Maestría en Explotación de Datos y Gestión del Conocimiento

Análisis de las características de los pacientes y su relación con los tiempos de atención en consultorios médicos

---

Tesis de Maestría

PAUL,PABLO HERNAN

Director: Martín Volpacchio

Agradecimientos

A mi director, Martín Volpacchio, por su tiempo y su paciencia.

A Gustavo, por sus más que atinadas observaciones.

A Héctor, por su inestimable ayuda.

A mi tío Beto, por mostrarme desde hace años todo lo que hoy me gusta.

A mi mamá, Cristina, por todo.

A Ceci, a Emi y a Abril, por hacerme formar parte de una hermosa familia.

Contenido

Resumen

Gran cantidad de habitantes de los grandes centros urbanos pasan muchas horas esperando en el consultorio médico. Es habitual que una persona aguarde ser atendida por más tiempo del deseado y es muy poco frecuente ser atendido en el horario estipulado. Peor aún, la consulta suele reducirse a unos breves minutos.

Los turnos suelen ser estándar, es decir, a todos se les asigna la misma duración. Se otorgan sin tener en cuenta las condiciones en las que el paciente arriba a la consulta ni sus antecedentes, repercutiendo en la satisfacción de los mismos, en los tiempos de espera y en el estrés laboral de los profesionales de la salud [21] [22].

La visita al médico resulta estresante para la mayoría de las personas, más aún para aquellos estudios que requieren preparación previa. Esta situación, sumada a las largas esperas, aumenta la incomodidad de quien aguarda ser atendido.

El foco de este trabajo se pone en los pacientes: se busca entender por qué el tiempo de atención de algunos es más extenso que el de otros. Se analiza la existencia de patrones que permitan determinar por qué ciertos pacientes pasan más tiempo que otros en los consultorios siendo atendidos. Se toman como base distintas características de los mencionados pacientes, tanto sociales como médicas y, de esta manera, se pretende poder asignar turnos acorde a las condiciones de cada persona.

El trabajo es realizado con datos de pacientes atendidos en el gran Buenos Aires, pudiendo los mismos ser residentes de otras partes de Argentina.

Como resultado de este trabajo se ha encontrado que determinadas características de la persona y el momento en que se atiende en el centro médico (duración de los turnos anteriores, día y mes de atención, edad, cantidad de veces que concurre al centro médico, medicaciones que consume, tipo de estudio que se realiza, la zona donde reside, entre otras) son influyentes a la hora de determinar la duración del tiempo que pasan cara a cara con el médico.

Abstract

Everyday lots of people in every big city spend hours waiting to be served in medical practices. Although almost every clinic and doctor's office works with fixed appointment times, it is unlikely to be seen on schedule.

Every clinic and every doctor works in a particular way. However, the appointments are given without taking into account neither the particular conditions in which the patient arrives to the practice nor its medical record. This has an impact on patients satisfaction and physicians stress. Consequently, the visit to the doctor can become stressful for many people, even more so in the case of studies that require a previous preparation.

The objective of this work is to focus on the patient: try to understand why some patient requires a longer time inside the doctor's office than others. The existence of patterns in patient medical or social characteristics that explain the delays in medical attention will be analyzed. The present work is done with data from patients seen in a practice in Buenos Aires.

As a result, it has been found that certain characteristics as previous appointment times, day and month of the appointment, age, area of residence, medication the patient consumes, type of study, among others, are important when setting the appointment time.

Capítulo 1. Conceptos Generales y Antecedentes

1.1 Definición del Problema

En el ámbito de la salud en Argentina se está comenzando a aplicar Data Mining a estudios sobre enfermedades o vacunas. Hay escasos antecedentes de trabajos realizados respecto a la administración de los recursos, menos aún sobre estudios de los pacientes y la duración de la atención de los mismos. Por ende, la motivación de este trabajo es múltiple:

- Explorar un territorio hasta ahora virgen
- Obtener resultados que permitan mejorar la atención y satisfacción de los pacientes
- Mejorar el ambiente laboral de los profesionales de la salud
- Cambiar el paradigma de turnos estándar por turnos de duración variable

Esto implica un beneficio tanto para los pacientes como para los médicos y centros de salud, dado que podrán administrar más eficazmente sus recursos.
La relevancia social que esto puede tener es enorme, debido a que se podrá minimizar el tiempo que las personas pasan esperando al médico asignado e implementar medidas que incrementen su satisfacción, factor fundamental a la hora de elegir un profesional con quien atenderse.

1.1.1 Introducción

Big Data, Data Science, Business Intelligence, Data Mining, Científico de Datos, son palabras muy en boga hoy día. Sin embargo no siempre son usadas como corresponde ni se comprende a que se apunta cuando se las usa.

Según Pang-Ning Tan [12] la minería de datos (Data Mining) es el proceso automático de descubrimiento de información útil en grandes repositorios de datos. Investigadores del ámbito de la medicina, ciencia e ingeniería están acumulando grandes cantidades de datos que son clave para nuevos descubrimientos. En salud el término Big Data hace referencia a volúmenes de información tan grandes y heterogéneos que no pueden ser manejados con el software y hardware tradicionales, ni fácilmente analizados con las herramientas convencionales de gestión de datos. En atención sanitaria interesa especialmente la información de la historia clínica, sistemas de introducción de órdenes y prescripciones médicas (recetas, laboratorio, derivaciones, etc), los sistemas de almacenamiento y comunicación de imágenes y bases de datos varias (altas hospitalarias, registros de mortalidad, de urgencias, de hospitalización a domicilio, etc) o de reembolso (dispensación farmacéutica, facturación de servicios cómo prótesis, ambulancias, etc). Lo llamativo de esta enorme cantidad de información es que todavía es vista como un subproducto de la atención sanitaria antes que como una herramienta central para mejorar la calidad, la seguridad y la eficiencia.

1.1.2 Data Mining en Salud

En el sector de la salud específicamente son varios los usos que se le están comenzando a dar al Data Mining: reconocimiento de imágenes en el cerebro, predicción y tratamiento de enfermedades de especialidad analizando la sintomatología, las enfermedades y los resultados de estos tratamientos [13]. También se aplica a la vigilancia de eventos, estudio epidemiológicos, planeación y evaluación de estrategias por zonas de salud mediante GIS (Geographic Information System – Sistemas de Información Geográfica) [14]. Incluso se están dando usos más puntuales como detección de cáncer cervical [15] y farmacovigilancia [16].

De acuerdo a Xuezhong Zhou *et al.* [17] la minería de datos aplicada a la medicina ha sido un tópico muy investigado en los últimos años. Aplicar la minería a datos médicos, de salud o

clínicos está considerado como el dominio más difícil para el Data Mining. Esto se ve justificado por la enorme cantidad de información generada a partir de los nuevos dispositivos electrónicos que se utilizan para recolectar datos sobre la salud de las personas. La gigantesca cantidad de datos almacenada en bases de datos hace extremadamente difícil, sino imposible, para los seres humanos analizarlos y obtener información [18]. Las fuentes de datos podrían resumirse a las siguientes:

Fuentes de Datos	**Externas**	Datos provenientes de empresas aseguradoras, tecnológicas, bancos, censos, obras sociales, etc	Mensajes de redes sociales
	Internas	Historias clínicas electrónicas, resultados de laboratorio	Fichas en papel, notas manuscritas, radiografías
		Estructurados	**Desestructurados**
		Tipos de dato	

Cuadro 1. Fuentes de datos en Salud

A pesar de lo mencionado anteriormente, la explotación de datos en el ámbito de la salud es un campo todavía virgen a nivel mundial. Si bien se han hecho algunas incursiones, en general está recién dando sus primeros pasos. En España, por ejemplo, el Consejo Asesor de Sanidad dependiente del Ministerio de Sanidad, Servicios Sociales e Igualdad, ha publicado en el año 2014 el informe "E-Salud: prioridad estratégica para el sistema sanitario", un documento que aborda el impacto positivo de la incorporación de las nuevas tecnologías sobre el Sistema Sanitario y su impacto en la mejora de la eficacia, calidad, accesibilidad y seguridad de los servicios de salud.

1.1.3 Estado del Arte

La relevancia que el Data Mining tendrá a futuro en la salud se puede ver reflejada en como los grandes jugadores de la industria hacen foco en el tema. En el año 2009 IBM anunció el lanzamiento del Health Analytics Solution Center, una red global de centros que buscaban dar respuesta a la creciente demanda de analítica avanzada para asistir a los centros médicos y a los profesionales de la salud para ayudar en la toma de decisiones y proveer cuidados de más alta calidad. Estos centros fueron abiertos en Dallas, Berlín, Beijing, Tokio, Nueva York, Londres y Washington DC. Solo en 2011 IBM invirtió más de U$S 6.000 millones en investigación y desarrollo en tópicos relacionados a big data y salud [19]. Además lanzaron al mercado Watson Health, una herramienta que da la "bienvenida a la era cognitiva de la salud". Esta aplicación tiene un producto especialmente destacado: Watson Oncology Advisor, una solución que permite hacer un tratamiento personalizado del cáncer, permitiendo comparar un caso con otros de todo el mundo. Reflejo del potencial del sector de la salud es la inversión de capital de riesgo que se vuelca al mismo: durante el 2014 el Silicon Valley Bank contabilizó cerca de un 20% de las inyecciones de capital en el segmento. Un año antes el share era de apenas 12%. Deloitte calcula que para 2017 los ingresos que generan apps móviles de salud totalizarán U$S 26.000 millones. Hace dos años sumaban U$S 2.400 millones.

No hay que olvidarse de aquellas empresas que diseñan las herramientas para recolectar datos. Apple, por ejemplo, incluye en sus teléfonos una app que permite cargar más de 70 datos de salud. Samsung, por su parte, también desarrolló una aplicación para celulares que ayuda a crear hábitos saludables y monitorear el estado de salud del usuario. Por otro lado, varios de los estudios que se realizan actualmente, aún los más rutinarios y comunes (monitoreo ambulatorio de presión arterial, monitoreo electrocardiográfico continuo con sistema holter, marcapasos, radiografías, resonancias magnéticas, etc) generan infinitas cantidades de datos que, tal como se mencionaba anteriormente, son imposibles de evaluar en su totalidad por una persona. Muchas empresas tecnológicas de primera línea (Fitbit, Apple, Google, Garmin, por ejemplo) han desarrollado wearables, dispositivos móviles para controlar constantes vitales, para que cada persona pueda monitorear su propia salud.

1.1.4 Aspectos Legales

Está claro que la salud es un diamante en bruto para la minería de datos. Sin embargo existen ciertas cuestiones que deben ser tenidas en cuenta, cómo lo referido a la privacidad de los datos. En el año 2000 el Congreso Argentino sanciona la ley 25.326 de Protección de Datos Personales. En su artículo 2° define como dato sensible, entre otros, a la información referente a la salud de una persona. En su artículo 7° reza "Los datos sensibles sólo pueden ser recolectados y objeto de tratamiento cuando medien razones de interés general autorizadas por ley. También podrán ser tratados con finalidades estadísticas o científicas cuando no puedan ser identificados sus titulares.". Mientras que en el artículo 8° detalla que "Los establecimientos sanitarios públicos o privados y los profesionales vinculados a las ciencias de la salud pueden recolectar y tratar los datos personales relativos a la salud física o mental de los pacientes que acudan a los mismos o que estén o hubieren estado bajo tratamiento de aquellos, respetando los principios del secreto profesional.". Muchos hospitales, tanto a nivel internacional como nacional, han creado redes interconectadas para generar y compartir sus historias clínicas electrónicas. Y ahí es cuando surgen dos inconvenientes: el primero es que la mayoría de la información sanitaria se encuentra en soporte papel (recetas, historias clínicas, informes, etc). La segunda es no superar el límite de lo permitido por la ley. Se podría agregar un tercer inconveniente que es la necesidad de digitalizar y estandarizar toda la información existente en soporte papel para poder explotarla. En Argentina la única institución que está empezando a trabajar este tipo de tecnologías es el Hospital Italiano, referencia obligada al momento de hablar de tecnologías de la información aplicadas a la medicina: cuenta con un departamento, una residencia y una maestría de Informática en Salud. Dicha entidad, junto con otras 11, han participado de la creación del MAIS (Marco Argentino de Interoperabilidad en Salud) mediante el cual se intercambia información de las historias clínicas bajo el estándar HL7 CDA (Clinical Document Architecture [45]) con sanatorios y hospitales del país y el exterior.

1.1.5 Contexto Local

La situación actual del sistema de salud en la Argentina dista de ser la ideal, tanto desde la perspectiva de los médicos como de los pacientes. Los primeros sufren una creciente frustración por la brecha que existe entre los intentos por brindar una atención ideal y las restricciones que impone el actual sistema de atención médica. Limitaciones en su capacidad para tomar decisiones clínicas independientes, incremento de tareas administrativas, restricciones de las instituciones pagadoras o de las Obras Sociales y Prepagas, son algunos de los problemas a los que deben hacer frente. Ante esta situación la mayoría de los médicos deben trabajar más horas y, además, aceptar la asignación de mayor cantidad de pacientes en detrimento de su vida personal y su tiempo de capacitación. Vale aclarar que muchos de los recintos donde los

profesionales se desempeñan son alquilados como poli-consultorios por entidades definidas como centros médicos. El valor de dicho alquiler no se encuentra regulado y lo determina el dueño del centro médico. Entre los principales gastos que poseen los médicos se encuentran los siguientes:

- Matrícula (provincial y/o nacional)
- Caja de Médicos
- Seguro Mala Praxis
- Alquiler del Consultorio
- Porcentaje de la facturación a abonar a la Caja de Médicos
- Porcentaje de la facturación a abonar al Círculo Médico
- Impuestos Provinciales y Nacionales (Ingresos Brutos, Iva, Ganancias, etc)

Es decir que todos los meses un profesional comienza con un saldo negativo de unos $15.000 aproximados (según datos de 2018).

Además son observados constantemente por Obras Sociales y Prepagas por un lado y empleadores por otro, quienes analizan la cantidad de derivaciones, de estudios o prestaciones que cada profesional realiza en determinado período de tiempo [21]. Según Gottschalk y Flocke [24] el 55% de las horas de trabajo de un médico transcurren en el "cara a cara" con el paciente. El resto se va en trabajo administrativo fuera del consultorio.

Mientras tanto, los pacientes bregan por conseguir un turno dentro de plazos razonables (dependiendo si la institución es privada o pública se suelen otorgar turnos a 20, 40 o más días), deben padecer largas esperas en los consultorios y, cuando son atendidos, el tiempo que pasan cara a cara con el médico es muy corto. Según una encuesta realizada en México en el año 2006 los principales factores asociados a una percepción de un servicio de salud de baja calidad son el tipo de institución, el tiempo de espera, el mejoramiento del cuadro luego de la consulta y un tiempo de consulta menor a 20 minutos [22]. La Organización Mundial de la Salud define a los turnos demasiado breves como uno de los factores que impide la realización de una buena promoción de la salud [25].

En el extremo opuesto se encuentra el ausentismo a las consultas. Un paciente que pide un turno puede llegar a conseguir recién para 3 o 4 meses más adelante y al llegar el día no concurrir. Si se pudiese gestionar eficazmente la cancelación de los turnos de aquellos pacientes que no van a asistir se podría optimizar el uso de los recursos. Las tasas de ausentismo en las clínicas y consultorios consultados superan el 35% (porcentaje que se incrementa mucho si el día de la cita el clima no es favorable). Vale aclarar que esta temática, si bien se menciona ocasionalmente durante el marco introductorio, queda fuera del alcance del presente trabajo.

La amplia mayoría de los profesionales, tanto en consultorios particulares o sanatorios privados, otorgan turnos de 10, 15 o 20 minutos. Según Outomuro y Actis [28], la duración promedio de turnos en clínica médica en Buenos Aires es de 15 minutos. La duración del turno es una exigencia hacia el médico por parte del empleador y de la Prepaga/Obra Social. Solo ciertas prácticas (por ejemplo Psiquiatría, Psicología) y en determinadas Prepagas u Obras Sociales tienen una duración fija estipulada. En la gran mayoría de los casos la agenda es manejada por la administración del consultorio y no por el médico. Esto genera que el médico no decida cuánto tiempo puede dedicarle a cada paciente, sino que se ve obligado a actuar en base a la agenda que cada administración le prepara. Y ahí debe encontrarse el equilibrio entre la agenda formal, los sobre-turnos y las potenciales ausencias de pacientes (los profesionales entrevistados

coinciden que la tasa de ausentismo es muy alta). La Sociedad Argentina de Cardiología realizó una encuesta en la que el 70% de los profesionales indicó que se les exige ofrecer turnos de 10 a 15 minutos, en tanto que ellos creían que lo correcto sería que las consultas tuvieran una duración de entre 20 y 30 minutos. En la práctica es muy poco probable ser atendido en el horario pautado y que la duración del turno sea la estipulada por agenda. Sauceda-Valenzuela *et al.* [22] encontraron en su estudio que los tiempos de espera pueden llegar a ser de hasta 120 minutos, mientras que los pacientes consideran aceptable una espera de 30 minutos. En tanto que la duración de la consulta se espera que sea no menor a 20 minutos para considerarla aceptable. Dicho estudio fue realizado tomando como fuente la encuesta Nacional de Salud y Nutrición de México del año 2006. Esta encuesta es realizada cada 6 años desde 1988 a nivel nacional en aquel país. Con estos datos han confeccionado un modelo de regresión logística binomial con múltiples variables para clasificar el nivel de calidad de atención de los distintos servicios de salud de México.

Desde la perspectiva de los profesionales hemos visto que la ecuación es complicada. Todos los meses deben cubrir un importante "saldo negativo" inicial para después empezar a obtener ganancias. Por ende, la necesidad por atender más pacientes en menos tiempo es imperiosa. Del otro lado, las Obras Sociales y Prepagas auditan que los tiempos de atención sean coherentes, caso contrario no abonan al profesional la consulta. Y en última instancia está el paciente, quien desea una atención de calidad.

1.2 Antecedentes Bibliográficos

No existen antecedentes bibliográficos en el país respecto de esta temática. A nivel mundial la persona que más se ha dedicado es Linda LaGanga, investigadora en la Universidad de Colorado Boulder y Vicepresidente de Calidad e Informática de Mental Health Partners, en Colorado, Estados Unidos. Ella ha escrito varios papers y realizado varias investigaciones referentes a la mejora en las citas médicas, el otorgamiento de sobreturnos (overbooking) y el ausentismo de los pacientes [29] [30] [31] [32].

Si bien hay otros trabajos realizados por distintos autores que han comenzado a abordar el tema, la bibliografía al respecto es escasa y analiza el tema desde otra perspectiva [33] [34] [35].

Capítulo 2. El Problema

2.1 Objetivo

El uso de la información mencionada hasta el momento debiera ser utilizada para la toma de decisiones clínicas y de gestión. Este trabajo se dirige en esa dirección: el paciente y los tiempos de atención y, consecuentemente, los de espera. El foco se pone en el paciente, sobre el tiempo que el mismo pasa cara a cara con el profesional médico. El objetivo es inferir la existencia de patrones en las características sociales, médicas y de filiación de los pacientes que expliquen por qué algunos de ellos pasan más tiempo en el consultorio médico que otros. Esto afecta directamente a la satisfacción de los pacientes, al tiempo transcurrido entre la fecha de solicitud del turno y la fecha de atención propiamente dicha, a las tasas de ausentismo, a la calificación que los pacientes otorgan a la atención recibida, a la forma de trabajar de los médicos y a la optimización de la ecuación de ingresos-recursos de los centros de salud.

Ligado al objetivo anterior podemos mencionar otro: que los profesionales cuenten con tiempo suficiente para poder realizar las tareas administrativas y de capacitación necesarias. Entre estas tareas se pueden incluir las que permitan tener una base de datos de pacientes más amplia, completa y con un nivel de detalle apropiado.

2.2 Hipótesis

Según indican las fuentes consultadas [4][5][6][10][11], y tal como se mencionará más adelante en el documento, se espera que ciertas características de los pacientes sean determinantes a la hora de calcular el tiempo que pasan dentro cara a cara con el médico. Entre estos factores se encuentran la edad y el sexo del paciente, los antecedentes médicos, si se trata de la primera visita, si consume fármacos, si tiene alguna patología previa, entre otros.

También se estima que el modelo a desarrollar sea aplicable a clínicas similares a la que está siendo estudiada, dado que ciertas características son propias de los estudios que se realicen y pacientes que allí concurren.

2.3 Marco Teórico

En el presente trabajo se busca desarrollar un modelo usando árboles de decisión. Los árboles de decisión son una de las técnicas más usadas en Data Mining. En la comunidad científica, su popularidad se debe a su simplicidad y transparencia: los árboles de decisión se explican por si solos, no hace falta ser un experto en explotación de datos para entender la lógica de los árboles. Son usualmente representados gráficamente como estructuras jerárquicas, haciéndolos muy fáciles de entender. Si bien puede haber algoritmos que son más performantes (Redes Neuronales, Máquinas de Vectores Soporte, Algoritmos Genéticos, etc) no son tan usados en ámbitos comerciales ya que los resultados que brindan son muy difíciles de explicar para el usuario regular.

Un árbol de decisión es un clasificador expresado como una partición recursiva del espacio de instancias. Un árbol consiste en nodos que forman una estructura con una raíz, es decir, un nodo que no tiene ramas entrantes. Los nodos restantes tienen exactamente una entrada. Un nodo que tiene ramas de salida será un nodo interno. Aquellos que no tengan ramas de salida serán nodos hoja. En los árboles de decisión cada nodo interno divide el espacio de instancias en dos o más subespacios de acuerdo con cierta función de separación.

Cada nodo hoja es asignado a una clase que representa el target u etiqueta más apropiado. De cada nodo interno pueden salir dos o más ramas. Cada nodo se corresponde con cierta

característica y las ramas se corresponden con un rango de valores. Este rango de valores debe generar una división del set de valores de la mencionada característica.

Las instancias son clasificadas partiendo desde la raíz hacia las hojas, de acuerdo a la salida de cada test a medida que se recorre el árbol. Específicamente, se comienza en el nodo raíz, se considera la característica que corresponde a dicho nodo y se define por qué rama seguir de acuerdo al valor de la característica. Luego se itera la misma operación en el nuevo nodo hasta que se alcance un nodo hoja.

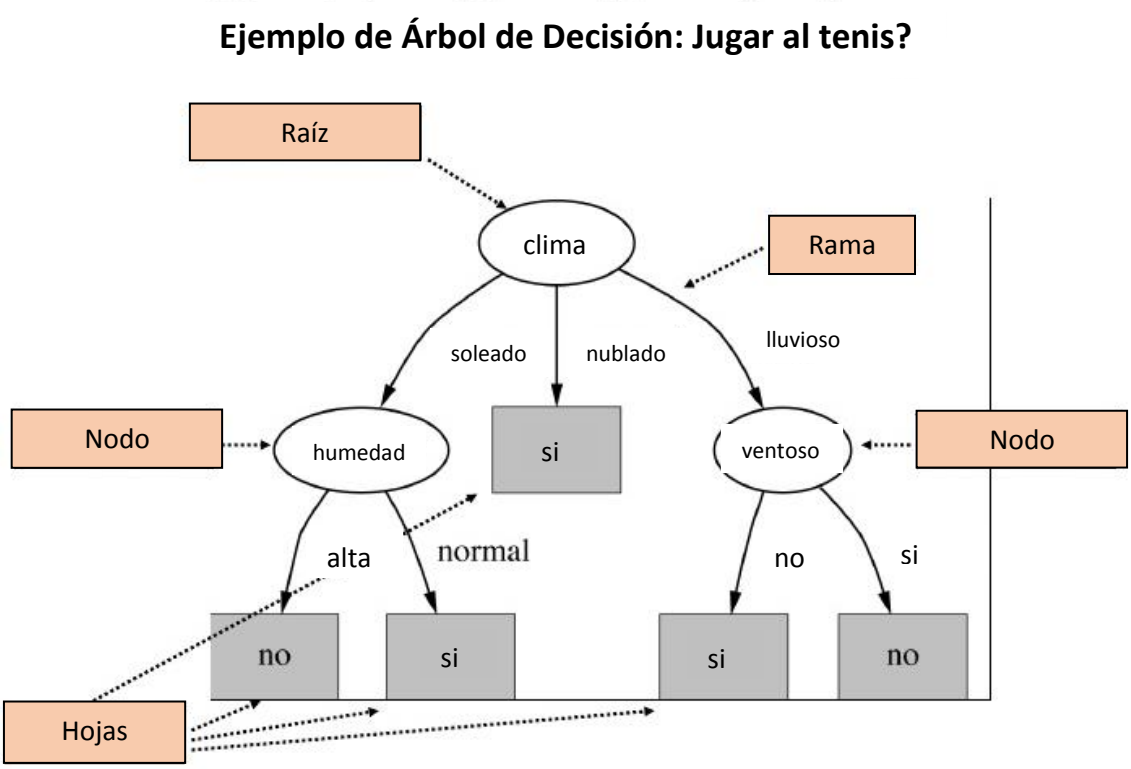


Figura 1. Ejemplo de árbol de decisión

Los algoritmos que crean los árboles de decisión buscan construir el mismo a partir de reglas que minimicen el error de generalización. Sin embargo, se puede buscar también reducir el número de nodos o la profundidad (cantidad de niveles) del árbol.

Pseudocódigo típico de un árbol de decisión

```
Procedimiento CrecerArbol(S,A,CriterioParada,CriterioDivision)
        S <– {s1,s2,s3,…,sn}        //comentario: conjunto de datos de entrenamiento//
        A <– {a1,a2,a3,….,an}       //comentario: conjunto de datos de entrada//
        y <– {t1,t2,t3,…,tn}        //comentario: variable objetivo//
        CriterioDivisión <– {CD} //comentario: método para evaluar una separación//
        CriterioParada <– {CP}   //comentario: criterio para frenar el crecimiento del árbol//

        SI CriterioParada(S) ENTONCES
                Marcar T como un nodo hoja con el valor más común de y en S como etiqueta
        SINO
                para todo ai ϵ A encontrar el a que obtenga el mejor CriterioDivision(ai, S).
        Etiquetar t con a
        PARA cada salida vi de a:
                Setear Subárbol_i = CrecerArbol(σ_a=vi S,A,y)
                Conectar el nodo raíz de tT a Subárbol_i con un borde etiquetado como vi
        FIN PARA
        FIN SI
FIN PROCEDIMIENTO
```

Los algoritmos más conocidos son ID3 (Quinlan J.R., 1986) y su sucesor C4.5 (Quinlan, 1993), CART (Breiman, Friedman, Stone & Olshen, 1984) y CHAID (Kass, 1980).

