicas de MNHC (6.3). Finalmente, una tercera parte, también práctica, donde se diseña e implementa un prototipo para una aplicación de overbooking inteligente usando los resultados del modelo anterior (6.4).

6.1. Estado del Arte

Antes de entrar de lleno en el desarrollo del proyecto, es conveniente introducir a alto nivel el estado actual de las tecnologías utilizadas. En esta sección se define el aprendizaje automático o Machine Learning y se introducen los principales algoritmos predictivos y las métricas para su evaluación. Posteriormente se introduce R, el principal lenguaje de programación R usado en el proyecto.

6.1.1. Machine Learning.

El aprendizaje automático, o machine learning en inglés, es un subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Se pretende por tanto, a partir del estudio de los patrones encontrados en los conjuntos de datos de ejemplo, inducir conocimiento.

Atendiendo a una primera clasificación básica, existe un tipo de aprendizaje denominado no supervisado donde un modelo se crea a partir de variables de entrada sin existir un conocimiento a priori del outcome (o valor a predecir). Aplicaciones como la compresión de datos o el clustering son típicas de este tipo de algoritmos. Por otro lado está el aprendizaje supervisado, en el cual un algoritmo es entrenado proporcionando

información sobre los valores reales observados además de las variables de entrada. El objetivo es crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada no visto previamente, generalizando a partir de la información recogida en el entrenamiento. Si el valor a predecir es una variable continua hablamos de un problema de regresión, mientras que si el outcome a predecir es una variable categórica el problema a resolver se denomina clasificación.

El desarrollo de algoritmos predictivos está en plena actualizad y existe un gran auge en este campo y el número de algoritmos disponible para resolver problemas de clasificación es cuantioso. A continuación, se van a describir algunos de los más importantes utilizados en el proyecto. Por no extender demasiado el tamaño de la memoria las descripciones serán concisas y de alto nivel, remitiendo a los lectores a la bibliografía para una detallada explicación matemática de los algoritmos, el libro de Trevor Hastie [43] es un recurso interesante. Se va a seguir la clasificación propuesta en el libro de Max Kuhn, Applied Predictive Modeling [42] que agrupa los algoritmos en Clasificadores Lineales, clasificadores No Lineales y clasificadores basados en modelos de árboles y reglas. El libro anterior, así como artículos publicados en la web [44] son usados como referencia para la descripción de los algoritmos en los siguientes apartados.

6.1.2. Algoritmos de Clasificación Lineal.

En esta sección se describen aquellos modelos en los que las funciones de dependencia entre los factores usadas para tomar la decisión de clasificación son lineales.

- **Regresión Logística.**

Es uno de los modelos lineales básicos y más populares debidos a su simplicidad y a la habilidad para producir intervalos de inferencia sobre los factores del modelo. La regresión logística se usa tanto para problemas de regresión como de clasificación, especialmente en variables categóricas binarias. El algoritmo de regresión logística trata de encontrar los parámetros que maximizan la función de probabilidad binomial. Una función sigmoidal es usada para establecer las probabilidades y limitarlas entre 0 y 1. Existen numerosas técnicas de regresión logística llamadas en su conjunto Generalized Linear Models (GLM).

Otro popular método es el LDA Linear Discriminant Analysis Fisher (1936) and Welch (1939). Para el problema de clasificación, Welch intento minimizar la probabilidad total de

subclasificación, la cual depende de las probabilidades de las clases y de distribución multivariable de los predictores. Fisher por su parte, formulo el problema de otra forma encontrando combinaciones lineales de los predictores tal que la varianza entre grupos fuera maximizada.

Como se ha comentado, la regresión logística permite realizar test de hipótesis sobre si los factores asociados a los predictores son estadísticamente significativos, pudiéndose establecer un ranking de la importancia de los predictores.

- **Modelos de penalización.**

Algunos modelos de clasificación utilizan penalización o regularización para mejorar el ajuste a los datos. Uno de los métodos más populares es el glmnet, Lasso y Elastic-Net Regularized Generalized Lineal Models. Este modelo usa regresión ridge y lasso simultáneamente y una estructura de penalizaciones. El parámetro α determina la proporción de cada modelo, así cuando tiene valor uno el modelo es de lasso puro, mientras que cuando tiene valor cero es regresión ridge. Otro parámetro importante del modelo es λ, el cual controla la cantidad de penalización en cada modelo.

- **Nearest Shrunken Centroids**.

The Nearest Shrunken Centroids, también conocido como PAM por predictive análysis for microarrays, es un modelo de clasificación lineal que funciona bien para modelos con alta dimensionalidad. Para cada clase, un centroide de los datos es calculado tomando el valor medio de cada predictor (por clase) en el set de datos de entrenamiento. Un centroide general es calculado usando los datos de todas las clases. Un método de clasificación de nuevos datos seria encontrar el centroide más cercano en todo el espacio dimensional y usar esa clase como predicción.

6.1.3. Algoritmos No lineales.

Existen numerosos modelos que son inherentemente no lineales en la naturaleza, cuando se usan estos modelos la forma exacta de no linealidad no suele ser conocida a priori. En esta sección se describen modelos capaces de aprender patrones no lineales entre los datos como son las Redes neuronales, regresión adaptativa multivariante (MARS), Support Vector Machine (SVM) o Nearest Neighbors (KNN) o Naive Bayes.

- **Redes Neuronales.**

Las redes neuronales son un potente método de regresión no lineal inspiradas en cómo funciona el cerebro. Las clases a predecir son codificadas como variables binarias y estas son modeladas por un conjunto intermedio de variables no observadas llamadas capas ocultas. Cada una de esta capa oculta es una combinación lineal de todas o de algunas de las variables utilizadas para predecir el outcome. Existen distintos tipos de redes neuronales en función de la estructura de sus capas ocultas, existiendo redes multicapa o incluso redes donde las capas se retroalimentan formando bucles bidireccionales.

Las redes neuronales en problemas de clasificación tienen gran potencial para el overfitting. Para combatir este problema se usan diferentes estrategias: al optimizar la entropía, se usa un parámetro de penalización (wight decay) para conseguir límites de clasificación más suaves; otra forma es realizar la estimación de la probabilidad media entre redes y usar estas para predecir las clases.

- **Support vector machines (SVM)**

Los Support vector machine son un conjunto de modelos estadísticos desarrollados a mediados de los 60 por Vladimir Vapnik. En años posteriores, el modelo se ha convertido en una de las herramientas de aprendizaje automático más flexible y utilizadas actualmente. El funcionamiento del SVM es construir un hiperplano en un espacio dimensional superior para lograr la separación de las clases, se logra una buena separación mediante el hiperplano que tiene la mayor distancia a los puntos de datos de entrenamiento más cercanos de cualquier clase; cuanto mayor sea el margen, menor será el error de generalización del clasificador.

- **K-Nearest Neighbors (KNN)**

K-Nearest Neighbors usa las distancias entre muestras para realizar predicciones. Para clasificar nuevas muestras KNN utiliza las K muestras más cercanas del conjunto del entrenamiento. El concepto de proximidad está determinado por una métrica de distancia (ej. Euclídea) que se determina antes de entrenar el modelo y de la que depende las características del predictor. Es importante tener en cuenta que las escalas de los predictores pueden impactar en los cálculos de las distancias y en la importancia de las variables.

- **Naive Bayes.**

El modelo de clasificación de Naive Bayes se basa en el teorema de la probabilidad condicional de Bayes y determina la probabilidad de una clase en función de la probabilidad de los predictores observados. El modelo utiliza los datos de entrenamiento para estimar las probabilidades condicionadas y con estas establecer predicciones sobre nuevas muestras.

El modelo Naive Bayes simplifica las probabilidades de los valores de los predictores al suponer que todos los predictores son independientes de los otros. Para la mayoría de las aplicaciones, sería difícil afirmar que esta suposición es realista. Sin embargo, el supuesto de independencia produce una reducción significativa en la complejidad de los cálculos, sin impactar gravemente en el rendimiento de los modelos.

6.1.4. Algoritmos basados en árbol.

Algunos de los algoritmos de clasificación más usados actualmente son de la familia de los basados en arboles de decisión y consisten básicamente en estructuras if-then más o menos complejas. Algunos beneficios de estos algoritmos es que pueden ser altamente interpretables y son robustos a diversidad de formatos y alteraciones en los datos de entrada. Por el contrario, en ocasiones no son muy estables o pueden concurrir en soluciones subóptimas.

- **Classification trees (CART, and C4.5)**

Un árbol de decisión o de clasificación es un árbol en el que cada nodo interno (no hoja) está etiquetado con una función de las variables de entrada, mientras que cada hoja del árbol se marca con una clase. El objetivo de los árboles de decisión es hacer particiones más pequeñas de los datos en grupos más homogéneos, repitiéndose el proceso generalmente de forma recursiva. Homogeneidad en este contexto significa que los nodos de las divisiones son más puros, es decir contienen la máxima proporción de una clase en cada nodo. Dos medidas alternativas, el índice de Gini y la entropía cruzada se usan para maximizar la pureza de los nodos.

Gini es el criterio utilizado por el modelo de Árboles de Clasificación (CART), en este modelo la poda se aplica a través del costo de la complejidad. El criterio de entropía cruzada es utilizado por el modelo C4.5. Ambos modelos son propensos al overfitting, para evitarlo CART utiliza una función de coste como método de poda, y C4.5 usa eliminación simple de subárboles.

- **Bagged Trees.**

Bootstrap aggregating o árboles bagged, son un método de clasificación que construye múltiples arboles sin podar para modelar las clases, donde cada modelo en el conjunto es usado para predecir la clase de la nueva muestra. Cada modelo presenta el mismo peso en el conjunto y se puede considerar que emite un voto para la clase a la que estima pertenecerá la nueva muestra. El número total de votos se divide luego por el número total de modelos del conjunto para producir un vector de probabilidad para predecir la muestra. La muestra se clasifica en el grupo que tiene más votos y, por lo tanto, la probabilidad más alta.

El número de árboles es un parámetro a definir en el entrenamiento del modelo. Los bagged Trees son capaces de reducir la varianza a través de conjunto y producir unos parámetros de rendimiento generalmente mejores que los arboles de decisión aislados.

- **Random Forests.**

El algoritmo para inducir un random forest fue desarrollado por Leo Breiman1 y Adele Cutler. El método combina la idea de bagging y la selección aleatoria de atributos para construir una colección de árboles de decisión con variación controlada. Los bosques aleatorios son bastante similares a los bagging, ya que cada árbol en el bosque emite un voto para la clasificación de una nueva muestra, y la proporción de votos en cada clase a través del conjunto es el vector de probabilidad predicho. Sin embargo, el tipo de árbol cambia en el algoritmo y hay un parámetro (mtry) para controlar el número de predictores seleccionados aleatoriamente en cada división. La idea detrás del muestreo aleatorio de predictores durante el entrenamiento es disminuir la correlación entre árboles en el bosque.

- **Boosting.**

El boosting consiste en combinar los resultados de varios clasificadores débiles para obtener un clasificador robusto. Cuando se añaden estos clasificadores débiles, se hace de modo que estos tengan diferentes pesos en función de la exactitud de sus predicciones. Después de añadir un clasificador débil, los datos cambian su estructura de pesos: los casos que son mal clasificados ganan peso y los que son clasificados correctamente pierden peso. Así, los clasificadores débiles se centran de mayor manera en los casos que fueron mal clasificados por los clasificadores débiles.

El boosting puede aplicarse a cualquier técnica de clasificación, pero los árboles de clasificación son el método más popular, ya que se pueden convertir en clasificadores débiles

al restringir la profundidad del árbol para crear árboles con pocas divisiones. Dado que los árboles de clasificación son una técnica de bajo sesgo y alta varianza, el conjunto de árboles ayuda a disminuir la varianza, produciendo un resultado que tiene un sesgo bajo y una varianza baja. Hay muchas clases de algoritmos de boosting, dos de los principales son Estocásticos Gradient Boosting y C5.0

- **C5.0.**

C5.0 es una versión más avanzada del modelo de clasificación C4.5 y tiene características adicionales. Se dice que es más rápido, usa la memoria de manera más eficiente, obtiene resultados similares con árboles más pequeños, ofrece soporte para boosting, le permite al usuario asignar diferentes ponderaciones a las clases y ofrece selección automática de predictores.

6.1.5. Evaluación de los modelos.

Existen diversas técnicas y métricas para evaluar y comparar los resultados de los algoritmos predictivos. Existen métricas como RMSE o R^2 apropiadas para problemas de regresión, otras como precisión (accuracy), sensibilidad (sensitivity), especificidad (specificity), F1 score o ROC son más apropiadas para problemas y algoritmos de clasificación. En la Tabla 2 se resumen los principales métodos de evaluación para problemas de clasificación.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">True Condition</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Condition Positive</td><td>Condition Negative</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted Condition Positive</td><td>True positive (TP)</td><td>False positive (FP)</td><td>PPV (or Precision) $\frac{TP}{TP+FP}$</td><td></td></tr>
<tr><td>Predicted Condition Negative</td><td>False negative (FN)</td><td>True negative (TN)</td><td colspan="2"></td></tr>
<tr><td colspan="2" rowspan="2"></td><td rowspan="2">Sensitivity (or Recall) $\frac{TP}{TP+FN}$</td><td rowspan="2">Specificity $\frac{TN}{TN+FP}$</td><td></td><td>Balanced Accuracy $\frac{1}{2} * \left(\frac{Sensitivity}{+Specificty}\right)$</td></tr>
<tr><td></td><td>F1 score $2 * \frac{PPV * Sensitivity}{PPV + Sensitivity}$</td><td></td></tr>
</table>

Tabla 2.- Resumen métricas evaluación algoritmos [42].

- **Matriz de confusión.**

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo de clasificación supervisado. Cada fila de la matriz representa el número de predicciones de cada clase, mientras que cada columna representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

- **Precisión** (Accuracy).

Es una de las métricas más simples y refleja grado de correlación entre los resultados observados y las predicciones. En problemas donde las frecuencias de la variable a predecir son muy asimétricas, o cuando los costes de cometer un tipo de error u otro son diferentes, no es la mejor métrica a seguir. Un ejemplo de esto es el problema tratado en este proyecto, donde la ratio de no asistencia suele estar en el 20%, de esta forma solo prediciendo siempre cada cita como asistencia conseguiríamos un 80% de exactitud, aunque no logremos en absoluto predecir el absentismo.

- **Sensibilidad** (sensitivity)

La sensibilidad de un modelo es la ratio en el cual el evento de interés se está prediciendo correctamente en relación todos los eventos producidos. La sensibilidad nos indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente observados como positivos; es decir, la proporción de positivos correctamente identificados. En nuestro proyecto, sería la ratio entre los pacientes clasificados como no_shows que realmente no asisten a su cita.

- **Especificidad** (specificity)

Al contrario que el caso anterior la especificidad es el número de no eventos, clasificados como no eventos. En nuestro caso sería la ratio entre los pacientes clasificados como asistencia entre todos los pacientes que realmente asisten a sus citas.

- **Curvas ROC y AUC**

Las curvas ROC (acrónimo de Receiver Operating Characteristic), es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador

según se varía el umbral de discriminación. Es creada evaluando las probabilidades de las clases a través de un umbral de discriminación continuo, calculándose en cada punto sensibilidad y especificidad y dibujándose uno contra el otro.

Las curvas ROC pueden usarse para valorar y comparar cuantitativamente un modelo, calculando el área bajo la curva AUC conseguida. Modelos perfectos alcanzarían un AUC de 1, mientras que modelos completamente inefectivos tendrían un AUC de 0.5. Esta métrica resulta más efectiva para caracterizar modelos, que precisión, sensibilidad o especificad puesto que no es sensible a proporciones asimétricas de las clases o a cambios en el umbral de decisión clasificatorio.

6.1.6. Lenguaje R y RStudio.

R es un entorno y lenguaje de programación diseñado por Ross Ihaka y Robert Gentleman. R es una implementación libre de lenguaje S, pero con un enfoque al análisis estadístico. Actualmente R es uno de los lenguajes más utilizados en investigación y en el desarrollo de proyectos de aprendizaje automático. Al tratarse de un proyecto abierto y colaborativo, dispone de una gran cantidad de bibliotecas o paquetes que los usuarios pueden publicar, con funcionalidades extra como puede ser la creación de gráficos. El repositorio oficial de paquetes cuenta actualmente con más de 9000 paquetes. R es parte del sistema GNU, pero también está disponible para sistemas operativos Windows o Macintosh. La última versión utilizada para el proyecto es la 3.5.1 [45].

IDE RStudio [46]. Para facilitar el trabajo con el lenguaje R se usó el entorno de desarrollo integrado (IDE) RStudio. RStudio está disponible de forma gratuita en y funciona en sistemas operativos Windows, Macintosh y Linux.

6.2. Legislación.

La privacidad en los datos de salud es un campo de máxima actualidad que esta suponiendo un reto a nivel de administraciones, organizaciones sanitarias y empresas privadas. La informatización de las prestaciones sanitarias, historias clínicas de pacientes, prescripción de medicamentos, datos financieros, etc se está extendiendo de forma generalizada, de la misma forma la interoperabilidad entre sistemas es cada vez más común.

Este hecho está propiciando la aparición de grandes bases de datos con información sensible, y cada vez preocupa más la protección y el uso que se hace de esta. En casi todos los países desarrollados existen normativas que protegen la privacidad de los datos de salud y limitan su uso.

En estados unidos, país donde se ha desarrollado el estudio, la ley de privacidad de datos se denomina "Health Insurance Portability and Accountability" HIPAA [48] [49] y determina el marco jurídico relativo a los datos sanitarios. La ley tiene 2 normativas básicas: la normativa de privacidad orientada a regular el uso que se hace de la información protegida de los pacientes por parte de organizaciones sanitarias. Por otro lado, la normativa de seguridad establece los requisitos exigidos a las empresas sanitarias que usan y comparten información de pacientes en formato electrónico, como la encriptación, control de acceso, almacenamiento de información, uso, etc.

Existen algunos conceptos claves como PHI (Protected Health Information) que se refiere a toda información pasada, presente o futura referente a una condición de salud física o mental relacionada con la provisión de servicio sanitarios o pago de los mismos almacenada de forma electrónica. Relacionado con este concepto está el Personally identifiable information (PII) siendo esta cualquier información anteriormente definida que incluye algún indicador que hace posible la identificación de forma univoca de un paciente.

En el marco del presente proyecto y con el fin de adecuarse a la legislación se tendrán en cuenta las siguientes consideraciones:

- Los datos usados serán sociodemográficos y de citación no conteniendo en ningún caso información de salud del paciente como diagnósticos, pruebas diagnósticas, medicaciones, códigos de facturación, etc.

- Los datos no se harán públicos en ningún momento, ni se compartirán con ninguna organización.

- El conjunto de datos será convenientemente anonimizado, de forma que no sea posible identificar los registros con un paciente de forma unívoca.

6.3 Creación del sistema predictivo.

En la presente sección se pretende explicar todo el proceso desarrollado para la construcción de los sistemas predictivos, siendo por tanto una de las más importantes de la memoria. En los capítulos previos se ha establecido el marco teórico del problema a resolver, a partir de aquí comienza la aportación real del proyecto a la resolución del problema de negocio concreto.

A diferencia del ciclo de vida estándar del desarrollo del software, son varios los autores que recomiendan seguir un proceso distinto para los proyectos de modelado predictivo y machine learning. Field Candy en su libro "Data Science Handbook" [50] describe seis etapas de un proyecto de Data Science: Establecer el marco del problema, Entender los datos, Extraer características, Modelar y Analizar, Presentar los resultados y Entregar el código. Se ha seguido una estructura similar para la construcción del modelo en este trabajo.

6.3.1. Creación del Data set.

Los datos son la materia prima esencial a la hora de construir sistemas predictivos. Por tanto, el primer paso en el proceso de construcción del modelo será recolectar, conocer y entender los datos necesarios para construirlo, entrenarlo y validarlo. En el presente proyecto la fuente principal de datos será la base de datos del sistema de historias clínicas de Mission Neighborhood Health Center (MNHC).

El sistema de historias clínicas (EHR) y citación (PM) en MNHC es gestionado por la aplicación comercial Nextgen [51], la cual soporta las operaciones clínicas y administrativas de la clínica. Este sistema de historia clínica electrónica fue implementado en 2014 y contiene todos los datos relativos a información clínica de pacientes, demográficos, citación y facturación.

El conjunto de datos necesario para entrenar y evaluar el sistema predictivo ha sido extraído directamente de la base de datos transaccional de Nextgen mediante un script de SQL (Anexo 1). El script fue ejecutado directamente en la base de datos SQL server de producción y el resultado almacenado en un texto plano. Se extraen un total de 150.000 registros correspondientes a datos de citación comprendidos entre 2016 y 2018, procedentes de los 3 edificios clínicos que tiene la compañía. Se incluyen todas las especialidades clínicas (pediatría, medicina de adultos, ginecología, obstetricia, salud mental y VIH) y todos los proveedores de salud. Cada una de las citas, organizadas en filas en el archivo, cuenta con los datos de sociodemográficos de pacientes, así como datos propios de la cita.

En cumplimiento de la legislación norteamericana sobre protección de datos (HIPAA), el set de datos extraído no contiene datos clínicos de los pacientes, no se hará público en ningún momento y es convenientemente anonimizado. Para anonimizar los datos, se eliminó cualquier campo capaz de identificar de forma única a un paciente, es decir, nombres, número de seguridad social, número de paciente y de historia clínica.

6.3.2. Análisis exploratorio de los datos.

Una vez extraídos los datos de la base de datos de Nextgen, el siguiente paso es realizar un análisis exploratorio de los mismos con el fin de ganar conocimiento y comprender mejor las características de los datos. Para realizar este proceso, los datos almacenados en el fichero de texto plano fueron cargados en RStudio, aplicación que se ha utilizado en los siguientes pasos de la construcción del modelo. Al final de la sección se presenta la Tabla 4 que resume las variables estudiadas.

Las variables fueron analizadas de forma individual para estudiar la presencia de valores perdidos, inapropiadamente altos o bajos (outliers), y los parámetros estadísticos más significativos. A su vez se evaluaron características conjuntas de variables como la colinealidad y la respuesta del outcome (asistencia o no) ante las variables. En el presente capitulo se resumen los resultados más relevantes del análisis de los datos.

- **Outcome**: Es la variable a predecir y la que determina la supervisión en el entrenamiento de los algoritmos de Machine Learning. Es una variable categórica que presenta dos niveles “No_show” y “Kept” de su nomenclatura en inglés para “no atención” y “atención”. No_show será la variable objetivo a predecir. La principal característica de esta variable es su asimetría, esto puede determinar negativamente los resultados de las predicciones y se tratará en detalle en siguientes capítulos. Las citas con No_show presentan una ratio del 21,2% sobre el total, por un 78,8% de citas atendidas (Kept)
- **Location:** Esta variable categórica representa cada una de las subdivisiones clínicas de la compañía. Se puede entender como las diferentes subespecialidades de atención que se desarrollan en MNHC. No se encontraron valores perdidos, pero el número de citas se distribuye de forma muy dispar entre las diferentes clínicas, esto podría ser un problema para algunos algoritmos predictivos, por lo que se ha de tener en cuenta. La cantidad de pacientes que no atienden a su cita y la ratio entre citas y no atención es variable entre clínicas.