Algoritmo	Criterio de División	Tipos de Atributos	Valores Faltantes	Estrategia de Poda	Detección de Valores Extremos
ID3	Ganancia de Información	Acepta solo valores categóricos	No maneja valores faltantes	No realiza poda	Susceptible a valores extremos
CART	Indice de diversidad de Gini	Maneja valores categóricos y numéricos	Maneja valores faltantes	Usa poda de costo-complejidad	Maneja valores extremos
C4.5	Indice de Ganancia	Maneja valores categóricos y numéricos	Maneja valores faltantes	Poda basada en tasa de error	* Maneja valores extremos

Cuadro 2. Comparación de los algoritmos ID3, CART y C4.5

* Algunos autores indican que este algoritmo sí es susceptible a valores extremos (ver punto siguiente).

2.4 Aspectos Metodológicos

La clínica en la cual se realiza el estudio utiliza un sistema de administración general implementado en el año 2004 (anterior a esta fecha la carga de datos se hacía en forma manual). El mismo abarca desde la administración de los turnos hasta la facturación, si bien una parte de esta última aún se lleva de forma manual. Dicho sistema fue adquirido a una empresa que se

dedica a vender esta clase de productos enlatados para instituciones de salud. La aplicación usa MS Access como base de datos. La misma se exporta a Oracle 10g para su manipulación. Posteriormente se exporta a R para realizar las tareas de explotación de datos.

Un conjunto de datos (conocido también por el anglicismo: *dataset*, comúnmente utilizado en algunos países hispanohablantes) es una colección de datos habitualmente tabulada. En adelante se utiliza "dataset" como sinónimo de "conjunto de datos".

Como se mencionó anteriormente, para el modelado se usan árboles de decisión. Los árboles de decisión son una de las clases de técnicas de data mining que tiene sus raíces en disciplinas estadísticas tradicionales como la regresión lineal. También comparten raíces en el mismo campo de las ciencias cognitivas que produjeron las redes neuronales.

Son varios los motivos por los que se eligen los árboles por sobre otros métodos:

- Son fáciles de entender y de interpretar: producen resultados que son muy simples de comunicar en términos simbólicos y visuales. Son fáciles de producir, de entender y de usar
- Tienen gran flexibilidad para trabajar con una amplia variedad de datos: nominal, numérico y de texto: los árboles de decisión fácilmente incorporan varios niveles de medida, incluyendo cualitativas y cuantitativas (ordinales e intervalos)
- Poseen mucha adaptabilidad para procesar datasets que poseen errores o valores ausentes (esto último ocurre en el dataset bajo estudio)
- Son no paramétricos y muy robustos y producen resultados similares sin importar la escala o la unidad de medida de los campos que se están usando

Dentro de todos los algoritmos disponibles para la construcción de árboles se opta por CART debido a que el dataset contiene varios campos con muchos valores nulos y otros en los que ocasionalmente pueden presentarse valores extremos. El mencionado algoritmo es el único que cumple, según todos los autores consultados, con las características necesarias para trabajar con este dataset [36] [37]. Este algoritmo se contrastará con otro llamado Random Forest. Random Forest es una combinación de árboles de forma tal que cada árbol depende del valor de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles [48]. Se elige este algoritmo dado que es conocido por ser un clasificador muy certero y robusto para datasets, aunque no posee una representación gráfica de fácil interpretación. De ahí que se use solamente como parámetro para comparar.

La presencia de valores extremos y valores nulos es algo esperable dada la naturaleza de los datos: si bien los datos básicos de cada paciente se ingresan en todos los casos, dependiendo del estudio que se le realice a cada paciente se ingresarán ciertos datos solamente. Asimismo en determinadas prácticas algunos valores pueden ser extremos (por ejemplo, cirugías).

2.5 Contribución de la Tesis

Lo desarrollado en esta tesis pretende, en primera instancia, comprender el motivo por el cual hay pacientes que tienen tiempos de atención notoriamente más largos o más cortos que otros. Se busca entender desde el momento de la reserva del turno que tipo de paciente va a concurrir.

Como segunda contribución se pretende introducir mejoras en la administración general de los turnos médicos, tanto del punto de vista de la clínica / profesional de la salud como del paciente.

Desde la perspectiva de la clínica y del profesional de la salud se busca mejorar la administración de los recursos, especialmente los humanos. Planificar en forma dinámica los turnos en base a las condiciones de los pacientes, reducir el stress de los médicos, otorgar a cada profesional el tiempo necesario para atender a cada paciente de la mejor manera. Un mejor aprovechamiento de los recursos redundará en mejoras económicas.

Desde el punto de vista de los pacientes se busca que pasen el menor tiempo posible en las salas de espera y una cantidad de tiempo óptima dentro del consultorio. Mejorar la calidad de la atención para así mejorar el proceso de curación y fidelizar a los pacientes.

Capítulo 3 – El modelo

3.1 Base de Datos

3.1.1 Descripción de los Datos

La base de datos cuenta con 84.365 pacientes, de los cuales 58.486 tienen historial de admisiones. El total de pacientes con historial activo han realizado 182.103 admisiones.

La primera admisión registrada en esta base de datos data del 13/12/2004 (antes de esta fecha la registración de admisiones era realizada en papel) y la última el 04/07/2015 (fecha de obtención de datos para este trabajo).

La base de datos cuenta con 126 tablas, algunas de las cuales no contienen datos.

Varias tablas tienen valores ausentes o valores extremos. Los datos son confiables y parte de ellos son validados por las empresas de medicina u obras sociales.

La clave que identifica a cada fila (clave primaria o primary key) está conformada por el número de historia clínica, el código de médico y la fecha de atención.

3.1.2 Análisis Inicial

El presente trabajo se realiza con los datos provistos por una clínica ubicada en la zona Sur del Gran Buenos Aires. La misma se encuentra en la región central de la cabecera del partido. Por su relevancia atiende pacientes en su mayoría del Gran Buenos Aires y de la Ciudad Autónoma de Buenos Aires y, en menor medida, del resto de la Argentina. Allí se atiende una especialidad médica en particular en todas sus variantes (pediatría, estudios específicos, cirugías, entre otros). Se atienden tanto obras sociales, prepagas y particulares.

Los datos abarcan desde Diciembre de 2004 hasta Julio de 2015. Este es el período a analizar en el presente trabajo.

Para poder generar el dataset, los datos se exportan desde el sistema de origen a Oracle 10g. Armado el dataset los datos se exportan a R para su explotación.

3.1.3 Origen y Limpieza de los Datos

El primer paso consiste en analizar la base de datos. Tal como se mencionó en el punto 3.1.1, la base cuenta con 126 tablas. De ese total 30 tablas están vacías, por lo que son descartadas de inmediato. Se confecciona un esquema de entidad-relación con el cual se analizan detalladamente todas las tablas de la base. Cada una de las 96 tablas restantes se examina en forma individual, controlando el contenido y su relación con el resto de las tablas. Por ejemplo, existe una tabla VADEMECUM que contiene apenas un puñado de medicamentos cargado. Al estar desactualizada se deja de lado.

De las 126 iniciales el universo de tablas que se utilizarán para el análisis se reduce a 26. De este nuevo total, casi el ciento por ciento es modificado:

1) Varias tablas poseen un campo OBSERVACIONES que debió ser discretizado dado que contiene texto libre en el cual se describe la enfermedad del paciente, medicamento que se receta, etc. Dependiendo del caso se opta por cambiarlo a PATOLOGICO o NO PATOLOGICO o bien se utiliza para generar nuevas tablas (por ejemplo, para el caso de los medicamentos se crea la clasificación ATC, lo cual se explicará en detalle más adelante).

Para esto se realiza mediante la aplicación RapidMiner 5.3 un algoritmo que permite, mediante técnicas de text mining (minería de texto), la extracción de palabras claves de las distintas columnas Observaciones existentes en cada tabla para poder clasificar el contenido y discretizarlo en dos valores: PATOLOGICO o NO PATOLOGICO. Si bien se escapa del alcance de la tesis, a continuación se detalla brevemente el tratamiento dado a estas columnas: los campos pueden estar vacíos, pueden indicar que el paciente estaba sano o pueden detallar la medicación que tomaban o el detalle de lo encontrado durante la exploración del paciente. Para los dos primeros casos se modifican los valores a NO PATOLOGICO. Si el contenido del campo menciona alguna medicación se extrae el dato para incorporar a las columnas de ATC (este punto se explicará en el apartado 3.2.8) y se completa el campo con el valor PATOLOGICO dado que se entiende que el paciente posee alguna enfermedad de base. Si contiene datos referidos a alguna enfermedad pre-existente o anomalías detectadas durante el estudio también se completa el campo con el valor PATOLOGICO. Este procedimiento se realiza en forma individual para cada una de las columnas denominadas Observaciones (casi el 95% de las tablas poseen dicha columna).

Para corroborar la eficacia del algoritmo se toma una muestra de 400 casos y se validan experimentalmente en conjunto con los profesionales del centro bajo estudio. La tasa de error encontrada es menor al 3% por lo que se considera correcto lo realizado por el algoritmo.

Los datos ya validados son exportados al conjunto de datos original.

2) Se calcula el tiempo que cada paciente permanece siendo atendido en el consultorio de la siguiente manera: hora_de_atencion(t+1) - hora_de_atencion(t), siendo t el primer paciente que ingresa y t+1 el paciente que ingresa a continuación. Con estos datos se crea una columna que contiene el tiempo de atención (tiempo, en minutos, que el paciente pasa dentro del consultorio cara a cara con el médico). El procedimiento se realiza en Oracle SQL Developer. Se deben contemplar 2 casos particulares: se informa 9999 si no se calcula por no contar con datos y -888 si se trata del último paciente del día y no se puede calcular por no haber un paciente a continuación.

   Vale la pena aclarar que la hora de atención de los pacientes no se empezó a ingresar al sistema desde sus orígenes, por lo que varios de los registros más viejos no cuentan con este dato (estos registros se completaron con 9999).

   La hora de atención se refiere a la hora en la que el paciente es recibido por el médico y es distinta a la hora de admisión que es la hora en la que es registrado en la administración apenas llega el paciente a la clínica.

   Cabe destacar que cuando se habla de tiempos se hace referencia a tiempos efectivos, es decir, no se consideran caídas de equipos, fallas de sistemas, cortes de energía eléctrica ni cualquier otro inconveniente que pudiera generar demoras mayores de lo normal.

3) Todas las tablas cuentan con tres campos que funcionan como identificador único de cada fila (primary key): Número de Historia Clínica, Fecha y ID de Médico. Por los tipos de práctica que se realizan en la clínica se encuentran registros duplicados, lo cual es correcto (algunas prácticas deben repetirse en el mismo día). Esto se corrige agregando columnas a las tablas y unificando los datos en un solo registro mediante una función programada en R. De esta manera se unifican en un solo registro todas las prácticas realizadas por día a cada paciente.

3.2 Factores Relevantes (Variables Explicativas)

3.2.1 Nuevas variables

Uno de los factores que se presume significativo a la hora de determinar el tiempo que demoran los pacientes en el consultorio es la cantidad de consultas totales que tiene el paciente, la cantidad de consultas parciales (cantidad de consultas hasta la fecha de ese registro), y la cantidad de consultas en el mes, semestre y año . Dado que estos campos no existen se crean mediante una función en R. Funciona de la siguiente manera: si entre la actual visita y la anterior hay 30 días o menos, adiciona 1 al valor del contador de la visita anterior (si la anterior era la primera visita y esta la segunda el valor estará en 1 para la actual, 0 para la anterior). Si al menos va al consultorio una vez cada 6 meses hace lo mismo para el semestre. Si va 1 vez cada 360 días, lo mismo para año.

Otro factor a priori determinante a la hora de calcular el tiempo de atención son las enfermedades o trastornos preexistentes [4] [5]. Para esto, a partir de la medicación que los pacientes declaran consumir al momento de confección de la historia clínica (nombre comercial o droga genérica) se obtiene su correspondiente clasificación ATC mediante el cruzamiento de los datos con un Vademecum on-line (ver punto 3.2.8). El código ATC o Sistema de Clasificación Anatómica, Terapéutica, Química (ATC: acrónimo de Anatomical, Therapeutic, Chemical classification system) es un índice de sustancias farmacológicas y medicamentos instituido por la Organización Mundial de la Salud . Está dividido en 5 niveles de los cuales, para este trabajo, se usan los dos primeros:
1.- Nivel (anatómico): Órgano o sistema en el cual actúa el fármaco.
2.- Nivel: Subgrupo terapéutico, identificado por un número de dos cifras.
De esta manera se busca determinar qué sistema u órgano del paciente está bajo tratamiento.

Cada área anatómica de estudio del paciente posee una tabla particular en el sistema en la cual se detalla la patología encontrada. Muchas de estas tablas poseían una excesiva cantidad de niveles, por lo que se procede a analizar la frecuencia de cada patología en particular, mantener las de mayor frecuencia y agrupar en la categoría OTROS a las menos frecuentes.
Además se genera una variable dummy para cada área anatómica en la que se marca si cada paciente tuvo o no alguna patología en dicha área.

Para poder trabajar con datos geográficos se analizan los campos Localidad, Provincia y Código Postal. Se detectan datos faltantes (por ejemplo existía la localidad pero no el código postal), los cuales se corrigen mediante una función en R. También hubo casos en los que la localidad se escribía de varias maneras, por ejemplo: L. de Zamora, Lomas, Lomas de Zamora. En estos casos se ejecutó otra función en R que corregía la mayor parte de los errores. Sin embargo, una cantidad menor de fallas debieron ser reparadas a mano.

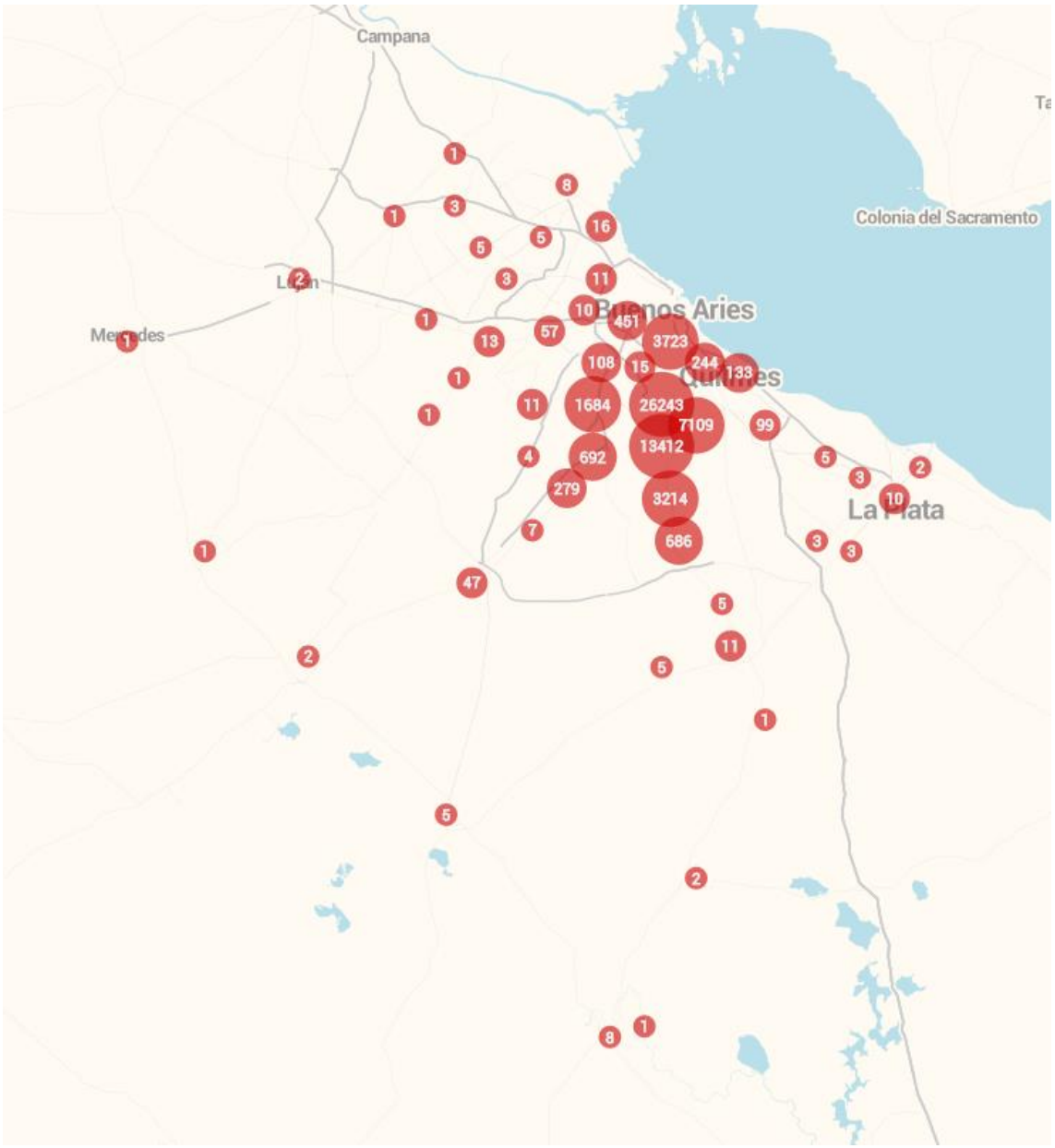
Para garantizar la privacidad de los datos se eliminaron todos aquellos campos que pudieran identificar al paciente: DNI, dirección, nombre, apellido y CUIL. Solo se mantuvo el código postal, la localidad y la provincia de la dirección declarada por el paciente.

3.2.2 Análisis Geográfico

Respecto de los pacientes activos, en los mapas 1 y 2 se puede observar la procedencia de los mismos. En primer mapa se puede observar un panorama a nivel nacional y en el segundo se amplía la zona de la Ciudad Autónoma de Buenos Aires (CABA) y alrededores para poder ver más en detalle la zona de influencia de la clínica. El mapa 1 muestra que, si bien hay pacientes que provienen de distintas zonas del país (incluso del exterior), la gran mayoría está radicado en el AMBA (Área Metropolitana de Buenos Aires).

En el mapa 2 se aprecia que la mayoría de los pacientes están domiciliados en la zona Sur del Gran Buenos Aires, justamente la zona donde se radica la clínica.
Amén de lo mencionado anteriormente, los mapas sirven para mostrar la importancia que tiene la clínica dada su amplia zona de influencia.

Mapa 1. Distribución de los pacientes a nivel nacional



Mapa 2. Distribución de los pacientes: Ciudad Autónoma de Buenos Aires y alrededores

En la tabla 1 se discrimina la cantidad de pacientes por partido y zona del Gran Buenos Aires (GBA) a la cual pertenecen.

Municipio	Cant.	Porcentaje	Zona GBA
LOMAS DE ZAMORA	35785	61,19%	Sur
ALMIRANTE BROWN	10477	17,91%	Sur
LANUS	5379	9,20%	Sur
ESTEBAN ECHEVERRÍA	2374	4,06%	Sur
EZEIZA	929	1,59%	Sur
PRESIDENTE PERON	810	1,38%	Sur
SAN VICENTE	685	1,17%	Sur
CAPITAL FEDERAL	451	0,77%	No Aplica
QUILMES	295	0,50%	Sur
AVELLANEDA	286	0,49%	Sur
FLORENCIO VARELA	196	0,34%	Sur
BERAZATEGUI	80	0,14%	Sur
OTROS	739	1,26%	No Aplica
total	**58486**	**100,00%**	

Tabla 1. Cantidad de pacientes por partido

Se aprecia claramente que más del 98% de los pacientes pertenecen a la zona Sur del Gran Buenos Aires.

3.2.3 Sexo del Paciente

En la figuras 2.a y 2.b se observa que la cantidad de pacientes de sexo femenino y masculino que han concurrido al menos en una ocasión a la clínica es bastante similar y acorde a la población por sexo publicada en el Censo Nacional de Población, Hogares y Viviendas 2010 de la República Argentina [46] para los 24 partidos del Gran Buenos Aires (48,57% de hombres y 51,43% de mujeres).

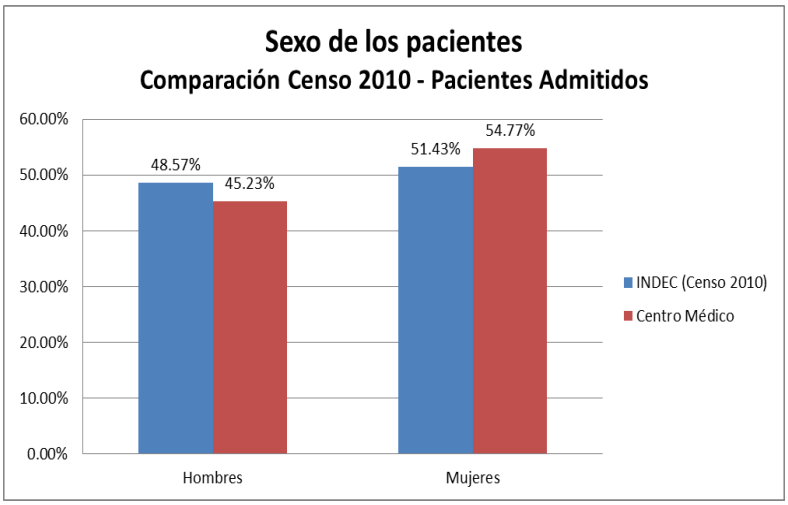


Figura 2.a. Porcentaje de pacientes según sexo –Comparación Censo-Admisiones

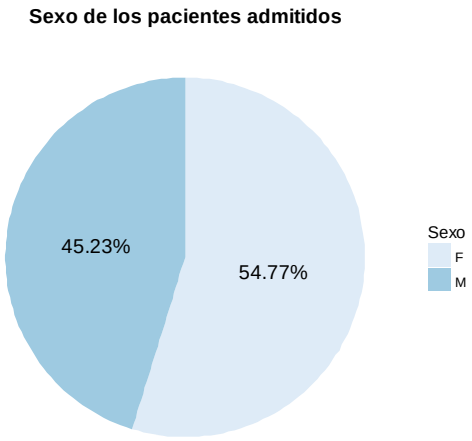


Figura 2.b. Porcentaje del sexo de los pacientes admitidos

Según indica Myriam Deveugele *et al.* [4], los tiempos de consulta son mayores en las mujeres que en los hombres. Los pacientes de sexo femenino son descriptos como más locuaces que los de sexo masculino, y están más dispuestos a discutir sobre temas psicosociales. También Hava Tabenkin *et al.* detalla la diferencia en los tiempos de atención según el sexo del paciente [5].

Si se analizan la cantidad de admisiones por sexo y por edad se puede ver que son más las mujeres admitidas en todo el rango de edad, salvo en los primeros años de vida de los pacientes (figura 3).

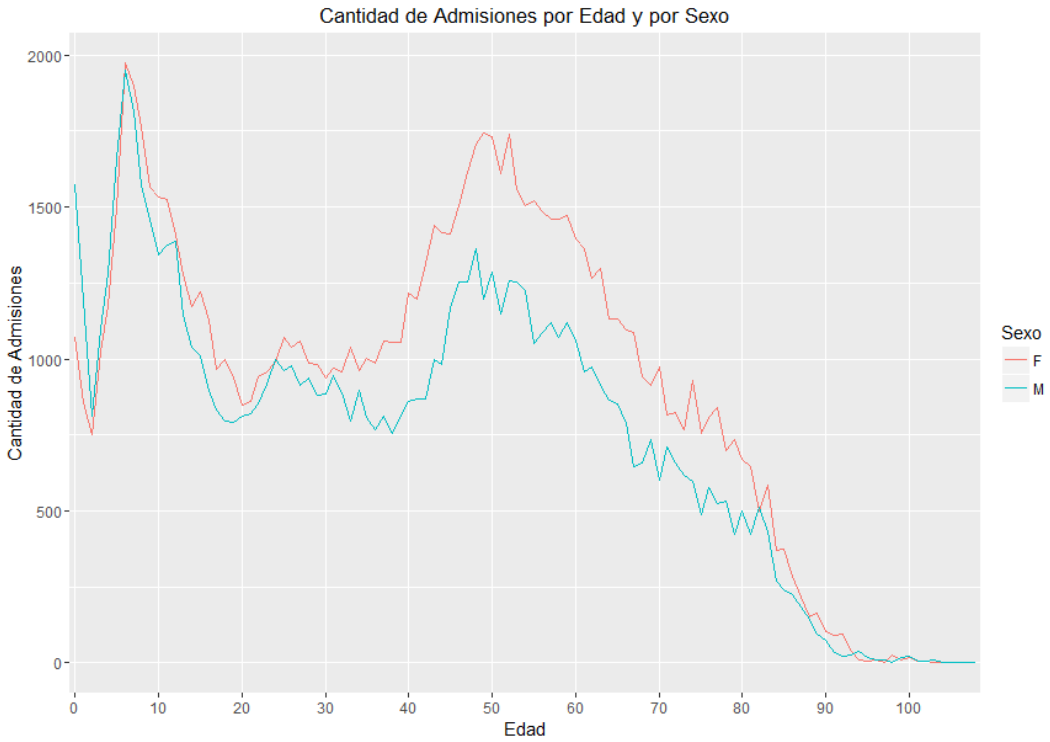


Figura 3. Admisiones por edad y sexo

3.2.4 Cantidad de Admisiones

En la figura 4 se aprecia la cantidad de admisiones por año. Vale la pena aclarar que un mismo paciente puede ser admitido más de una vez por cada período. Esto es relevante porque, tal como menciona Alice Migongo [6], en aquellos pacientes que ya han visto previamente a su médico de cabecera, o bien al mismo profesional en reiteradas oportunidades, se ve afectado notoriamente el tiempo de atención.

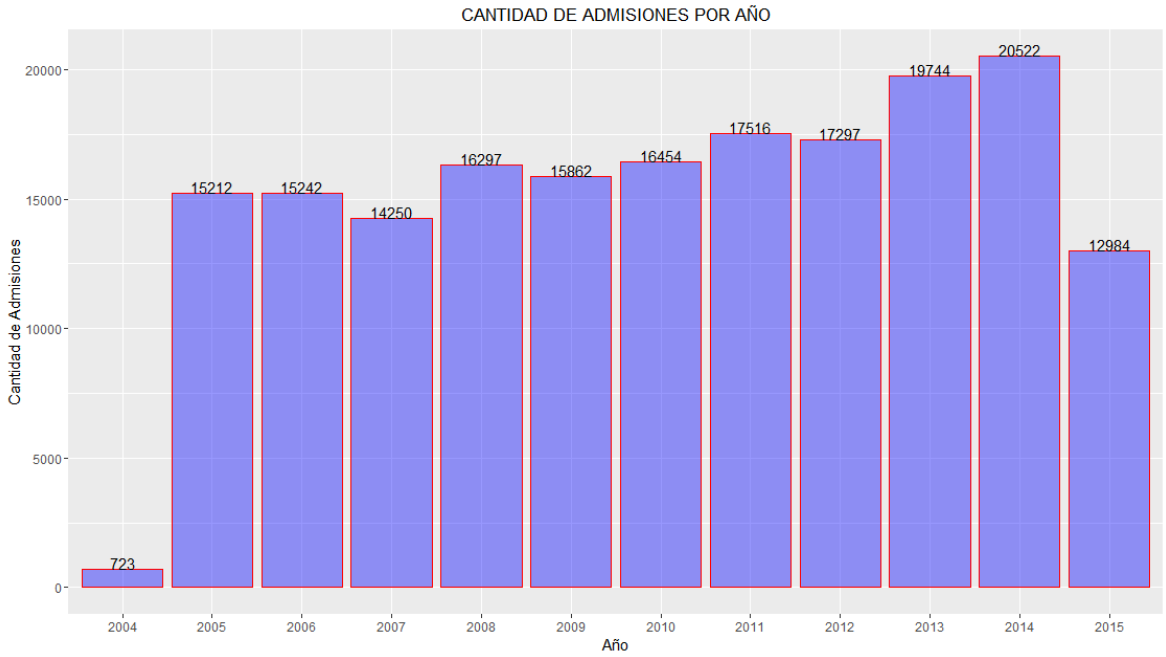


Figura 4. Cantidad de admisiones por año entre el 13/12/2004 y el 04/07/2015

Los valores para el año 2004 y para el 2015 son parciales debido a la digitalización del sistema y la fecha de obtención de los datos, respectivamente.

En la figura 5 se detallan la cantidad de admisiones únicas, es decir, considerando único el par paciente-año.

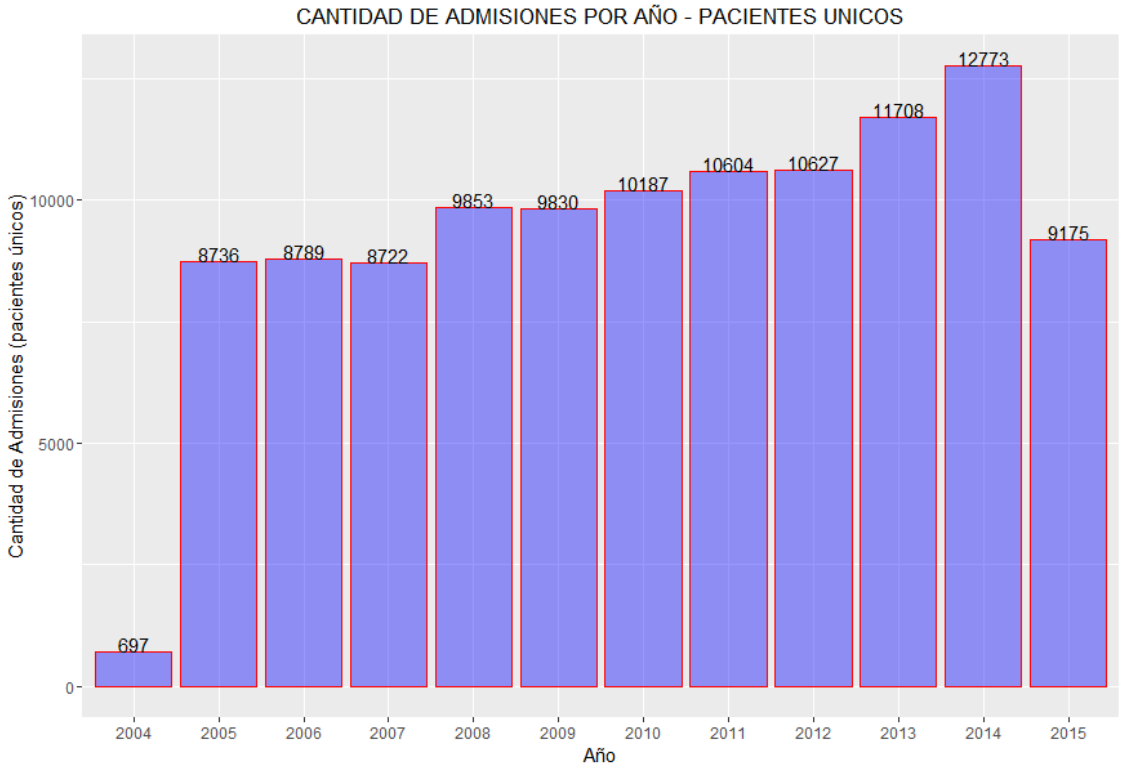


Figura 5. Cantidad de pacientes únicos admitidos por año

La cantidad de pacientes únicos admitidos en la clínica tuvo un incremento para el período 2005-2014 del 46%. De los 8.736 pacientes de 2005 se pasó a recibir 12.773 pacientes en 2014 (figura 5).
La cantidad total de admisiones en el 2014 se incrementó en un 34,9% respecto de 2005. En este último año se realizaron 15.212 admisiones mientras que en 2014 se hicieron 20.522 (figura 4).

La diferencia entre los valores del gráfico de la figura 4 y de la 5 la constituyen aquellos pacientes que acudieron más de una vez por año a la clínica. Se puede apreciar que, para todos los años, alrededor del 60% son pacientes que concurrieron 1 sola vez a la clínica, mientras que cerca del 40% son pacientes que concurrieron 2 o más veces (tabla 2).

Pacientes que concurren más de una vez a la clínica

Año	Cantidad_de_Adm	Adm_Pac_Unico	Diferencia	Porcentaje
2004	723	697	26	3.596127
2005	15212	8736	6476	42.571654
2006	15242	8789	6453	42.336964
2007	14250	8722	5528	38.792982
2008	16297	9853	6444	39.541020
2009	15862	9830	6032	38.027991
2010	16454	10187	6267	38.088003
2011	17516	10604	6912	39.461064
2012	17297	10627	6670	38.561600
2013	19744	11708	8036	40.700972
2014	20522	12773	7749	37.759478
2015	12984	9175	3809	29.336106

Tabla 2. Porcentaje de pacientes por año que concurre más de una vez a la clínica

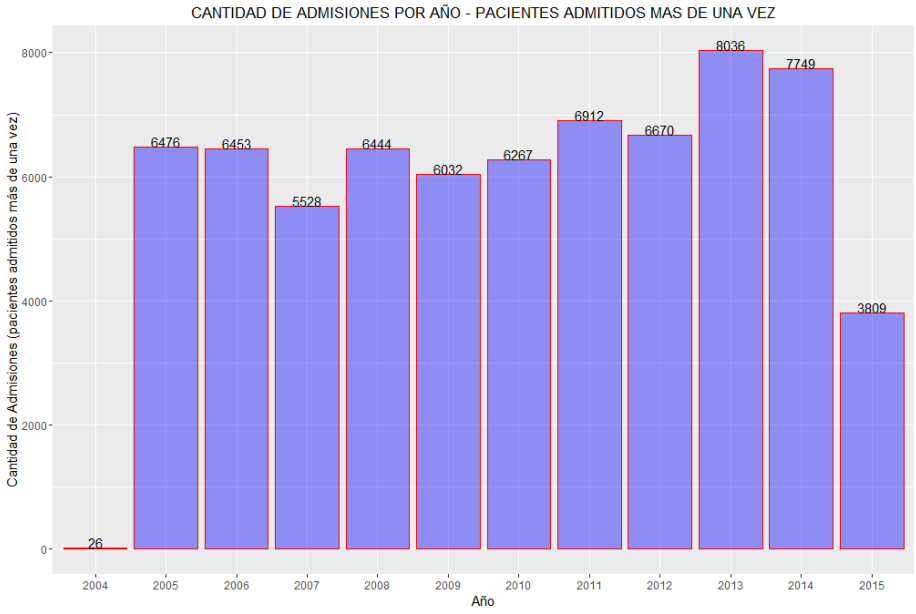


Figura 6. Cantidad de pacientes por año que son admitidos más de una vez en la clínica

3.2.5 Revisitas

Cada paciente puede ser admitido 1 o más veces dependiendo del motivo de consulta. Por ejemplo, los pacientes que concurren en busca de un apto físico no necesariamente serán atendidos nuevamente en la clínica, a diferencia de un paciente operado que volverá en reiteradas ocasiones para control. La tabla 3 y la figura 7 detallan la cantidad de pacientes según la cantidad de admisiones.

Cant. de Admisiones	Cant. de Pacientes	Porcentaje
1	27338	46.74%
2	10859	18.57%
3	6005	10.27%
4	3713	6.35%
5	2488	4.25%
6	1756	3.00%
7	1268	2.17%
8	961	1.64%
9	747	1.28%
10	584	1.00%
11	477	0.82%
12	377	0.64%
13	290	0.50%
14	251	0.43%
15	198	0.34%
16	162	0.28%
17	133	0.23%
18	141	0.24%
19	90	0.15%
20	76	0.13%
más de 20	572	0.98%
TOTAL	**58486**	**100%**

Tabla 3. Frecuencia de pacientes según la cantidad de admisiones

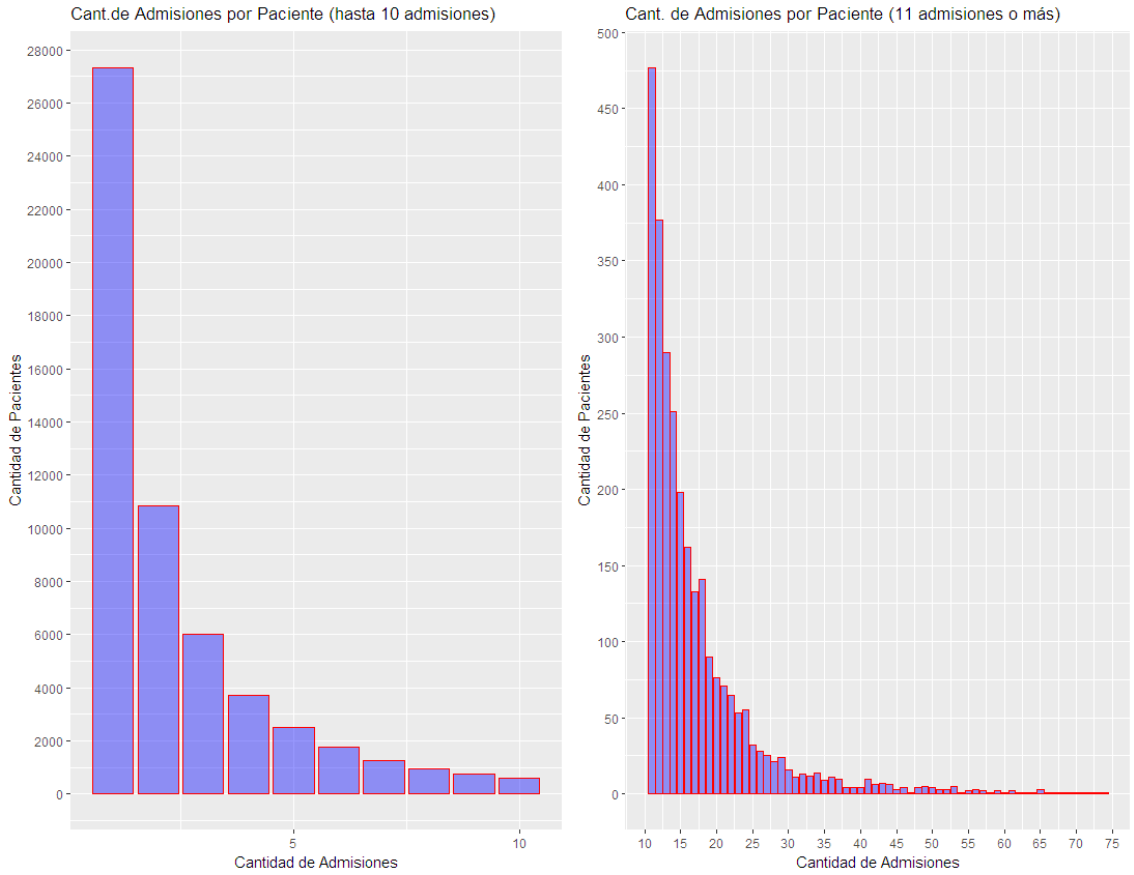


Figura 7. Frecuencia de pacientes según la cantidad de admisiones (izq.: hasta 10 admisiones, der.: 11 admisiones o más)

Tanto la tabla 3 como la figura 7 muestran que gran parte de los pacientes (46,74%) concurrieron 1 sola vez a la clínica durante el período analizado. El restante 53,26% concurrieron más de una vez. Este es un dato no menor dado que como menciona Migongo et at. [6], la continuidad en el cuidado y atención de un paciente no solo contribuye a la satisfacción y a la salud del paciente sino que también es una de las claves en el manejo eficiente del tiempo de atención en las prácticas médicas.

3.2.6 Estacionalidad

Al analizar la estacionalidad de la cantidad de admisiones de pacientes se detecta un incremento durante los meses de Febrero y Marzo respecto del resto del año (figura 8).

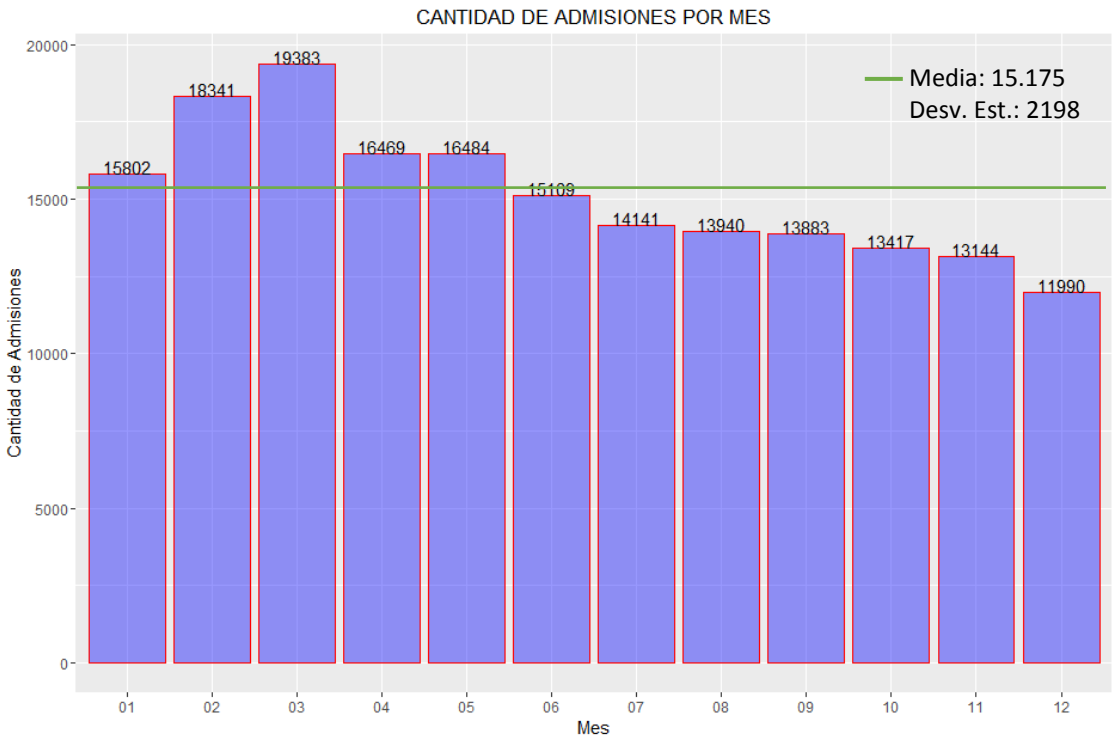


Figura 8. Distribución de admisiones por mes

Con el fin de profundizar en el análisis del pico de admisiones en los meses de Febrero y Marzo se analiza la edad de los pacientes al momento de ser admitidos.

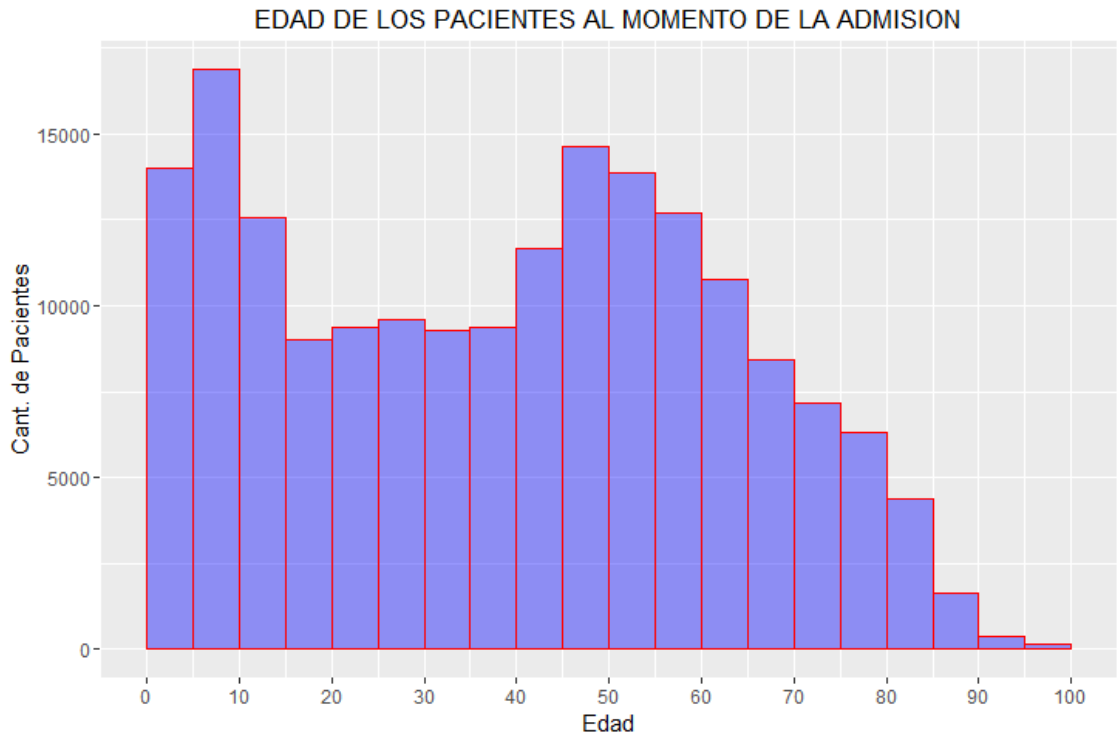


Figura 9. Cantidad de pacientes por edad al momento de ser admitidos

3.2.7 Grupos Etarios

En la figura 10 se destacan 4 grandes grupos etarios:

- De 0 a 15 años
- De 16 a 40 años
- De 41 a 65 años
- Mayores de 65 años

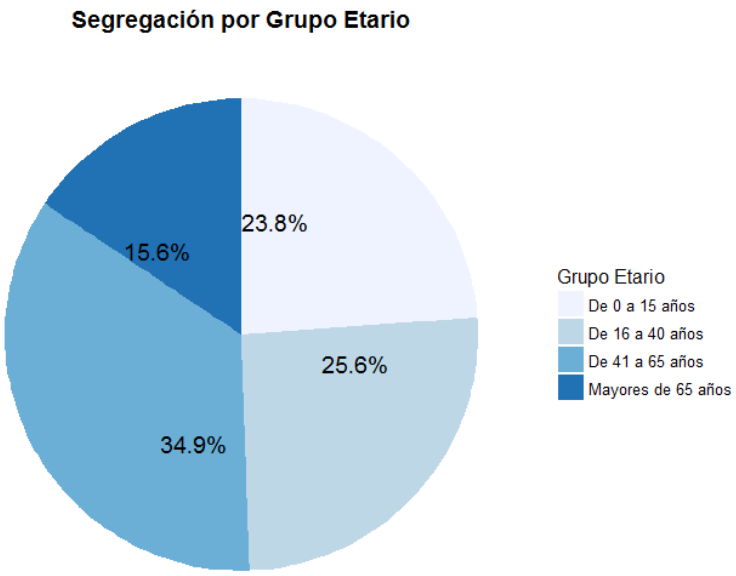


Figura 10. Agrupación de admisiones por edad de los pacientes

El primer grupo posee los picos más altos en el histograma (figura 9) y representa el 23,8% del total de admisiones. Esto se explica por la poca cantidad de especialistas pediátricos que existen para la rama de la medicina que se practica en esta clínica. Para ilustrar más aún este dato se analiza la cartilla de una de las prepagas más importantes. Esta posee, para un plan medio en la Ciudad de Buenos Aires, 179 prestadores para adultos y solamente 27 para pediatría. Gran parte de los establecimientos educativos solicitan al inicio de clases un apto médico del alumno, por lo que resulta lógico el pico antes indicado (es obligatorio en la provincia de Buenos Aires la confección de una ficha de salud, de acuerdo con lo normado por la Dirección General de Cultura y Educación). En la figura 11 se puede observar que el pico de admisiones de pacientes pediátricos se da durante los primeros meses del año.

Para el resto de las franjas etarias la cantidad de admisiones durante el año es mucho más pareja, sin observarse picos como en el caso antes mencionado.

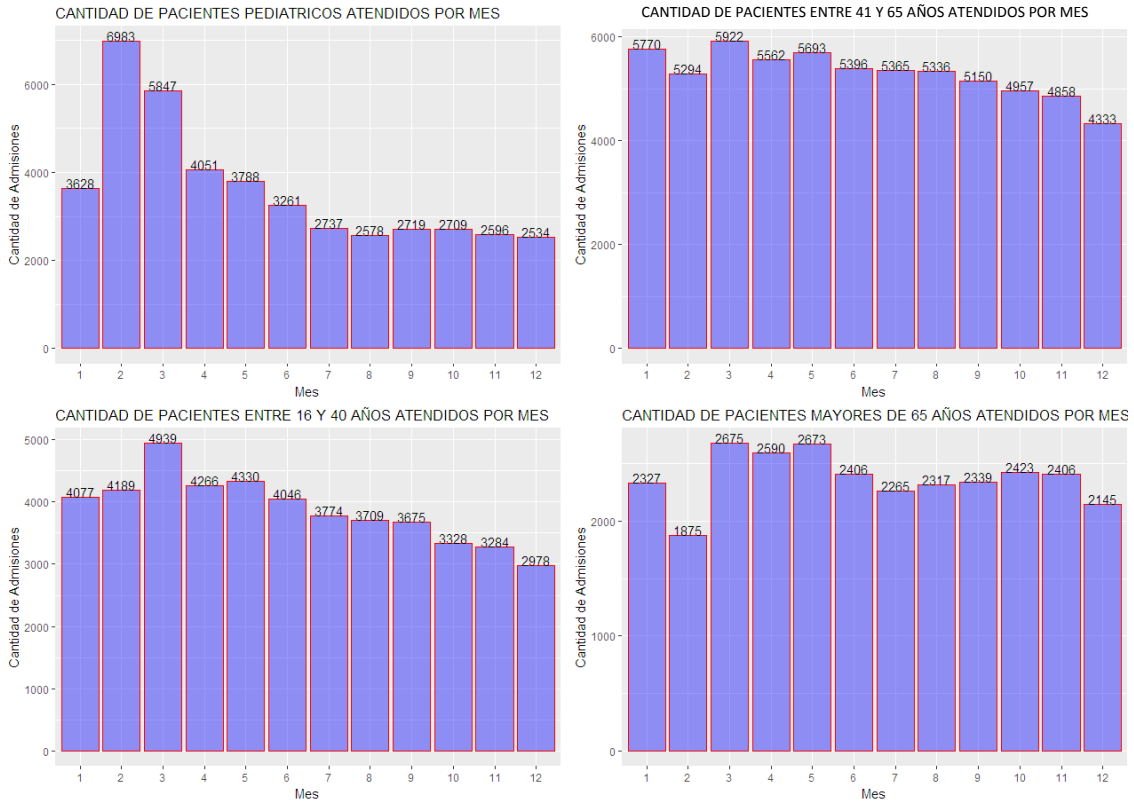


Figura 11. Distribución de pacientes según edad admitidos durante los distintos meses para el total de años analizados

En cuanto al tercer grupo (entre 41 y 65 años), coincide con la edad en la cual, según la bibliografía, los pacientes comienzan a mostrar deterioros en la salud que ameritan consultar al profesional médico [1] [2].

Respecto del último grupo (más de 65 años), resulta coherente que sea la porción más pequeña de la figura 10 debido a la edad y expectativa de vida [3]. Se nota una bajante muy importante en el mes de Marzo, justamente cuando se da el pico de pacientes pediátricos. Se puede suponer que por ser gente de mayor edad prefiere ir en momentos en los que la concurrencia es menor.

Las figuras 12, 13 y 14 detallan la cantidad mensual de admisiones para el período bajo estudio para el total de pacientes, para pacientes adultos y para pacientes pediátricos, respectivamente. De los mismos se puede concluir que la cantidad de pacientes admitidos ha aumentado durante los últimos 10 años y que los picos ocurridos durante los meses de Febrero y Marzo corresponden a la mayor concurrencia de pacientes pediátricos.

Los pacientes adultos muestran, dentro de ciertas variaciones, una estabilidad a través del tiempo. Lo mismo sucede con los pacientes pediátricos luego del pico previo al comienzo del ciclo escolar.

Se destaca también una baja de la cantidad de admisiones cada Diciembre, en algunos casos muy importante. La explicación está en la cantidad de feriados por las fiestas navideñas, de fin de año y el comienzo de las vacaciones de verano.