- **Resource name**: Es una variable categórica que contiene la información sobre el proveedor de salud que realiza los servicios en la cita. No se encuentran valores perdidos, destaca la dispersión de las citas donde algunos proveedores acumulan muchas visitas, mientras en otros la presencia de citas es testimonial. Existe una variabilidad importante en el outcome con respecto al proveedor de servicios. El problema que presenta esta variable, es la cantidad de niveles incluidos, por lo que se han de estudiar formas de reducir la dimensionalidad de la variable.
- **Event Type**: Variable categórica que almacena el tipo de visita. Puesto que se ha decidido, por motivos de protección de datos, no incluir información clínica relevante de los pacientes en el ámbito del proyecto, esta variable será la que más información aporte sobre la urgencia y el motivo de la visita. Al igual que el caso anterior el reto consiste en gestionar la alta dimensionalidad de la variable.
- **App_date**: Fecha de realización de la cita. Es una variable temporal interesante para conocer si el absentismo presenta patrones temporales. Para facilitar su uso por los algoritmos de predicción se realizará una factorización de los valores en día de la semana, mes y año. (ver sección siguiente).
- **Begintime**: Hora de inicio de la cita. Aunque la naturaleza de la variable es continua, en nuestro sistema la modelamos de forma categórica por simplificación. Se observa asimetría en la hora de inicio de las citas, siendo entre 09-11 y 14-15 las horas con más ocupación. Existen un número muy reducido de citas en horas atípicas (21pm – 7am) que se excluirán del análisis al tratarse de posibles errores, ya que los centros no ofrecen servicio a esas horas. Con respecto al outcome el índice de absentismo es mayor a primera hora de la mañana y a ultima de la tarde, siendo las diferencias estadísticamente significativas.
- **Duration**: Duración de la cita. Igual que en el caso anterior se transforma una variable continua en categórica, ya que las citas se realizan en espacios de tiempo predeterminados (30min, 40, 1 hora etc.). Existen algunos outliers con una frecuencia muy baja, estos son tratados aproximando a una duración máxima de la cita de 120 minutos.
- **Sex:** Sexo del Paciente, variable binomial categórica (M, F). Existe una mayor asistencia a la clínica por parte de Mujeres, hecho posiblemente derivado de la oferta de servicios de Ginecología y Obstetricia. Las ratios de pacientes que no asisten a sus citas es algo

mayor en hombres que en mujeres (22,4 % vs 20,0 %). Existen 2 valores perdidos que se eliminan del análisis.

- **City:** Cuidad de residencia del paciente. Esta variable ha sido recolectada durante años como texto libre, por lo que existen una gran cantidad de valores perdidos y de diferentes categorías producidas por errores en las entradas manuales de los datos. Tras analizar el contenido se decide no incluir la variable en el modelo por su alto nivel de ruido.
- **State**: Estado. Al igual que el caso anterior la variable contiene más ruido que información.
- **ZIP:** Es una variable categórica que representa la dirección postal del paciente, no se incluye directamente en el modelo, pero es utilizada para crear una variable continua de distancia de la residencia del paciente a la clínica donde es atendido (ver siguiente sección).
- **Sex_orient y gen_ident**: Orientación sexual e identificación de género. Hace dos años comenzó a ser de obligado cumplimiento colectar estos datos de los pacientes. Los datos recogidos aún no son consistentes, apareciendo una gran cantidad de datos desconocidos, por lo que no son tenidos en cuenta para nuestro modelo.
- **Age:** Es una variable numérica entera que representa la edad del paciente en años en el día de la cita. La distribución del número de citas es multimodal con dos picos uno en pacientes pediátricos de corta edad y otro entre los 30-50 años. En cuanto a la distribución del índice de absentismo, llama la atención los índices más bajos en edades tempranas y en pacientes senior, siendo mayor en adolescentes y personas adultas de mediana edad.
- **Homeless:** Variable categórica que almacena la situación de "sin techo" de algunos pacientes y su tipo. Casi un 12% de la población atendida en las clínicas son sin hogar, existiendo un porcentaje de absentismo mas elevado entre estos pacientes.
- **Lang_barrier:** Variable categórica que recoge si el paciente tiene barreras de idioma o no. Un alto porcentaje de pacientes atendidos en MNHC presenta barreras de idiomas, siendo este grupo el que tiene un índice de absentismo más bajo.
- **veteran_ind.** Variable categórica que indica si un paciente ha sido veterano en el ejercito estadounidense. Atendiendo a los datos solo un pequeño porcentaje de la población es "veterana" y la relación con el outcome no es significativa, por lo que no es incluida en nuestro análisis.

- **Ethnicity**: Variable categórica que agrupa a la población atendida en (hispana o latina, no latina o desconocida). Destaca un alto valor de hispanos en la población atendida por la clínica, siendo estos los pacientes los que más atienden a sus citas respecto a los "No latinos" (20%~25%). Destaca también que hay un número importante de valores desconocidos, que se mantendrán como una categoría independiente.
- **Race**: Esta variable categórica representa la raza del paciente. La distribución del número de citas entre razas es dispersa, así como el porcentaje de atención respecto a la raza.
- **Enc_payer**: Variable categórica que representa el tipo de seguro que tiene el paciente. El número de categorías de esta variable es muy elevado por lo que se creó una variable resumen del tipo de aseguramiento (ver ingeniería de factores). En los algoritmos predictivos sensibles a la alta dimensionalidad de las variables categóricas no será incluida esta variable.
- **Family_income**: Es una variable numérica que almacena los ingresos económicos anuales percibidos por el paciente. La variable presenta una distribución asimétrica siendo los valores bajos mucho más numerosos. En cuanto al incide de absentismo existe una correlación negativa entre ingresos y no asistencia. Se observan algunos valores extremos 9999999, esta fue una forma de representar valores desconocidos por parte de los usuarios, por lo que se han sido eliminados del conjunto de datos. También se eliminan las citas que no tienen registrada los ingresos del paciente.

Variables	Tipo	Descripcion
Outcome	**Categorica**	Es una variable categórica que presenta dos niveles "No_show" y "Kept" de su nomenclatura en inglés para "no atención" y "cumplido. Es la variable a predecir
Location	**Categorica**	Esta variable categórica representa cada una de las subdivisiones clínicas de la compañía.
Resource name	**Categorica**	Es una variable categórica que contiene la información sobre el proveedor de salud que realiza los servicios en la cita
Event Type	**Categorica**	Variable categórica que almacena el tipo de visita
App_date	**Time**	Fecha de realización de la cita
Begintime	**Categorica**	Hora de inicio de la cita
Duration	**Categorica**	Duración de la cita
Sex	**Categorica**	Sexo del Paciente
Age	**Numerica**	Es una variable numérica entera que representa la edad del paciente en años
Homeless	**Categorica**	Variable categórica que almacena la situación de "sin techo"
Lang_barrier	**Categorica**	Variable categórica que recoge si el paciente tiene barreras de idioma o no.
Ethnicity	**Categorica**	Variable categórica que agrupa a la población atendida en (hispana o latina, no latina o desconocida)
Race	**Categorica**	Esta variable categórica representa la raza del paciente
Enc_payer	**Categorica**	Variable categórica que representa el tipo de seguro que tiene el paciente.
Family_income	**Numerica**	Es una variable numérica que almacena los ingresos económicos anuales percibidos por el paciente.

Tabla 3.- Datos extraídos de la Base de Datos de Nextgen

6.3.3. Ingeniería de los factores.

Atendiendo a la revisión bibliográfica realizada en los primeros capítulos, existen factores determinantes para explicar el absentismo, algunos ejemplos son: la distancia a la clínica, el tiempo de espera, comportamientos previos del paciente, etc. Estos factores no están almacenados de forma explícita en la base de datos y es necesario realizar transformaciones mas o menos complejas para conseguirlos. De otro lado, el formato de algunos factores puede no ser apropiado según el algoritmo usado para predecir, por este motivo algunos han de ser adaptados. En la presente sección se detallan las nuevas variables construidas, resumiéndose al final en la Tabla 4.

- **Diff_date**. Es una variable continua que representa la distancia en días entre el día en que la cita es programada y el día de la cita, es decir, el tiempo de espera. La variable presenta una asimetría con predominancia de valores bajos. Según la literatura especializada, existe una correlación positiva entre tiempo de espera e índice de absentismo. Este hecho se confirma en nuestro conjunto de datos, siendo más alta la probabilidad de no asistencia para aquellas citas de larga espera.
- **Site:** Se crea una variable categórica que representara el edificio físico donde tiene lugar la cita. Esta variable se crea mediante la agrupación de la variable location. Existe variabilidad significativa de la ratio de absentismo en función de la variable Site. “Shotwell”, “Excelsior” y “Resource Center” son los posibles valores de esta variable.
- **Resource Type:** Es una variable categórica construida por composición de la variable resource_name. Esta variable contiene información sobre el la categoría del profesional que realiza el servicio de la cita. MD, NP, RN, BH, O y UNK son los posibles valores. La variación de la ratio del outcome entre los distintos tipos de profesionales es significativa.
- **Weekday**: Dia de la semana de la cita. Es una variable categórica construida a partir de la factorización del appt_date. Observando la variable, se aprecia una distribución asimétrica en el numero de citas respecto a día laborable o fin de semana, como era de esperar. También se aprecia una ratio de absentismo mayor en el lunes que en el resto, siendo la diferencia poco significativa. Es destacable un alto porcentaje de absentismo en domingo, pero como la cantidad de citas es mínima no es un indicador fiable, por lo que será considerado como un outlier y en este caso se decide eliminarlo del análisis.

- **Month**: Mes de realización de la cita. También se construye a partir de la variable app_date. El número de citas por mes se reparte de forma más o menos uniforme y es reseñable un ligero incremento en el absentismo para los dos últimos meses del año. Este incremento es consistente entre los diferentes años lo que es indicativo de una ligera estacionalidad mensual.
- **Year**: Año de realización de la cita. Última variable categórica construida a partir de la app_date (o día de citación). El número de citas anual es más o menos uniforme y se observa un decremento suave del porcentaje de absentismo gradualmente desde 2015 hasta 2018.
- **Homeless Factor:** Se ha realizado una agrupación de los distintos tipos de personas sin hogar para conseguir una variable dicotómica homeless "Si", "No". Es indicativo el valor de absentismo mayor para pacientes sin hogar.
- **Payor_type:** Es una variable categórica creada por agrupación del tipo de aseguramiento del paciente. Las categorías o niveles de la variable son (medical, medicare, private, uninsurance y desconocido (UNK). La cantidad de citas por tipo de pagador es variable predominando el aseguramiento de Medical y los pacientes no asegurados. Es destacable un aumento del porcentaje de absentismo para aquellos pacientes que no tienen o se desconoce su cobertura.
- **Dist:** Variable creada para recoger la distancia desde la vivienda del paciente al lugar de la cita. Para construir esta variable, primero se realiza un mapping entre los distintos ZIP codes donde el paciente tiene residencia y su geolocalización, creándose una tabla con (ZIP, longitud y latitud). Posteriormente se utiliza la librería "*geosphere*" y la función "*distHaversine*" para calcular la distancia en millas entre los distintos puntos de longitud y latitud, resultando una variable numérica discreta indicativa de la distancia del paciente al centro donde tenía la cita. Analizando las características de la variable, se observa una asimetría en los valores, siendo más comunes las citas más próximas a la respectiva clínica. Esta asimetría será tratada posteriormente con transformaciones logarítmicas o box-cox. En cuanto a los índices de absentismo estos aumentan de forma positiva según aumenta la distancia a la cita. Existen valores perdidos que se eliminan del set de datos.
- **ns_hist_df.** Variable numérica entera que representa en numero de no-show anteriores del paciente. De acuerdo a lo encontrado en la bibliografía, el comportamiento previo del paciente puede ser un factor relevante para predecir la futura asistencia del paciente a las citas. Para construir esta variable se ha realizado un script SQL que determina el

numero de no asistencias previas del paciente, realizándose posteriormente en R una carga del fichero y la oportuna combinación de los datos con el data set origen. El resultado es una variable con el numero de no asistencias del paciente anteriores a la fecha de la cita en cuestión. En consonancia con la bibliografía se aprecia un aumento positivo del absentismo para aquellos pacientes que presentan un comportamiento de no asistencia previo.

Variables	Tipo	Descripcion
Site	**Categorica**	Variable categórica que representara el edificio físico donde tiene lugar la cita
Resource Type	**Categorica**	Esta variable contiene información sobre el la categoría del profesional que realiza el servicio de la cita.
Weekday	**Categorica**	Dia de la semana de la cita
Month	**Categorica**	Mes de realización de la cita
Year	**Categorica**	Año de realización de la cita
Homeless Factor	**Categorica**	variable dicotómica homeless "Si", "No"
Payor_type	**Categorica**	Es una variable categórica creada por agrupación del tipo de aseguramiento del paciente
Dist	**Numerica**	Variable creada para recoger la distancia desde la vivienda del paciente al lugar de la cita
ns_hist_df	**Numerica**	Variable numérica entera que representa en numero de no-show anteriores del paciente.

Tabla 4.- Variables construidas mediante ingeniería de factores.

6.3.4. Set de entrenamiento, Validación y técnicas de Resampling.

Una vez extraídos y analizados los datos y realizado un primer tratamiento de los valores perdidos, outliers, asimetrías, formatos, etc, es el momento de construir los sets de datos de training y testing para entrenar y validar nuestros modelos predictivos. En esta fase se decidió la manera de particionar los datos, el formato más apropiado de los factores y las técnicas de resampling a usar para evitar el overfitting.

Es de vital importancia la adecuada preparación de los datos antes de comenzar a utilizarlos para entrenar los algoritmos de machine learning seleccionados. Algunos algoritmos como los basados en arboles de decisión suelen ser bastante consistentes a diferentes formas de presentar los datos, valores perdidos o outliers. Sin embargo, otros algoritmos son muy sensibles al formato de los datos de entrada, por ejemplo, los algoritmos de regresión logística y otros basados en métodos lineales, no soportan variables categóricas como factores de entrada. Una práctica común para solucionar este problema es la creación de dummies variables, donde se dividen las variables categóricas en variables lógicas con valores 0 ó 1, tantas como distintas categorías tenga la variable original. Puesto que en nuestro trabajo queremos evaluar distintos tipos de algoritmos, se ha decidido crear 2

conjuntos de datos, uno manteniendo las variables categóricas y otro con todas las variables convertidas a numéricas.

Antes de comenzar a entrenar los algoritmos, es práctica común realizar una partición de los datos, separando aquellos utilizados para entrenar el modelo y dejando un subconjunto únicamente para validar el rendimiento. En ocasiones se crea un tercer conjunto intermedio para realizar un refinamiento o tunning de los modelos. No hay un estándar en cuanto al porcentaje de datos utilizados para uno u otro fin y dependerá del conjunto de datos en particular. Esta tarea se vuelve critica en conjuntos de datos con pocas observaciones, más aún si el número de factores es elevado. Para nuestro modelo se decidió usar los datos de 2016, 2017 y una parte de los datos de 2018 para entrenar el modelo. El 25% de datos de 2018 se reservó para Test por su proximidad con los datos reales a predecir. La seleccion de datos se realizó mediante muestreo estraficado de los datos (stratified sampling) con el fin de mantener la proporción de "No_Shows" y "Kept".

El overfitting es un fenómeno, a evitar, que ocurre cuando un algoritmo predictivo es capaz de aprender tanto la señal o información que tienen los datos como su componente de ruido. Podemos decir que, debido a su potencia de cálculo, algunos de estos algoritmos son capaces de memorizar los datos, perdiendo así la capacidad para detectar patrones generales y por tanto de generar futuras predicciones. El resampling es un conjunto de técnicas para evitar el overfitting al entrenar los modelos y evaluar estos de forma más precisa. Al igual que en el apartado anterior no existe una técnica claramente superior al resto en términos de rendimiento y su elección dependerá del conjunto de datos y/o del algoritmo a utilizar. Para nuestro proyecto la técnica utilizada ha sido el resampling K-Fold siendo 5 el número de repeticiones elegido. Esta decisión esta motivada por el considerable tamaño de nuestro dataset que aseguran una variabilidad suficiente en las muestras, sin impactar sobremanera en el coste computacional ya de por sí alto por el tamaño de set de entrenamiento.

6.3.5. Preprocesamiento de los datos.

Hay diversas técnicas para preprocesar los datos y modificar sus componentes estadísticas potenciando su poder predictivo y mejorando el rendimiento de los algoritmos. Las técnicas más utilizadas son, el escalado y centrado, transformación espacial, transformación box-cox, o PCA (Partial Component Analysis) siendo más o menos útiles dependiendo del algoritmo y de los datos.

En el próximo capitulo se indicarán cada una de las técnicas utilizadas a la hora de entrenar el modelo. A continuación, vamos a describir dos técnicas que se han aplicado de forma genérica a nuestros conjuntos de datos: Reducción de colinealidad y No Zero Values.

Algunos algoritmos reducen de forma importante su rendimiento cuando encuentran variables que están altamente correladas, por este motivo antes de realizar el entrenamiento de los modelos se ha calculado la matriz de correlación (corrplot) de los dataset y se han eliminado de los mismos aquellas variables con un índice de correlación mayor al 95% (findCorrelation).

Igual que en el caso anterior, algunas técnicas de machine learning tienen dificultades cuando encuentran variables con escasa o nula varianza entre sus valores. Anteriormente en la descripción del dataset se mencionó la existencia de variables categóricas con escasos datos para algunos de sus niveles (ej domingo en el día de la semana, algunas aseguradoras con pocos pacientes o tipos de cita no utilizados a menudo). Este problema se agudiza al realizar técnicas de resampling sobre los datos, ya que al particionar aleatoriamente los mismos las probabilidades de reducir la varianza a cero aumenta. Por este motivo antes de entrenar los modelos se decidió eliminar aquellos factores con varianza cercana a zero mediante la fórmula nearZeroVar.

Además de las técnicas mencionadas, hemos de asegurarnos que tratamos los valores nulos o perdidos en nuestros datos. En nuestro caso, encontramos valores perdidos en las variables dist y family income. Aunque existen técnicas avanzadas de imputación de datos, mediante algoritmos como k-neighbordhood que calculan los datos perdidos por proximidad con sus vecinos en el dataset, en nuestro caso hemos simplificando eliminando las observaciones con valores perdidos de nuestro conjunto de datos. Esta decisión está justificada por ser el data set disponible suficientemente grande y no existir correlación de los valores perdidos con el outcome.

6.3.6. *Entrenamiento y selección de algoritmos.*

En los capítulos previos se ha trabajado en la preparación y procesamiento del conjunto de datos. Ahora es el momento de emplear los algoritmos de machine learning en nuestro set de entrenamiento para construir los modelos. Se han seleccionado para su evaluación 17 modelos divididos según su naturaleza en modelos lineales, modelos no lineales y modelos basados en árbol, todos ellos permiten resolver problemas de clasificación como el que nos ocupa. Una vez entrenados todos los modelos se procedió a la evaluación

de su rendimiento y se eligieron los mejores en cada categoría para su posterior implementación. En las 3 secciones siguientes se realiza el refinado de los hyper-parámetros (o tunning) de los modelos seleccionados, su calibrado para las predicciones y se analiza la importancia de los predictores en cada uno. En el anexo 2 adjunto esta el archivo de código que realiza todo el proceso de construcción de los modelos.

Para el entrenamiento de los modelos se ha utilizado como herramienta básica el paquete CARET[1] (classification and regression training). Este paquete incluye una serie de funciones que facilitan el uso de decenas de métodos complejos de clasificación y regresión. Por otro lado, debido al tamaño del conjunto de datos y a la complejidad de los algoritmos, el coste computacional de entrenarlos para crear los modelos es muy alto, por este motivo se han usado técnicas de paralelización implementadas en el paquete "doParallel" para mejorar el tiempo se entrenamiento de los modelos.

El primer paso es definir los parámetros comunes que aplican al entrenamiento de los algoritmos, esto se realiza con las variables de la estructura de datos de train_control donde se especifica el método de resampling, que como se señaló será un k-fold con 5 repeticiones, también se indica el parámetro de validación a optimizar por parte de los modelos que será el ROC. Otros parámetros son el conjunto de datos utilizado vara validar los modelos en el resampling "index", el tipo de predicción "classprob", si queremos salida del estado de entrenamiento "Verbose" o si almacenamos las predicciones también son indicados.

- **Algoritmos Lineales**.

Una de las consideraciones más importantes para entrenar métodos lineales es que no toleran bien las variables categóricas, por lo que se usará el data set de entrenamiento donde estas variables son convertidas a numéricas (dummies variables). Otro factor a tener en cuenta es la sensibilidad a valores perdidos y de varianza nula, por lo que se recomienda usar técnicas de preprocesado como non-zero-value. Por último, hay que señalar que también son sensibles a la colinealidad.

Se realiza el entrenamiento con 6 algoritmos lineales: Regresion Lineal (glm), Linear discriminant Analysis (lda), Partial Least Squares Discriminant Analysis (pls), Penalized Models (glmnet), Sparse logistic regression (sparseLDA) y Nearest Shrunken Centroids

[1] Max Kuhn. CARET Manual. http://topepo.github.io/caret/index.html.

(pam). A continuación, la Tabla 5 resume los parámetros de entrenamiento y técnicas de reprocesado para cada modelo.

Modelos Lineales				
Modelo	**Metodo**	**Data_set**	**Parametros**	**Preprocesado**
Logistic Regression	glm	training_SAM[,reducedSet_SAM]	None	center, scale,zv
Linear Discriminant Analysis	lda	training_SAM[,reducedSet_SAM]	None	center, scale,zv
Partial Least Squares Discrimir	pls	training_SAM[,reducedSet_SAM]	ncomp	center, scale,zv
Penalized Models	glmnet	training_SAM[,fullSet_SAM]	alpha, lambda	center, scale,zv
Sparse logistic regression	sparseLDA	training_SAM[,fullSet_SAM]	NumVars, lambda	center, scale,zv
Nearest Shrunken Centroids	pam	training_SAM[,fullSet_SAM]	threshold	center, scale,zv

Tabla 5.- Características Algoritmos lineales

Las Figuras 3 y 4 muestran un análisis del rendimiento, AUC, sensibilidad y especificidad y la comparación de curvas ROC.