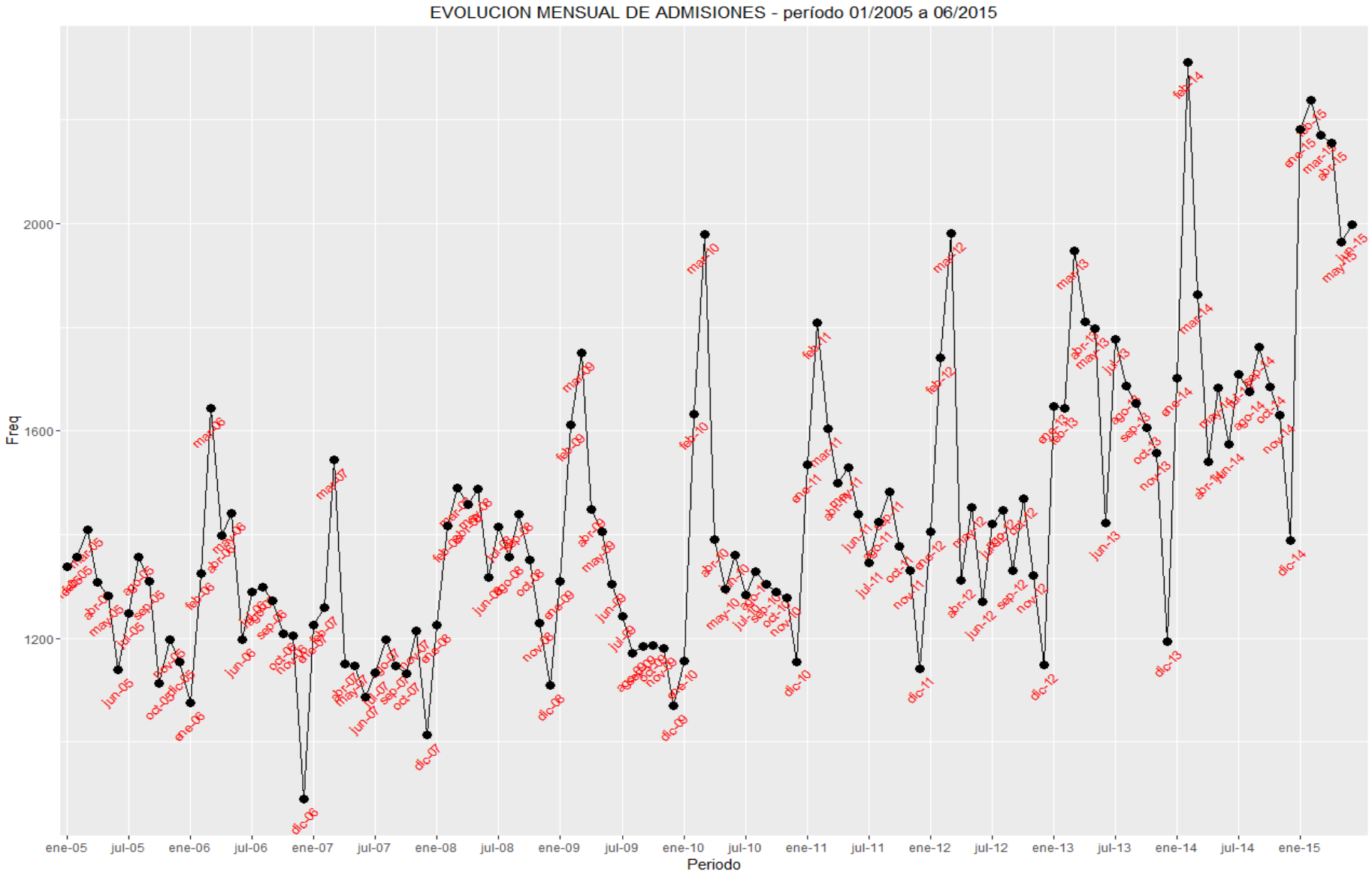


Figura 12. Evolución mensual de la cantidad de admisiones desde Enero 2005 a Junio 2015

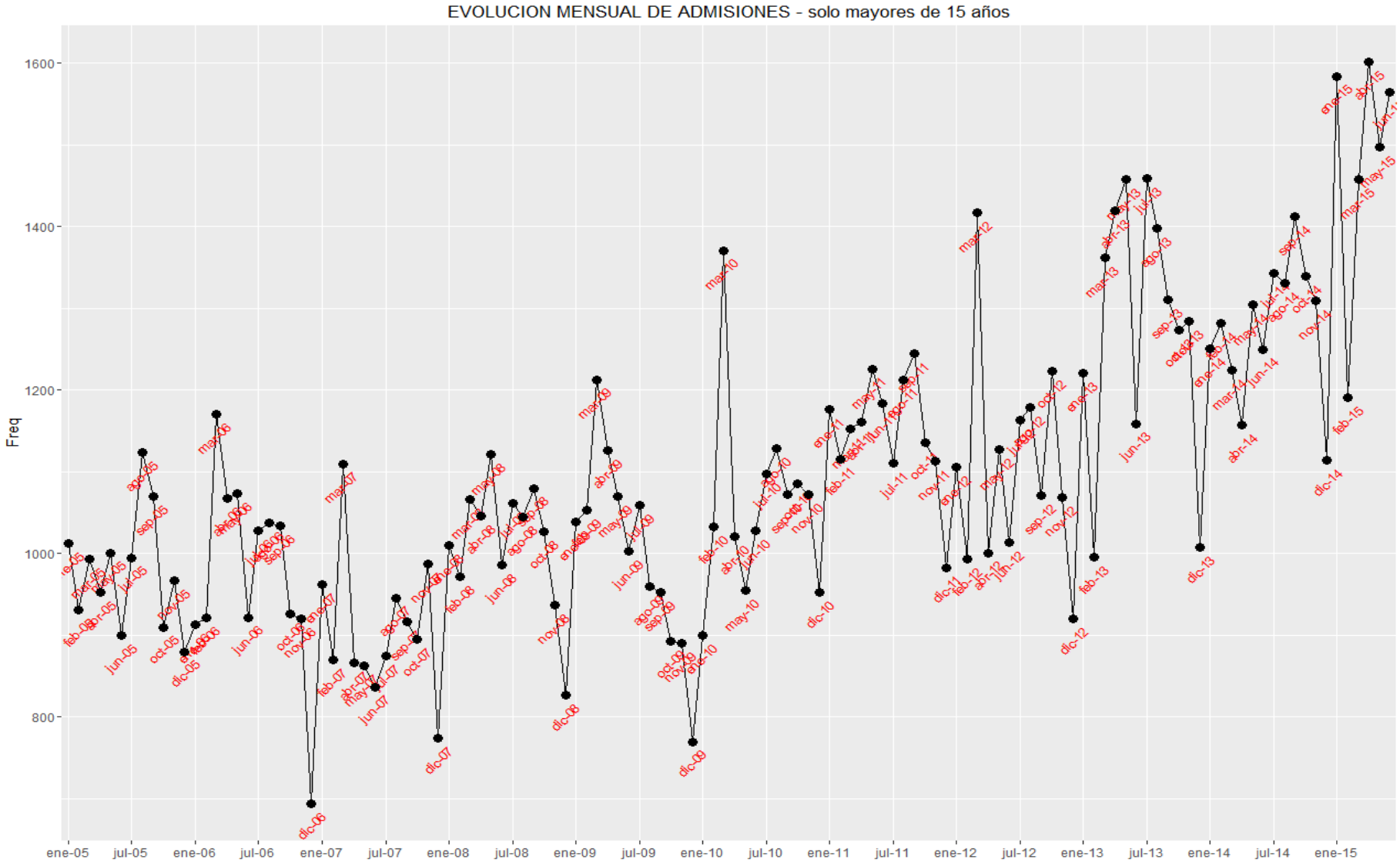


Figura 13. Evolución mensual de la cantidad de admisiones desde Enero 2005 a Junio 2015 – Pacientes adultos (mayores a 15 años)

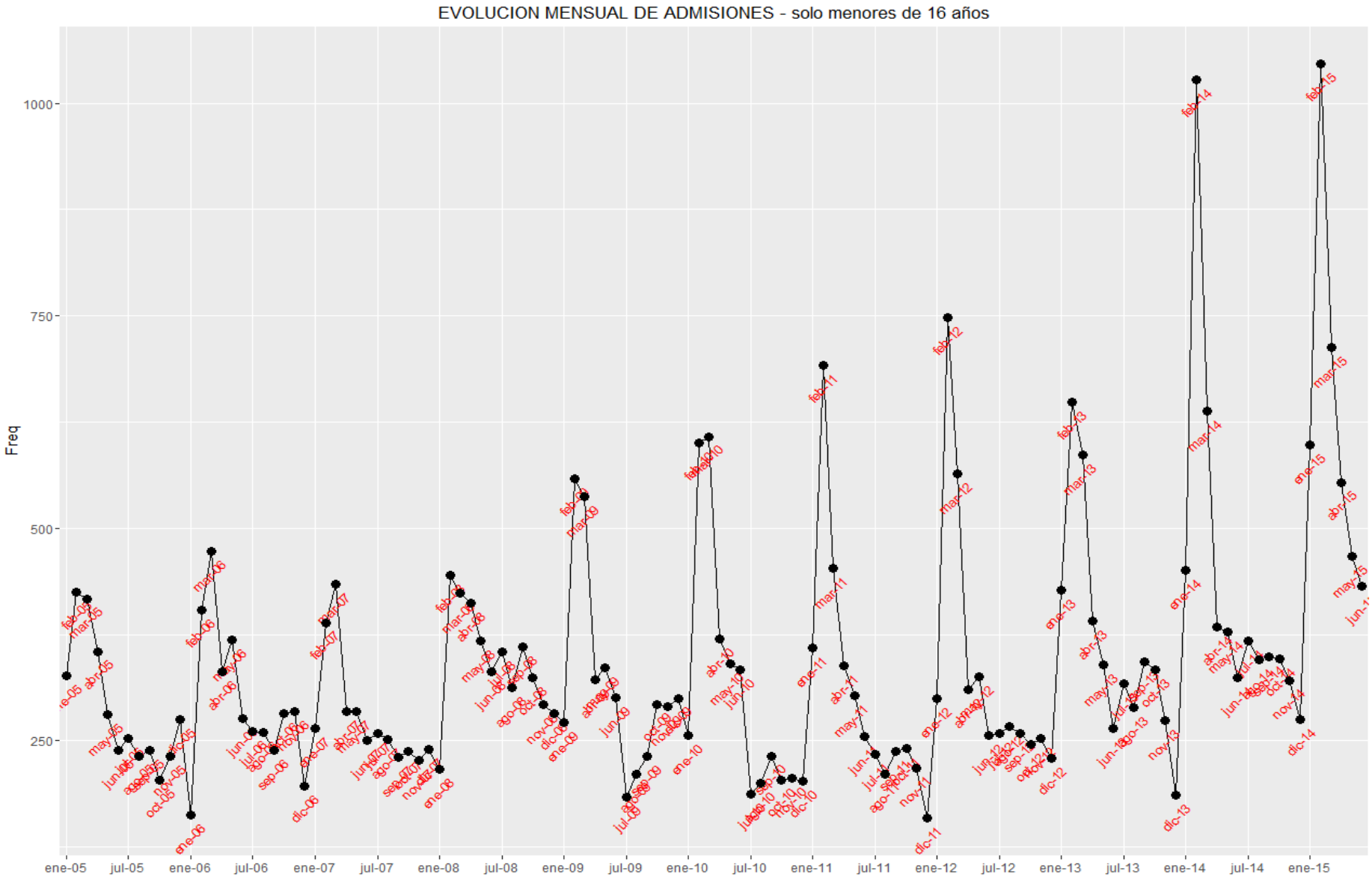


Figura 14. Evolución mensual de la cantidad de admisiones desde Enero 2005 a Junio 2015 – Pacientes pediátricos (menores a 15 años)

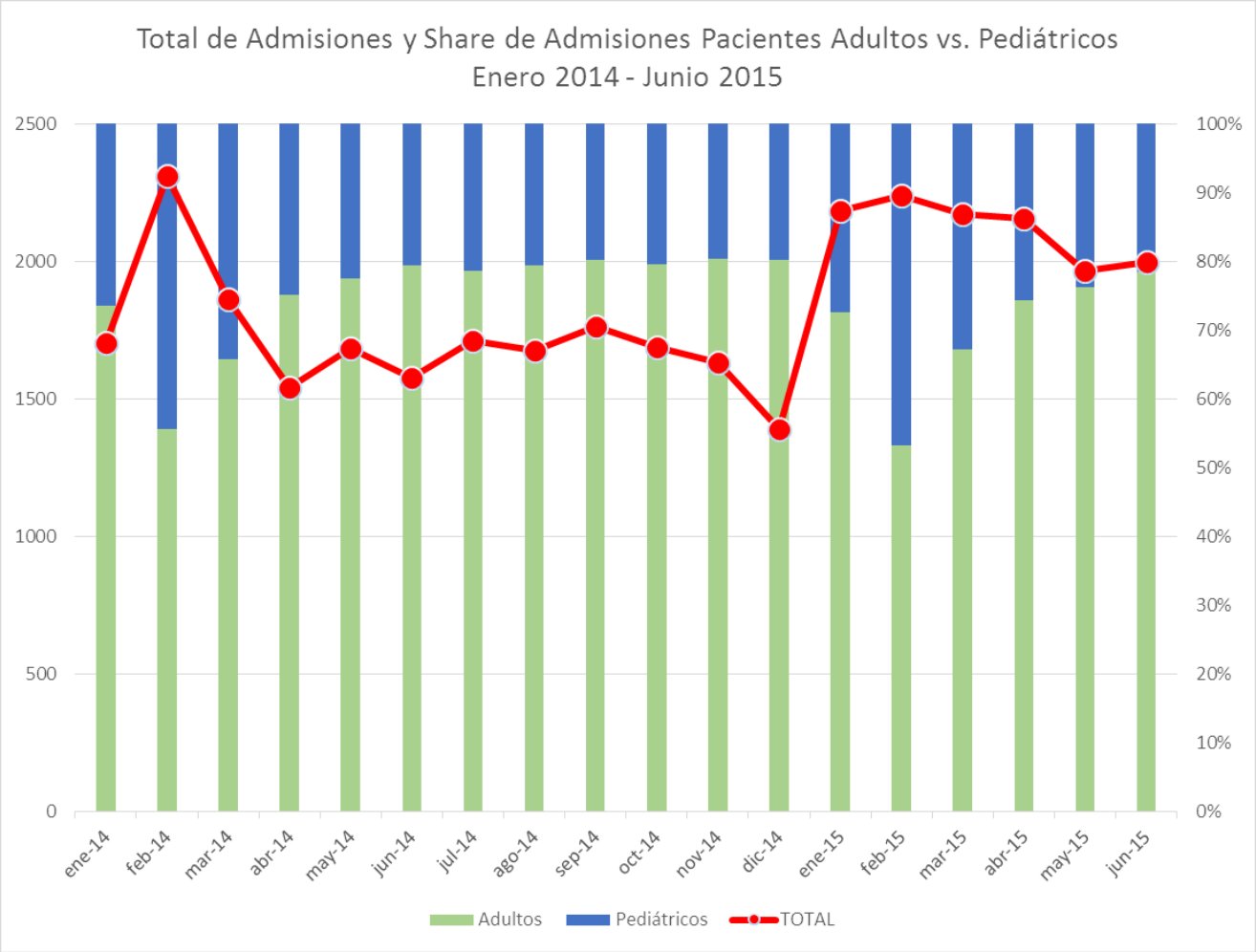


Figura 15. Admisiones totales y Distribución de Admisiones Pacientes Adultos-Pediátricos (Enero 2014 – Junio 2015)

En el gráfico de la figura 15 se compara el porcentaje de admisiones de pacientes adultos y de pacientes pediátricos. Además se muestra la cantidad de admisiones totales para el mismo período. Tal lo manifestado en los párrafos anteriores, se observa que en los meses de Enero, Febrero y Marzo hay un mayor porcentaje de admisiones de pacientes pediátricos que empieza a descender en Abril, para estabilizarse a partir de Julio. Por otro lado, la cantidad máxima de admisiones de pacientes pediátricos coincide el pico del total de admisiones, que se da en Febrero, y se estabiliza en Abril.

Se observa una baja importante en la cantidad de admisiones durante el mes de Diciembre, la cual se revierte en Enero, justamente cuando comienza a incrementarse la cantidad de pacientes pediátricos. Estos datos, si bien fuera de lo que se pretende analizar en este trabajo, son de vital importancia dado que permite identificar aquellos períodos en los que debiera preverse una mayor demanda de atención para determinado tipo de pacientes.

3.2.8 Antecedentes Clínicos del Paciente

Otro punto importante según Deveugele *et al.*[4], Westcott [10] y Raynes [11] es el diagnóstico que presenta el paciente, ya sea precisado en ese momento o previamente por otro profesional. Por ejemplo, los autores antes detallados mencionan que cuando el paciente presenta una afección psiquiátrica o psicológica se ve incrementado el tiempo que el mismo pasa en el consultorio médico. En la clínica bajo estudio se le consulta a cada paciente los antecedentes de salud personales y familiares. También se pregunta si está en ese momento consumiendo alguna clase de medicación. Este dato se deja detallado en el sistema. De todos modos se debe ser cuidadosos dado que muchos pacientes por vergüenza, olvido o alguna otra razón no indican que poseen alguna afección o que están medicados. No obstante, según los datos de la clínica el 20% de los pacientes activos declaran tomar alguna medicación (se considera cada paciente en forma única independientemente de la cantidad de veces que haya sido admitido).

Si se compara con otros estudios [41] [42] los porcentajes son bajos. Se estima que alrededor del 12% de la población en Argentina toma algún medicamento para tratar enfermedades sicológicas/siquiátricas, el 7% tiene diabetes, el 6% EPOC (Enfermedad Pulmonar Obstructiva Crónica). Por otro lado, más del 80% de los adultos mayores consumen 1 o más medicamentos [42]. Por ende, se puede inferir que son muchos más los pacientes que están medicados pero que no lo han indicado.

A pesar de esta falencia, usando esta información y los antecedentes médicos del paciente, los cuales declarará o no de la misma forma que sucede con la medicación que consumen, se puede en cierta medida determinar qué clase de afección de base posee cada persona que es atendida.

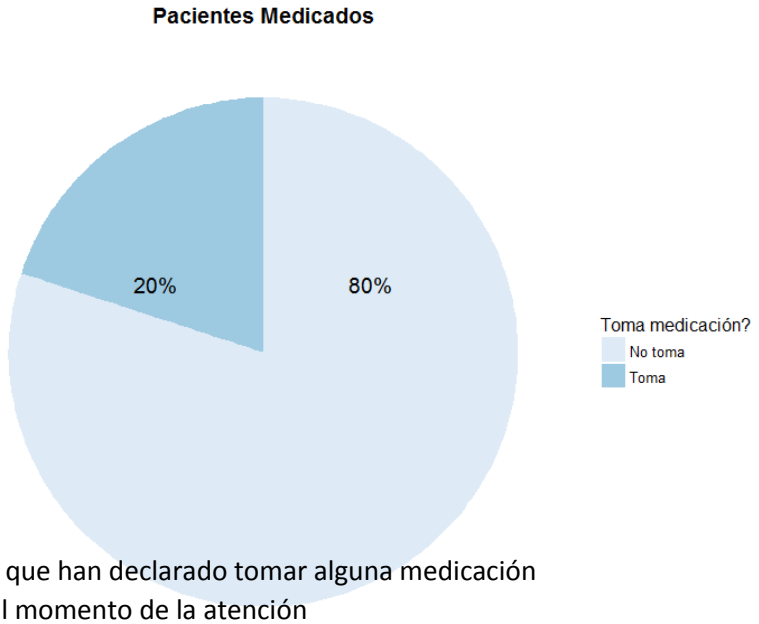


Figura 16. Pacientes que han declarado tomar alguna medicación al momento de la atención

En la tabla 5 y en la figura 17 se describen los grupos ATC de medicamentos que toman aquellos pacientes que declararon consumir algún fármaco. La clasificación ATC codifica los medicamentos en distintos niveles según el sistema u órgano sobre el que actúa, el efecto farmacológico, las indicaciones terapéuticas y la estructura química del fármaco. El primer nivel de la clasificación, el anatómico, el cual indica el órgano o sistema en el cual actúa el fármaco. Existen 14 grupos en total:

Grupo	Descripción
A	Sistema Digestivo Y Metabolismo
B	Sangre Y Órganos Hematopoyéticos
C	Sistema Cardiovascular
D	Medicamentos Dermatológicos
G	Aparato Genitourinario Y Hormonas Sexuales
H	Preparados Hormonales Sistémicos, Excl. Hormonas Sexuales
J	Antiinfecciosos En General Para Uso Sistémico
L	Agentes Antineoplásicos E Inmunomoduladores
M	Sistema Musculoesquelético
N	Sistema Nervioso
P	Productos Antiparasitarios, Insecticidas Y Repelentes
R	Sistema Respiratorio
S	Órganos De Los Sentidos
V	Varios

Tabla 4. Grupos ATC – 1er Nivel

Los datos de la medicación que toma cada paciente es unificada en una serie de campos especialmente diseñados a partir de la información obtenida desde las columnas OBSERVACIONES de cada tabla del sistema (ver punto 3.1.3) y los antecedentes médicos del paciente (registrados en una tabla en particular). Para determinar los niveles de ATC de cada una de las medicaciones que los pacientes declaran consumir en forma regular se desarrolla un algoritmo en R. Este cruza los datos de los nombres comerciales de los medicamentos con un vademécum online para así obtener los principios activos (drogas que lo componen). Una vez que se cuenta con el detalle de las drogas que contiene cada medicación el algoritmo busca automáticamente en la web de la ANMAT (http://www.anmat.gov.ar/atc/CodigosATC.asp?letra=A) el correspondiente código ATC para cada principio activo. La ANMAT posee el código completo (5 niveles), por lo que se extraen los primeros 3 caracteres de dicho código, que son los que se usan en el presente trabajo (dos primeros niveles: Nivel Anatómico y Subgrupo Terapéutico). Si la medicación estuviera conformada por dos o más drogas se toma solo la principal.

Esta información resulta de vital importancia para esta tesis dado que sirve para saber las enfermedades de base que tiene cada persona que se atiende en el consultorio.

En la tabla 5 y en la figura 17 se ve la cantidad y el porcentaje de Pacientes, respectivamente, según cada Nivel Anatómico (Nivel 1) de la clasificación ATC. El color más oscuro en la columna Frecuencia (Freq) indica un valor más alto.

ATC_Clase	Freq	Porcentaje
A	1851	11.9
B	382	2.4
C	4127	26.5
D	167	1.1
G	278	1.8
H	1235	7.9
J	238	1.5
L	272	1.7
M	701	4.5
N	1285	8.2
P	63	0.4
R	1440	9.2
S	3541	22.7
V	14	0.1

Tabla 5. Tipo de medicamento que toman los pacientes según clasificación ATC – Nivel 1

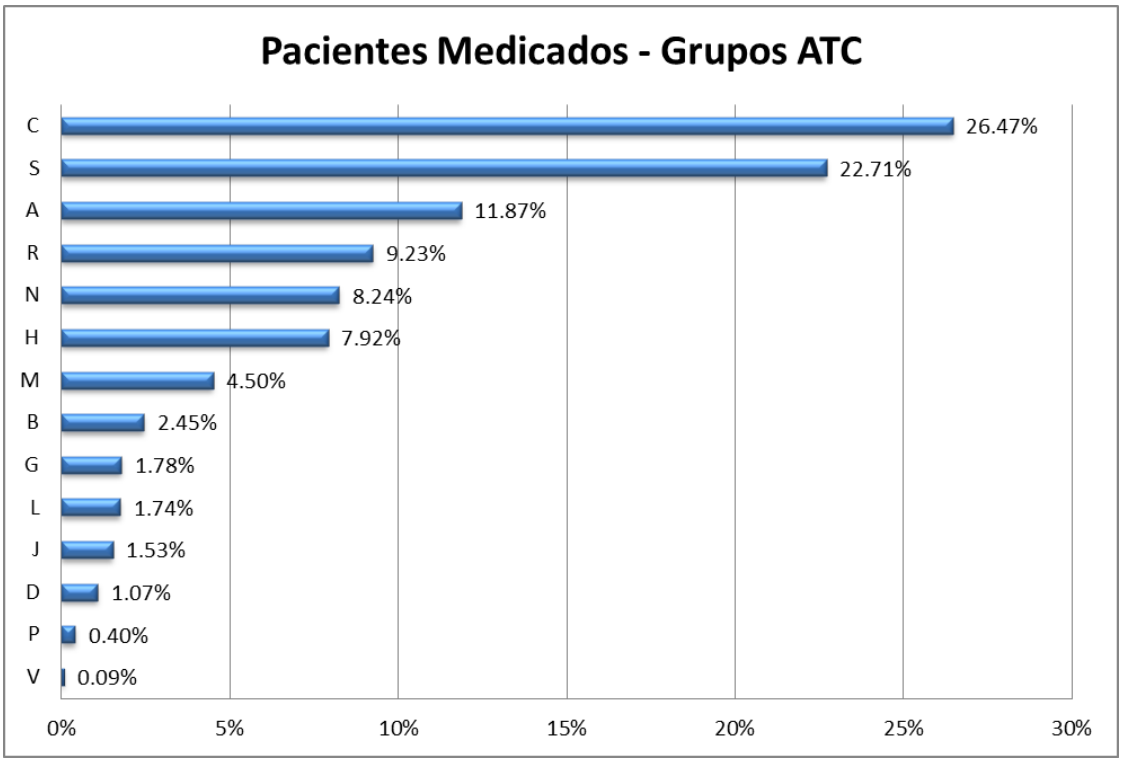


Figura 17. Tipo de medicamento que toman los pacientes según clasificación ATC – Nivel 1

Se aprecia que el tipo de medicamento consumido más declarado es el que afecta al sistema Cardiovascular (Grupo C). En segunda instancia aparece el Grupo S (Organos de los Sentidos), algo esperable dada la especialidad de la clínica.

La aparición del grupo C en el primer puesto es lógico si se tiene en cuenta que, tal como se mencionó en la figura 10, el 50,5% de los pacientes tienen más de 40 años, edad en la cual comienzan a aparecer este tipo de patologías [8] [9]. Este tipo de medicamentos incluye antihipertensivos, vasodilatadores, diuréticos, vasoprotectores, beta-bloqueantes, agentes reductores de lípidos y de terapia cardíaca general. La gran cantidad de pacientes que consumen medicación clase C del Nivel 1 de la clasificación ATC es concordante con lo informado por el Ministerio de Salud de la Nación, quien indica que la primera causa de muerte en Argentina son las enfermedades cardiovasculares [26].

Cada grupo del nivel 1 se abre en varios subgrupos del Nivel 2, llamado Subgrupo Terapéutico. En la figura 18 se puede observar la misma información que en la tabla 5 pero segregada por el nivel 2 de la clasificación ATC. Para este nivel se utilizan dos números. Por ejemplo, la amoxicilina es reconocida con el código J01, que se obtiene de la siguiente manera:

- J → Anti-infecciosos En General Para Uso Sistémico
- 01 → Antibacterianos para uso sistémico.

Por ejemplo el nivel 2 del Grupo Anatómico N es el siguiente:

Grupo ATC: N – Fármacos de uso sobre el sistema nervioso	
N01	Anestésicos
N02	Analgésicos
N03	Antiepilépticos
N04	Antiparkinsonianos
N05	Psicolépticos
N07	Otras drogas que actúan sobre el sistema nervioso

Tabla 6. Grupos ATC – 2do Nivel para el Nivel Anatómico N

Según el listado de la figura 18 los medicamentos más consumidos por los pacientes son los de los grupos:

- C09 - SISTEMA CARDIOVASCULAR - Agentes que actúan sobre el sistema renina-angiotensina
- S01 - ÓRGANOS DE LOS SENTIDOS - Oftalmológicos
- A10 - SISTEMA DIGESTIVO Y METABOLISMO - Fármacos usados en diabetes
- H03 - PREPARADOS HORMONALES SISTÉMICOS, EXCL. HORMONAS SEXUALES - Terapia tiroidea
- C07 - SISTEMA CARDIOVASCULAR - Agentes beta-bloqueantes
- C03 - SISTEMA CARDIOVASCULAR - Diuréticos
- R06 - SISTEMA RESPIRATORIO - Antihistamínicos para uso sistémico
- C10 - SISTEMA CARDIOVASCULAR - Agentes que reducen los lípidos séricos
- N05 - SISTEMA NERVIOSO - Psicolépticos

Entonces, sabiendo qué grupo anatómico del paciente está siendo tratado se puede conocer las enfermedades de base del paciente.

Grupo_ATC	Freq	Porcentaje
A02	1286	1.501
A03	164	0.191
A05	3	0.004
A06	1	0.001
A07	29	0.034
A08	6	0.007
A09	18	0.021
A10	6485	7.570
A11	685	0.800
A12	2202	2.570
A16	42	0.049
B01	2065	2.410
B03	350	0.409
C01	1722	2.010
C02	4	0.005
C03	4711	5.499
C04	152	0.177
C05	230	0.268
C07	5407	6.311
C08	2509	2.929
C09	14168	16.538
C10	4242	4.952
D06	312	0.364
D07	719	0.839
D10	23	0.027
D11	67	0.078
G02	14	0.016
G03	801	0.935
G04	425	0.496
H01	61	0.071
H02	399	0.466
H03	5957	6.953

Grupo_ATC	Freq	Porcentaje
J01	750	0.875
J02	47	0.055
J04	75	0.088
J05	254	0.296
L01	1236	1.443
L02	211	0.246
L03	60	0.070
L04	46	0.054
M01	3005	3.508
M03	71	0.083
M04	220	0.257
M05	1293	1.509
N01	9	0.011
N02	241	0.281
N03	2665	3.111
N04	191	0.223
N05	3509	4.096
N06	1110	1.296
N07	265	0.309
P01	474	0.553
P02	5	0.006
R03	1740	2.031
R05	185	0.216
R06	4401	5.137
R07	36	0.042
S01	8217	9.591
V01	60	0.070
V03	4	0.005
B02	3	0.004
R01	2	0.002
S02	26	0.030

Figura 18. Tipo de medicamento que toman los pacientes según clasificación ATC – Nivel 2 (color más oscuro indica valor más alto)

3.3 Variable Objetivo – Tiempos de Atención

3.3.1 Situación actual

La planificación de la agenda de turnos médicos debe ser, a la vez, eficiente y flexible. Debe satisfacer las necesidades de los pacientes, de los profesionales de la salud y del personal no médico. Según el país que se trate, existen grandes diferencias acerca del tiempo de atención óptimo que permitiría alcanzar los objetivos antes descriptos. En España los turnos varían entre 10 y 20 minutos; en Rusia está reglamentado en 10 minutos; en El Salvador cuentan con turnos de 10 minutos mientras que en Perú los turnos son de 12 minutos. En Estados Unidos, un estudio que recopiló información de más de 46.000 consultas médicas entre 1997 y 2005, demostró que el tiempo promedio de consulta había aumentado de 16 a 20,8 minutos. En Japón el promedio de duración de la consulta médica es de 6 minutos. En Etiopía una investigación puso de manifiesto que la consulta es de unos 6,26 +/- 2,55 minutos, pero los pacientes esperan que la misma dure unos 14,02 +/- 6,73 para considerarse bien atendidos. En Canadá hay dos sistemas

de salud: uno donde el médico cobra honorarios por cada consulta y otro en el cual tiene un salario mensual. Los tiempos de consulta en el primer caso varían entre 10 y 15 minutos mientras que en el segundo lo hacen entre 20 y 45 minutos.

Deveugele *et al.* [4] condujeron un estudio en Europa sobre la duración de las consultas. Tomaron grupos de profesionales de 6 países a los cuales les entregaron encuestas que debían ser completadas tanto por ellos como por los pacientes. Contaron con la participación de 190 profesionales y 3600 pacientes. La duración media de la consulta general fue de 10,7 minutos (tabla 7).

País	Tiempo Medio en minutos (Desv.Estand.)
Alemania	7,6 (4,3)
España	7,8 (4,0)
Reino Unido	9,4 (4,7)
Holanda	10,2 (4,9)
Bélgica	15,0 (7,2)
Suiza	15,6 (8,7)
General	10,7 (6,7)

Tabla 7. Duración de la consulta médica en Europa [4]

Según los autores, la variación entre los países se debe básicamente al sistema de atención. En Alemania y España cada profesional médico tiene un promedio de 200 consultas a la semana. Esto alta demanda de médicos hace que las consultas sean breves.

En Bélgica y Suiza los médicos operan en un mercado abierto, donde los pacientes tienen acceso a más de un médico y especialista. Esto implica que el profesional deba realizar cierta inversión en cada paciente para satisfacer sus necesidades y garantizarse su fidelidad.

En el Reino Unido y Holanda poseen servicios de salud bien organizados, con listas de pacientes limitadas en cuanto a cantidad. El estado les abona a cada médico una comisión por cada paciente atendido. De ahí que los tiempos de atención no sean ni tan bajos como en Alemania ni tan altos como en Suiza.

Por otro lado, Landau *et al.* [27] indican en su estudio (encuestas realizadas a 7100 pacientes en 4 clínicas de Israel; resultados analizados mediante análisis de correlación de Pearson o Spearman) que el promedio de la consulta médica es de 12,1 +/- 1,6 minutos. La expectativa por parte de los pacientes respecto de una consulta con duración óptima era de 15,4 +/- 8,4 minutos en promedio, con una mediana de 15 minutos.

De acuerdo a la bibliografía consultada, el promedio de duración de la consulta oscila entre 10 y 15 minutos, tiempo en apariencia insuficiente. Solamente alguna bibliografía proveniente de Europa propone destinar 60 minutos a la primera consulta y 20 minutos en las siguientes.

En Buenos Aires, Argentina, según detallan Outomuro y Actis [28], en clínica médica el tiempo de atención promedio es de 15 minutos y es algo mayor en otras especialidades como pediatría o salud mental. Según las autoras, durante la década de 1990-99 el tiempo de atención pasó de 7,5 a 10 minutos. Esto redituó en beneficio de los pacientes, quienes se sentían mejor atendidos, y de los médicos, que disminuyeron su nivel de stress.

La forma en la que se abona a los médicos en Argentina varía según la institución sea pública o privada. En el ámbito privado el médico cobra por paciente, mientras que en el público cobra un sueldo fijo por mes. Tal como se mencionaba anteriormente, esto es determinante para el profesional al momento de determinar cuánto tiempo dedica a cada paciente. También varía la forma de trabajo, dado que en el ámbito público los puestos tienen una determinada carga horaria fija, mejor capacitación y mayor estabilidad laboral, mientras que en el privado pueden tener varios trabajos sin carga horaria predeterminada, mejores sueldos aunque mayor incertidumbre laboral. Es común que muchos profesionales atiendan en sendos ámbitos para poder lograr el equilibrio entre un ingreso fijo, buena capacitación y mejores ingresos.

Vale aclarar que se consideran únicamente profesionales en ejercicio y no estudiantes ni participantes del Sistema de Residencias Médicas Universitarias de Argentina.

3.3.2 Situación en la clínica bajo estudio

En la clínica bajo estudio se abona a los profesionales por paciente atendido. El tiempo promedio que los médicos pasan con cada paciente es de 13,66 minutos, mientras que la moda es de 8 minutos *. Los deciles son los siguientes:

Tiempos de Atención en la Clínica - Deciles

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	4	5	7	8	10	12	14	18	25	91

Tabla 8. Deciles de los tiempos de atención en la clínica bajo estudio

** ver apartado 3.3.5 Valores Extremos*

Resumiendo, la mitad de los pacientes pasan 10 minutos o menos en el consultorio del médico.

En los gráficos de las figuras 19 a 23 se muestra la evolución de los tiempos de atención a través de los años de 5 médicos distintos tomados al azar. Se puede ver que en la totalidad de los casos el tiempo de atención disminuye en el último año respecto del primero.

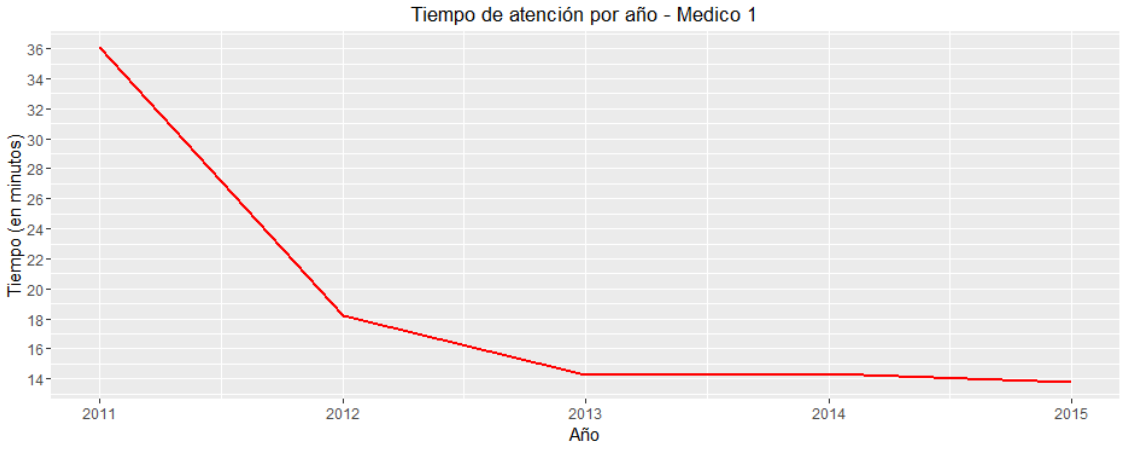


Figura 19. Tiempo de atención a través de los años

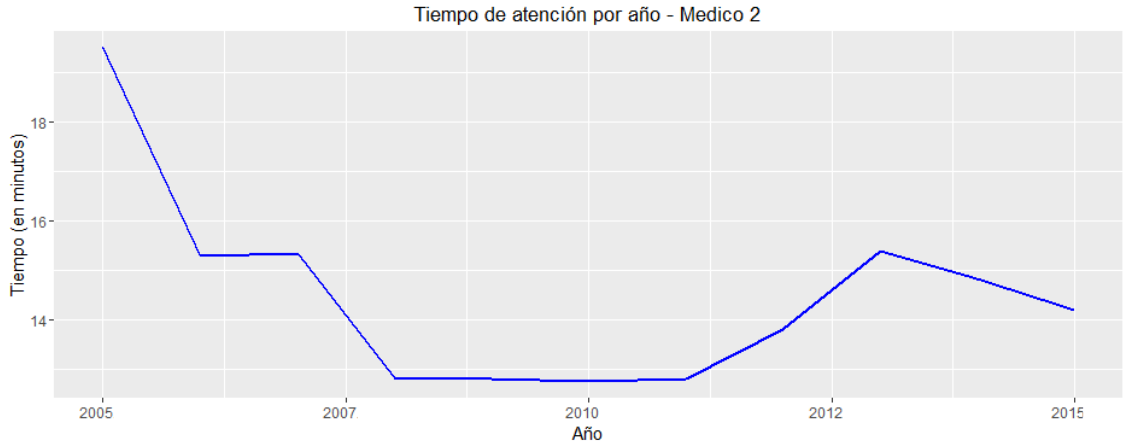


Figura 20. Tiempo de atención a través de los años

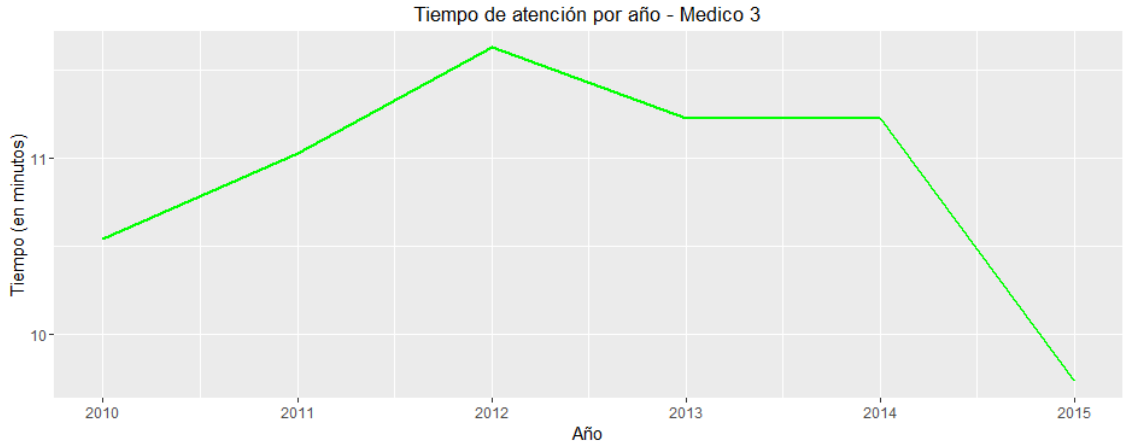


Figura 21. Tiempo de atención a través de los años

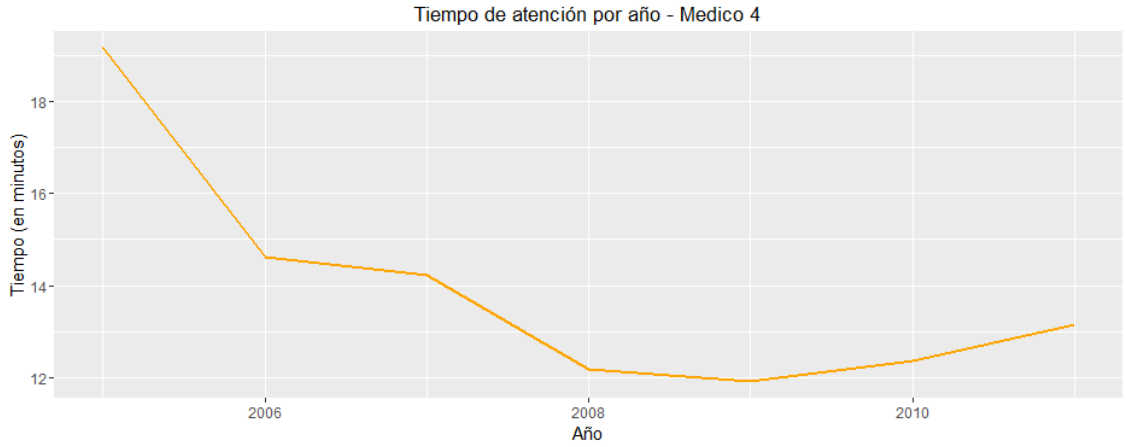


Figura 22. Tiempo de atención a través de los años

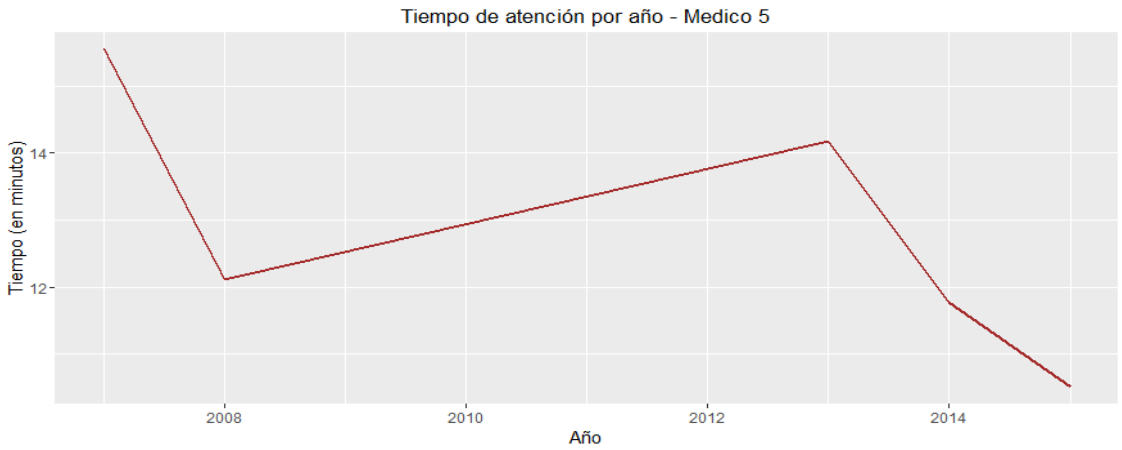


Figura 23. Tiempo de atención a través de los años

En los 5 gráficos se ve claramente que el tiempo de atención de cada uno de los profesionales disminuyó a través de los años. Si bien debe considerarse la cantidad de pacientes atendidos por cada profesional, las horas que cada uno atiende, que los médicos pueden atender en más de una clínica por día, si trabajan tiempo parcial o tiempo completo, los registros de los últimos años detallan tiempos de atención de entre 9 y 14 minutos. Si bien queda fuera del alcance del trabajo, se aprecia que existen motivos por los cuales los profesionales deben reducir el tiempo que pasan con cada paciente. Sin embargo hay que destacar que el tiempo que pasan con los pacientes decrece con el correr de los años y que, según lo mencionado anteriormente, está muy por debajo de las expectativas de los pacientes.

3.3.3 Tiempos de atención en la clínica bajo estudio

En la figura 24 se puede observar la distribución de los pacientes según el tiempo que pasaron siendo atendidos por el profesional médico. Cómo era de esperar muestra una distribución sesgada hacia la derecha.

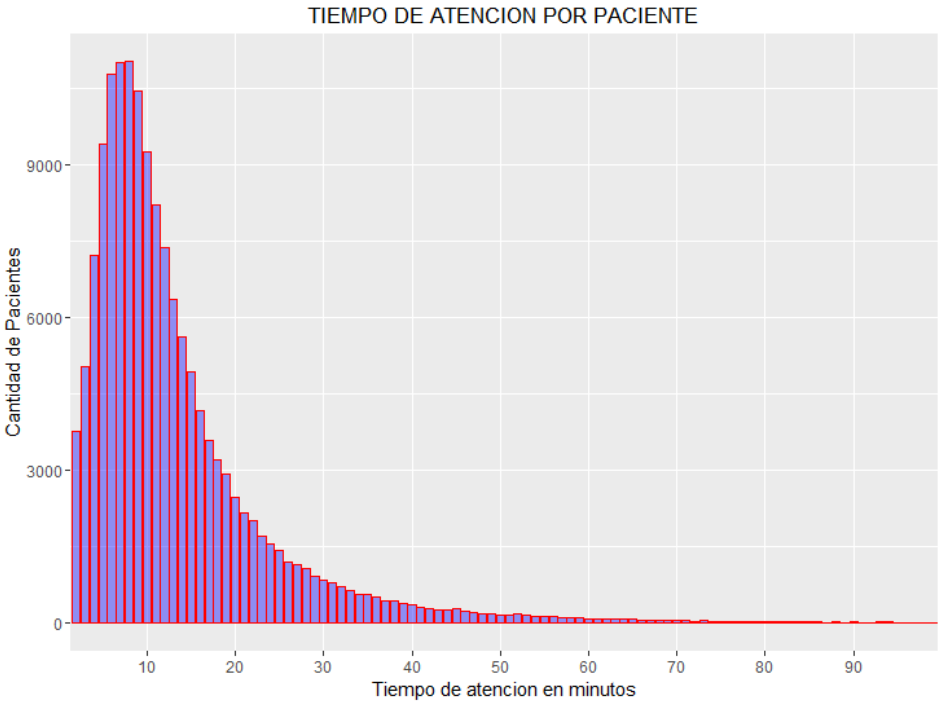


Figura 24. Cantidad de pacientes según tiempo de atención

En base a lo detallado en la bibliografía y a los datos obtenidos de esta clínica se tomará como tiempo de corte 10 minutos, dado que esa es la duración estipulada del turno en la clínica bajo estudio y en gran parte de los consultorios relevados. Entonces, como objetivo de este trabajo se evaluarán los motivos por los cuales los pacientes demoran más o menos de 10 minutos en el consultorio médico. Se realizará un análisis cualitativo a partir de definir una variable objetivo binaria con valores posibles “ALTO” (atención mayor a 10 minutos) y “BAJO” (atención menor a 10 minutos).

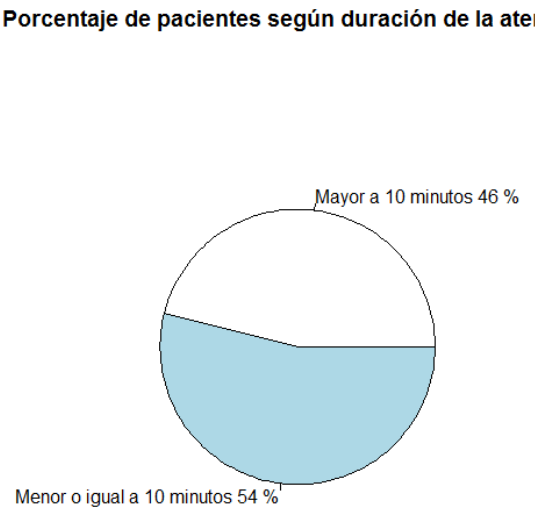


Figura 25. Proporción de pacientes según el tiempo de atención

La creación de este tipo de variable objetivo permite analizar por qué los pacientes demoran más o menos que la duración estándar de los turnos en la zona donde está ubicada la clínica (Ciudad de Buenos Aires y Gran Buenos Aires).

Transformar la variable objetivo de continua en binaria responde al objetivo de este trabajo, que es determinar cuáles son las características de los pacientes que hacen que demoren más que lo estipulado por las agendas médicas actuales en Buenos Aires. No se busca cuantificar (hacer un pronóstico de cuánto tardará cada paciente) sino hacer un análisis cualitativo de los pacientes en base a los tiempos vigentes en el sistema médico actual.

3.3.4 Situaciones Excepcionales

En muchas ocasiones ocurre que miembros de una familia se atienden en forma conjunta (por ejemplo, una madre con sus hijos). En esos casos todos los pacientes son admitidos al mismo tiempo, por lo que el tiempo de atención es 0 debido a la metodología de cálculo utilizada. Estos casos fueron eliminados, así como los primeros y los últimos pacientes de cada día, dado que la fórmula de cálculo del tiempo de atención necesita que haya un paciente antes o después. Ninguno de estos casos fue tenido en cuenta a la hora de procesar los datos.

3.3.5 Valores de tiempo de atención extremos (outliers)

Previo al armado de los modelos se analizan los tiempos de atención de los pacientes en busca de outliers. Los principales estadísticos descriptivos muestran lo siguiente:

Tiempos de atención – Estadísticos Descriptivos

Tiempo (en min.)	
Min.	0.00
1st Qu.	6.00
Median	10.00
Mean	12.74
3rd Qu.	16.00
Max.	752.00

Tabla 9. Tiempos de atención – Detección de valores extremos

A simple vista se puede apreciar que hay valores en 0 y que el máximo es de más de 12 horas y media (752 minutos), lo cual no resulta coherente.

Los tiempos altos responden a dos motivos:

- Errores en la carga del paciente
- Pacientes pre-quirúrgicos

Se detectaron casos en los que el paciente era mal admitido por sistema y no se concluía su atención en un tiempo prolongado, por lo que el tiempo de atención resultó alto. Estos casos se eliminan.

Por otro lado a algunos de los pacientes a los que se les practica una cirugía deben someterse a una serie de estudios previos que pueden durar entre 90 minutos y 4 horas aproximadamente. Siendo que el objetivo del trabajo es analizar el tiempo de los pacientes cara a cara con el médico, se eliminan aquellos registros con tiempo de atención mayor a 60 minutos (se eliminan todos los pacientes quirúrgicos).

En cuanto a los pacientes con tiempos de atención iguales a 0, tal como se mencionaba en el apartado 3.3.4, corresponde a miembros de una misma familia que se atienden juntos. Estos registros son eliminados a fin de evitar distorsiones en el análisis.

Con los datos ya regularizados se realiza un boxplot a fin de cotejar la presencia de outliers.

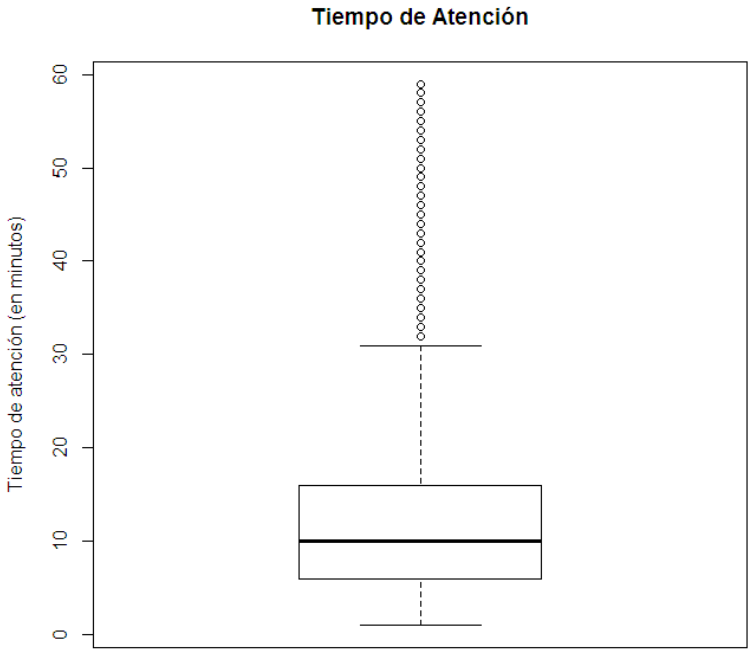


Figura 26. Diagrama de caja de los tiempos de atención

Además se calculan los cuartiles:

Cuartiles de los tiempos de atención (sin valores extremos)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	6.00	10.00	12.33	16.00	59.00

Tabla 10. Tiempos de atención sin valores extremos

Se puede observar gran cantidad de valores por encima de los 16 minutos (3er cuartil) y que el 50% de los valores se encuentran por sobre los 10 minutos.