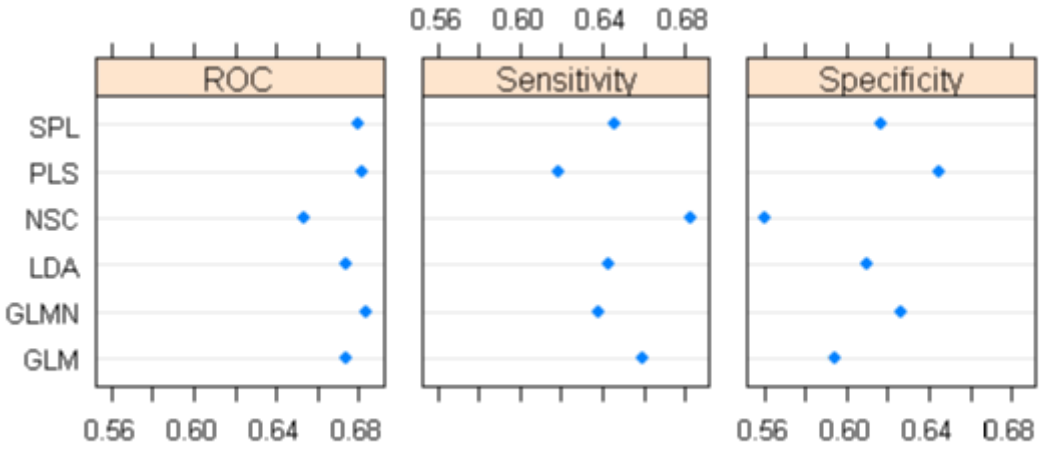


Figura 3.- Rendimiento comparado de los modelos lineales.

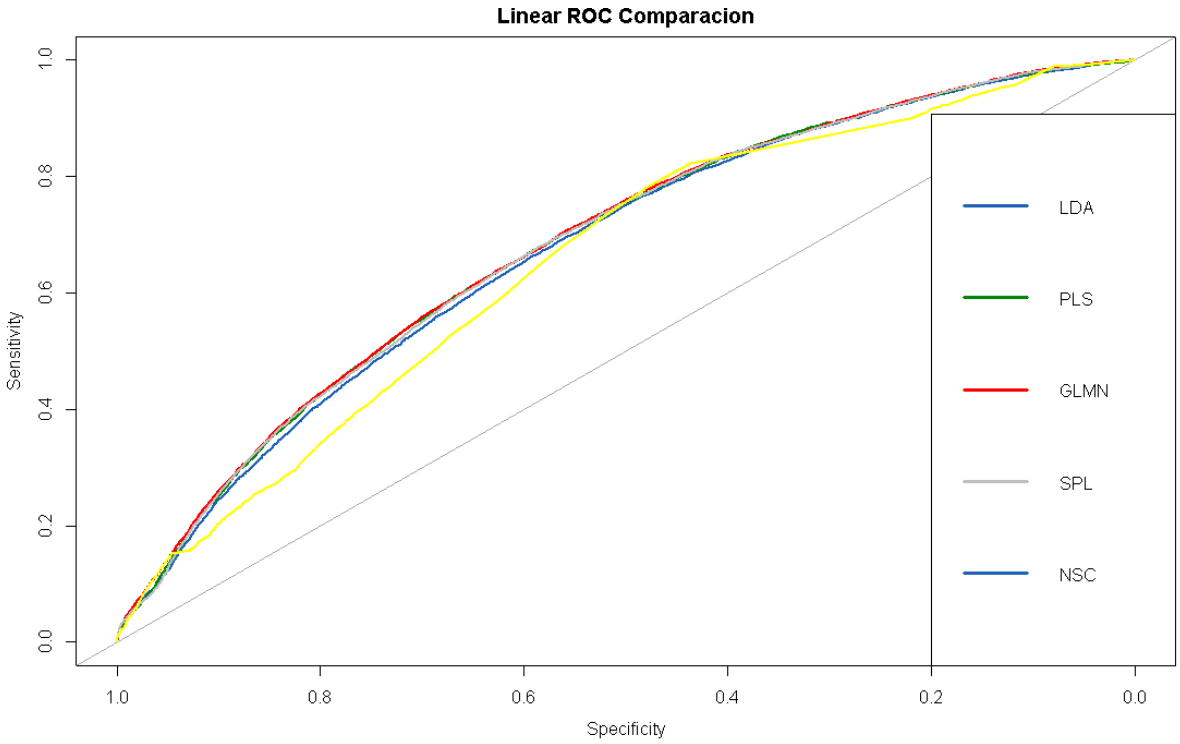


Figura 4.- Comparativa curvas ROC

Los modelos lineales presentan similares características de rendimiento para el set de datos entrenado. El rendimiento es homogéneo para todos los algoritmos y los resultados obtenidos son discretos para implementar en un entorno de producción. El modelo que mejor comportamiento presenta es el Penalized Models (glmnet), por lo que será uno de los candidatos a estudiar en detalle.

- **Algoritmos No lineales.**

En esta sección se describen modelos capaces de aprender patrones no lineales entre los datos. Al igual que en los métodos lineales, los algoritmos se comportan mejor en el set de entrenamiento numérico compuesto por dummies variables a excepción del modelo Naive Bayes. El centrado y escalado son las técnicas de preprocesado de datos aplicadas en los datos de entrenamiento, mientras que las Redes Neuronal presenta pequeñas mejoras al realizar también una proyección espacial de los datos previa al entrenamiento. Es de destacar el alto coste computacional requerido por estos algoritmos, algunos como el SVM han requerido días para completar la ejecución. También hay que destacar la complejidad de algunos modelos y el tamaño de los archivos generados que pueden impactar a la hora de su implementación en una aplicación en producción.

Se realizó el entrenamiento de 5 algoritmos NO Lineales: Neural Network (nnet), Flexible Discriminant Analysis (FDA), Support Vector Machines (SVM), K-Nearest Neighbors (KNN) y Naive Bayes (nb). A continuación, la Tabla 6 resume los parámetros de entrenamiento y técnicas de reprocesado para cada modelo.

Algoritmos NO Lineales				
Modelo	**Metodo**	**Data_set**	**Parametros**	**Preprocesado**
Neural Network	nnet	training_SAM[,reducedSet_SAM]	size, decay, bag	center, scale,spatialSign
Flexible Discriminant Analysis	fda	training_SAM[,reducedSet_SAM]	degree, nprune	
Support Vector Machines with	svmRadial	training_SAM[,reducedSet_SAM]	sigma, C	center, scale
K-Nearest Neighbors	knn	training_SAM[,reducedSet_SAM]	k	center, scale
Naive Bayes	nb	training_od_AM[,orig_df_fact_AM]	fL, usekernel, adjust	

Tabla 6.- Características Algoritmos lineales

Las Figuras 5 y 6 muestran un análisis del rendimiento, AUC, sensibilidad y especificidad y la comparación de curvas ROC.

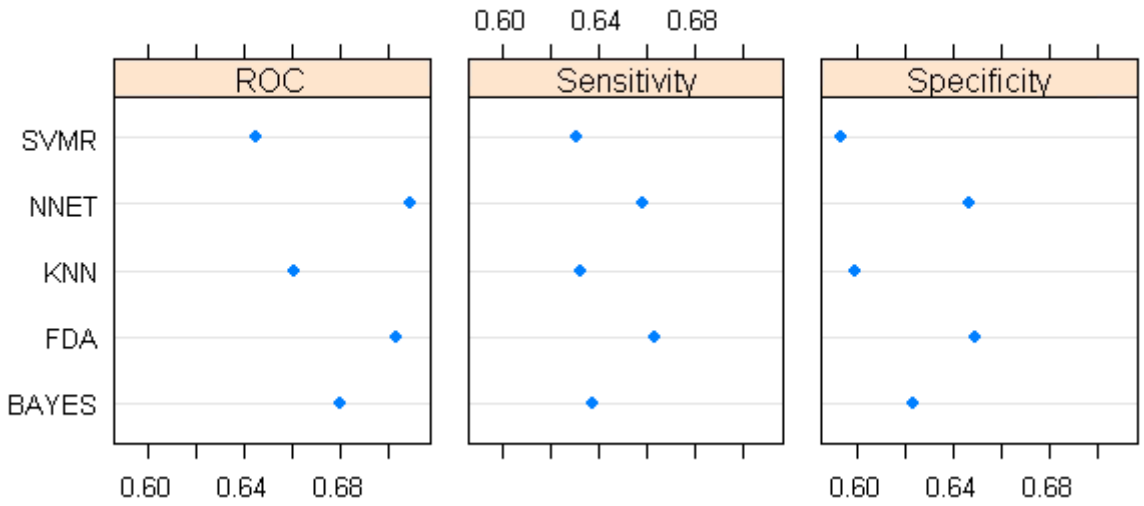


Figura 5.- Rendimiento comparado modelos NO lineales.

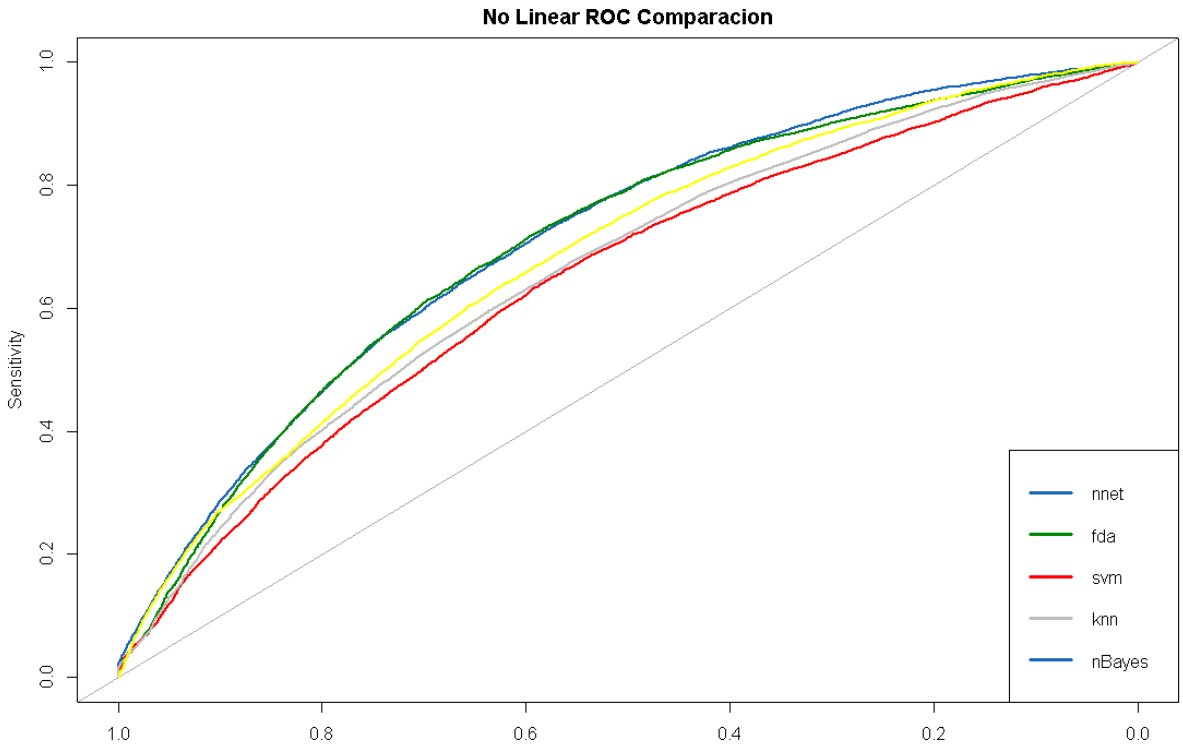


Figura 6.- Comparativa curvas ROC Modelos No lineales.

Observando las métricas de clasificación y las curvas ROC podemos concluir que los modelos que tienen un mejor comportamiento para los datos del problema son la red neuronal (nnet) y el FDA. Optaremos por el segundo al tratarse de un modelo mucho más simple lo que facilitará su posterior implementación.

- **Algoritmos basados en Árbol.**

Algunos de los algoritmos de clasificación más usados actualmente son de la familia de los basados en árboles. Debido a su tolerancia en los factores de entrada es posible utilizar cualquiera de los dos sets de datos creados, teniendo más rendimiento en la práctica el set numérico con las dummies variables. No se han utilizado técnicas adicionales de preprocesado de datos que las comentadas en capítulos anteriores. El coste computacional de entrenar los algoritmos ha sido alto en la mayoría de los casos, del orden de varias horas paralelizando con 4 cores.

Se realizó el entrenamiento de 6 algoritmos basados en árbol: Stochastic Gradient Boosting (gbm), CART, C4.5-like Trees, Bagged CART, Random Forest y C5.0.

A continuación, la Tabla 7 resume los parámetros de entrenamiento y técnicas de reprocesado para cada modelo.

Modelos basados en Arbol			
Modelo	**Metodo**	**Data_set**	**Parametros**
Stochastic Gradient Boosting	gbm	training_SAM[,reducedSet_SAM]	n.trees, interaction.depth, shrinkage, n.minobsinnode
CART	rpart	training_SAM[,reducedSet_SAM]	cp
C4.5-like Trees	J48	training_SAM[,reducedSet_SAM]	C,M
Bagged CART	treebag	training_SAM[,reducedSet_SAM]	nbagg
Random Forest	rf	training_SAM[,reducedSet_SAM]	mtry
C5.0	C5.0	training_SAM[,reducedSet_SAM]	trials, model, winnow

Tabla 7.- Características Algoritmos basados en árbol.

Las Figuras 7 y 9 muestran un análisis del rendimiento, AUC, sensibilidad y especificidad y la comparación de curvas ROC.

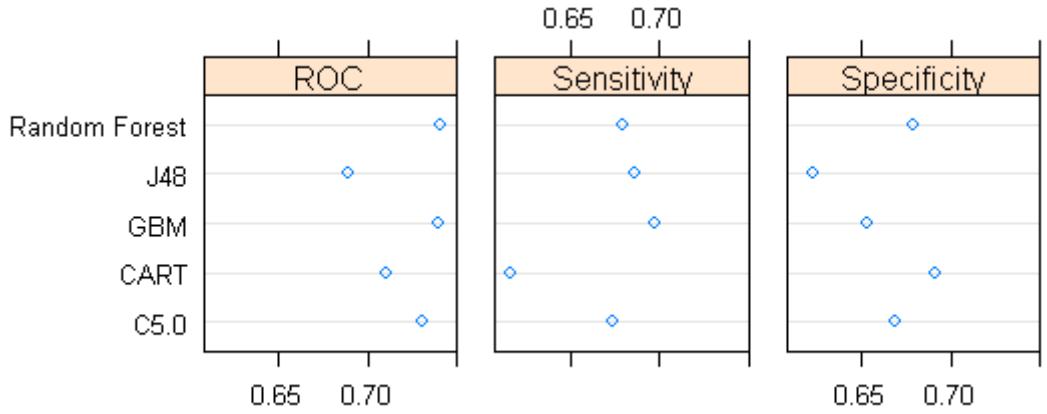


Figura 7.- Rendimiento comparado de modelos basados en árbol.

*Se ha presentado el modelo Bagged en la Figura 8 por separado debido que el volumen del modelo final presentaba problemas para su almacenado y carga con el resto de modelos.

```
Bagged CART

146454 samples
   300 predictor
     2 classes: 'No_show', 'Kept'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 110540
Addtional sampling using down-sampling

Resampling results:

  ROC        Sens       Spec
  0.7287223  0.6704579  0.6622195
```

```
Cross-Validated (5 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

Confusion Matrix and Statistics

          Reference
Prediction No_show  Kept
   No_show    4480  9874
   Kept       2202 19358

               Accuracy : 0.6638
                 95% CI : (0.6588, 0.6686)
    No Information Rate : 0.8139
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2306
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6705
            Specificity : 0.6622
         Pos Pred Value : 0.3121
         Neg Pred Value : 0.8979
             Prevalence : 0.1861
         Detection Rate : 0.1247
   Detection Prevalence : 0.3997
      Balanced Accuracy : 0.6663

       'Positive' Class : No_show
```

*Figura 8.-Rendimiento del modelo Bagged. (se añade aparte por problemas de memoria)

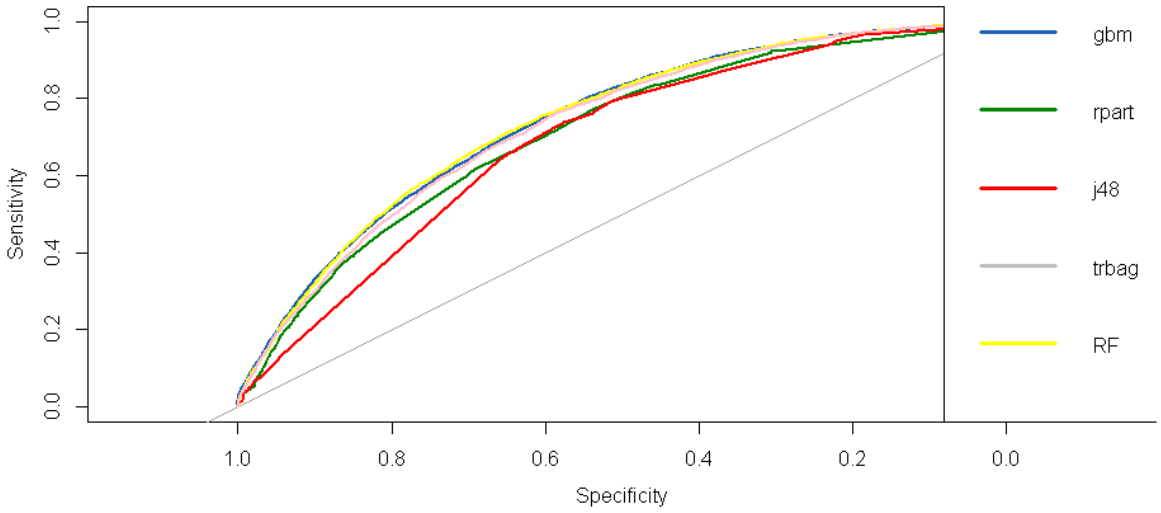


Figura 9.- Comparativa curvas ROC Modelos Árbol

Los modelos basados en árbol son los que mejor rendimiento presenta de todos los evaluados, se ha elegido el Random Forest como modelo a implementar en la aplicación de overbooking. La elección del Random Forest en vez del GBM ha sido debido a su sencillez para su implementación y mejor rendimiento posterior en datos nuevos de Test. En la práctica, el GBM presentaba algunos indicios de problemas de overfitting.

6.3.7. Tunning de los parámetros.

Cada uno de los algoritmos entrenados en el capítulo anterior presenta una serie de parámetros que permiten controlar ciertas características del algoritmo con el fin de alcanzar sus máximas prestaciones para el set de datos y/o de evitar el overfitting. Para entrenar cada uno de los algoritmos hemos seleccionado combinaciones estándar de parámetros, siguiendo las recomendaciones de Max Kuhn [42], creando una serie de tablas introducidas en la estructura de training mediante tuneGrid. Los modelos para cada posible combinación son entrenados según se refleja en la Figura 10.

```
Define sets of model parameter values to evaluate
for each parameter set do
    for each resampling iteration do
        Hold–out specific samples
        [Optional] Pre–process the data
        Fit the model on the remainder
        Predict the hold–out samples
    end
    Calculate the average performance across hold–out predictions
end
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

Figura10.- Proceso de tunning y resampling con CARET

A lo largo de la sección vamos a describir y justificar los parámetros seleccionados para los modelos candidatos a implementarse seleccionados en el capítulo anterior.

Modelo lineal de penalización (glmn)

Glmn usa regresión ridge y lasso simultáneamente y una estructura de penalizaciones. Tiene 2 parámetros principales para controlar el modelo: α y λ. El parámetro α determina la proporción de cada modelo, así cuando tiene valor=1 el modelo es de lasso puro, mientras que cuando tiene valor = 0 es regresión ridge. Otro parámetro importante del modelo es λ, el cual controla la cantidad de penalización en cada modelo. La matriz de parametrización utilizada es la siguiente:

```
glmnGrid <- expand.grid(alpha = c(0,  .1,  .2, .4, .6, .8, 1),
                        lambda = seq(.01, .2, length = 40))
```

Las figuras 11 y 12 muestran el rendimiento del modelo para cada combinación de los parámetros.

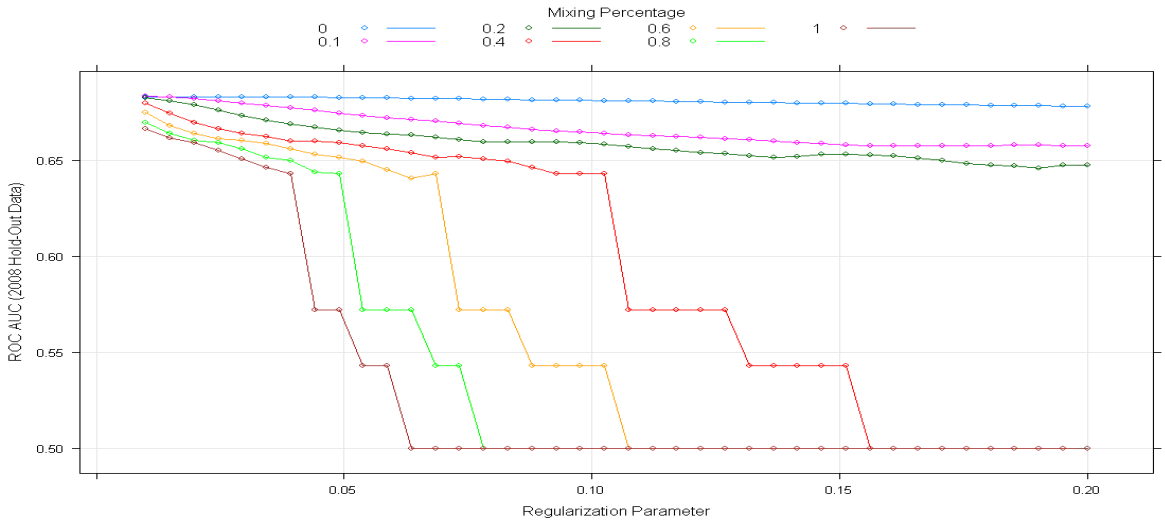


Figura 11.- AUC por combinación de parámetros.

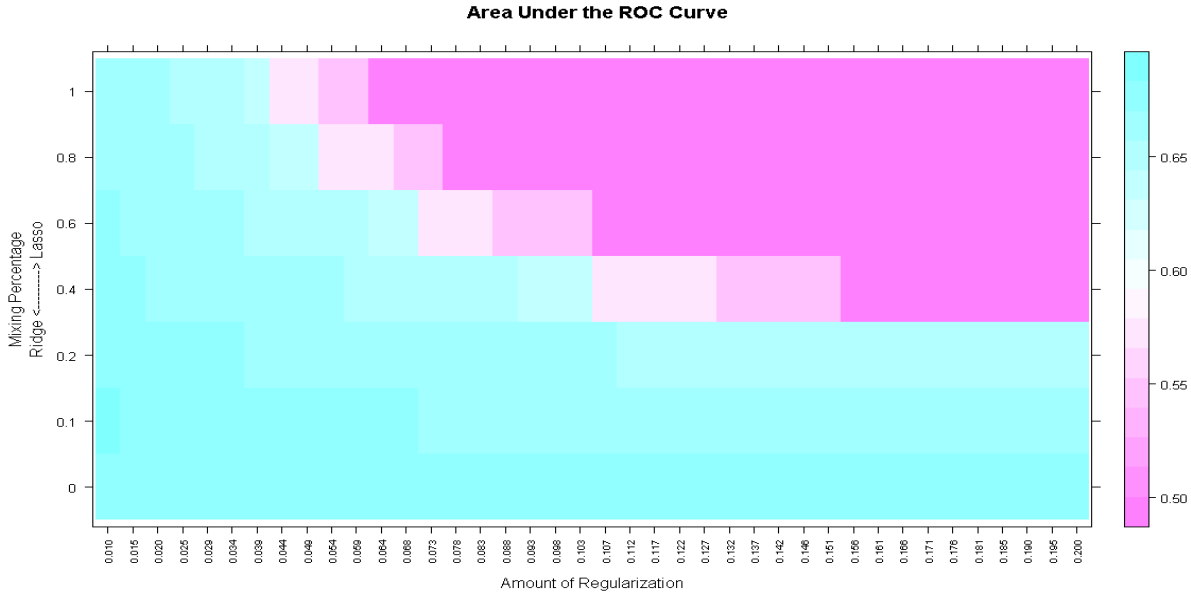


Figura 12.- Heatmap de rendimiento por parámetros α y λ.

En ambas graficas puede verse como el resultado es óptimo para α=0.1 y λ=0.01. Siendo el AUC de 0.6833.