3.3.5 Tiempos de Espera

Amén que es un dato que queda fuera del alcance de este trabajo, los tiempos de espera de los pacientes en el centro médico bajo estudio son muy altos. Para este trabajo se considera tiempo de espera al tiempo transcurrido entre la admisión del paciente y la atención. Si bien existe la posibilidad de que el paciente llegue mucho tiempo antes del turno, del relevamiento realizado sobre 300 atenciones se constata que más del 97% de los pacientes llega entre 15 minutos antes a 5 minutos después del horario de su turno.

Los estadísticos descriptivos de los tiempos de espera son:

Tiempos de Espera en la Clínica Bajo Estudio

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	28.00	51.00	58.17	81.00	238.00

Tabla 11. Tiempos de Espera en la Clínica bajo Estudio - Estadísticos Descriptivos

Con un tiempo promedio de espera de poco más de 58 minutos y más de la mitad de los pacientes con demoras para ser atendidos de entre 51 y 238 minutos podemos afirmar que los tiempos de espera son altos

Capítulo 4 – Resultados y Discusión

4.1 Configuración del algoritmo

Para desarrollar los modelos se generan dos sub-datasets: uno que contiene el 70% de los registros del dataset original, que se usará para entrenar al modelo y otro con el 30% restante para prueba. También se mantendrá el dataset original para realizar corridas de k-fold cross validation.

Tal como se mencionaba anteriormente, el algoritmo seleccionado para trabajar es el CART. Esta decisión es arbitraria, sino que el mismo tiene las características apropiadas para procesar el dataset (manejo de valores faltantes y/o nulos, tratamiento de variables categóricas y numéricas). Además este algoritmo es de fácil interpretación.

Existen muchas opciones para la implementación del mencionado algoritmo. Para este trabajo se usará la librería RPART [38] de R.

Los algoritmos de árboles de decisión permiten cambiar varios parámetros para generar modelos alternativos, lo que los hace muy flexibles. En el caso del algoritmo CART que implementa la librería RPART permite modificar, entre otros, el nivel de significancia, la profundidad del árbol y el tamaño de la hoja.

- Mínima cantidad de casos por hoja

Este parámetro determina el tamaño mínimo que tiene que tener una hoja para poder ser dividida. Suele ser conocido como tamaño de la hoja. Se usa esta regla de parada para evitar el sobreajuste de los modelos. El valor que toma este parámetro es arbitrario, no existe un criterio general para calcular el tamaño óptimo de una hoja.

- Máxima profundidad del árbol

Es otra regla de detención de crecimiento del árbol. Establece un límite a la generación de nodos hijos con el objetivo de evitar el sobreajuste y disminuir la complejidad de las reglas resultantes.

- Nivel de significancia

El parámetro de complejidad o de significancia define qué tanto debe mejorar el ajuste general del modelo una división para llevarse a cabo.

El sobreajuste (en inglés overfitting) es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. Cuando un sistema se entrena demasiado, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento y no funcionar correctamente ante datos nuevos (no generaliza).

Se seleccionaron estos tres hiperparámetros del árbol dado que son los más usados por los analistas de data mining (no se considera la cantidad máxima de ramas debido a que este algoritmo solo crea árboles binarios, es decir, con dos ramas de salida).
Para configurarlos se procedió a la generación automática de árboles mediante una función escrita en R. Se fijan los valores mínimos y máximos de cada hiperparámetro además del incremento deseado para cada uno. En la tabla 12 se puede apreciar la configuración definida para cada uno. El total de combinaciones posibles, es decir, la cantidad de modelos que se ejecutan, es de 450.

Parámetros	Valor Mínimo	Valor Máximo	Incremento	Valores Posibles
Nivel de Significación	0.00005	0.05	0.00355	15
Profundidad Máxima	5	30	5	6
Tamaño Mínimo de la Hoja	10	50	10	5
Combinaciones Totales				**450**

Tabla 12. Configuración de los parámetros del algoritmo CART

4.2 Evaluación de Modelos

4.2.1 Métricas Tradicionales

Existen varias opciones para evaluar un modelo de este tipo. Entre las más conocidas están las medidas resultantes de la matriz de confusión.

La matriz de confusión se crea analizando los 4 resultados posibles de un clasificador. Si la instancia es positiva y es clasificada como tal, se contabiliza como un verdadero positivo. Si es clasificada como negativa una instancia negativa se considera un verdadero negativo. Si la instancia es positiva y se clasifica como negativa se considera un falso negativo. El caso inverso, una instancia negativa que se clasifica como positiva, es un falso positivo.

		Clase Verdadera	
		Positiva	Negativa
Clase Predicha	Positiva	TP (Verdaderos Positivos)	FP (Falsos Positivos)
	Negativa	FN (Falsos Negativos)	TN (Verdaderos Negativos)
		P (Total Positivos)	**N (Total Negativos)**

Tabla 13. Matriz de Confusión

Los números en la diagonal principal (TP y TN) representan las decisiones correctas, mientras que la diagonal inversa representa los errores. Esta matriz deriva en las métricas básicas de un clasificador:

- Sensibilidad o Recall = $\frac{\mathrm{TP}}{TP+FN}$
- Negative Predicted Value = $\frac{\mathrm{TN}}{TN+FN}$
- Especificidad = $\frac{\mathrm{TN}}{TN+FP}$
- Accuracy (Exactitud)=$\frac{\mathrm{TP+TN}}{TP+TN+FP+FN}$
- Precisión = $\frac{\mathrm{TP}}{TP+FP}$
- F1 = $\frac{2*\mathrm{TP}}{2*TP+FN+FP}$

Se pueden encontrar más métricas, aunque estas son las más comunes.

En el ámbito de este trabajo se considera un POSITIVO a un paciente con duración mayor a 10 minutos dentro del consultorio médico (ALTA DURACION), y NEGATIVO a aquel paciente que demora 10 o menos minutos en ser atendido por el profesional médico (BAJA DURACION).

4.2.2 Matriz de Costos

Por defecto todos los errores (FN y FP) tienen el mismo peso. Sin embargo, cualquiera de ellos puede tener un valor bajo pero un efecto más severo. Por ejemplo, no diagnosticar correctamente a un paciente enfermo puede derivar en la muerte del mismo, pero hacer estudios redundantes a un paciente sano solamente costará más dinero. Queda claro que el primer tipo de error es mucho más grave que el segundo: la diferencia entre un falso positivo y un falso negativo en el ámbito de la salud es crítico.

En muchos escenarios el costo de un FP y de un FN es diferente. Por ese motivo se construye una matriz de costos que permita castigar más a determinado tipo de error a la hora de armar el modelo.

En el caso que nos ocupa los errores pueden ser:

- Falsos Negativos, es decir, pacientes que son catalogados como que van a tener una baja duración (turno normal), que se agendan como tal, pero finalmente demoran más tiempo en el consultorio.
- Falsos Positivos, pacientes que se catalogan como que van a tener turnos largos y finalmente tardan 10 minutos o menos.

De acuerdo a lo detallado se aprecia que el trabajo tiene una capacidad de optimización multiobjetivo: minimizar los tiempos de espera de los pacientes o maximizar los beneficios económicos de la clínica. La confección de la matriz de costos óptima es un tema propio de otro trabajo de tesis. Por motivos que se destacan en los puntos subsiguientes y en el objetivo de este trabajo, el foco se pone los pacientes y en minimizar los tiempos de espera.

Según lo expuesto los Falsos Negativos son los más costosos de los dos. Si muchos pacientes son catalogados de duración baja, se le otorgan turnos de hasta 10 minutos y finalmente demoran más, se atrasarían las demás consultas. Esto podría ocasionar malestar en los pacientes, deterioro en la percepción de la calidad y baja fidelización. En tanto que en los profesionales aumentaría el nivel de stress y acumulación de tareas administrativas.

En la segunda situación, si muchos pacientes son catalogados para ser atendidos en turnos largos y finalmente la atención es más corta produciría tiempos ociosos, es decir, poca o nula espera de los pacientes para ser atendidos pero tiempo ocioso para los profesionales médicos dado que la cola de pacientes está vacía. Sin embargo, ese tiempo sin pacientes podría ser usado para tareas administrativas (según lo indicado por Gottschalk y Flocke casi la mitad de las horas de trabajo de los profesionales médicos se usa para tareas administrativas [24]) o capacitación.

El costo para los Verdaderos Positivos (TP) y los Verdaderos Negativos (TN) es 0 dado que no acarrea mayor peso la decisión los aciertos.

		Clase Verdadera	
		Positiva	Negativa
Clase Predicha	Positiva	0	2
	Negativa	4	0
		P (Total Positivos)	**N (Total Negativos)**

Tabla 14. Matriz de Costos

El paquete RPART que implementa el algoritmo CART en R permite la configuración de una matriz de costos mediante la configuración de parámetros al momento de modelar. Gracias a

esta facilidad se automatiza también la ejecución de varios modelos con distintos valores en la matriz de costo. Se configuran mínimos en 2 y máximos en 10 con incrementos de 2 unidades por corrida. Dejando la diagonal principal siempre en 0 y variando solo la diagonal invertida (valores de FP y FN) da un total de 5x5 = 25 combinaciones. Teniendo en cuenta las 450 combinaciones posibles con los parámetros que fijan el tamaño de la hoja, la profundidad máxima y el nivel de significancia, resulta en un total de 11250 modelos.

4.2.3 Fidelidad

Se ha mencionado antes en este trabajo que gran cantidad de los pacientes concurren solamente una vez a la clínica. También se destacó el hecho que es una de los pocos centros de la zona que atienden la especialidad de Pediatría. Considerando ambos hechos se aprecia que el porcentaje de pacientes que concurrió dos o menos veces a la clínica es el siguiente:

Cantidad y Porcentaje de Pacientes que concurrieron 1 o 2 veces

Edades	Cant. de Pacientes	Porcentaje (sobre total de pacientes)
Menores a 16	11076	18,94%
Entre 16 y 40	12327	21,08%
Mayores a 40	14794	25,29%
Total Pacientes con 1 o 2 visitas	**38197**	**65,31%**
Total de Pacientes: 58486		

Tabla 15. Fidelidad de Pacientes

Se advierte que, dependiendo de la franja de edad, entre el 18% y el 25% de los pacientes no vuelven a la clínica luego de la 1era o 2da consulta. Considerando el total de pacientes que son atendidos, el 65,31% no regresan a la clínica luego de la 1era o la 2da visita. Teniendo en cuenta que el Ministerio de Salud de la Nación recomienda una visita anual o bianual al médico según la edad [39], estos valores de "fuga" de pacientes son una alerta a tener presente. Por este motivo y por lo indicado en el punto 4.2.2, al armar la matriz de costo mencionada en el punto anterior se "castiga" más a los falsos negativos que a los falsos positivos. En otras palabras, se le da un mayor peso a los errores cometidos por clasificar a paciente que tendría un tiempo de atención alto y es clasificado como bajo. De este modo se fija la matriz de costos para que los valores ubicados en el cuadrante inferior izquierdo (FN) sean mayores a los del cuadrante superior derecho (FP). La cantidad de combinaciones posibles queda reducida a 10, por lo que el total de modelos a ejecutar será de 450 x 10 = 4500.

Los modelos fueron ejecutados en una Notebook con 5 procesadores y 12Gb de memoria. El tiempo aproximado de ejecución ronda las 96hs (por corrida).

4.2.4 Curva ROC y AUC

La curva ROC (Receiver Operating Characteristics) es un gráfico que permite visualizar y comparar los resultados de varios modelos predictivos [40]. Es un gráfico de dos ejes entre 0 y 1, en el cual se encuentran los casos positivos y los casos negativos ordenados por la probabilidad calculada por el modelo. Al graficar los casos ordenados por la probabilidad cuanto más se alejen de la diagonal entre (0;0) y (1;1), mayor será la capacidad de predicción del modelo. Este gráfico resulta una herramienta muy eficiente para poder comparar modelos entre sí.

En el eje X del gráfico se le muestra el resultado de restar 1 a la cantidad de verdaderos negativos (TN) dividido la cantidad total de negativos (coeficiente llamado False Positive Rate o tasa de falsos positivos), mientras que en el eje Y se muestra el True Positive Rate (tasa de verdaderos positivos), resultante de dividir TP (verdaderos positivos) por la cantidad total de positivos (fig. 27).

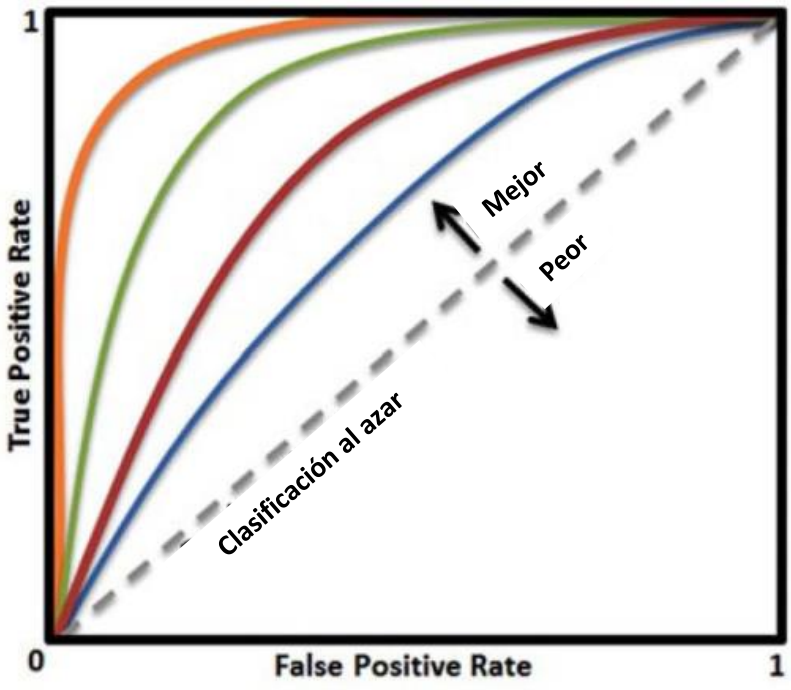


Figura 27. Ejemplo de Curva ROC

Al momento de evaluar distintos modelos con la curva ROC conviene calcular el área bajo la curva (AUC: Area Under an ROC Curve) para poder medir numéricamente los resultados de los modelos. Así se puede comparar entre varios modelos cual es el que posee mayor área y por lo tanto mayor capacidad de predicción.

Este valor es llamado AUC y su valor se encontrará entre 0 y 1. A mayor AUC, mejor modelo.

4.3 Ejecución de Modelos

4.3.1 Primeros Modelos usando AUC

Los primeros modelos ejecutados cuentan con algunas variables que hacen referencia a los médicos (edad, sexo y tipo de profesional) y estas características mostraban ser importantes. Resulta evidente entonces que existen demoras que se originan en los pacientes y otras que nacen en los médicos.

Queda claro que los pacientes dependen de los médicos y viceversa. El sistema de salud involucra a los dos actores. La exclusión de los médicos para este trabajo responde solamente a hipótesis planteada. Ergo, los tiempos de servicio dependen tanto de unos como de otros. Sin embargo, como el presente trabajo se centra en los pacientes se optó por eliminar todas aquellas variables que hagan referencia a las características de los profesionales de la salud.

Tal como se mencionó anteriormente, la ejecución de los modelos fue automatizada modificando 3 parámetros: Nivel de Significación, Profundidad Máxima y Tamaño Mínimo de la

Hoja. La elección de valores mínimos y máximos de estos parámetros es arbitraria. A estos parámetros se agrega la configuración de la matriz de costos, tomando en consideración la necesidad de que los FN sean mayores a los FP (ver punto 4.2.3). En esta etapa se toma como medida de evaluación el AUC de cada modelo.

Los valores de AUC de los 4500 modelos no dan buenos resultados: el mayor AUC encontrado es de 0.57, un valor muy cercano al azar. Por este motivo se analizan, en los mejores modelos, las variables de mayor importancia a fin de intentar refinar los modelos.

Al no obtener mejoras significativas se crean distintas variables a partir de las ya existentes:

- Dummy de estudio realizado, para saber qué estudio se realizó un paciente
- Cantidad de consultas en el mes / edad en la atención
- Cantidad de consultas en el semestre / edad en la atención
- Cantidad de consultas en el año / edad en la atención
- Cantidad parcial de consultas / edad en la atención
- Cantidad total de consultas / edad en la atención
- Suma de variables dummy de estudio realizado, para conocer el total de estudios por paciente
- Variable lógica que indica si el paciente vive en el Gran Buenos Aires o en la Ciudad Autónoma de Buenos Aires
- Variable que indica en qué zona del Gran Buenos Aires vive el paciente
- Partido del Gran Buenos Aires en el que vive el paciente
- Sumatoria de las dummy de medicación, para conocer la cantidad total de medicamentos consumidos por paciente
- Dummy de medicación / edad en la atención
- Dummy de estudio / edad en la atención
- Dummy de medicación / Sumatoria dummy estudios

Finalmente se obtiene una matriz de 182103 filas por 341 columnas, siendo cada fila un registro de admisión de paciente y cada columna una variable. Si bien son una gran cantidad de variables, dada la naturaleza del trabajo a realizar y la cantidad de estudios que se llevan a cabo en la clínica, dicho valor se considera aceptable.

En un esfuerzo denodado por encontrar mejores modelos se ejecutan, de forma complementaria, modelos de Ensemble y algunas corridas de Random Forest. Sin embargo los resultados no muestran mejorías notorias.

Los métodos de Ensemble combinan múltiples modelos (en este caso árboles de decisión) en uno solo generalmente más preciso que el resto de los componentes por separado. En cuanto al algoritmo de Random Forest busca promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto, reducir la variación. Ambos algoritmos fueron ejecutados en R (librerías Caret y RandomForest, respectivamente). Para Random Forest se utiliza la misma metodología que se emplea con Rpart: se definen valores máximos, mínimos y un intervalo de incremento para cada hiperparámetro y se automatiza la ejecución del algoritmo buscando la mejor combinación de los mismos. Solo de Random Forest se ejecutaron 1001 modelos tomando como hiperparámetros el tamaño mínimo de los nodos terminales (nodesize), el número de árboles a usar para armar el modelo (ntree) y el número de variables elegidas aleatoriamente como candidatas para cada división de ramas (mtry).

Parámetros	Valor Mínimo	Valor Máximo	Incremento	Valores Posibles
Tamaño Mínimo del Nodo	500	1100	50	13
Número de Árboles	5	55	5	11
Variables Elegidas	2	14	2	7
Combinaciones Totales				1001

Tabla 16 Configuración de los parámetros del algoritmo Random Forest

El problema encontrado en todos los casos es similar: los modelos tienden a predecir la mayoría de los casos como positivos o como negativos, dependiendo de la configuración del algoritmo que se ejecute.

- Ejemplo de corrida de algoritmo CART (Rpart) con minsplit = 50 y cp = 1.289858e-03 (minsplit: tamaño mínimo de hoja; cp: nivel de significancia).

modelo.test	alto	bajo
alto	4003	3254
bajo	11203	14499

Tabla 17. Matriz de Confusión – Algoritmo CART

- Ejemplo de corrida de Random Forest con ntree=50 y nodesize=50 (ntree: cantidad de árboles a modelar para armar el modelo final; nodesize: tamaño mínimo de la hoja).

rf.modelo.test	alto	bajo
alto	13420	15703
bajo	1786	2050

Tabla 18. Matriz de Confusión – Algoritmo Random Forest

En ambos casos el valor del AUC es cercano al de la clasificación al azar (0.5), por ende se puede concluir que los modelos no son buenos.

En primera instancia, al observar los modelos resultantes y las variables que se determinan más importantes, se intuye que el inconveniente puede subsanarse armando modelos diferentes por franjas etarias. Teniendo presente lo mencionado en el punto 3.2.7 respecto de los edades de los pacientes, se ejecutan 4 modelos según las bandas etarias allí detalladas (menores de 16 años, entre 16 y 40 años, entre 41 y 65 años y mayores a 65 años). Los resultados no muestran mejorías respecto del modelo general.

4.3.2 Feature Engineering (Ingeniería de características)

El aprendizaje automático ajusta los modelos matemáticos a los datos para descubrir información o hacer predicciones. Estos modelos toman características (variables) como materia prima para crear los modelos. Una característica es una representación de un aspecto de los

datos crudos. La ingeniería de características es el acto de extraer características de los datos sin procesar y transformarlos de forma tal que ayuden a los algoritmos en la predicción.

A raíz de los magros resultados obtenidos hasta el momento se procede a hacer feature engineering con las variables existentes.

a) Variables etarias
   - Bandarización de edades: se crean 9 variables agrupando pacientes por bandas de 10 años (0 a 10, 11 a 20, 21 a 30, 31 a 40, 41 a 50, 51 a 60, 61 a 70, 71 a 80 y más de 80 años)
   - Menores de edad: se crean tres variables que identifican a los menores de 5 años, a los menores de 10 años y a los menores de 18 años. Estas variables se crean dado que durante el período previo al inicio de clases (meses de Enero a Marzo) hay una gran afluencia de pacientes pediátricos que concurren solo para obtener un apto físico
   - Mayores de edad: se introducen dos variables que marquen aquellos pacientes que son mayores de 40 y de 60 años. La mayor cantidad de pacientes que se atienden durante los meses de Abril y Diciembre tienen más de 40 años
b) Variables de fecha
   - Verano: se crea una variable binaria que indica si un paciente fue atendido durante los meses de Enero a Abril, meses en los cuales la concurrencia de pacientes pediátricos es mayor. Por ser consultas en busca de un apto físico, se presumen consultas de corta duración
   - Mes: se extrae el mes de la fecha de atención
   - Día: se crea una variable que marca qué día de la semana concurrió el paciente. Además, durante el análisis se identifican dos grupos de días: Lunes, Martes y Jueves por un lado, Miércoles, Viernes y Sábado por el otro. Se arman variables binarias para estos grupos
c) Variables horarias
   - Hora de atención: se crea una variable que indica la hora de atención normalizada (hora / 24)
   - Franja horaria: se generan 3 bandas de atención según el horario (mañana, tarde y noche)
   - Franjas específicas: se detectan particularidades en los tiempos de atención en 2 bandas horarias (10 a 16hs y 18 a 22hs). Se crean variables binarias que indican si el paciente se atendió en esas bandas
d) Variables mixtas
   - Edad y Meses: al haberse detectado gran cantidad de pacientes pediátricos que se atienden en verano, se decide crear 4 variables binarias que indiquen si el paciente es menor / mayor de edad y fue atendido en verano / invierno
   - Mes, Hora y Día: se crean 10 variables binarias que determinan si un paciente fue atendido en verano, en alguna de las franjas horarias específicas mencionadas en el apartado c y en algún día de los detectados en el punto b

e) Variables relacionadas con toma de medicamentos
   En base al análisis realizado en la relación de la cantidad de medicamentos consumidos y el tiempo de consulta se generan las siguientes variables:
   - Pacientes sin medicaciones: se genera una variable binaria que indica si un paciente no toma ninguna medicación
   - Pacientes que toman hasta dos medicamentos: se genera una variable binaria que indica si un paciente toma hasta 2 medicamentos
   - Franjas de medicamentos: se genera una variable que discretiza la cantidad de medicamentos consumidos en 3 niveles (bajo, medio, alto). Esto es útil para inferir el estado de salud de los pacientes
   - Clustering de Medicamentos: mediante análisis de clustering de variables se agruparon aquellos tipos de medicamentos con características similares
f) Variables relacionadas con la atención
   - Orden de atención: se indica el orden de atención para cada paciente (número de paciente atendido por un determinado profesional en un día determinado)
   - Total de pacientes: ratio del orden de atención y total de pacientes atendidos por el profesional en esa fecha
   - Orden de atención y Hora: resultado de multiplicar el orden de atención y la hora estandarizada (punto c)
   - Día, Orden y Hora: resultado de multiplicar los Pacientes atendidos los días Lunes, en franja de 18 a 22hs y el orden de atención. Durante el análisis se detectaron particularidades en esa franja horaria de ese día. Se crea esta variable para marcarla
   - Turno anterior: se generan 3 variables indicando los tiempos de atención (si existiesen) de las 3 consultas anteriores
   - Tendencia: se crean 3 variables indicando la diferencia de duración cualitativa de la n-1, n-2 y n-3 consultas anteriores (incremento, decremento, igualdad)
g) Variables Climáticas
   Se obtienen datos climatológicos del Servicio Meteorológico Nacional referidos a la temperatura máxima, mínima, cantidad de lluvia caída, cantidad de días con lluvia por mes y año, humedad máxima y mínima. Se agregan estas variables al dataset
h) Otras variables
   - Pacientes sin obra social: se crea una variable binaria que marca aquellos pacientes atendidos siempre de forma particular.
   - Pacientes con Prepagas: se genera una variable binaria que indica la pertenencia de los pacientes a determinadas prepagas. Mediante clustering de variables se detectaron grupos que fueron aglutinados en el dataset mediante esta variable.
   - Clustering de Estudios Realizados: mediante análisis de clustering de variables se agruparon aquellos estudios con características similares.
   - Clustering de Antecedentes Médicos: mediante análisis de clustering de variables se agruparon aquellos antecedentes con características similares.
   - Pacientes con presumible mala salud: se calcula un índice indicador de presumible mala salud. Resulta de multiplicar la cantidad de medicamentos que toma un paciente, la cantidad de consultas realizadas en la clínica y la cantidad de estudios realizados.

Con estas variables se ejecutan nuevamente los modelos según lo detallado anteriormente en los puntos 4.1 y 4.3.1.

4.3.3 Nuevos Modelos y Resultados

Los modelos se ejecutan con los algoritmos RPART y Random Forest. En una primera etapa se buscan los mejores hiperparámetros y luego se ejecutan corridas de validación cruzada con 10 cortes (10-folds Cross Validation) para corroborar los resultados.

El mejor modelo con RPART se obtiene con los siguientes parámetros:

- Minsplit = 50
- CP = 0.00005
- Maxdepth = 20
- Matriz de Costo = $\begin{bmatrix} 0 & 2 \\ 4 & 0 \end{bmatrix}$

El AUC obtenido con este modelo es de 0. 7199652

La matriz obtenida es la siguiente:

modelo.test.c	Alta	Baja
Alta	7209	4499
Baja	1226	2438

Tabla 19. Matriz de Confusión – Algoritmo CART

La curva ROC obtenida se muestra a continuación:

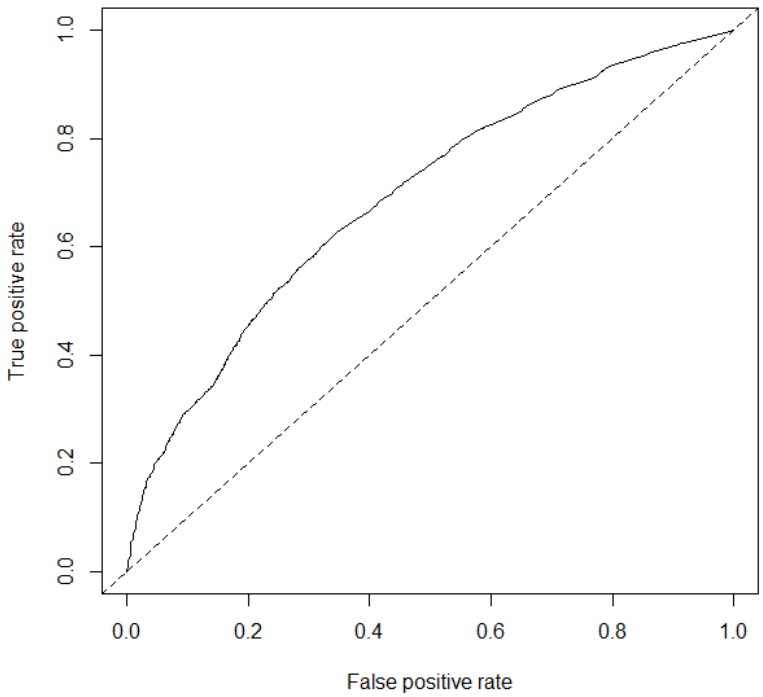


Figura 28. Curva ROC algoritmo CART

El modelo muestra un buen rendimiento, a pesar de haber gran cantidad de registros (4.499) predichos como de ALTA duración que finalmente fueron BAJA. Aun así, se cumplió la premisa impuesta en la matriz de costo y son mayores que los falsos negativos.

Las variables más importantes a la hora de determinar la duración de la atención según el modelo objetivo son las detalladas a continuación:

	modelo.train.variable.importance
Hora_atencion_estandarizada	885.51400828
ordenXhoraestand	701.07980531
Orden.de.Atencion	633.51631642
Paci_Total	610.02577480
menor_verano	382.93142201
T_max	210.44367859
T_min	207.32801838
Mes	179.65675641
Estud...Edad	178.71308329
Dia_Lunes_1	170.76706836
EDAD_EN_LA_ATENCION	167.37867584
CTC	161.24097107
CPC	152.64059748
lluv_mm	147.87655727
lluv_dias	121.18157649
CCA	108.15012491
CCS	102.80493645
hum_min	102.67693413
Lu_1822_Orden	97.42488487
CANT_PARCIAL_CONSULTAS	92.52112401
CANT_CONS_ANIO	89.84965295
CANT_TOTAL_CONSULTAS	85.48740145
Estudio	78.28961443
verano	74.28171950
CCM	70.96978837
CANT_CONS_SEMESTRE	68.82627705
clus_dummy	68.54749411
mayor_invierno	67.19600333
CANT_CONS_MES	65.97683167
Lunes	65.67122192
LuMaJu	62.81505136
sum_dummy	55.11272624
mayor_verano	54.84175299
Martes	53.53189755
malasalud	53.42568867
atc3	52.10285710
REFRACCIONDUMMY	50.68778807
tiempo.1	47.98363785
Jueves	47.15012523
ABC_SUM...Estud	45.55354256

Tabla 20. Variables más importantes según algoritmo CART (RPart)

Esta salida muestra cuán importante es cada variable a la hora de generar cada partición (nueva rama) del árbol. Esta es una medida general de importancia de las variables: la suma de la bondad de las medidas de división para cada división para la que fue la variable principal, más la bondad (acuerdo ajustado) para todas las divisiones en las que fue un sustituto.

Detalle de las 20 primeras principales variables:

1. Hora_atención_estandarizada: se toma solamente la hora de atención (sin minutos) y se estandariza a valores entre 0 y 1 (HH / 24)
2. OrdenXhoraestand: resultante de multiplicar el orden de atención de cada paciente respecto del profesional que lo atiende y la hora estandarizada
3. Orden.de.Atencion: ubicación en la cola de atención del profesional
4. Paci_Total: posición del paciente sobre el total de pacientes atendidos por el médico en el día
5. Menor_verano: flag que indica si se trata de un paciente pediátrico atendido durante los meses de Enero, Febrero, Marzo o Abril
6. T_max: temperatura máxima media registrada por el Servicio Meteorológico Nacional durante el mes de atención
7. T_min: temperatura mínima media registrada por el Servicio Meteorológico Nacional durante el mes de atención
8. Mes: mes de atención
9. Estud…Edad: ratio de la cantidad total de estudios y la edad del paciente
10. Dia_Lunes_1: Día de la semana, siendo Lunes=1
11. EDAD_EN_LA_ATENCION: edad del paciente al momento de ser atendido
12. CTC: ratio de cantidad total de consultas y edad en la atención
13. CPC: ratio de cantidad parcial de consultas del paciente (hasta el momento de atención) y la edad
14. Lluv_mm: milímetros de lluvia caída durante el mes de atención
15. Lluv_dias: cantidad de días con lluvia durante el mes de atención
16. CCA: ratio de la cantidad de consultas en el año (últimos 12 meses) y la edad
17. CCS: ratio de la cantidad de consultas en el año (últimos 6 meses) y la edad
18. Hum_min: humedad mínima promedio para el mes de atención
19. Lu_1822_Orden: resultante de multiplicar Lunes (flag que indica si el día de atención es Lunes), Franja_1822 (flag que indica se la hora de atención es entre las 18hs y las 22hs) y la edad del paciente al momento de la atención
20. Cant_Parcial_Consultas: Sumatoria de la cantidad de consultas realizadas por el paciente hasta el día de atención (se considera una consulta cada ocasión que el paciente es atendido en el consultorio)

Para verificar el resultado, se ejecuta una corrida de 10-Fold Cross Validation, con los siguientes resultados:

I. 0.712264
II. 0.7077359
III. 0.7021634
IV. 0.7080522
V. 0.6985422
VI. 0.7152381
VII. 0.7069064
VIII. 0.7011781
IX. 0.7163363
X. 0.7182822

Los resultados se muestran consistentes con el AUC obtenido por el modelo.

El árbol queda definido tal como se muestra en la figura 29 (por una cuestión de espacio se muestra una versión podada del mismo).

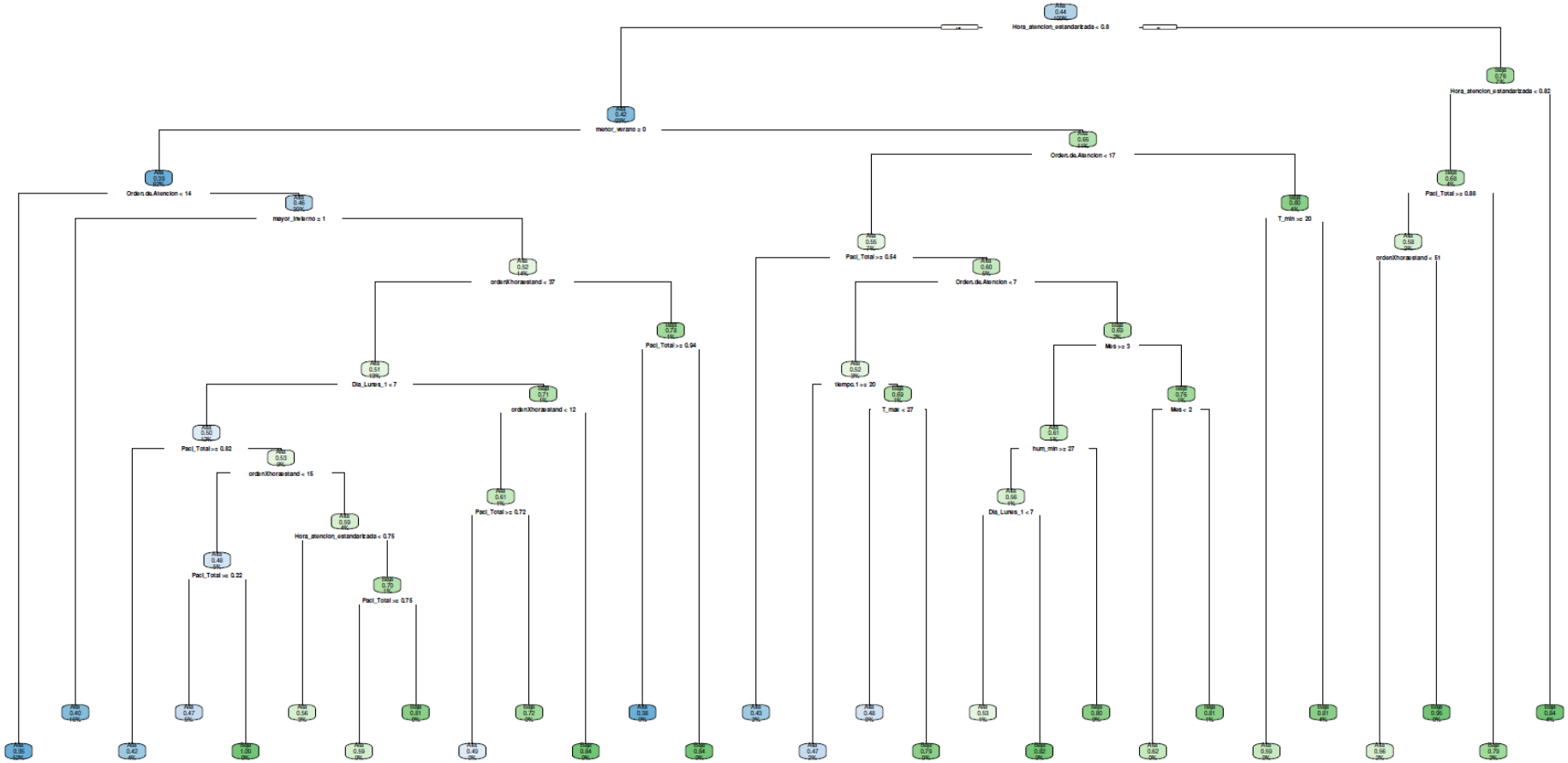


Figura 29. Porción superior del árbol Rpart

El mejor modelo con Random Forest se obtiene con los siguientes parámetros:

- ntree = 55
- nodesize = 50
- mtry = 14

El AUC obtenido con este modelo es de 0.7343583.

La matriz obtenida es la siguiente:

rf.modelo.test	Alta	Baja
Alta	5177	2399
Baja	3201	4490

Tabla 21. Matriz de Confusión – Algoritmo Random Forest

En este caso, a diferencia de lo obtenido con Rpart, la cantidad de falsos negativos es mayor que los falsos positivos. Además la cantidad de registros ALTA bien predichos es menor y la de registros BAJA es mayor. Aun así el rendimiento del modelo es bueno.

La curva ROC obtenida se muestra a continuación:

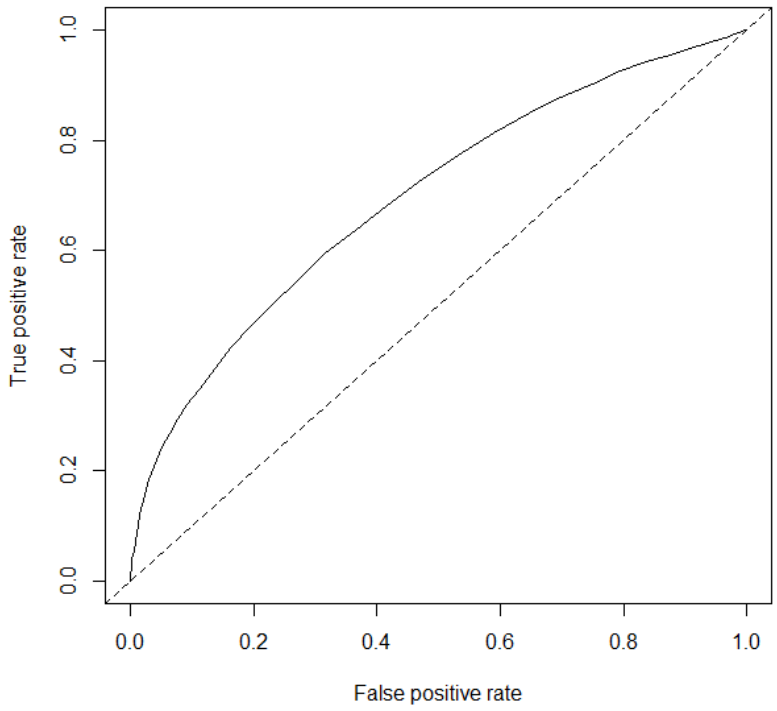


Figura 30. Curva ROC algoritmo Random Forest

Las variables más importantes a la hora de determinar la duración de la atención según el modelo objetivo son las detalladas a continuación:

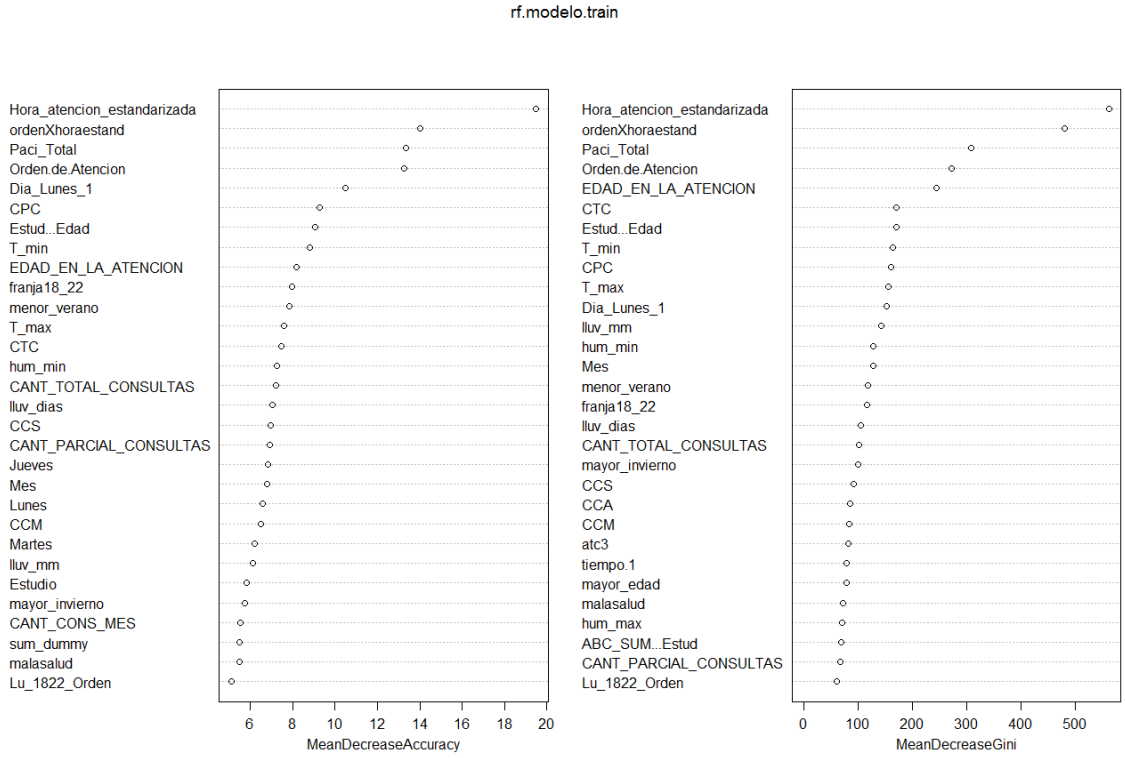


Tabla 22. Variables más importantes según algoritmo Random Forest

Dos medidas de importancia se muestran en la tabla 22. Accuracy (exactitud) mide cuanto peor es la performance del modelo sin cada variable, por lo que un gran descenso en la exactitud es esperable para variables muy predictivas. En cuanto a Gini tiene como base los cálculos matemáticos existentes detrás de la creación de los árboles. Básicamente mide cuán puros son los nodos al final de cada árbol. Igual que en el caso anterior, calcula cómo se ven afectado los modelos si la variable es eliminada del modelo. Mientras más alto el valor, más importante es la variable. Por ejemplo, si se saca del modelo la variable Hora_atencion_estandarizada la impureza de los nodos decrecerá en promedio 500 puntos y la exactitud descenderá en promedio 20 unidades.

Las variables de mayor importancia para Accuracy y para Gini son (no se detallan aquellas ya descriptas para el algoritmo RPART):

1. Franja18_22: variable binaria que indica si un paciente se atendió entre las 18hs y las 22hs
2. CANT_TOTAL_CONSULTAS: cantidad total de consultas realizadas por el paciente
3. Jueves: Flag que indica si el paciente se atendió un día jueves
4. Lunes: Flag que indica si el paciente se atendió un día lunes
5. CCM: ratio de la cantidad de consultas en el mes y la edad
6. Mayor_invierno: Flag que indica si el paciente es mayor (>=18 años) y fue atendido en invierno

Para verificar el resultado, se ejecuta una corrida de 10-Fold Cross Validation, con los siguientes resultados:

I. 0.7236555
II. 0.7131933
III. 0.7133002
IV. 0.7207447
V. 0.7063564
VI. 0.7192589
VII. 0.7202181
VIII. 0.7156649
IX. 0.7246471
X. 0.7304501

Los resultados se muestran consistentes con el AUC obtenido por el modelo.

4.4 Análisis de los Resultados

Ambos algoritmos muestran desempeños similares. A pesar de haber tenido un peor rendimiento inicialmente, el modelo con Random Forest muestra mejores resultados en las corridas de 10-Fold Cross Validation, aunque apenas superiores a las realizadas con Rpart.

La gran mayoría de las variables que resultaron relevantes para un algoritmo lo fueron para el otro. Las pocas que registraron cambios fueron Franja18_22, CANT_TOTAL_CONSULTAS, Jueves y Lunes (aparecen en Random Forest pero no en Rpart).

En la figura 31 se muestra una comparación de las curvas ROC expuestas anteriormente.

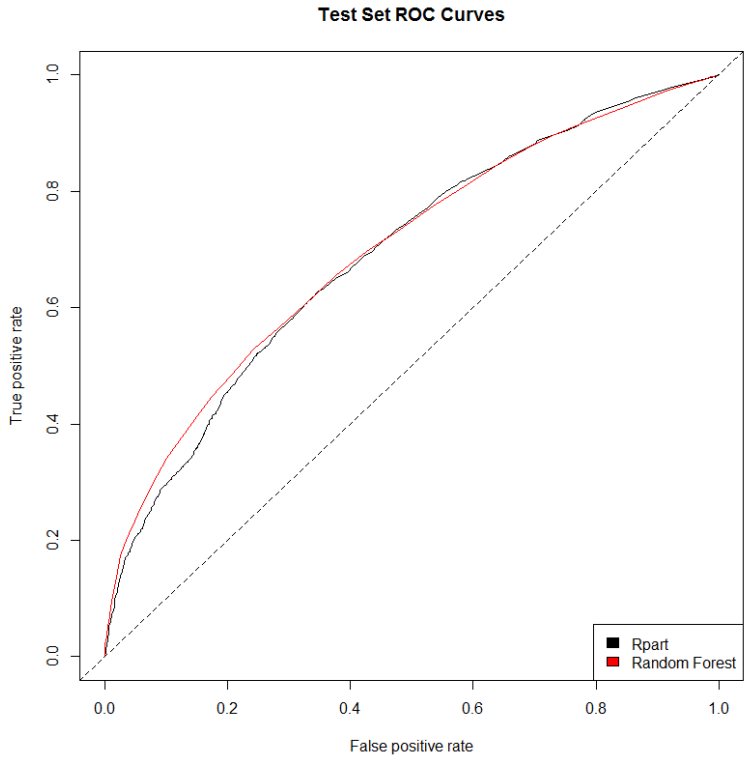


Figura 31. Comparación Curvas ROC

A excepción de unos pequeños tramos donde las curvas se solapan (principalmente en el centro del gráfico), el modelo obtenido con Rpart se muestra superior a lo largo del tramo final de la curva mientras que el modelo Random Forest supera a Rpart en el inicio. Considerando las similitudes de rendimiento de ambos modelos y el mejor resultado obtenido en falsos negativos (lo que redunda en reducción de tiempos de espera) y en registros ALTA, podemos afirmar que el realizado con Rpart es el mejor modelo obtenido.

Respecto de las variables, aquellas que muestran mayor peso a la hora de determinar la duración son las relacionadas con la hora y el orden de atención: la hora en la que fue atendido el paciente, el orden en la cola de atención del médico a consultar, la cantidad total de pacientes que el médico atiende ese día determinan cuál será el tiempo de la próxima consulta. En un segundo nivel se aprecian variables estacionales (si es un paciente menor de edad que se atiende en verano o, dicho de otra manera, paciente pediátrico atendido entre Enero y Abril) y variables climáticas, estacionales o que indican la cantidad de consultas previas del paciente. Estas últimas, que indican la cantidad de visitas que han realizado previamente los pacientes independientemente del plazo que se tome en cuenta, fueron también indicadas por varios autores como relevantes.

Una variable quizás indicadora del estado de salud del paciente, Estud…Edad, resultante de dividir la cantidad de estudios por la edad del paciente, se muestra muy determinante para el modelo. Esto es interesante dado que no existen variables que manifiesten claramente el estado de salud previo del paciente (una especie de "triage"** en la base de datos).

Un dato que surge del análisis, y no mencionan los autores descriptos durante el desarrollo de este trabajo, es que el día de la semana, el mes del año y las condiciones climáticas influyen también en el tiempo de atención.

Algunos puntos interesantes:

- Los pacientes pediátricos atendidos en verano tienden a tener consultas cortas.
- Los pacientes adultos atendidos en verano suelen tener consultas más largas.
- Mientras más tarde es atendido un paciente, más corta es la consulta*.
- La cantidad total de pacientes atendidos por el profesional también influye: a mayor cantidad de pacientes menor tiempo de atención.
- Los pacientes que son atendidos más temprano y en primer orden tienden a tener tiempos de consultas más altos.
- A mayor cantidad total de consultas, mayor cantidad de estudios. Si además no se poseen datos de la historia clínica del paciente, el tiempo de atención es alto. Si se cuenta con datos previos de la historia del paciente, el tiempo de atención es bajo. Esto es coincidente con lo detallado en el punto 3.2.4.
- El sexo del paciente no muestra ser determinante. Si bien en algunas ramas aparece, lo hace muy abajo en el árbol. Contrariamente a lo especificado en el punto 3.2.3, el sexo del paciente no parece definir la duración de la atención.
- Muchos estudios son determinantes a la hora de indicar la duración del turno. Cómo era de esperar, ciertos estudios más complejos demoran más la atención que aquellos más simples o rutinarios (por ejemplo, toma de presión).
- Las medicaciones que toma el paciente no muestran ser importantes. Contrariamente a lo indicado por otros autores, ninguna de las variables relacionadas con la toma de medicamentos aparece en la lista de aquellas variables más importantes.

*(Entiéndase consulta por visita al médico)

** El triaje o protocolo de intervención es un método de selección y clasificación de pacientes empleado en la medicina de emergencias y desastres. Evalúa las prioridades de atención, privilegiando la posibilidad de supervivencia, de acuerdo con las necesidades terapéuticas y los recursos disponibles

Capítulo 5 – Conclusiones

A lo largo de este trabajo se han planteado, en base a la revisión bibliográfica realizada, distintos supuestos sobre variables que permitirían la identificación de aquellos pacientes que demoran más en la atención cara a cara con el médico. También se ha visto que dicha bibliografía no es abundante y que en muchos casos es contradictoria, planteando en un caso ciertas variables como relevantes mientras que otros autores indican lo contrario.

En este documento se han analizado muchas de esas variables mencionadas en la bibliografía como otras tantas más específicas de la clínica bajo estudio. No fue posible encontrar un modelo lo suficientemente bueno tan solo con características de los pacientes. Fue necesario crear variables con datos internos y externos para lograr un buen rendimiento.

Según los resultados obtenidos, las variables que indican la hora de atención, el orden de atención, la cantidad total de pacientes que un médico ve por día son las variables más importantes de los modelos obtenidos.

La creación de variables alternativas (feature engineering) resulta fundamental. Ciertas variables que no fueron tenidas en cuenta al momento del desarrollo o implementación del sistema en la clínica, tal vez por no ser necesarias para el negocio, resultan críticas a la hora de modelar. No solo aquellas resultantes de algún cálculo matemático sino también las que funcionan como un indicador (variables binarias o flags) que determinan, por ejemplo, si un paciente consume o no medicación, si es mayor de edad o si se ha realizado estudios previamente en la clínica con anterioridad.

Se detectó que la fidelización de los pacientes es muy baja y los tiempos de espera muy altos. Ambos temas pueden impactar directamente en esta clínica en particular ante la potencial apertura de competidores en cierto rubro (cómo pediatría). El riesgo reputacional y su correlación con los tiempos de atención o de espera (que si bien quedó fuera del alcance de este trabajo tiene relación directa con la asignación de turnos y los tiempos de atención) va de la mano del “paciente cautivo”, es decir, de la ausencia de otras opciones donde atenderse.

La estacionalidad de la cantidad de admisiones es un punto a tener presente. Se han observado claros picos estacionales que se reiteran a lo largo de los años. Si bien los profesionales médicos no fueron analizados por quedar fuera del alcance, estos picos deberán ser considerados para la mejor administración de los recursos profesionales, lo que, indirectamente, impactan en los tiempos de atención. A esto debe sumarse la edad de los pacientes que concurren en determinada época del año. En verano hay mayor cantidad de pacientes pediátricos con turnos cortos y el resto del año hay mayor cantidad de pacientes adultos con turnos cortos.

También resultaron importantes las franjas en los horarios de atención, los días de la semana y los meses. Se detectaron horarios, días y horarios en días específicos en los cuales el tiempo de atención es más prolongado. Sin embargo, no se puede aseverar que esto se repita de igual manera en otras clínicas, sino que debiera ser tenido en cuenta y estudiado.

De igual manera, los primeros y los últimos turnos del día y de cada médico marcaron diferencias. Los matinales y/o primeros turnos resultaron ser más largos que los últimos turnos y/o nocturnos.

Si bien externo a los pacientes, el factor climático influye en la duración de los turnos. La temperatura, la humedad y la lluvia fueron variables que figuraron entre las más importantes en los modelos obtenidos. Es lógico pensar que en época estival las altas temperaturas de

Buenos Aires inhiban al público de concurrir al consultorio y quienes lo hagan pasen más tiempo con el médico al no haber tanta cola de pacientes.

En cuanto al sexo del paciente y los antecedentes clínicos son dos de las variables marcadas por otros autores como importantes pero que no resultaron relevantes en este trabajo. Por el contrario, muchas variables referidas a consultas anteriores en la clínica (cantidad de visitas en el semestre o en los últimos 12 meses, cantidad parcial y total de consultas, entre otras) sí fueron de importancia.