```
> glmnFit_SAM$bestTune
   alpha lambda
41   0.1   0.01
> auc( glmnFit_SAM$roc)
Area under the curve: 0.6833
> ci.auc( glmnFit_SAM$roc)
95% CI: 0.6764-0.6902 (DeLong)
```

Random Forest (RF).

El algoritmo random forest combina la idea de bagging y la selección aleatoria de atributos, para construir conjuntos de árboles de decisión o bosques. Tiene un parámetro principal de configuración (mtry) que controla el número de predictores seleccionados aleatoriamente. El número de árboles (ntree) de estableció en 500 de forma constante.

La matriz de parámetros se expreso de la siguiente forma:

```
mtryValues <- c(5, 10, 20, 32, 50, 100, 250, 500, 1000)
ntree = 500
```

De forma gráfica se puede ver en la figura 13 el rendimiento del área bajo la curva ROC (AUC) para combinaciones de parámetros.

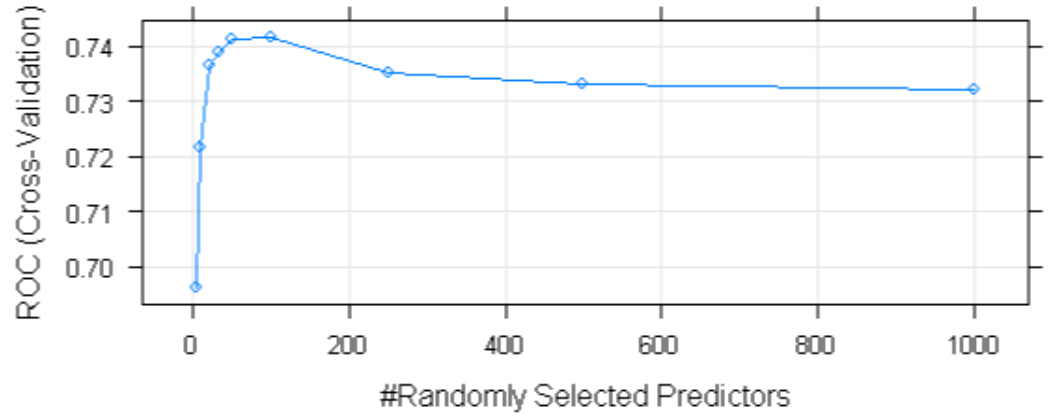


Figura 13.- Rendimiento del RF en función del parámetro mtry.

El mejor rendimiento se alcanza con los parámetros señalados a continuación.

```
> rfFit_SAM$bestTune
  mtry
6  100
> auc(rfFit_SAM$roc)
Area under the curve: 0.7414
```

El valor optimo se alcanza con un valor de 100, añadir más términos al algoritmo solo incrementaría la complejidad del mismo. El valor final del AUC es 0.741

Flexible Discriminant Analysis (FDA)

Aunque hemos usado el paquete CARET para trabajar con los modelos, este internamente requiere el paquete "earth". Trabajamos con 2 parámetros en este modelo, Maximum degree of interaction (degree), que se establece a 1 para determinar que es un modelo aditivo, y (nprune) que determina el máximo número de términos en el modelo podado.

La matriz de parámetros es la siguiente:

```
tuneGrid = expand.grid(degree = 1, nprune = 2:25)
```

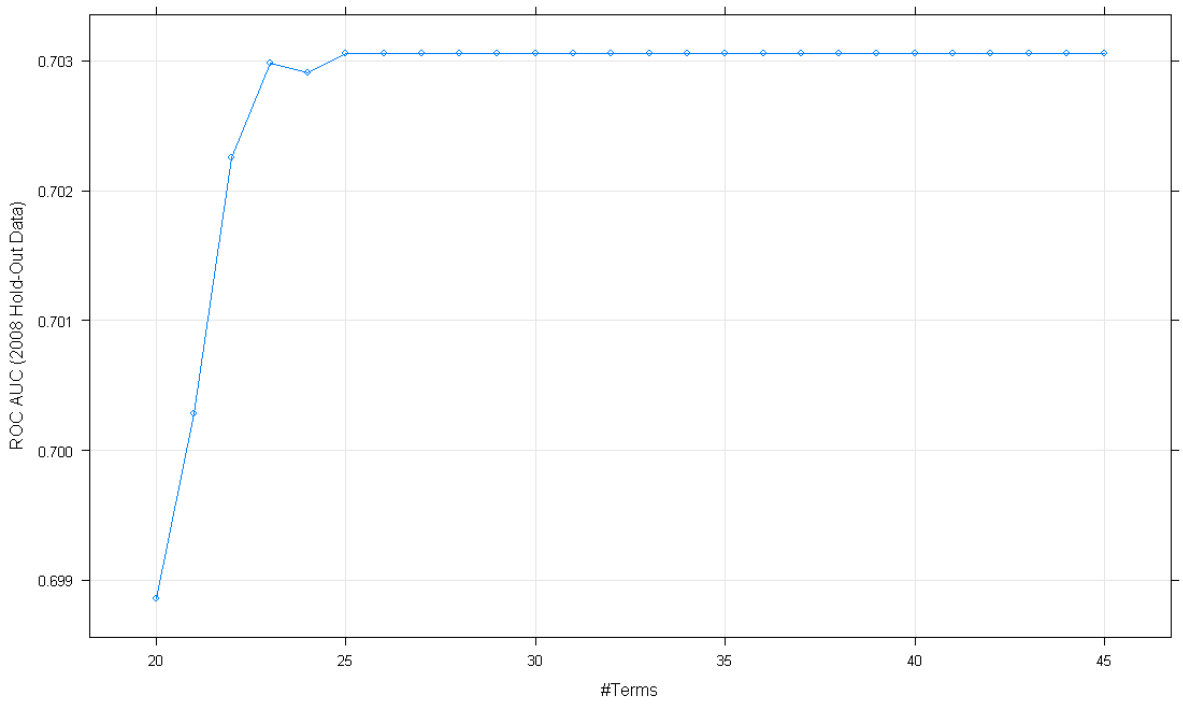


Figura 14.- ROC por parámetro [nprune]

```
> fdaFit_SAM$bestTune
  degree nprune
6      1     25
> auc( fdaFit_SAM$roc)
Area under the curve: 0.7031
> ci.auc( fdaFit_SAM$roc)
95% CI: 0.6963-0.7098 (DeLong)
```

En la Figura 14 se observa que el valor optimo se alcanza con un valor entre 24 y 25. Añadir más términos al algoritmo solo incrementa la complejidad del mismo sin aportar mejoras en el rendimiento. El valor final del AUC es ligeramente superior a 0.7.

6.3.8. Corregir la asimetría del outcome.

Cuando se modelan clases discretas, las frecuencias relativas de las clases pueden tener un importante impacto en la efectividad del modelo. La asimetría (o class imbalance en inglés) ocurre cuando una clase tiene muy poca proporción de apariciones en los datos de entrenamiento en comparación con las demás. En nuestro set de datos la asimetría de clases es bastante importante, debido a que el porcentaje de pacientes que no atienden a la cita, evento de interés, tiene una proporción alrededor del 20%, mientras el de los pacientes que asisten es de un 80%.

El efecto de la asimetría en los modelos resulta en que los patrones útiles para predecir la clase minoritaria se ven minimizados por el gran porcentaje de la clase mayoritaria. Así, las principales métricas del modelo se ven alteradas, por ejemplo, definiendo el absentismo como evento a predecir, es muy común encontrar que el modelo presenta una buena especificad, ya que casi la mayoría de pacientes asisten a su cita, pero tienen una pobre sensibilidad, es decir una pobre predicción del absentismo.

Existen diferentes estrategias para paliar los efectos de la asimetría. En este capítulo vamos a aplicar 2 de ellas, las técnicas de sampling y usar alternativos puntos de corte o umbrales para generar las predicciones.

- **Sampling**

Las técnicas de sampling consisten en modificar los datos de entrenamiento para igualar la probabilidad de ocurrencia entre clases. Así tenemos la técnica del down-sampling donde aleatoriamente se eliminan datos de la clase mayoritaria para igualar las proporciones y el up-sampling donde aleatoriamente se repiten datos de la clase minoritaria para igualarla. Por último, existen métodos híbridos o sintéticos como ROSE o SMOTE que sintetizan nuevos datos por proximidad a los datos existentes en el modelo además de realizar up o down-sampling.

Debido al volumen de nuestro set de datos las técnicas de up-sampling o SMOTE resultan prohibitivas con la capacidad de computo disponible. Por tanto, se aplicó la técnica de down-sampling a la hora de entrenar los modelos para los algoritmos seleccionados en el capitulo anterior. Aunque se consiguió suavizar la asimetría entre sensibilidad y especificidad, los valores de AUC fueron inferiores para los modelos de LDA y Glmn, por lo que se descartó la técnica para la implementación. En cambio, en el RF el valor de AUC se mantiene constante, por lo que se seleccionó el modelo conseguido a través de esta técnica.

- **Cambiar los niveles de clasificación (Alternate Cutoffs)**

Esta técnica consiste en usar la curva ROC, que representa los valores de sensibilidad y especificidad para los posibles valores umbrales de probabilidad en la clasificación, y seleccionar el punto de corte óptimo para el problema en cuestión.

En nuestro trabajo se muestra el estudio de la modificación del umbral de predicción para el algoritmo LDA. Primero observamos la Matriz de confusión con el umbral estándar

0.5. En la Figura 15 podemos ver como los valores de sensibilidad y especificidad no son operativos para su implementación.

```
> (fdaFit_SAM$CM <-fdaFit_SAMCM)
Cross-Validated (5 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

Confusion Matrix and Statistics

          Reference
Prediction No_show  Kept
   No_show     528   905
   Kept       6154 28327

               Accuracy : 0.8034
                 95% CI : (0.7993, 0.8075)
    No Information Rate : 0.8139
    P-Value [Acc > NIR] : 1

                  Kappa : 0.069
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.07902
            Specificity : 0.96904
         Pos Pred Value : 0.36846
         Neg Pred Value : 0.82152
             Prevalence : 0.18606
         Detection Rate : 0.01470
   Detection Prevalence : 0.03990
      Balanced Accuracy : 0.52403

       'Positive' Class : No_show
```

Figura 15.- Matriz de confusión antes de adaptar el umbral.

Ajustamos el umbral observando la ROC en la Figura 16.

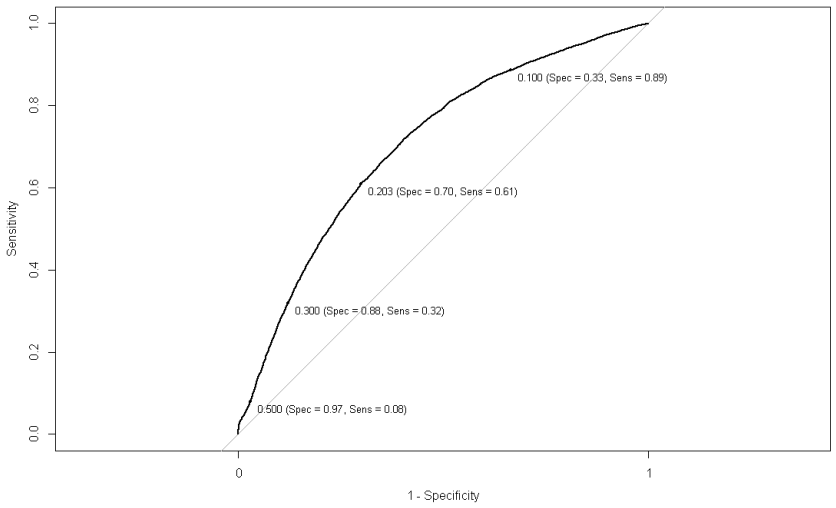


Figura 16.- Curva ROC con posibles valores de corte

```
> #Ajustando el umbral de prediccion
>     rfThresh <- coords(fdaFit_SAM_roc, x = "best", ret="threshold",
+                        best.method="closest.topleft")
> rfThresh
[1] 0.186851
```

Volvemos a clasificar las muestras con el nuevo umbral y observamos la matriz de confusión en la Figura 17.

```
Confusion Matrix and Statistics

          Reference
Prediction No_show  Kept
   No_show    4435 10256
   Kept       2247 18976

               Accuracy : 0.6519
                 95% CI : (0.6469, 0.6568)
    No Information Rate : 0.8139
    P-Value [Acc > NIR] : 1

                  Kappa : 0.214
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6637
            Specificity : 0.6492
         Pos Pred Value : 0.3019
         Neg Pred Value : 0.8941
             Prevalence : 0.1861
         Detection Rate : 0.1235
   Detection Prevalence : 0.4091
      Balanced Accuracy : 0.6564

       'Positive' Class : No_show
```

Figura 17.- Matriz de confusión después de clasificar con el nuevo umbral

Se puede observar cómo las métricas de predicción se vuelven más simétricas y se ajustan más a lo esperado. Es importante comentar que este método no mejora el rendimiento general del modelo, que está determinado por su UAC, pero si es posible ajustar las predicciones optimizándolas para un problema concreto. En nuestro caso podemos acentuar la importancia de predecir los pacientes que no asisten a sus citas.

6.3.9. Importancia de los predictores en los modelos.

Algunos algoritmos además de realizar predicciones resultan interpretables, es decir, podemos conocer las relaciones entre los factores predictivos para llegar a la conclusión del valor a predecir, además de la importancia y el peso de las variables en la predicción. En el presente capitulo analizamos la importancia de los predictores en los algoritmos seleccionados. El paquete de datos CARET presenta una función para construir un ranking con la importancia de los predictores.

La Figura 18 muestra la importancia de los predictores en el algoritmo Glmn.

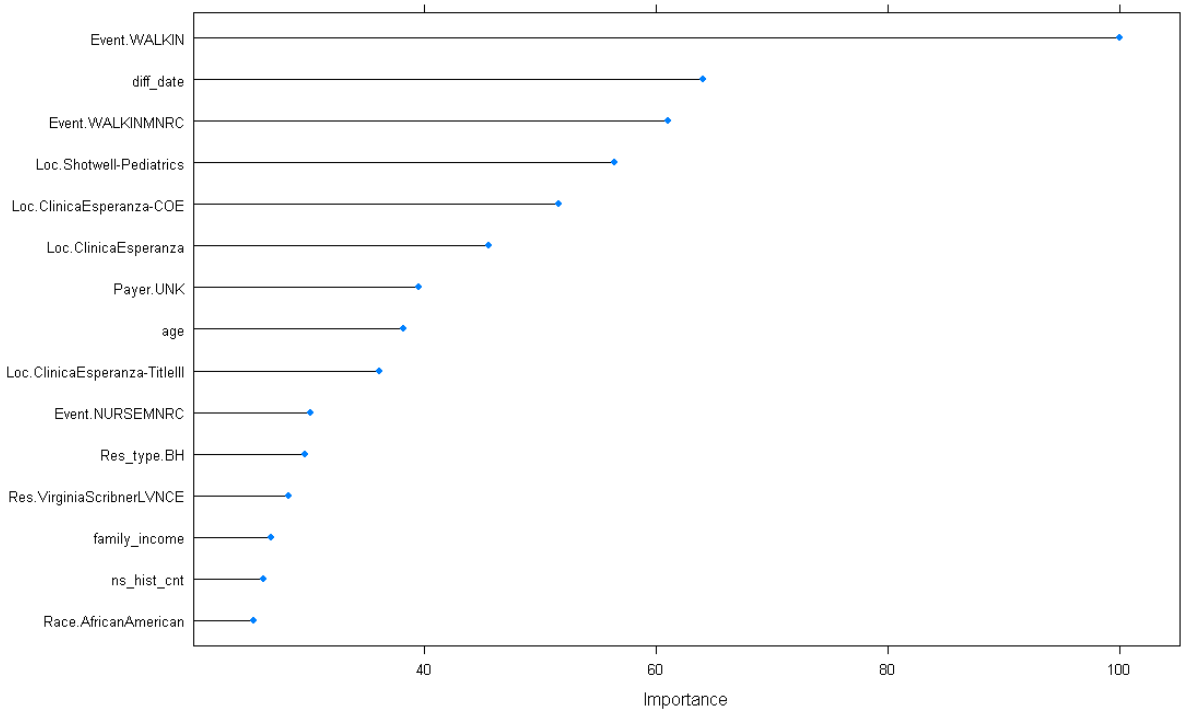


Figura 18.- Ranking predictores algoritmo Gglmn

En la Figura 19 se observa el ranking de los predictores para el Random Forest.

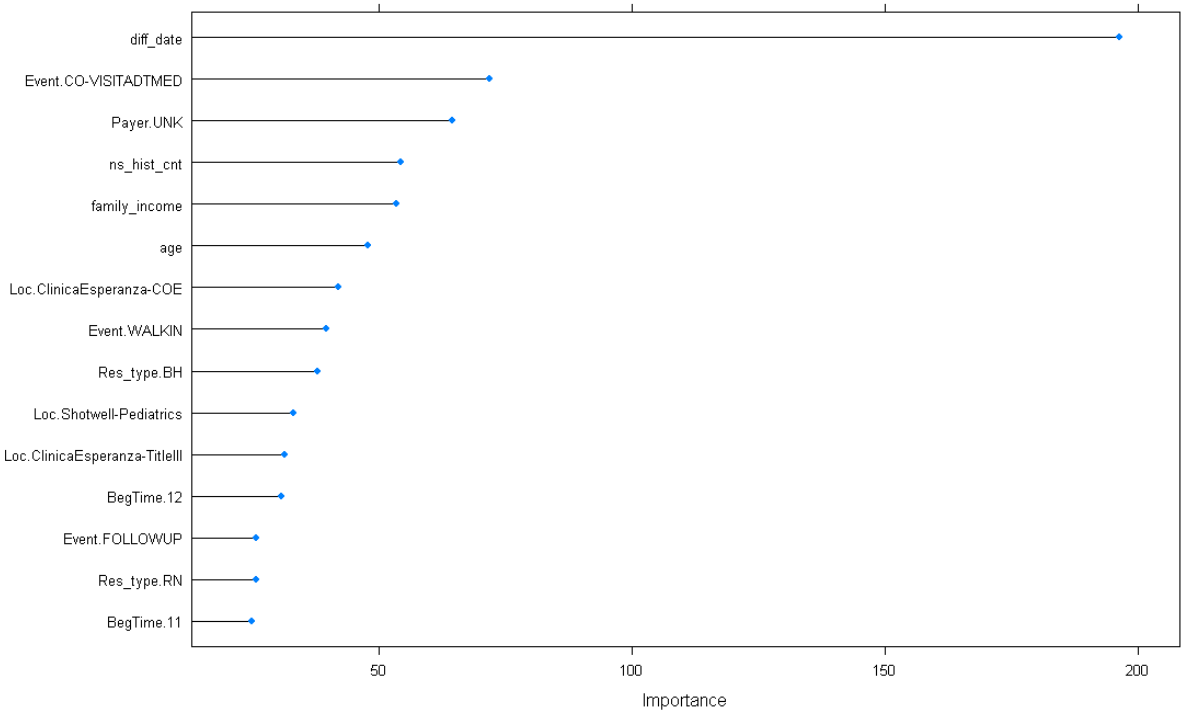


Figura19.- Ranking predictores algoritmo Random Forest

En la Figura 20 se observa el ranking de los predictores para el algoritmo FDA.

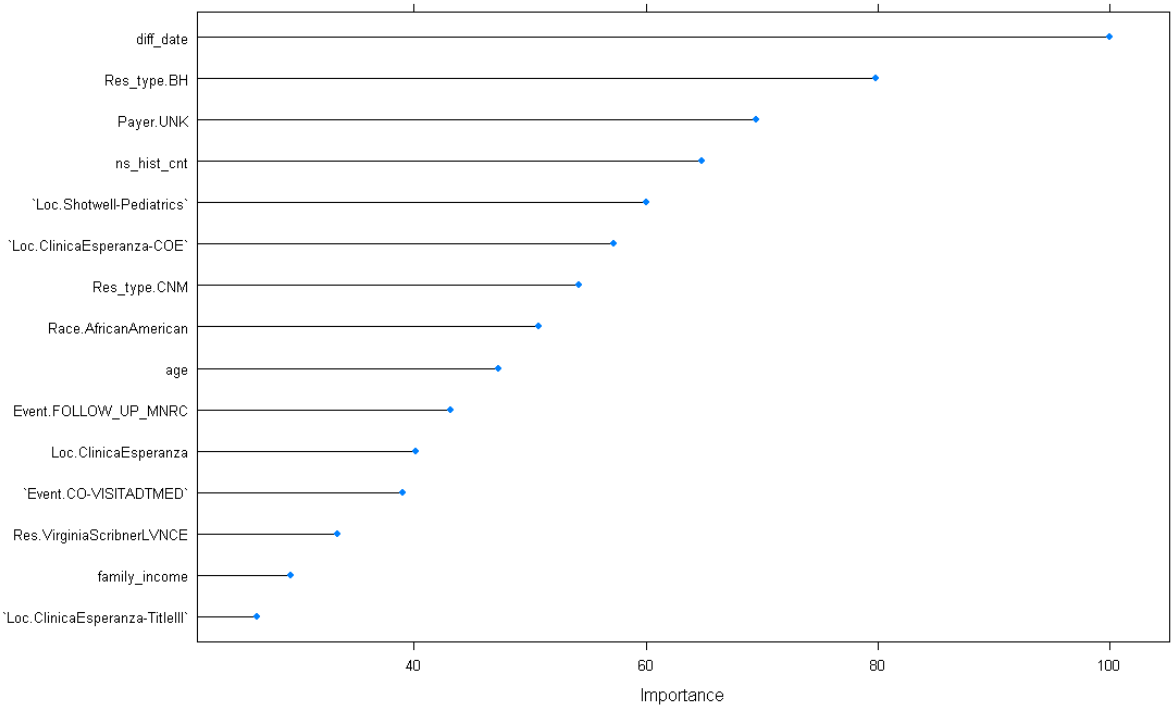


Figura 20.- Ranking de predictores algoritmos FDA.

Las conclusiones más notables observando los gráficos es que en todos los modelos, las variables creadas en el apartado de ingeniería de factores aparecen en el top 10, este hecho pone de manifiesto la importancia de crear variables con alto poder predictivo y la relevancia de esta fase en el proceso de construcción del modelo para conseguir valores óptimos de rendimiento.

Otro dato importante es la correlación de la mayoría de los factores resaltados como importantes por los algoritmos con lo hallado en el estudio bibliografía. Así, factores como el tiempo de espera (diff_date), los ingresos familiares, la edad, raza, aseguramiento (Payer), historia antigua de no atención (ns_hist_cont) aparecen como los más importantes a la hora de explicar el absentismo tanto en los algoritmos construidos como en la literatura experta.

6.4. Aplicación para realizar el overbooking.

Uno de los principales objetivos de este trabajo es realizar una aproximación practica aplicando las tecnologías de aprendizaje automático estudiadas para mejorar la productividad clínica en un entorno real. Para alcanzar este objetivo se ha diseñado e

implementado un prototipo de aplicación capaz de realizar recomendaciones de overbooking en tiempo real a partir de las predicciones de algoritmos de machine learning y los datos de planificación de los centros médicos. Se pretende evaluar si es posible con dichas recomendaciones paliar los efectos económicos y de productividad motivados por la no asistencia de pacientes, sin impactar sobre la carga de trabajo de los profesionales o en las esperas de los pacientes.

6.4.1. Análisis.

El objetivo principal de la aplicación será realizar recomendaciones del número de pacientes a citar sobre la capacidad normal diaria (overbooking) para un proveedor de salud. Para conseguirlo, el sistema ha de ser capaz de recoger datos de planificación del centro de salud, realizar predicciones utilizando los modelos predictivos creados durante el trabajo y establecer recomendaciones del número de pacientes a citar a partir de estas.

La aplicación presentada en este apartado no pretende ser una aplicación final preparada para su implementación y puesta en marcha en un entorno de producción. Mas bien pretende ser un prototipo con el que evaluar los modelos y las recomendaciones realizadas, por lo que ha de tener la suficiente funcionalidad para tomar futuras decisiones de viabilidad, desarrollo y mejora, sin tener demasiado en cuenta parámetros formales de diseño, codificación o baterías de testing.

En esta sección se detallan los requerimientos funcionales y técnicos que ha de tener la aplicación para lograr sus objetivos.

- **Requerimientos Funcionales:**

- El sistema ha de implementar la capacidad de obtener los datos de programación de citas, de los proveedores y los datos sociodemográficos de los pacientes citados.
- El sistema ha de ser capaz de utilizar los modelos predictivos construidos y generar predicciones sobre nuevos datos. Para ello se han de implementar las funciones de ingeniería de factores y preprocesamiento desarrolladas durante la construcción de los modelos.
- La aplicación ha de producir recomendaciones en lenguaje natural entendibles por el usuario final basadas en los resultados estadísticos de las predicciones.
- Se implementarán al menos 3 algoritmos predictivos distintos.

- La aplicación ha de implementar una forma de mostrar objetivamente el rendimiento general de los modelos predictivos seleccionados.
- La aplicación ha de proveer mecanismos objetivos, basados en métricas estándar para evaluar la exactitud de predicción sobre nuevas citas.
- Se ha de implementar un sistema grafico para evaluar por proveedor y turno las predicciones realizadas por el algoritmo, así como las recomendaciones finales que genera el sistema.
- La aplicación ha de ser distribuida, pudiendo ser accesible de forma concurrente por las distintas clínicas que componen MNHC.

- **Rendimiento del sistema.**

- La aplicación no ha de superar los 60 segundos de espera en la carga y el procesado inicial de los datos.
- La respuesta de los controles de la interfaz gráfica han de ser inferior a 5 segundos.
- La renderización de gráficos y la generación de recomendaciones serán menor a 15 segundos.

- **Interoperabilidad.**

- La aplicación deberá soportar la conexión con la base de datos de Nextgen para generar consultas SQL.
- La aplicación debe ser capaz de leer un conjunto de datos en texto plano .txt.
- La aplicación ha de ser interoperable con el lenguaje de programación R, el cual soporta toda la lógica de los modelos predictivos.

- **Usabilidad**.

- La interfaz ha de ser lo mas sencilla e intuitiva posible, ya que el usuario final será el personal de planificación clínica de centros de salud con un nivel de conocimiento tecnológico medio.
- La interfaz gráfica ha de proveer un sistema de selección de criterios (proveedor, rango de fechas y modelo predictivo).
- La interfaz integrara un sistema de carga de los datos por fichero.

- Los gráficos, evaluaciones y recomendaciones resultantes se mostrarán en marcos navegables implementados en un sistema de pestañas para facilitar su accesibilidad.

➢ **Seguridad.**

- Los datos utilizados y su acceso han de estar protegidos en la forma que establece la HIPAA.

➢ **Restricciones**

- Restricciones Hardware: La ejecución de algoritmos de predicción y las funciones de procesamiento de los datos tienen un alto coste computacional y de memoria. Para cumplir los requerimientos de rendimiento se recomienda al menos 8GB de RAM y intell core i5 64b 2.3 GB.

- Restricciones Software: La aplicación ha de implementarse un entorno web. Ha de implementar conexiones ODBC para ejecutar SQL querys de consulta a la Base de datos de Nextgen. Por último, debe ser compatible con las funciones de preprocesamiento de datos y algoritmos predictivos desarrolladas en lenguaje R.

6.4.2. Diseño

Una vez definidos los requerimientos que ha de cumplir la aplicación, el siguiente paso en el ciclo de desarrollo software es concretar el diseño. En los siguientes apartados se justifica la tecnología elegida, así como la arquitectura del sistema. Posteriormente se definen los distintos casos de uso implementados y la interacción usuario-sistema. Por último, se describe la interfaz gráfica.

➢ **Tecnología y arquitectura.**

Puesto que los modelos predictivos se implementaron usando el lenguaje R y la aplicación ha de ser distribuida y de entorno web, se decidió usar “Shiny” como framework para su desarrollo. Shiny es un paquete de R que permite un entorno de trabajo para desarrollar web programadas, soportando a su vez HTML5/CSS3 y Javascript+ Node.js

La arquitectura de la aplicación se compone principalmente de 3 partes:

- **ui.R:** Define la interfaz gráfica (UI) de la aplicación. Es la página web que muestra la interfaz y contiene una secuencia de comandos que controla el diseño y aspecto de la aplicación.
- **server.R:** Instrucciones que constituyen los componentes de R de la aplicación. Interactúa con la UI y contiene las funciones que el equipo necesita para construir la aplicación.
- **Funciones Externas:** Son las funciones de carga de datos, procesado y predicción. Son llamadas desde el servidor con los parámetros definidos por el usuario a través de la ui.

El código de la aplicación se entregará en un mismo archivo llamado anexo 3, adjunto con la memoria.

La aplicación puede ser ejecutada desde la línea de comando de R:

runApp("Test_Overbook")

También se puede ejecutar la aplicación desde cualquiera de los archivos en el editor de texto RStudio, mediante el botón Run App, como muestra la Figura 21.

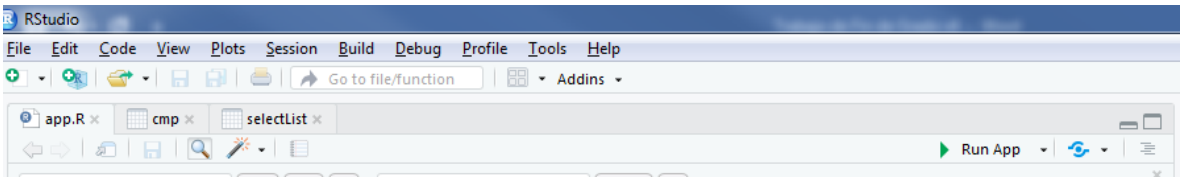


Figura 21.- Ejecución desde RStudio

➢ **Diseño de la funcionalidad.**

A continuación, se definen las funciones básicas que ha de implementar la aplicación. Estas son la carga de datos, selección de parámetros, procesado de datos, predicciones, resultados del modelo, predicciones individuales y recomendaciones diarias. La relación entre ellas se ilustra en la Figura 22.

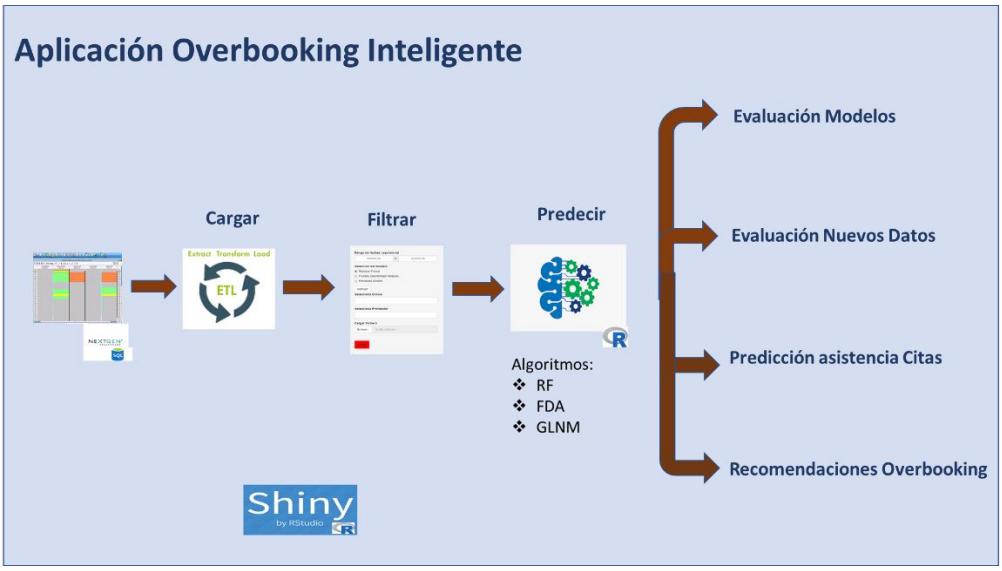


Figura 22.- Gráfico de funcionalidades de la aplicación.

- Cargar datos.

La aplicación ha de implementar un mecanismo para extraer, cargar y procesar los datos procedentes de la base de datos de citación de la clínica. Dentro del alcance de la primera versión de la aplicación la carga de datos se implementará por fichero plano, conteniendo este los datos necesarios para las predicciones. La función de carga será iniciada por el usuario final seleccionando el fichero desde la interfaz gráfica. Posteriormente la función ha de cargar el fichero seleccionado, procesar los datos y construir los factores necesarios para la entrada de los modelos.

- Seleccionar criterios y filtrar.

El usuario tendrá el control sobre los criterios de tiempo, clínica y proveedor por lo que ha de incluirse en la aplicación una función capaz de filtrar por estos parámetros antes de predecir y ofrecer los resultados. También se hace necesario incluir en la interfaz gráfica un panel de control para definir estos parámetros. Se implementará la capacidad de elegir el modelo usado para predecir entre los tres modelos implementados en la aplicación.

- Procesar datos y predicciones.

Esta funcionalidad es realizada automáticamente por la aplicación que recoge los datos preparados por la función de carga y los parámetros seleccionados por el usuario y calcula las predicciones individuales de cada una de las citas contenidas en el set de datos. La

función ofrecerá como salida los valores de probabilidad de la no asistencia y la clasificación final de las citas generadas por el modelo.

- Evaluar el Modelo.

La aplicación ha de permitir evaluar las características más importantes de los modelos predictivos desarrollados en el trabajo. Por esto, se ha de implementar una función que calcule la matriz de confusión, las curvas ROC, permita visualizar la matriz de parámetros de tunning y calcule el ranking de importancia de las variables predictores. Todas estas características del modelo serán accesibles en forma de tablas y graficas a través de la interfaz.

- Evaluar nuevos datos.

De la misma forma que es posible evaluar los modelos sobre los datos de entrenamiento, ha de ser posible evaluar cómo se comportan con los datos nuevos cargados y procesados por la aplicación. Esta funcionalidad devolverá una evaluación de los modelos usando los nuevos datos. Para ello se calculará la matriz de confusión, ROC, un grafico con el tipo de errores por probabilidad de no asistencia y unas graficas para comparar la bondad de las precisiones por día de la semana y hora.

- Predicciones Individuales

Esta función implementará un mecanismo de visualización de las predicciones de citas individuales para su revisión por los profesionales de programación de la clínica. La funcionalidad usará los valores de probabilidad y clasificación como datos de entrada y los integrará de forma gráfica por proveedor, día y hora. Ha de ser posible visualizar la clasificación entre asistencia o no, así como conocer la probabilidad de la misma por cada una de las citas.

- Consultar recomendaciones.

Por último, la aplicación implementará una funcionalidad que realice recomendaciones del número de pacientes a sobre citar por día y proveedor. Para realizar estas recomendaciones se usarán los valores de probabilidad y clasificación emitidos por los modelos, además de los niveles de confianza conocidos de las predicciones. Las

recomendaciones serán accesibles a los usuarios en forma grafica de calendario y en lenguaje natural.

- **Interfaz gráfica**

Ha de ser posible distinguir dos componentes principales:

- Panel de control a la izquierda, donde el usuario podrá seleccionar los parámetros para realizar las predicciones.
- Paneles de resultados, a la derecha. Se estructura en pestañas, donde cada una de ellas mostrara el gráfico o resultado de la predicción.

6.4.3. Implementación

El resultado de la implementación es una aplicación capaz de cargar y seleccionar los datos desde un fichero y aplicar filtros de selección sobre los mismos. Una vez los datos están en la aplicación, esta los procesa y utiliza el modelo predictivo seleccionado para generar las predicciones. Finalmente, como outcome, la herramienta permita evaluar los parámetros de modelo, evaluar el rendimiento predictivo sobre nuevos datos, consultar los datos cargados, examinar las predicciones por cita y explorar las recomendaciones de overbooking diarias. A continuación, se detallan las funcionalidades implementadas.

La Figura 21 muestra la disposición de la interfaz gráfica de la aplicación.

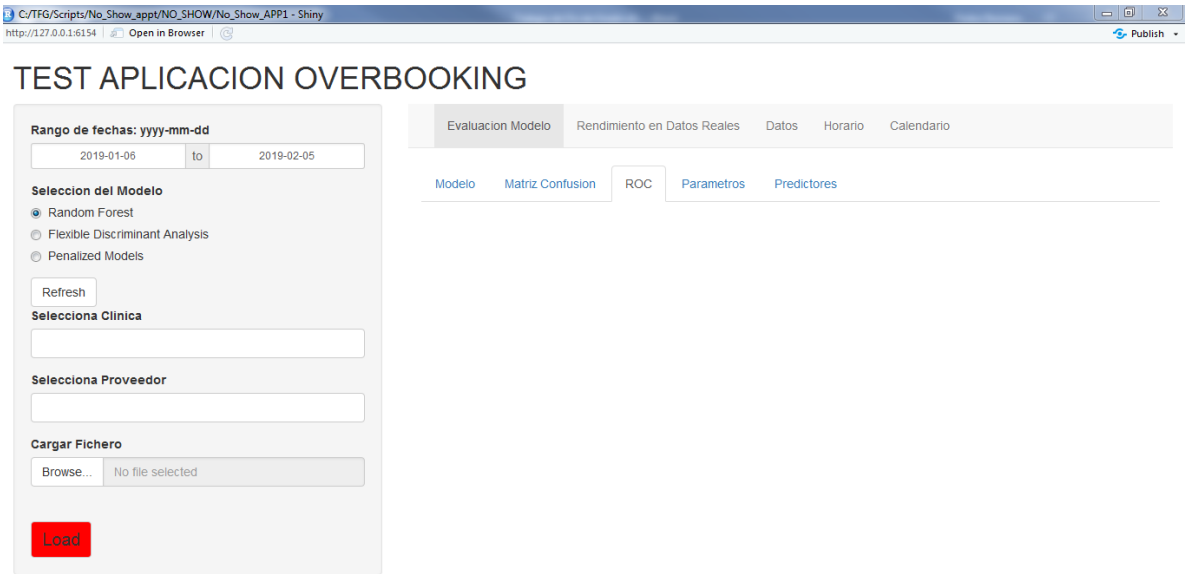


Figura 21.- Interfaz gráfica de la aplicación.

En el marco izquierdo están implementados los controles necesarios para cargar los datos y seleccionar los parámetros, pudiéndose controlar un rango de fechas, la clínica a evaluar y el médico o recurso que provee las prestaciones. También es posible elegir qué modelo de los 3 implementados se usará para predecir: Random Forest, Flexible Discriminant Analysis (FDA), Penalized Model (Glmn). La disposición de los controles se muestra en la Figura 22.

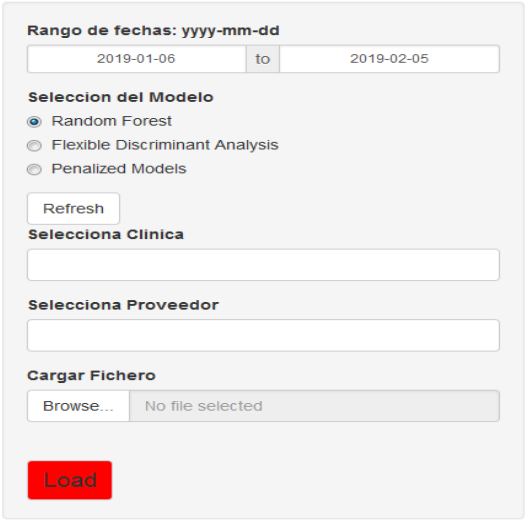


Figura 22.- Panel de control

En el Panel de la derecha se muestran los resultados de la aplicación y tiene 5 funciones diferenciadas, estructuradas en el menú horizontal superior como se puede ver en la Figura 23.

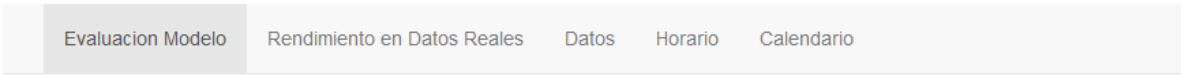


Figura 23.- Menú horizontal

En el apartado evaluación del modelo es posible revisar los parámetros de rendimiento que presenta cada uno de los modelos seleccionados. Es posible revisar las características del modelo, su matriz de confusión, la curva ROC, la configuración de parámetros para el entrenamiento y el ranking de importancia de los predictores. La Figura 24 muestra el sistema de pestañas para navegar entre las funcionalidades.

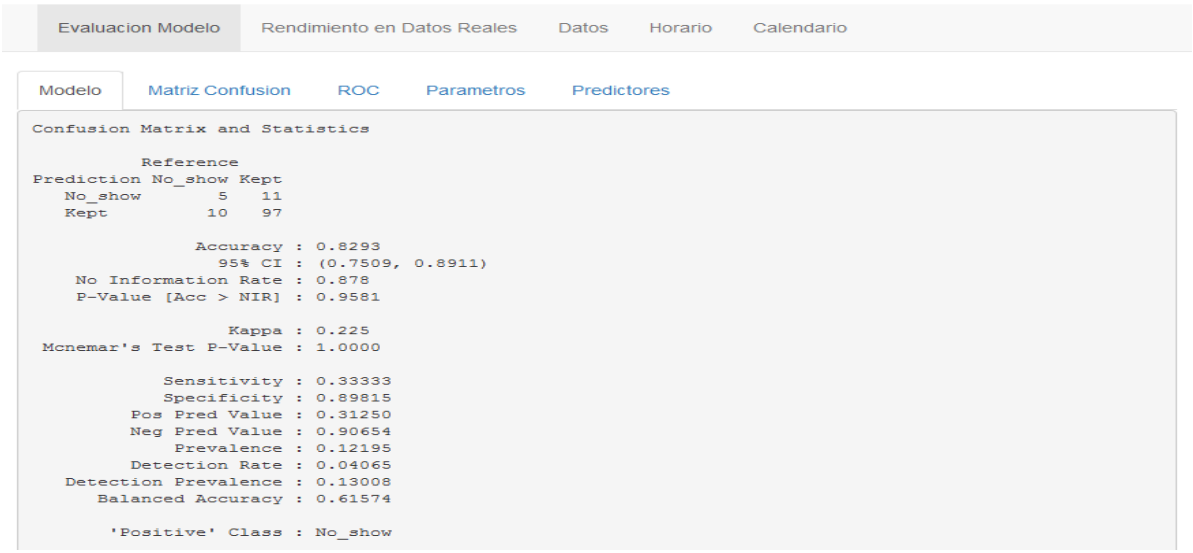


Figura 24.- Datos evaluación del modelo.

En la sección de "Rendimiento en Datos Reales" es posible evaluar de forma objetiva como el modelo se esta comportando ante los nuevos datos cargados y procesados por la aplicación. Hay 4 outcomes para interpretar los resultados.

- Matriz de confusión. Muestra el resultado de la clasificación del modelo y las métricas precisión, sensibilidad y especificidad, entre otras (Figura 25).

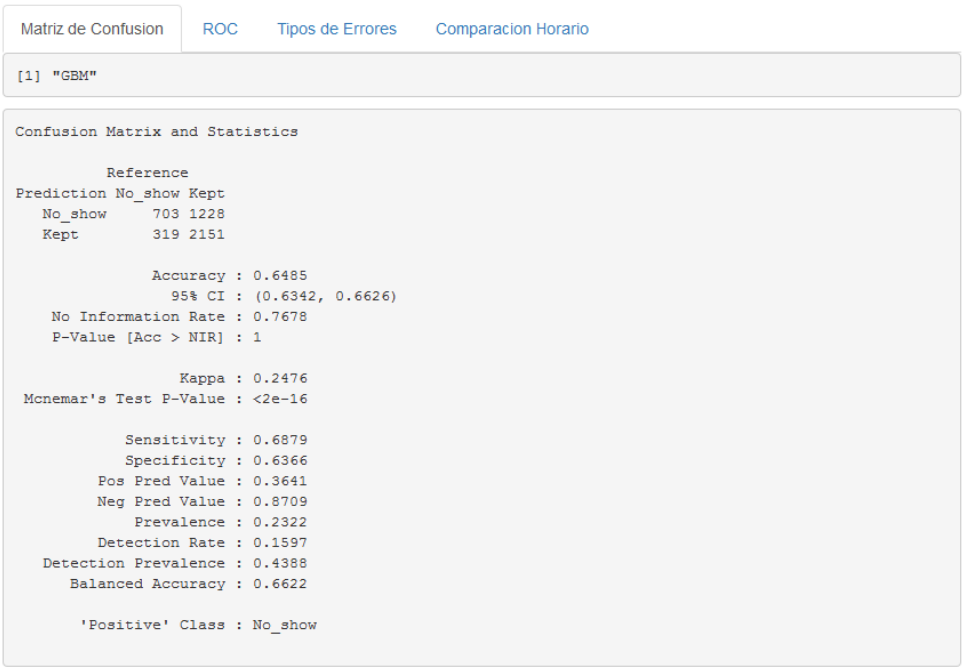


Figura 25.- Matriz confusión de los datos seleccionados.

- ROC. Dibuja la curva ROC y el área bajo la curva AUC (Figura 26).

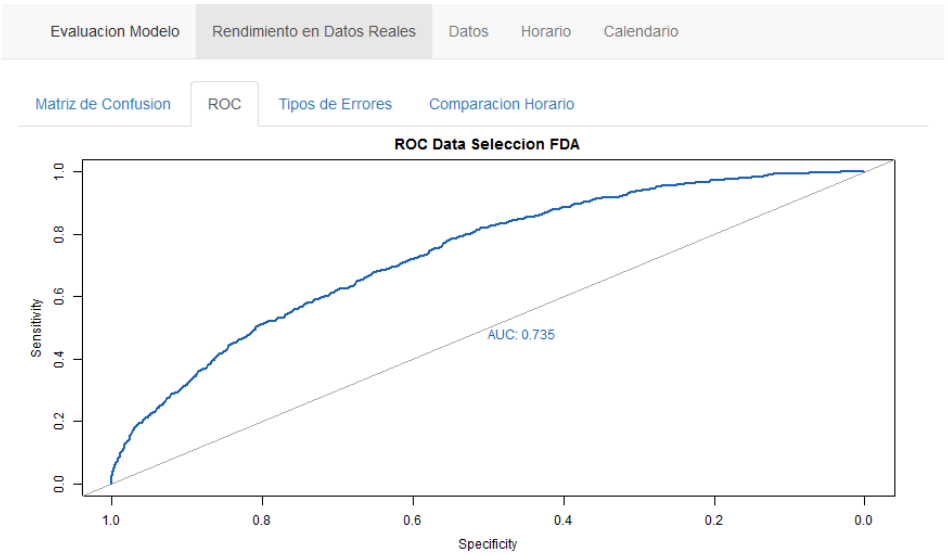
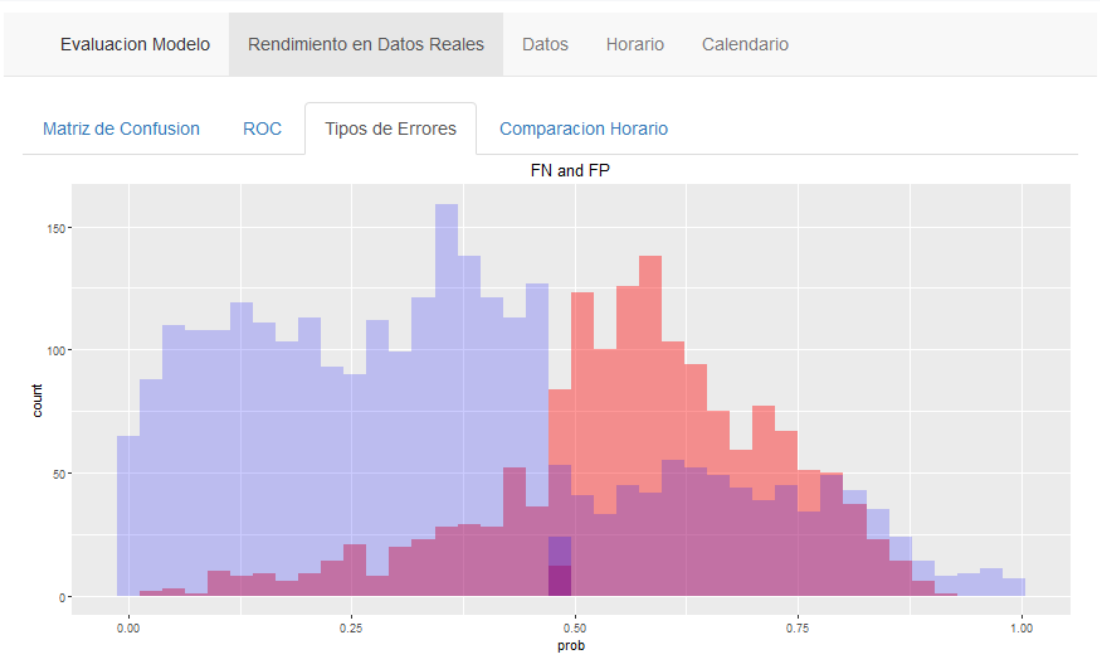


Figura 26.- Curva ROC de los datos seleccionados.

- Gráfico de tipos de errores. Muestra los falsos positivos y falsos negativos en función de la probabilidad de no asistencia de los pacientes. Esta información es útil para conocer las fortalezas y debilidades del modelo que impactaran a la hora de recomendar el overbooking en un entorno real (Figura 27).



-

Figura 27.- Distribución del tipo de error en función de la probabilidad de no asistencia.

- Comparación Horario. Comparación de dos heatmaps que permiten conocer el grado de ajuste de las predicciones a los valores reales observados por día de la semana y hora (Figura 28).

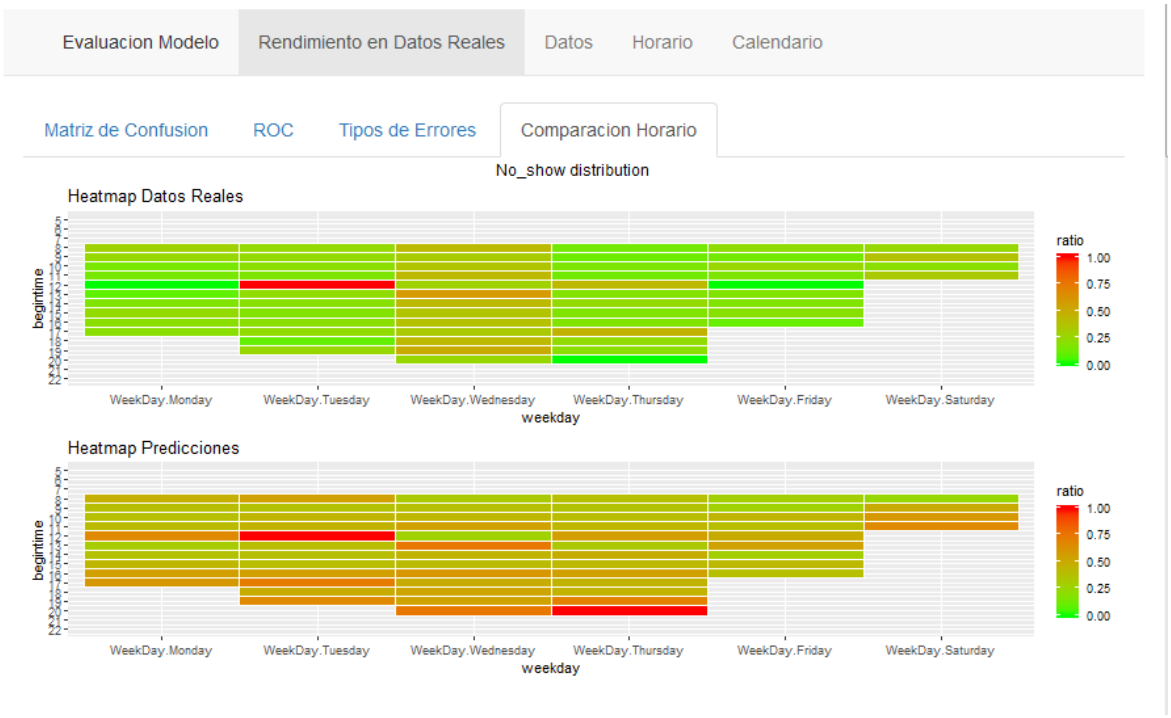
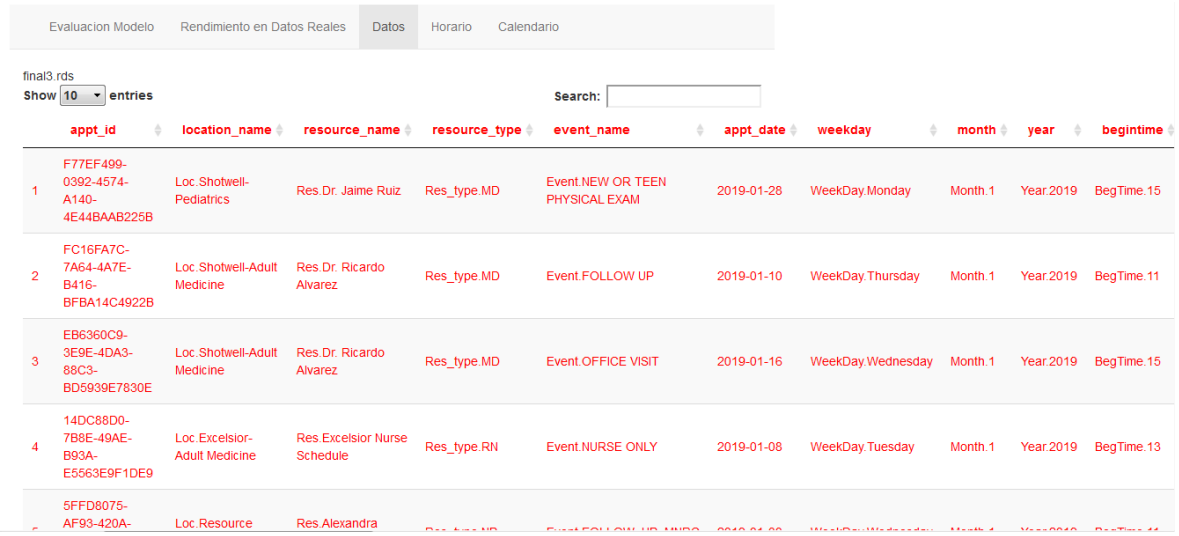


Figura 28.- Comparación de la ratio de absentismo por día y hora entre valores predicho y observados.

En la sección “Datos” del menú, Figura 29, es posible consultar una tabla con el detalle de las variables predictivas y las predicciones de cada modelo para los datos seleccionados.



	appt_id	location_name	resource_name	resource_type	event_name	appt_date	weekday	month	year	begintime
1	F77EF499-0392-4574-A140-4E44BAAB225B	Loc.Shotwell-Pediatrics	Res.Dr. Jaime Ruiz	Res_type.MD	Event.NEW OR TEEN PHYSICAL EXAM	2019-01-28	WeekDay.Monday	Month.1	Year.2019	BegTime.15
2	FC16FA7C-7A64-4A7E-B416-BFBA14C4922B	Loc.Shotwell-Adult Medicine	Res.Dr. Ricardo Alvarez	Res_type.MD	Event.FOLLOW UP	2019-01-10	WeekDay.Thursday	Month.1	Year.2019	BegTime.11
3	EB6360C9-3E9E-4DA3-88C3-BD5939E7830E	Loc.Shotwell-Adult Medicine	Res.Dr. Ricardo Alvarez	Res_type.MD	Event.OFFICE VISIT	2019-01-16	WeekDay.Wednesday	Month.1	Year.2019	BegTime.15
4	14DC88D0-7B8E-49AE-B93A-E5563E9F1DE9	Loc.Excelsior-Adult Medicine	Res.Excelsior Nurse Schedule	Res_type.RN	Event.NURSE ONLY	2019-01-08	WeekDay.Tuesday	Month.1	Year.2019	BegTime.13
5	5FFD8075-AF93-420A-	Loc.Resource	Res.Alexandra							

Figura 29.- Tabla detalle de los datos seleccionados.

Hasta el momento las funcionalidades comentadas hacen referencia a valorar el comportamiento de los modelos predictivos. Las dos siguientes tratan sobre cómo la aplicación aporta la información necesaria al personal de citación para establecer el overbooking.

La sección “Horario” permite la visualización de la agenda para un proveedor (o grupo de ellos) por día y horas. Para cada cita se muestra el resultado de la predicción asistencia (Kept) o absentismo (no_show) y la probabilidad de que el paciente no atienda a la cita. El resultado se muestra en la Figura 30. Este grafico está inspirado en la librería “Timevis” [53].

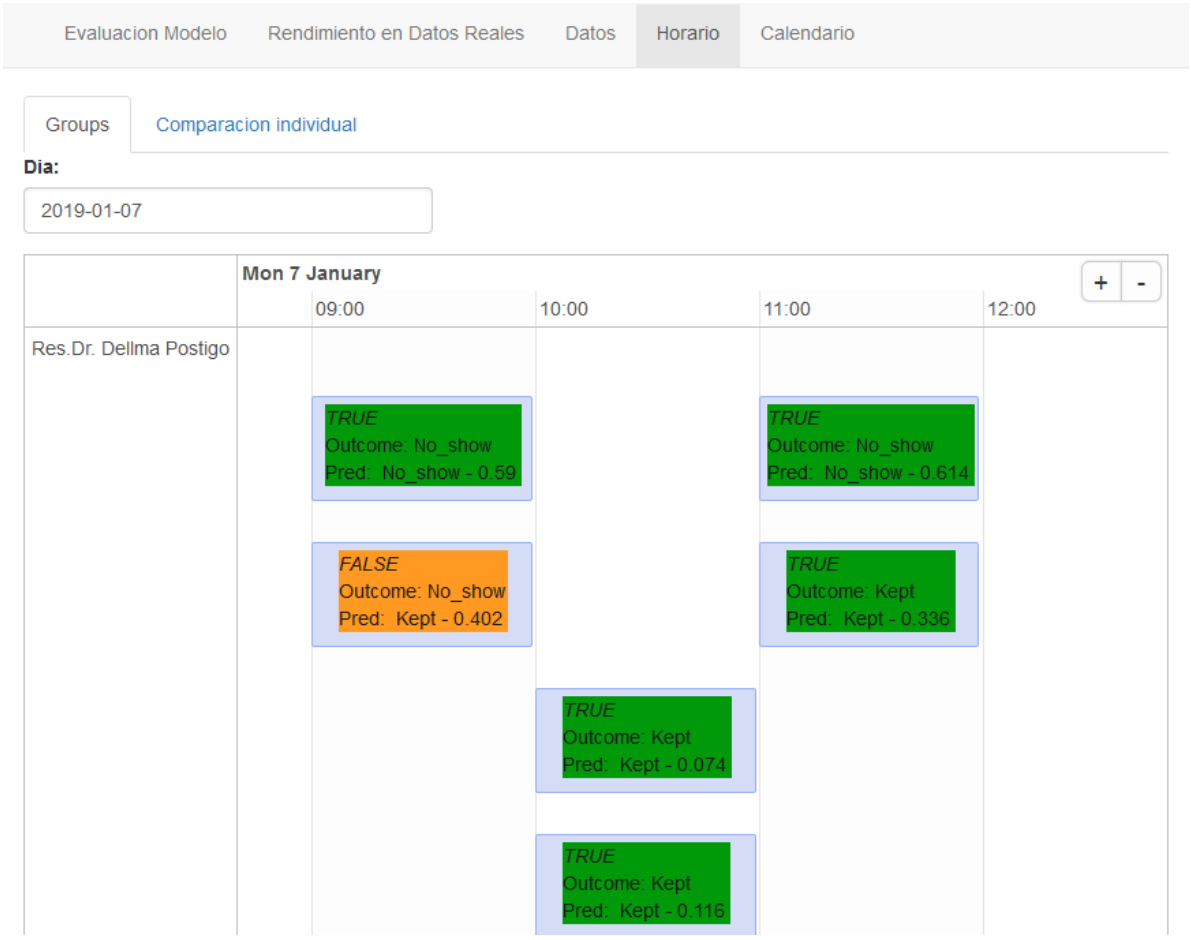


Figura 30.- Predicciones individuales de cada cita.

El subapartado comparación individual, Figura 30, aporta información del ajuste entre las predicciones por cita y los valores reales observados.

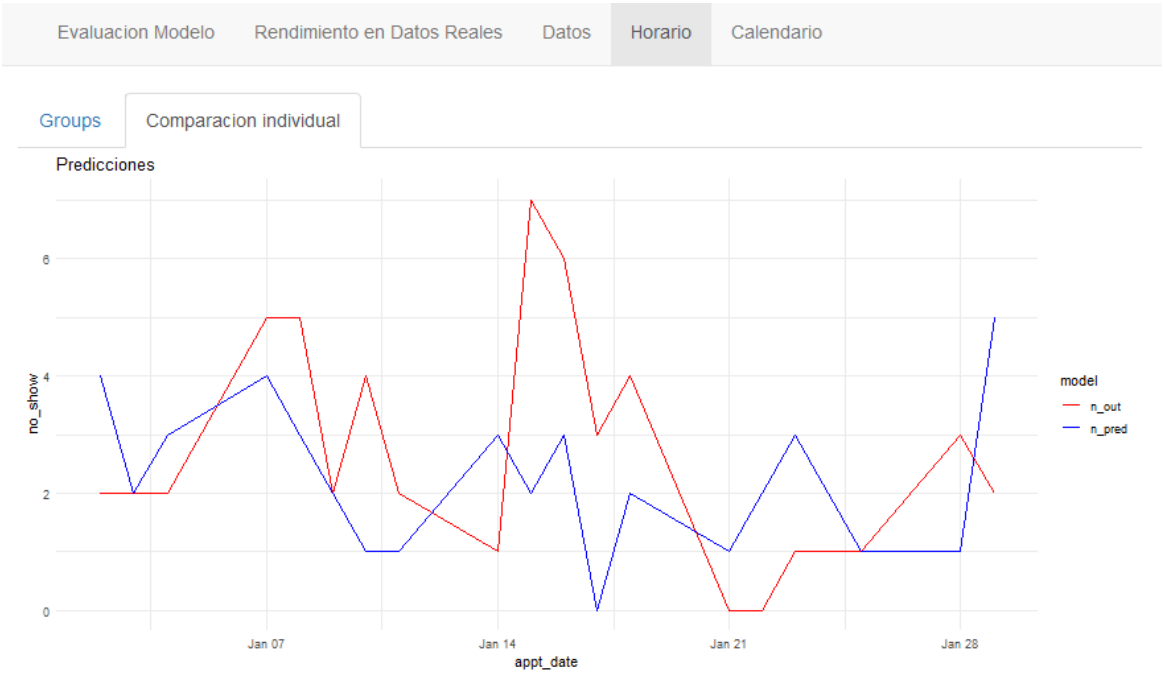
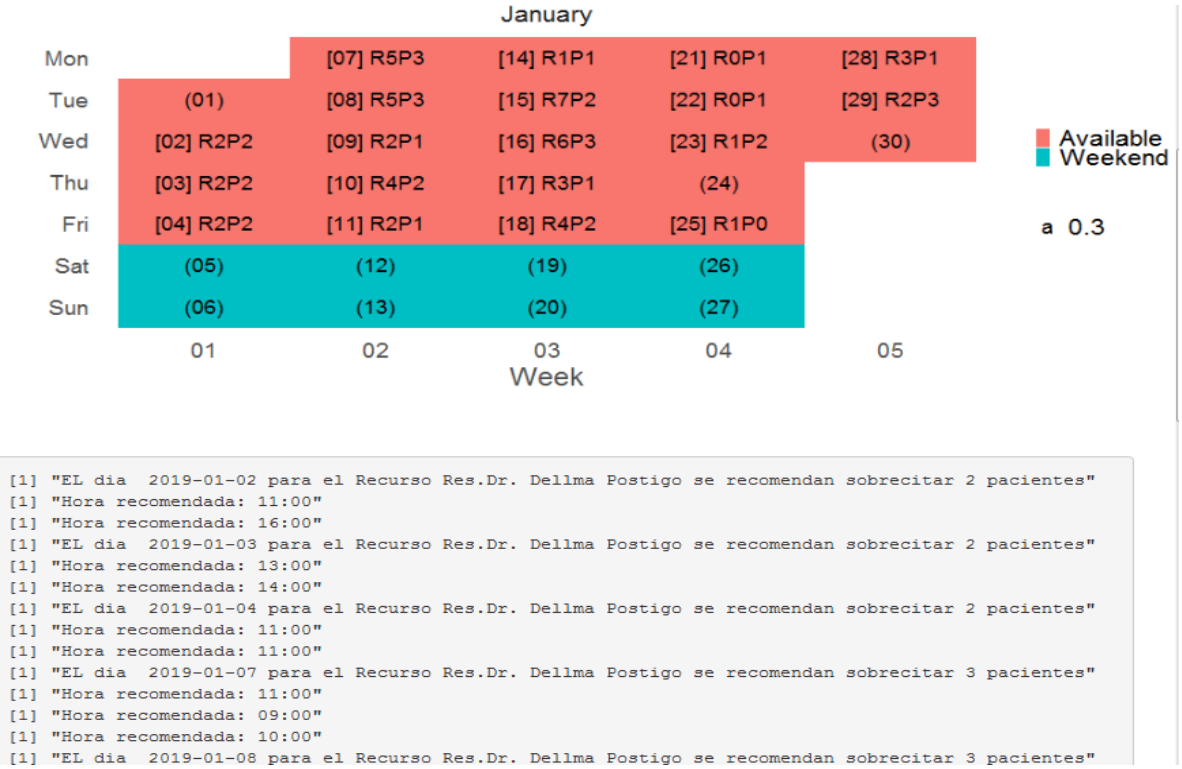


Figura 31.- Evaluación de las predicciones individuales.

La última pestaña del menú “Calendario”, es la mas importante a nivel práctico y muestra las recomendaciones de overbooking diarias por proveedor. Es posible visualizarlas de forma gráfica en un calendario o examinarlas en lenguaje natural donde además del numero de citas a sobre programar se establece una estimación de las mejores horas para hacerlo. El resultado se muestra en la Figura 32. La implementación del calendario está basada en el trabajo de Roy Francis [52].




```
[1] "EL dia  2019-01-02 para el Recurso Res.Dr. Dellma Postigo se recomendan sobrecitar 2 pacientes"
[1] "Hora recomendada: 11:00"
[1] "Hora recomendada: 16:00"
[1] "EL dia  2019-01-03 para el Recurso Res.Dr. Dellma Postigo se recomendan sobrecitar 2 pacientes"
[1] "Hora recomendada: 13:00"
[1] "Hora recomendada: 14:00"
[1] "EL dia  2019-01-04 para el Recurso Res.Dr. Dellma Postigo se recomendan sobrecitar 2 pacientes"
[1] "Hora recomendada: 11:00"
[1] "Hora recomendada: 11:00"
[1] "EL dia  2019-01-07 para el Recurso Res.Dr. Dellma Postigo se recomendan sobrecitar 3 pacientes"
[1] "Hora recomendada: 11:00"
[1] "Hora recomendada: 09:00"
[1] "Hora recomendada: 10:00"
[1] "EL dia  2019-01-08 para el Recurso Res.Dr. Dellma Postigo se recomendan sobrecitar 3 pacientes"
```

Figura 32.- Recomendaciones de overbooking por día y proveedor.

En el gráfico de calendario, las recomendaciones por día están representadas por un valor numérico y una P. En caso de conocerse los valores observados se representarán con una R.

Como en el apartado anterior, la subsección comparación mostrada en la Figura 32, permite valorar la precisión de los valores de predicción con los observados en la realidad.

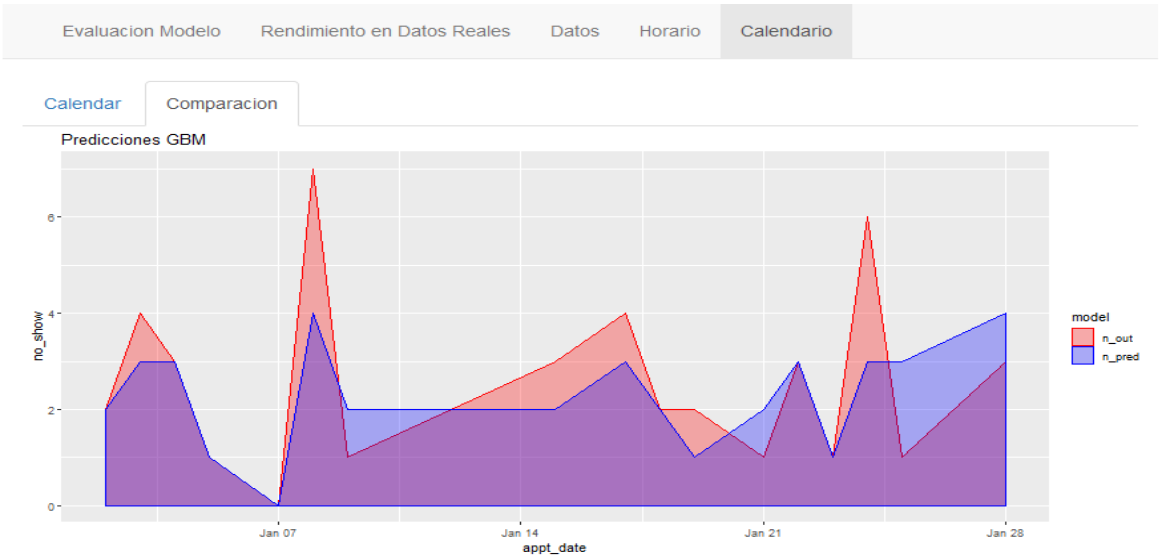


Figura 32.- Recomendaciones de overbooking por día y proveedor

7. RESULTADOS Y DISCUSION.

El absentismo de los pacientes a sus prestaciones de salud es un importante problema en el sector sanitario con impacto en los recursos económicos y en la salud de los pacientes. Es por ello que el desarrollo de nuevas herramientas e intervenciones que mejoren las actuales ratios de asistencia a las citas y minimicen sus efectos negativos se antojan fundamentales para la sostenibilidad de la sanidad. El objetivo principal del proyecto ha sido construir y evaluar un sistema predictivo con algoritmos de aprendizaje automático con el que predecir la asistencia de los pacientes a sus citas. A su vez, se ha desarrollado y evaluado en un entorno clínico real un prototipo de aplicación que utiliza las predicciones para implementar un sistema de overbooking inteligente, con el que trasladar la información en acciones concretas que mejoren los datos de asistencia. A continuación, se describen los resultados y se realiza un análisis crítico de los mismos.

Al comienzo del proyecto se hizo una revisión sistemática de la bibliografía de referencia con el fin de conocer en detalle el problema del absentismo, su impacto y los factores que lo causan. Entre los principales hallazgos encontrados destaca que el absentismo implica importantes pérdidas económicas en la sanidad pública y privada del orden de millones de Euros, reduce la productividad de los profesionales sanitarios, dificulta la planificación de recursos y aumenta las listas de espera. También se ha encontrado evidencia del impacto negativo en la salud de los pacientes, disminuyendo la adherencia a los tratamientos, el autocuidado y control de enfermedades crónicas. Atendiendo a los factores que determinan la no asistencia, tienen una importancia significativa los sociodemográficos como la edad, raza, sexo, nivel educativo, nivel ingresos, distancia a la clínica o dificultades de transporte. Otros factores como la lista de espera o el historial previo de no asistencia del paciente tienen alto valor explicativo. Factores clínicos como la gravedad del paciente, tipo de prestación programada, el proveedor de salud, antecedentes de enfermedades mentales o abuso de sustancias también determinan las ratios de asistencia.

Tras la revisión bibliográfica, se procedió a construir los sets de datos de Training y Test para entrenar y validar los modelos. Los datos fueron extraídos de la base de datos de Nextgen de MNHC, se crearon nuevos factores predictivos y se limpiaron y procesaron los datos para usar en los modelos. Con el fin de decidir qué algoritmo de clasificación usar en la aplicación final, se entrenaron y evaluaron una amplia variedad de algoritmos predictivos. En la Tabla 8 se describen los algoritmos y resultados encontrados.

Modelos Lineales						
Modelo	**Metodo**	**Data_set**	**Variable**	**ROC**	**Parametros**	**Preprocesado**
Linear Discriminant Analysis	lda	training_SAM[,reducedSet_SAM]	ldaFit_SAM	0.67	None	center, scale,zv
Partial Least Squares Discrimir	pls	training_SAM[,reducedSet_SAM]	plsFit_SAM	0.68	ncomp	center, scale,zv
Penalized Models	glmnet	training_SAM[,fullSet_SAM]	glmnFit_SAM	0.68	alpha, lambda	center, scale,zv
Sparse logistic regression	sparseLDA	training_SAM[,fullSet_SAM]	spLDAFit_SAM	0.68	NumVars, lambda	center, scale,zv
Nearest Shrunken Centroids	pam	training_SAM[,fullSet_SAM]	nscFit_SAM	0.65	threshold	center, scale,zv
Algoritmos NO Lineales						
Modelo	**Metodo**	**Data_set**	**Variable**	**ROC**	**Parametros**	**Preprocesado**
Neural Network	nnet	training_SAM[,reducedSet_SAM]	nnetFit_SAM	0.71	size, decay, bag	center, scale,spatialSign
Flexible Discriminant Analysis	fda	training_SAM[,reducedSet_SAM]	fdaFit_SAM	0.70	degree, nprune	
Support Vector Machines with	svmRadial	training_SAM[,reducedSet_SAM]	svmRFit_SAM_Reduced	0.64	sigma, C	center, scale
K-Nearest Neighbors	knn	training_SAM[,reducedSet_SAM]	knnFit_SAM	0.66	k	center, scale
Naive Bayes	nb	training_od_AM[,orig_df_fact_AM]	nBayesFit_od_AM	0.68	fL, usekernel, adjust	
Modelos basados en Arbol						
Modelo	**Metodo**	**Data_set**	**Variable**	**ROC**	**Parametros**	**Preprocesado**
Stochastic Gradient Boosting	gbm	training_SAM[,reducedSet_SAM]	gbmFit_SAM	0.74	n.trees, interaction.depth, shrinkage, n.minobsinnode	
CART	rpart	training_SAM[,reducedSet_SAM]	rpartFit_SAM	0.71	cp	
C4.5-like Trees	J48	training_SAM[,reducedSet_SAM]	j48Fit_SAM_down	0.69	C,M	
Bagged CART	treebag	training_SAM[,reducedSet_SAM]	treebagFit_SAM	0.73	nbagg	
Random Forest	rf	training_SAM[,reducedSet_SAM]	rfFit_SAM	0.74	mtry	
C5.0	C5.0	training_SAM[,reducedSet_SAM]	c50FactorFit_SAM	0.73	trials, model, winnow	

Tabla 8.- Descripción comparativa de los modelos evaluados.

La métrica elegida para comparar los modelos fue el área bajo la curva ROC (AUC), ya que nos permite la comparación de algoritmos en problemas de clasificación con dos variables. Una vez analizados los resultados sobre el set de datos de training, observamos que los modelos basados en árbol presentan el rendimiento más elevado situándose todos ellos por encima del AUC 0.7, excepto el J48 (AUC 0.68). El más destacado fue el algoritmo de Random Forest (AUC 0.74), siendo uno de los algoritmos seleccionados para implementar por su rendimiento y simplicidad.

Los algoritmos lineales presentaron un rendimiento más discreto, siendo muy homogéneos los resultados entre ellos, estando entre el 0.64 de AUC del NSC y el 0.68 de AUC del Glmn. Los valores de sensibilidad y especificidad alcanzados son demasiado limitados para su implementación en aplicaciones reales. Se seleccionó el algoritmo de penalización Glmn para implementar como línea base de referencia de un algoritmo lineal.

Otros tipos de algoritmos no lineales evaluados como las redes neuronales, Support Vector Machine (SVM), K- Neighborhood, Flexible Discriminant Analysis (FDA) o Naive Bayes, presentan un rendimiento mayor que los algoritmos lineales, aunque sin alcanzar las prestaciones de los basados en árbol. El modelo con mejor rendimiento fue el basado en red Neuronal (AUC 0,71), aunque su tamaño y complejidad hacía complicada su implementación. Por este motivo se seleccionó el algoritmo FDA (AUC 0,7) para su integración con la aplicación de overbooking.

Tras evaluar todos los algoritmos, se seleccionaron los mejores en cada categoría para realizar un tunning detallado de los parámetros y evaluar en el data set de testing. Por

su rendimiento (AUC) y viabilidad de implementación, el Random Forest, el algoritmo linear de penalización Glmn y el Flexible Discriminant Analysis (FDA) fueron los seleccionados.

Realizada una revisión en detalle de los parámetros óptimos de entrenamiento y aplicadas las técnicas para mejorar la asimetría en los datos, se obtuvieron valores definitivos de rendimiento. A continuación, en las Figuras 33 y 34 se muestra una comparativa con los valores de rendimiento y las curvas ROC.

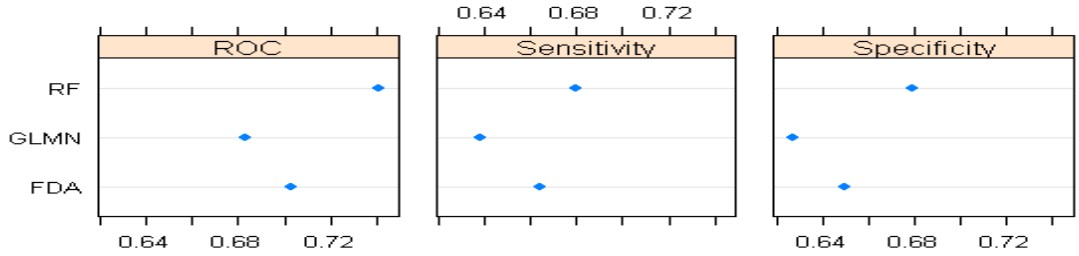


Figura 33.- Rendimiento comparado AUC, sensibilidad y especificidad de los modelos seleccionados para implementar la aplicación de Overbooking

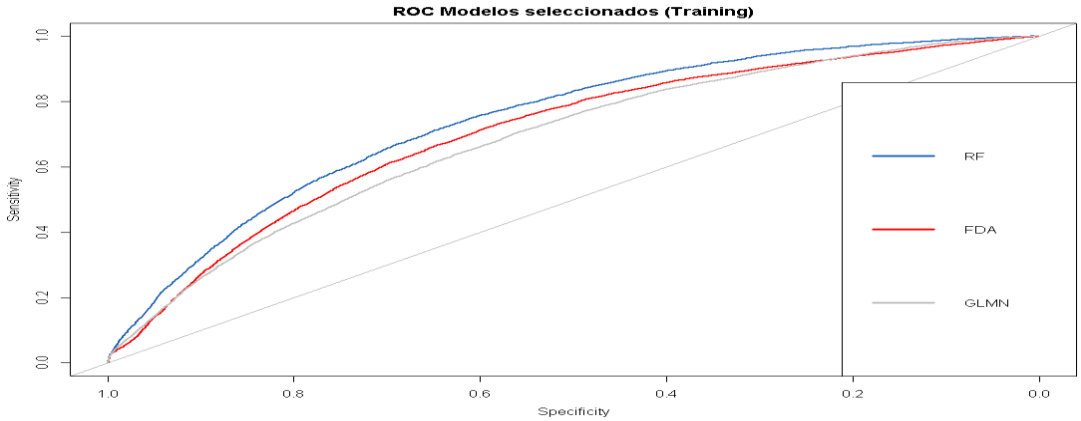
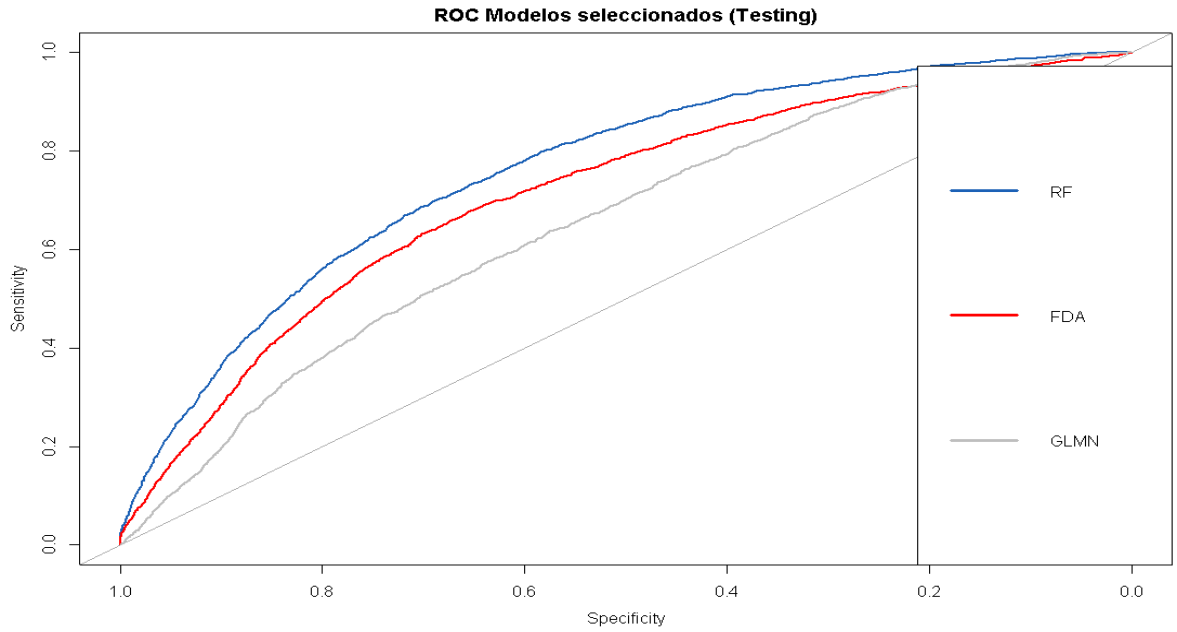


Figura 34.- Comparación ROC entre los modelos seleccionados.

Una vez completada la parametrización, los modelos fueron evaluados en el data set de Test, como prueba final para conocer el rendimiento que presentan con nuevos datos no usados para su entrenamiento. En la Figura 35 se muestran los resultados.




```
> auc(gbmTest_SAM_roc)# AUC RF
Area under the curve: 0.7601
> auc(fdaTest_SAM_roc)# AUC FDA
Area under the curve: 0.7111
> auc(glmnTest_SAM_roc)# AUC GLNM
Area under the curve: 0.6509
```

Figura 35.- Comparación ROC entre los modelos seleccionados en set de Test.

A la vista de los resultados, se puede concluir que los resultados en testing son coherentes con los obtenidos previamente al entrenar los modelos, descartando así problemas importantes de overfitting o en la metodología utilizada. El modelo con peor rendimiento es el de clasificación lineal Glmn presentando unos valores de (0.65 AUC), siendo estos demasiado limitados para implementar un modelo viable de aplicación real. El modelo Random Forest es que el que presenta mejores resultados en los datos de Testing (0.76 AUC), estos resultados, aunque son aún discretos, resultan prometedores como punto de partida, por lo que es muy interesante implementar los modelos en la aplicación para su evaluación y seguir trabajando en ciclos de mejora e incorporando nuevos predictores para su optimización.

Aunque el objetivo principal en la construcción del modelo fue el rendimiento, también ha sido posible interpretar los modelos y hacer un ranking de los factores con más peso a la hora de predecir cada uno de los modelos. En el capítulo 6 se muestran 3 gráficas (Figuras 18, 19 y 20) con los factores mas importantes por modelo ordenados por su relevancia a la hora de predecir.

Las conclusiones observando los gráficos son: Existe coherencia entre los factores a la hora de su importancia en la predicción entre los distintos modelos. Así factores como el

tiempo de espera, los ingresos familiares, la edad, raza, aseguramiento, historia antigua de no atención son importantes a la hora de explicar el absentismo en todos los modelos. Un gran porcentaje de las variables creadas en el apartado de ingeniería de factores aparecen el top 10 de importancia de todos los modelos, este hecho pone de manifiesto la relevancia de crear variables con alto poder predictivo y la gran importancia de esta fase en el proceso de construcción del modelo. Por último, es importante señalar que las variables clasificadas como importantes en los modelos, coinciden con lo encontrado en la revisión de la bibliografía inicial, por lo que el resultado de los algoritmos es acorde al conocimiento sobre el problema publicado.

Los modelos resultantes fueron implementados en una aplicación práctica y se procedió a analizar cómo estos son usados para recomendar el overbooking basado en las predicciones, evaluar su exactitud y valor en la práctica clínica real. En la aplicación se implementaron los tres modelos seleccionados anteriormente y se utilizan los datos de citación del mes de enero de 2019 de MNHC para evaluar el valor práctico de la aplicación. La aplicación permite valorar los parámetros globales de los modelos con nuevos datos, emitir predicciones de probabilidad de absentismo para citas individuales y establecer recomendaciones basadas en turnos por día, clínica y proveedor. A continuación, se muestran los resultados encontrados.

En las siguientes gráficas de la Figura 36 se muestra el comportamiento de los modelos con datos nuevos procesados de forma independiente por la aplicación para los modelos de Random Forest y FDA:

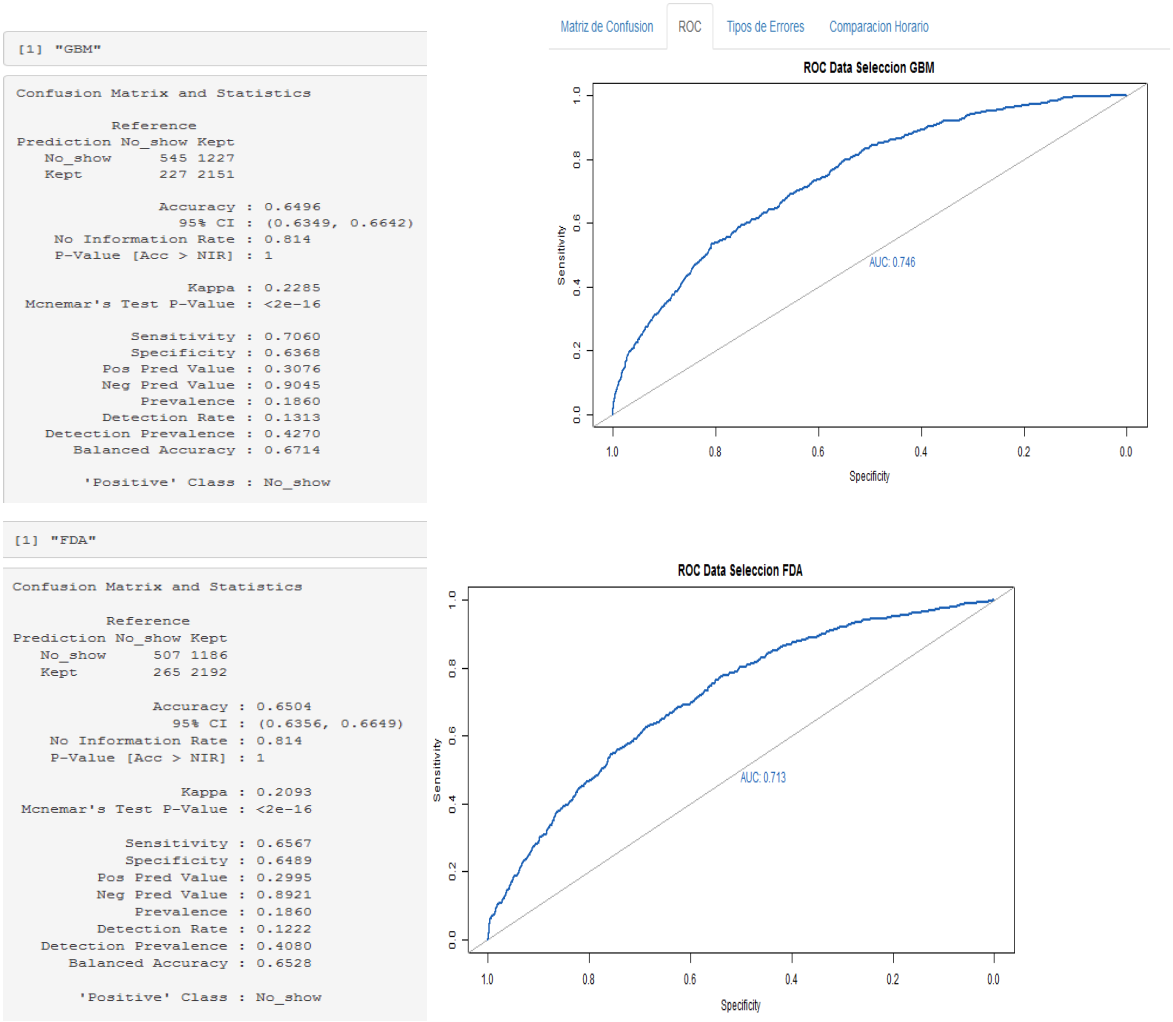


Figura 36.- Rendimiento de los modelos RF y FDA para predicciones con datos nuevos.

Se observó como los datos son coherentes con lo hallado tras entrenar y evaluar los modelos en los datos de Training y Testing. El modelo linear en cambio presenta un rendimiento mucho más bajo de lo esperado, pudiendo ser causado por alguna incoherencia en el procesado de los nuevos datos por la aplicación.

La aplicación permite realizar y evaluar predicciones de las citas a nivel individual por proveedor, día y hora existiendo una interfaz gráfica para interpretar los resultados. En la Figura 37 es posible observar el ajuste de las predicciones diarias por cita a lo que realmente ocurrió en la realidad para las clínicas con más volumen de pacientes. Los datos están referidos al modelo de RF, que es el que mejor rendimiento presenta.

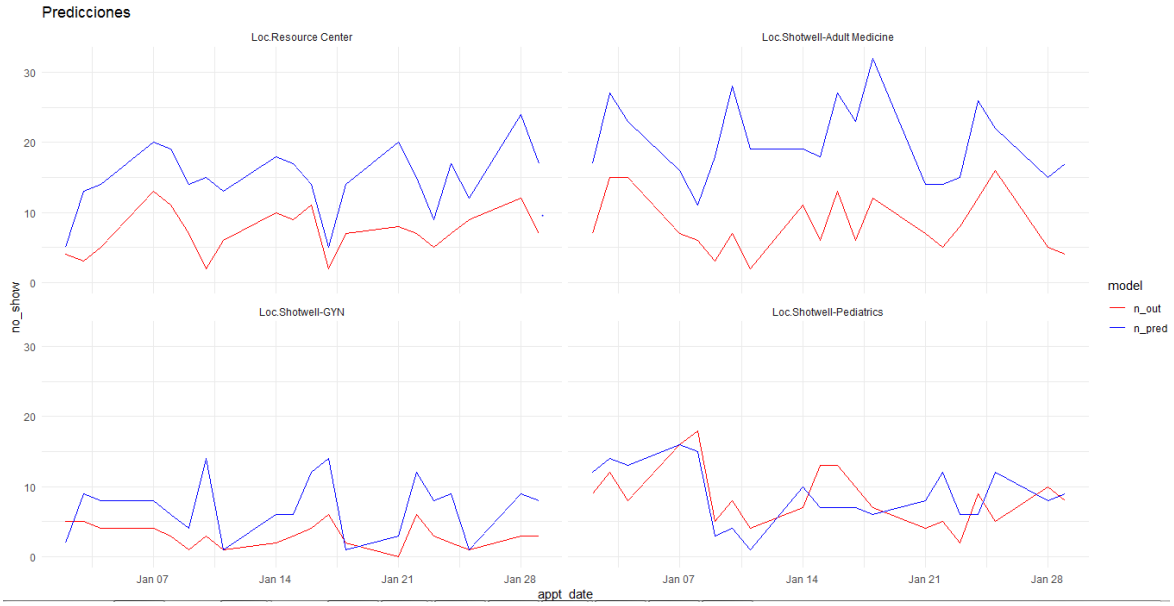
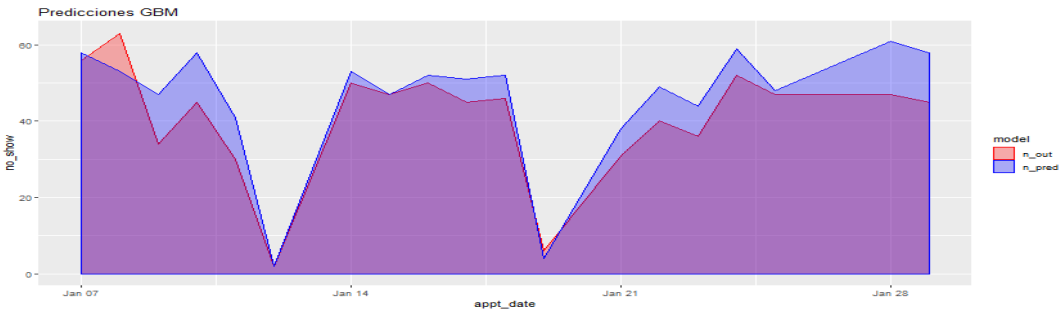


Figura 37.- Nivel de precisión de las predicciones por cita, día y clínica (rojo predicciones, azul valores reales).

Como se puede observar, los valores predichos por el RF no se ajustan con precisión a los observados. Las predicciones son heterogéneas dependiendo de la clínica, existiendo un problema de falsos positivos, es decir, predecir mas absentismo del real, en algunas clínicas como en Medicina de Adultos o el Resource Center. Para Pediatría o Ginecología los valores están más ajustados. Este resultado abre una posible vía de investigación para mejorar los resultados abordando el absentismo de cada clínica por separado y entrenado modelos solo con datos de una clínica a la vez. La conclusión de lo observado es que, si en el entorno real se siguen las predicciones individuales rigurosamente, puede llevar a una sobreutilización de recursos y por tanto al malestar de los proveedores que tienen que sobre trabajar y pacientes que tienen que esperar a ser atendidos.

La aplicación también permite realizar recomendaciones por día y proveedor ajustando las predicciones individuales a los intervalos de confianza conocidos. Los resultados para el modelo RF se describen gráficamente en las Figuras 38 y 39.



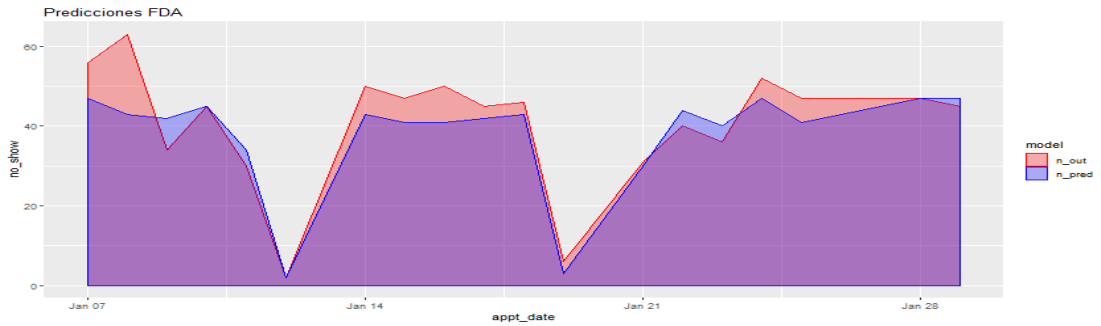


Figura 38.- Comparación entre predicciones (azul) y valores observados (rojo) agrupados por día para el mes de Enero en MNHC.

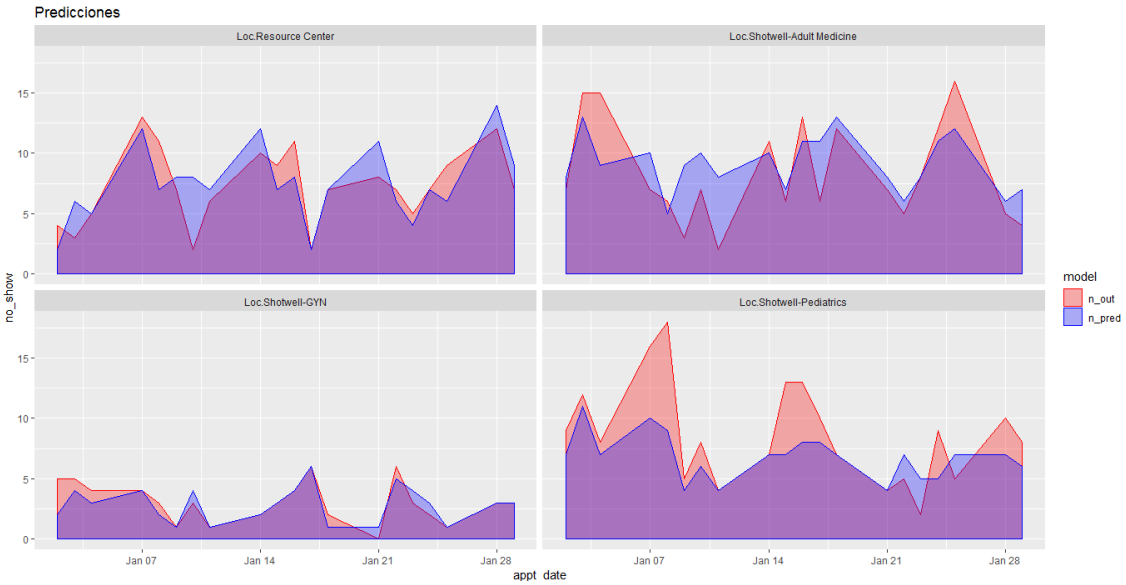


Figura 39.- Comparación entre predicciones (azul) y valores observados (rojo) agrupados por día para el mes de Enero separados por clínica.

Observando las gráficas es posible notar como las predicciones por turno se ajustan mucho mejor a las observadas en la realidad que las individuales. Los valores son ligeramente distintos por clínica existiendo problemas para predecir picos de absentismo en pediatría, así como para predecir valores mínimos en medicina de adultos. La conclusión que se puede extraer a la vista de los resultados es que las recomendaciones por turno pueden ser valiosas, por lo que se recomienda realizar un proyecto piloto de implantación en las clínicas. Aumentando la calidad de predicción de los modelos en sucesivos ciclos de mejora, añadiendo nuevos predictores y técnicas, los resultados son potencialmente prometedores.

8. CONCLUSIONES.

Durante el proyecto se ha construido y evaluado un sistema predictivo con algoritmos de aprendizaje automático con el que predecir la asistencia de los pacientes a sus citas. A su vez, se ha desarrollado en un entorno clínico real un prototipo de aplicación que utiliza las predicciones para implementar un sistema de overbooking inteligente.

Atendiendo a los resultados descritos en el capítulo anterior, la conclusión principal del trabajo es que, aunque existe aún margen de mejora en la precisión de las predicciones, el uso de las recomendaciones de overbooking de la aplicación puede aumentar la eficiencia en las clínicas, mejorar el numero de pacientes que asisten a la cita sin sobrecargar a los proveedores, aumentar la productividad diaria en las clínicas y evitar costes.

Respondiendo a los objetivos planteados podemos concluir en función de los resultados obtenidos lo siguiente:

- La realización del trabajo ha contribuido al conocimiento del problema y de los factores implicados en el absentismo.
- La predicción de la asistencia individual de los pacientes es posible, pero aún con limitaciones importantes ya que los modelos de predicción tienen un rendimiento limitado. Las técnicas de predicción con mejor rendimiento han sido los algoritmos basados en árbol.
- El sistema de recomendaciones de overbooking producidos por la aplicación puede tener valor en la práctica diaria de MNHC y es recomendable su implementación de forma controlada en un proyecto piloto.

Son varias las líneas futuras a desarrollar como complemento a este trabajo. En cuanto a los modelos predictivos es recomendable realizar ciclos de mejora de los mismos para aumentar la exactitud en las predicciones. Confeccionar nuevas variables que aumenten la capacidad predictiva, añadiendo datos clínicos de pacientes puede ser útil. Disponer de un entorno computacional más potente que permita el uso de técnicas de alto coste y mejorar los tiempos de entrenamiento, podría mejorar el proceso de desarrollo. Por último, utilizar una función con penalizaciones en función del coste de la no asistencia de un lado y el desgaste o “burnout” de los proveedores y pacientes de otro, entrenando los modelos optimizando esta función podría aumentar el valor de las predicciones adecuándolas al problema concreto.

En cuanto a la aplicación de Overbooking, el actual desarrollo permitiría hacer un estudio piloto en producción para valorar su utilidad real. Si el resultado es positivo, antes del despliegue sería necesario un nuevo proyecto con el que mejorar las prestaciones y usabilidad del software. Se recomienda usar una plataforma en cloud escalable en la que ubicar los modelos predictivos, los cuales tienen un alto consumo computacional y de memoria. Por otro lado, con el feedback de los usuarios finales seria necesario mejorar la usabilidad y el sistema de recomendaciones para optimizar su uso en el entorno de trabajo.

Agradecimientos

Quiero terminar el trabajo agradeciendo a mi Tutora Carmen Martínez sus recomendaciones y apoyo sobre todo en la recta final, sin ella no lo hubiese conseguido. A Mission Neighborhood Health Center por permitirme hacer el estudio, en especial al Dr. Ruiz por facilitarlo, apoyarme y regalarme su valioso feedback. Finalmente, muy en especial a Carol por la paciencia y cariño a pesar de la cantidad de horas robadas delante de la pantalla.

Declaración de conflictos e Intereses.

Actualmente trabajo como analista de negocio en la clínica MNHC, en la cual se enmarca el proyecto. MNHC ha permitido explícitamente el uso de los datos anónimos de citación para el presente proyecto de investigación siguiendo la normativa HIPAA. No se ha recibido ninguna financiación o ayuda para la realización del proyecto, siendo elaborado de forma totalmente individual y fuera de horas laborales.

9. BIBLIOGRAFÍA

[1] Kheirkhah P, Feng Q, Travis LM, Tavakoli-Tabasi S, Sharafkhaneh A. Prevalence, predictors and economic consequences of no-shows. BMC Health Serv Res, 2016 Jan 14, 16-13.

[2] Berg BP, Murr M, Chermak D, Woodall J, Pignone M, Sandler RS, Denton BT. Estimating the cost of no-shows and evaluating the effects of mitigation strategies. Med Decis Making, 2013 Nov, 33(8) 976-85.

[3] Jabalera Mesa ML, Morales Asencio JM, Rivas Ruiz F. Determinants and economic cost of patient absenteeism in outpatient departments of the Costa del Sol Health Agency. An Sist Sanit Navar, 2015 May-Aug, 38(2) 235-45.

[4] Beecham L. Missed GP appointments cost NHS money. *Br Med J* 1999; 319**:** 536.

[5] Hirani N, Karafillakis EN, Majeed A. Why children do not attend their appointments: is there a need for an interface between general practitioners and hospitals allowing for the exchange of patients' contact details? JRSM Open. 2016 Aug 17, (8)

[6] Chang JT, Sewell JL, Day LW. Prevalence and predictors of patient no-shows to outpatient endoscopic procedures scheduled with anesthesia. BMC Gastroenterol. 2015 Sep, 3015:123.

[7] Mani J, Franklin L, Pall H. Impact of Pre-Procedure Interventions on No-Show Rate in Pediatric Endoscopy. Children (Basel). 2015 Mar 17, 2(1):89-97.

[8] McMullen MJ, Netland PA. Lead time for appointment and the no-show rate in an ophthalmology clinic. Clin Ophthalmol. 2015 Mar, 18(9):513-6.

[9] Davies ML, Goffman RM, May JH, Monte RJ, Rodriguez KL, Tjader YC, Vargas DL. Large-Scale No-Show Patterns and Distributions for Clinic Operational Research. Healthcare (Basel). 2016 Feb, 164(1).

[10] Nancarrow S, Bradbury J, Avila C. Factors associated with non-attendance in a general practice super clinic population in regional Australia: A retrospective cohort study. Australas Med J. 2014 Aug, 317(8):323-33.

[11] Filippidou M, Lingwood S, Mirza I. Reducing non-attendance rates in a community mental health team. BMJ Qual Improv Rep. 2014 Sep, 123(1).

[12] Nuti LA, Lawley M, Turkcan A, Tian Z, Zhang L, Chang K, Willis DR, Sands LP. No-shows to primary care appointments: subsequent acute care utilization among diabetic patients. BMC Health Serv Res, 2012 Sep, 612:304.

[13] Rost L, Jenkins LS, Emmink B. Improving access to health care in a rural regional hospital in South Africa: Why do patients miss their appointments?. Afr J Prim Health Care Fam Med, 2017 Mar, 309(1)

[14] Theuring S, Jefferys LF, Nchimbi P, Mbezi P, Sewangi J. Increasing Partner Attendance in Antenatal Care and HIV Testing Services: Comparable Outcomes Using Written versus Verbal Invitations in an Urban Facility-Based Controlled Intervention Trial in Mbeya, Tanzania. PLoS One. 2016 Apr 4, 11(4)

[15] Boos EM, Bittner MJ, Kramer MR. A Profile of Patients Who Fail to Keep Appointments in a Veterans Affairs Primary Care Clinic. WMJ. 2016 Aug11, 5(4):185-90.

[16] Thongsai S. Do illness perceptions predict the attendance rate at diabetic outpatient clinic? Glob J Health Sci. 2014 Nov 16, 7(2):254-62.

[17] Gurol-Urganci I, de Jongh T, Vodopivec-Jamsek V, Atun R, Car J. Mobile phone messaging reminders for attendance at healthcare appointments. Cochrane Database of Systematic Reviews 2013, Issue 12. Art. No.: CD007458.

[18] Peterson K, McCleery E, Anderson J,Waldrip K, Helfand M. Evidence Brief: Comparative Effectiveness of Appointment Recall Reminder Procedures for Follow-up Appointments. VA ESP Project #09-199; 2015.

[19] Robotham D, Satkunanathan S, Reynolds J, et al. Using digital notifications to improve attendance in clinic: systematic review and metaanalysis. BMJ Open 2016.

[20] Sionnadh McLean, Melanie Gee, Andrew Booth, Sarah Salway, Susan Nancarrow, Mark Cobb and Sadiq Bhanbhro. Targeting the Use of Reminders and Notifications for Uptake by Populations (TURNUP): a systematic review and evidence synthesis

[21] Jayaram M, Rattehalli RD, Kader I. Prompt letters to reduce non-attendance: applying evidence based practice. BMC Psychiatry. 2008 Nov, (16) 8:90.

[22] Shah SJ, Cronin P, Hong CS, Hwang AS, Ashburner JM, Bearnot BI, Richardson CA, Fosburgh BW, Kimball AB. Targeted Reminder Phone Calls to Patients at High Risk of No-Show for Primary Care Appointment: A Randomized Trial. Gen Intern Med. 2016 Dec 31, (12):1460-1466

[23] Perron NJ, Dao MD, Kossovsky MP, Miserez V, Chuard C, Calmy A, Gaspoz JM. Reduction of missed appointments at an urban primary care clinic: a randomised controlled study. BMC Fam Pract. 2010 Oct, 2511:79.

[24] Lin CL, Mistry N, Boneh J, Li H, Lazebnik R. Text Message Reminders Increase Appointment Adherence in a Pediatric Clinic: A Randomized Controlled Trial. Int J Pediatr. 2016 Dec, 2016:8487378.

[25] Liew SM, Tong SF, Lee VK, Ng CJ, Leong KC, Teng CL Text messaging reminders to reduce non-attendance in chronic disease follow-up: a clinical trial. Br J Gen Pract. 2009 Dec, 59(569), 916-20.

[26] Koshy E, Car J, Majeed A. Effectiveness of mobile-phone short message service (SMS) reminders for ophthalmology outpatient appointments: observational study. BMC Ophthalmol. 2008 May, 318:9.

[27] Hallsworth M, Berry D, Sanders M, Sallis A, King D, Vlaev I, Darzi A. Stating Appointment Costs in SMS Reminders Reduces Missed Hospital Appointments: Findings from Two Randomised Controlled Trials. PLoS One. 2015 Sep, 1410(9)

[28] Narasimhan K. Text message appointment reminders. Am Fam Physician. 2013 Jul 1;88(1):20-1.

[29] Chiara Anna Parente, Domenico Salvatore, Giampiero Maria Gallo, Fabrizio Cipollini. Using overbooking to manage no-shows in an Italian healthcare center. BMC Health Serv Res. 2018 Mar 15;18(1):185.

[30] Cronin PR, Kimball AB. Success of automated algorithmic scheduling in an outpatient setting. Am J Manag Care. 2014 Jul, 20(7):570-6.

[31] Percac-Lima S, Cronin PR, Ryan DP, Chabner BA, Daly EA, Kimball AB. Patient navigation based on predictive modeling decreases no-show rates in cancer care. Cancer. 2015 May, 15121(10):1662-70.

[32] Reid MW, Cohen S, Wang H, Kaung A, Patel A, Tashjian V, Williams DL Jr, Martinez B, Spiegel BM. Preventing patient absenteeism: validation of a predictive overbooking model. Am J Manag Care. 2015 Dec, 21(12):902-10."

[33] Woodward B, Person A, Rebeiro P, Kheshti A, Raffanti S, Pettit A. Risk Prediction Tool for Medical Appointment Attendance Among HIV-Infected Persons with Unsuppressed Viremia. AIDS Patient Care STDS. 2015 May, 29(5):240-7

[34] Blumenthal DM, Singal G, Mangla SS, Macklin EA, Chung DC. Predicting Non-Adherence with Outpatient Colonoscopy Using a Novel Electronic Tool that Measures Prior Non-Adherence. J Gen Intern Med. 2015 Jun, 30(6):724-31.

[35] Devasahay SR, Karpagam S, Ma NL. Predicting appointment misses in hospitals using data analytics. Mhealth. 2017 Apr, 173:12.

[36] Goffman RM, Harris SL, May JH, Milicevic AS, Monte RJ, Myaskovsky L, Rodriguez KL, Tjader YC, Vargas DL. Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior in the Veterans Health Administration. Mil Med, 2017 May, 182(5)

[37] Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, Noguchi T, Ohe K. Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients With Diabetes. J Diabetes Sci Technol. 2016 May, 310(3):730-6.

[38] Daggy J, Lawley M, Willis D, Thayer D, Suelzer C, DeLaurentis PC, Turkcan A, Chakraborty S, Sands L. Using no-show modeling to improve clinic performance. Health Informatics J. 2010 Dec, 16(4):246-59.

[39] Adel Alaeddini, Kai Yang, Pamela Reeves & Chandan K. Reddy. A hybrid prediction model for no-shows and cancellations of outpatient appointments, IIE Transactions on Healthcare Systems Engineering. 2015, 5:1, 14-32,

[40] C. Elvira, A. Ochoa, J. C. Gonzálvez, F. Mochón. Machine-Learning-Based No Show Prediction inOutpatient Visits. International Journal of Interactive Multimedia and Artificial Intelligence. 2018, 4(7), 20-34

[41] Nang Laik, Khataniar; WU, Dan; and NG, Serene Seng Ying. Predictive Analytics for Outpatient Appointments. Research Collection School Of Information Systems (2014).

[42] Max Kuhn, Kjell. Johnson. Applied Predictive Modeling. New York: 2013.

[43] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction Stanford, California: August 2008

[44] Brett Lantz. Machine Learning with R. Birmingham: 2015.

[45] Web oficial de R. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[46] Webs oficiales de RStudio https://www.rstudio.com/ https://blog.rstudio.org/

[47] Web Oficial shiny. http://shiny.rstudio.com/ http://www.shinyapps.io/

[48] HIPAA for Professionals, https://www.hhs.gov/hipaa/index.html

[49] https://www.hipaajournal.com/

[50] Field Candy. Data Science Handbook. 2017

[51] Aplicación Nextgen https://www.nextgen.com/

[52] Roy Francis, Calendar Planer. http://www.roymfrancis.com/calendar-plot-shiny-app-and-dynamic-ui/

[53] Timevis code examples.-https://daattali.com/shiny/timevis-demo/