No estuvieron entre las principales variables aquellas que pueden servir de indicador del estado de salud. Solamente el ratio entre cantidad de estudios realizados y la edad del paciente apareció en la lista.

Ha quedado evidenciado que es fundamental contar con buenos datos. Los sistemas debieran pensarse no solo para la atención sino para la explotación de los datos en pos de lograr mejoras en todos los procesos. Esta también forma parte del negocio.

Capítulo 6 – Trabajos a futuro

Cómo corolario del trabajo y de las conclusiones del capítulo 5 se desprenden los trabajos a futuro.

En un tiempo próximo será interesante poder trabajar con instituciones que cuenten con historias clínicas electrónicas que garanticen en mayor medida la calidad de los datos, que permitan comparar el comportamiento de los pacientes en distintas instituciones y que cuenten con mayor cantidad de datos de los pacientes (por ejemplo en lo referido a datos que pueden resultar sensibles para el paciente, cómo la medicación, enfermedades previas, historial médico familiar, etc).

Tampoco se han tenido en cuenta los datos de los médicos: si bien el trabajo se centra en los pacientes, quedó evidenciado que la duración de los turnos varía de paciente en paciente según el profesional que lo atienda. Sería importante considerar las variables que hagan referencias a los médicos en futuros trabajos.

Será importante también poder realizar este tipo de análisis en distintos tipos de instituciones, públicas y privadas, para poder evaluar las potenciales diferencias entre las mismas y los pacientes que acuden a cada una de ellas.

Se considera menester poder realizar estudios similares al presente pero segregado por tipo de estudio. Dado que un mismo paciente puede realizarse varios estudios en el mismo día el tiempo de atención puede verse alterado. Por ende, la correcta identificación de los estudios que se realice cada paciente y el tiempo que demora en cada uno podrán mejorar los futuros modelos.

Es menester profundizar en el análisis pormenorizado de las obras sociales, prepagas y pacientes particulares. Se ha detectado en este trabajo que la cobertura que posee cada paciente incide en el tiempo de atención, por lo que deberá tenerse en cuenta para nuevos trabajos.

Se mencionó en el documento que el ausentismo no fue tenido en cuenta a la hora de generar los modelos. Dado que es un punto relevante para las empresas prestadoras de salud y los profesionales que atienden en forma particular, resultará conveniente realizar el mismo análisis aquí desarrollado tomando este dato en cuenta. Incluso se podría desarrollar un modelo que permita detectar estas situaciones. Si se redujeran los tiempos de espera, la calidad de la atención probablemente mejoraría y los turnos podrían ser otorgados en plazos más cortos y así reducir las tasas de ausentismo.

Otro dato que no fue tenido en cuenta por la metodología de cálculo adoptada a la hora de determinar los tiempos de atención fue el de las familias que ingresan juntas a ser atendidas (por ejemplo, una madre que lleva a atender a sus hijos). Podría generarse una variable que defina estas “atenciones grupales” y considerarlas en la duración de la atención.

Un interesante objetivo a futuro será realizar mining (minería) de imágenes: mediante determinados algoritmos poder generar datos que puedan ser ingresados al dataset para ser evaluados con las restantes variables. Algo similar a lo que se realizó en los campos Observaciones de las tablas (ver punto 3.1.3) pero aplicado a imágenes [44].

En un grado más avanzado de análisis sería ideal trabajar con variables objetivo numéricas que permitan predecir el tiempo que el paciente permanecerá en el consultorio. De esa manera se podrá pasar de la asignación de turnos estáticos a turnos totalmente dinámicos en base a las características del paciente. De esta manera el aprovechamiento del tiempo sería óptimo.

Independientemente de la duración de los turnos que se asignen hoy día en cada clínica, del análisis efectuado se desprende gran cantidad de información que indica que se podrían armar turnos de duración variable. Sin embargo, surgen muchos interrogantes. ¿Es factible implementar una determinada cantidad de turnos de distinta duración? ¿Cuántos? ¿Cómo se procede a organizar la agenda? ¿Cómo deberían trabajar los profesionales? ¿Se debería atender un día a aquellos pacientes de bajo tiempo de atención y otro día a los de largo tiempo de atención? ¿Cómo se mantiene la lógica administrativa? Todas estas preguntas escapan al alcance de este trabajo. Si bien mediante lo expuesto se ha achicado el abismo y la situación mejorará, queda mucho aún por resolver.

Apéndice

Librerías de R

A continuación se detallan las librerías de R usadas durante el presente trabajo:

- Caret
- Dplyr
- Formattable
- Ggplot2
- party
- randomForest
- rattle
- RColorBrewer
- ROCR
- rpart
- rpart.plot
- Scales
- Sp
- SQLDF
- Stringr
- Tree

Listado de variables utilizadas

Se detallan en la siguiente lista las variables utilizadas en el análisis. No se muestran aquellas que pueden vulnerar la confidencialidad de los pacientes.

Campo	Descripción
ID	Código de paciente (uso solo para este trabajo)
COD_POST	Código postal del domicilio del paciente
PROVINCIA	Provincia del domicilio del paciente
SEXO	Sexo del paciente
TIPO_DEUDOR	Particular, Obra Social o Prepaga
COD_DEUDOR	Denominación de la Obra Social o Prepaga
COD_PLAN	Plan de la Obra Social o Prepaga
PRIMERAVEZ	Binario que indica primera vez que se atiende
TIEMPO_DE_ATENCION	Tiempo de atención en consultorio
CANT_CONS_MES	Cantidad de consultas del paciente en el mes
CANT_CONS_SEMESTRE	Cantidad de consultas del paciente en el semestre
CANT_CONS_ANIO	Cantidad de consultas del paciente en el año
CANT_PARCIAL_CONSULTAS	Cantidad de consultas del paciente hasta esa admisión
CANT_TOTAL_CONSULTAS	Cantidad de consultas del paciente totales en su historial
EDAD_EN_LA_ATENCION	Edad del paciente al momento de ser atendido
Franja	Franja horaria en que fue atendido (mañana, tarde, noche)
Hora_atención_estandarizada	Resultado de dividir la hora de atención del paciente por 24
Franja	Franja horaria de atención del paciente (mañana, tarde o noche)
Orden.de.Atencion	Orden relativo del paciente en la cola de atención del médico
Paci_Total	Total de pacientes atendidos por el médico ese d
Mes	Mes en el que se atendió el paciente
Dia_Lunes_1	Día en el que se atendió el paciente (Lunes = 1)
Orden_de_atención	Número de paciente atendido por el médico designado
Hora Normalizada	Hora en la que fue atendido
DIA_lunes_1	Día de la semana
VITREODUMMY	Dummy que indica si la afección es en el VITREO
CONJUNTIVADUMMY	Dummy que indica si la afección es en la CONJUNTIVA
CORNEADUMMY	Dummy que indica si la afección es en la CORNEA
PUPILASDUMMY	Dummy que indica si la afección es en la PUPILA
PAPILASDUMMY	Dummy que indica si la afección es en la PAPILA
GLOBOOCULARDUMMY	Dummy que indica si la afección es en el GLOBO OCULAR
IRISDUMMY	Dummy que indica si la afección es en el IRIS
LAGRIMALDUMMY	Dummy que indica si la afección es en el LAGRIMAL
PRESIONOCULARDUMMY	Dummy que indica si la afección es en la PRESION OCULAR
CISTALINODUMMY	Dummy que indica si la afección es en el CISTALINO
CAMARAANTERIORDUMMY	Dummy que indica si la afección es en la CAMARA ANTERIOR
PARPADOSDUMMY	Dummy que indica si la afección es en el PARPADO
ORBITADUMMY	Dummy que indica si la afección es en la ORBITA
ESCLERADUMMY	Dummy que indica si la afección es en la ESCLERA
RETINADUMMY	Dummy que indica si la afección es en la RETINA
Clus_dummy	Clustering de las variables dummy relativas a afecciones

ant.diabetes_p	Binario que indica antecedentes de diabetes
ant.hipertension_p	Binario que indica antecedentes de hipertensión
ant.alergias_p	Binario que indica antecedentes de alergias
ant.glaucoma_p	Binario que indica antecedentes de glaucoma
ant.desprendimientos_p	Binario que indica antecedentes de desprendimientos
ant.estrabismo_p	Binario que indica antecedentes de estrabismo
ant.traumas_p	Binario que indica antecedentes de traumas
ant.refraccion_p	Binario que indica antecedentes de refracción
ant.diabetes_f	Binario que indica antecedentes de diabetes
ant.hipertension_f	Binario que indica antecedentes de hipertensión
ant.alergias_f	Binario que indica antecedentes de alergias
ant.glaucoma_f	Binario que indica antecedentes de glaucoma
ant.desprendimientos_f	Binario que indica antecedentes de desprendimientos
ant.estrabismo_f	Binario que indica antecedentes de estrabismo
ant.refraccion_f	Binario que indica antecedentes de refracción
ant.gpo_sanguineo	Grupo sanguíneo del paciente
ant.fact_sanguineo	Factor sanguíneo del paciente
Clus_ant	Clustering de las variables relativas a antecedentes
mco.tiempo_dolencia	Indicador del tiempo de dolencia
mco.paciente_sano	Binario que marca si el paciente es sano
mco.dificultad_ver_lejos	Indicador si el paciente tiene dificultad de ver de lejos
mco.dificultad_ver_cerca	Indicador si el paciente tiene dificultad de ver de cerca
mco.dificultad_dist_media	Indicador si el paciente tiene dificultad de ver dist. media
mco.dolor	Indicador si el motivo de consulta es por dolor
mco.pinchazo	Indicador si el motivo de consulta es por pinchazo
mco.lloran	Indicador si el motivo de consulta es por lloran
mco.ardor	Indicador si el motivo de consulta es por ardor
mco.picazon	Indicador si el motivo de consulta es por picazón
mco.cuerpo_extraneo	Indicador si el motivo de consulta es por cuerpo extraño
mco.molesta_luz	Indicador si el motivo de consulta es por molesta luz
mco.dolor_de_cabeza	Indicador si el motivo de consulta es por dolor de cabeza
mco.desvia_ojo	Indicador si el motivo de consulta es por desvía ojo
mco.ve_moscas	Indicador si el motivo de consulta es por ver moscas
mco.perdio_vision_mom	Indicador si el motivo de consulta es por perdió visión momentáneamente
mco.recibe_medicacion	Indicador si el motivo de consulta es por recibe medicación
mco.tiene_distorsion	Indicador si el motivo de consulta es por tiene distorsión
mco.ve_halos	Indicador si el motivo de consulta es por ver halos
mco.vision_doble	Indicador si el motivo de consulta es por visión doble
mco.ojos_secos	Indicador si el motivo de consulta es por ojos secos
REFRACCIONDUMMY	Dummy que indica si ha sido estudiado la refracción
TLO_CODIGO_TRAT_LOCAL	Código de tratamiento otorgado
TLO_DROGA	Código de droga prescripta
TLO_CODIGO_TRAT_LOCAL1	Código de tratamiento otorgado 2
TLO_DROGA1	Código de droga prescripta 2
TLO_CODIGO_TRAT_LOCAL2	Código de tratamiento otorgado 3

TLO_DROGA2	Código de droga prescripta 3
TRATLOCALDUMMY	Dummy de tratamiento local otorgado
TAB_QUE_COD_PRACT1	Código de práctica ejecutada 1
TAB_QUE_COD_PRACT2	Código de práctica ejecutada 2
TAB_QUE_COD_PRACT3	Código de práctica ejecutada 3
TAB_QUE_COD_PRACT4	Código de práctica ejecutada 4
TAB_QUE_COD_PRACT5	Código de práctica ejecutada 5
TAB_QUE_COD_PRACT6	Código de práctica ejecutada 6
TAB_QUE_COD_PRACT7	Código de práctica ejecutada 7
TAB_QUE_COD_PRACT8	Código de práctica ejecutada 8
TAB_QUE_COD_PRACT9	Código de práctica ejecutada 9
TAB_QUE_COD_PRACT10	Código de práctica ejecutada 10
TAB_QUE_COD_PRACT11	Código de práctica ejecutada 11
TAB_QUE_COD_PRACT12	Código de práctica ejecutada 12
TAB_QUE_COD_PRACT13	Código de práctica ejecutada 13
TAB_QUE_COD_PRACT14	Código de práctica ejecutada 14
TAB_QUE_COD_PRACT15	Código de práctica ejecutada 15
TABLAQUEDUMMY	Dummy de práctica ejecutada
CAMPOVISUALDUMMY	Dummy que indica si ha sido estudiado el campo visual
A	Binario que indica si toma medicación clase A según ATC
B	Binario que indica si toma medicación clase B según ATC
C	Binario que indica si toma medicación clase C según ATC
D	Binario que indica si toma medicación clase D según ATC
G	Binario que indica si toma medicación clase G según ATC
H	Binario que indica si toma medicación clase H según ATC
J	Binario que indica si toma medicación clase J según ATC
L	Binario que indica si toma medicación clase L según ATC
M	Binario que indica si toma medicación clase M según ATC
N	Binario que indica si toma medicación clase N según ATC
P	Binario que indica si toma medicación clase P según ATC
R	Binario que indica si toma medicación clase R según ATC
S	Binario que indica si toma medicación clase S según ATC
V	Binario que indica si toma medicación clase V según ATC
Clus_ABC	Clustering de las variables relativas a la toma de medicación
ABCDGHJLMNPRSV_DUMMY	Dummy que indica si toma medicación
atc1	Clasificación ATC del medicamento que toma 1
atc2	Clasificación ATC del medicamento que toma 2
atc3	Clasificación ATC del medicamento que toma 3
atc4	Clasificación ATC del medicamento que toma 4
atc5	Clasificación ATC del medicamento que toma 5
atc6	Clasificación ATC del medicamento que toma 6
atc7	Clasificación ATC del medicamento que toma 7
atc8	Clasificación ATC del medicamento que toma 8
atc9	Clasificación ATC del medicamento que toma 9
Clus_ATC	Clustering de las variables referentes a la clasificación ATC
diagn_ojo	Ojo izquierdo, derecho o ambos

diagn_codigo	Código de diagnóstico
diagn_obs	Notas del diagnóstico
diagn_ojo1	Ojo 1
diagn_codigo1	Código de diagnóstico 1
diagn_obs1	Notas del diagnóstico 1
diagn_ojo2	Ojo 2
diagn_codigo2	Código de diagnóstico 2
target	Target definido: ALTO - BAJO
CCM_EDAD	Ratio entre Cant. Cons. Del Mes y Edad
CCS_EDAD	Ratio entre Cant. Cons. Del Semestre y Edad
CCA_EDAD	Ratio entre Cant. Cons. Del Año y Edad
CPC_EDAD	Ratio entre Cant. Parcial de Cons. y Edad
CTC_EDAD	Ratio entre Cant. Total del Cons. y Edad
GBA	Binario que indica si paciente vive en Gran Buenos Aires o Ciudad de Bs As
ZONA_GBA	Zona del Gran Buenos Aires en la que vive el paciente (CABA, Norte, Sur, Oeste)
PARTIDO	Partido de la Provincia de Buenos Aires
sum_dummy	Suma de todas las variables dummy
ant_estudio	Dummy que indica si se posee antecedentes cargados
CAMARA_ANTERIOR_estudio	Dummy que indica si se realizó estudios de CAMARA_ANTERIOR
CONJUNTIVA_estudio	Dummy que indica si se realizó estudios de CONJUNTIVA
CORNEA_estudio	Dummy que indica si se realizó estudios de CORNEA
cvi_estudio	Dummy que indica si se realizó estudios de cvi
diagn_estudio	Dummy que indica si se realizó estudios de diagnósticos
ESCLERA_estudio	Dummy que indica si se realizó estudios de ESCLERA
GLOBO_OCULAR_estudio	Dummy que indica si se realizó estudios de GLOBO_OCULAR
LAGRIMAL_estudio	Dummy que indica si se realizó estudios de LAGRIMAL
ORBITA_estudio	Dummy que indica si se realizó estudios de ORBITA
PAPILAS_estudio	Dummy que indica si se realizó estudios de PAPILAS
PARPADOS_estudio	Dummy que indica si se realizó estudios de PARPADOS
PRESION_OCULAR_estudio	Dummy que indica si se realizó estudios de PRESION_OCULAR
PUPILAS_estudio	Dummy que indica si se realizó estudios de PUPILAS
RET_estudio	Dummy que indica si se realizó estudios de RET
rfr_estudio	Dummy que indica si se realizó estudios de refracción
TAB_QUE_estudio	Dummy que indica si se realizó estudios de TAB_QUE
VITREO_estudio	Dummy que indica si se realizó estudios de VITREO
estudio	Sumatoria del total de dummys de estudio
ABC_sum	Ratio de dummy de medicación y sumatoria de dummys
ABCsum_EDAD	Ratio de dummy de medicación y edad
estud_EDAD	Ratio de dummy de estudios y edad
ABCsum_estud	Ratio de dummy de medicación y estudios
Mayor_edad	Binario que indica si el paciente es mayor de edad
Verano	Binario que indica si el paciente fue atendido entre Enero y Abril
Malasalud	Resultado de multiplicar la cantidad de estudios realizados, la cantidad de medicamentos que toma y la cantidad total de consultas en la clínica

Menor10anos	Flag que indica si el paciente es menor de 10 años
Menor5anos	Flag que indica si el paciente es menor de 5 años
Mayor40anos	Flag que indica si el paciente es mayor de 40 años
Mayor60anos	Flag que indica si el paciente es mayor de 60 años
Franja10_16	Flag que indica si el horario de atención es entre las 10 y las 16hs
Franja18_22	Flag que indica si el horario de atención es entre las 18 y las 22hs
LuMaJu	Binaria que marca si el día de atención es Lunes, Martes o Jueves
Lunes	Binaria que marca si el día de atención es Lunes
Martes	Binaria que marca si el día de atención es Martes
Jueves	Binaria que marca si el día de atención es Jueves
Sin_medic	Marca de paciente que no toma medicación
Franja_medic	Bandarización de pacientes según cantidad de medicación que toma (hasta 2, entre 2 y 5, más de 5)
Hasta2_medic	Flag que indica si el paciente toma hasta 2 medicamentos
ordenXhoraestand	Resultado de multiplicar el orden de atención x la hora estandarizada
Vera_franja2	Marca de paciente atendido en verano en el horario de 18 a 22hs
Vera_franja2_lu	Marca de paciente atendido en verano, un día lunes, en el horario de 18 a 22hs
Vera_franja2_ma	Marca de paciente atendido en verano, un día martes, en el horario de 18 a 22hs
Vera_franja2_ju	Marca de paciente atendido en verano, un día jueves en el horario de 18 a 22hs
Vera_franja2_lumaju	Marca de paciente atendido en verano, un día lunes, martes o jueves, en el horario de 18 a 22hs
Vera_franja1	Marca de paciente atendido en verano en el horario de 10 a 16hs
Vera_franja1_lu	Marca de paciente atendido en verano, un día lunes, en el horario de 10 a 16hs
Vera_franja1_ma	Marca de paciente atendido en verano, un día martes, en el horario de 10 a 16hs
Vera_franja1_ju	Marca de paciente atendido en verano, un día jueves, en el horario de 10 a 16hs
Vera_franja1_lumaju	Marca de paciente atendido en verano, un día lunes, martes o jueves, en el horario de 10 a 16hs
T_max	Temperatura máxima media registrada durante el mes de atención
T_min	Temperatura mínima media registrada durante el mes de atención
Lluv_mm	Milímetros caídos de lluvia durante el mes de atención
Lluv_dias	Cantidad de días con lluvia durante el mes de atención
Hum_max	Humedad máxima media durante el mes de atención
Hum_min	Humedad mínima media durante el mes de atención
De0_10anos	Flag que indica si el paciente tiene entre 0 y 10 años
De11_20anos	Flag que indica si el paciente tiene entre 11 y 20 años
De21_30anos	Flag que indica si el paciente tiene entre 21 y 30 años
De31_40anos	Flag que indica si el paciente tiene entre 31 y 40 años
De41_50anos	Flag que indica si el paciente tiene entre 41 y 50 años
De51_60anos	Flag que indica si el paciente tiene entre 51 y 60 años
De61_70anos	Flag que indica si el paciente tiene entre 61 y 70 años

De71_80anos	Flag que indica si el paciente tiene entre 71 y 80 años
Masde80anos	Flag que indica si el paciente tiene más de 80 años
Mayor_verano	Binario que indica si el paciente es mayor de edad y se atendió en verano
Menor_invierno	Binario que indica si el paciente es menor de edad y se atendió en invierno
Mayor_invierno	Binario que indica si paciente es mayor de edad y se atención en invierno
Menor_verano	Binario que indica si el paciente es menor de edad y se atendió en verano
Lu_1822_Orden	Resultado de multiplicar las variables Lunes, Franja 18_22 y el Orden de Atención
Tiempo-1	Tiempo de atención de la anteúltima consulta
Tiempo-2	Tiempo de atención de la 2da consulta anterior
Tiempo-3	Tiempo de atención de la 3er consulta anterior
TendenciaQL_12	Comparación entre la duración de la consulta anterior (consulta-1) y la antepenúltima (consulta-2). Determina si la primera fue más larga, igual o más corta
TendenciaQL_13	Comparación entre la duración de la consulta antepenúltima (consulta-1) y la previa a la antepenúltima (consulta-3). Determina si la primera fue más larga, igual o más corta
TendenciaQL_23	Comparación entre la duración de la anterior antepenúltima (consulta-2) y la previa a la antepenúltima (consulta-3). Determina si la primera fue más larga, igual o más corta

Bibliografía

- Decisión Trees for Business Intelligence and Data Mining – Barry de Ville – SAS Press Series – 2006
- Data Mining with Decision Trees – Lior Rokach, Oded Maimon – World Scientific – 2008
- Data Science for Business – Foster Provost and Tom Fawcett – O'Reilly – 2013
- The R Book – Michael J. Crawley – Wiley – 2013
- R Programming for Data Science – Roger D. Peng – 2016
- Nuevo Algoritmo de clasificación supervisado sin parámetros, no afectado por el desbalanceo y overfitting (Tesis de Maestría) – Pablo A. Poloni – 2014
- Comparación interactiva de modelos de minería de datos utilizando técnicas de visualización (Tesis de Maestría) – Luciana M. Padua – 2014
- C4.5: Programs For Machine Learning - J. Ross Quinlan – 1993
- Feature Engineering for Machine Learning – A. Zheng & A. Casari - 2018

Referencias

[1] Queirós A, González-Meijome JM, Jorge J, Parafita MA - Presbicia. Análisis de la prevalencia y perfil de una población présbita en el norte de Portugal - Ver y Oír 2006; 407-412.

[2] Sergio Bonafonte, E. Milla - Esquemas Clinico-Visuales En Oftalmología (3ª ED.) - Masson (2006)

[3] Global Health Observatory data repository – World Health Organization (OMS)

[4] Myriam Deveugele, Anselm Derese, Atie van den Brink-Muinen, Jozien Bensing, Jan De Maeseneer - Consultation length in general practice: cross sectional study in six European countries – BMJ - 2002;325:472

[5] Hava Tabenkin, Meredith A. Goodwin, Stephen J. Zyzanski, Kurt C. Stange, and Jack H. Medalie - Journal of Women's Health - July 2004, 13(3): 341-349.

[6] Alice W. Migongo, MPH, Richard Charnigo, PhD, Margaret M. Love, PhD, Richard Kryscio, PhD, Steven T. Fleming, PhD, Kevin A. Pearce, MD, MPH - Factors Relating to Patient Visit Time With a Physician - Medical Decision Making/Jan–Feb 2012

[7] Organización Mundial de la Salud - Purpose of the ATC/DDD system - www.whocc.no/atc_ddd_methodology/purpose_of_the_atc_ddd_system/

[8] Susanna Sans Menéndez – Enfermedades Cardiovasculares - Institut d' Estudis de la Salut, Barcelona - 2007

[9] M. Gloria Icaza, M. Loreto Núñez – Atlas de Mortalidad por Enfermedades Cardiovasculares – Universidad de Talca – 2006

[10] Westcott R. – The length of consultations in general practice – J R Coll Gen Pract – 1977;27:552-5

[11] Raynes NV, Cairns V – Factors contributing to the length of general practice consultations – J R Coll Gen Pract – 1980;30:496-8

[12] Pang-Ning Tan, Michael Steinbach, Vipin Kumar – Introduction to Data Mining – Pearson Education – 2006

[13] Carlos Vega, Genoveva Rosano, Juan M. López, José L. Cendejas, Heberto Ferreira – Data Mining aplicado a la Predicción y Tratamiento de Enfermedades – Conferencia Iberoamericana en Sistemas, Cibernética e Informática – 2012

[14] Liset González Polanco, Yadian Guillermo Pérez Betancourt – Spatial data mining and its application in health and epidemiology studies – Revista Cubana de Información en Ciencias de la Salud – 2013; 24(4):482-489

[15] Kuttiannan Thangavel, P. Palanichamy Jaganathan and P.O. Easmi – Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique – Asian Journal of Information Technology 5 (4) – 2006;413-417

[16] Andrew M. Wilson, Lehana Thabane, Anne Holbrook – Application of data mining techniques in pharmacovigilance – British Journal of Clinical Pharmacology – 2013; 57:2; 127-134

[17] Xuezhong Zhou, Shibo Chen, Baoyan Liu, Runsun Zhang, Yinghui Wangd, Ping Li, Yufeng Guo, Hua Zhang, Zhuye Gao, Xiufeng Yan - Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support – Artificial Intelligence in Medicine 48 – 2010; 139-152

[18] Ruben D. Canlas Jr. – Data Mining in Healthcare: Current Applications and Issues – Carnegie Mellon University – 2009

[19] IBM Launches Health Analytics Center - http://www-03.ibm.com/press/us/en/pressrelease/28757.wss - 2009

[20] IBM Expands Health Analytics Solution Center to Address Explosive Growth of Medical Information - http://www-03.ibm.com/press/us/en/pressrelease/34610.wss

[21] Hernán C. Doval – Malestar en la medicina. Insatisfacción y descontento en los médicos – Revista Argentina de Cardiología – 2007; Vol.75 N° 4

[22] Alma Lucila Sauceda-Valenzuela, Veronika J. Wirtz, Yared Santa-Ana-Téllez and María de la Luz Kageyama-Escobar - Ambulatory health service users's experience of waiting time and expenditure and factors associated with the perception of low quality of care in Mexico – BMC Health Services Research – 2010; 10:178

[23] https://www.ibm.com/watson/health/

[24] Andrew Gottschalk, Susan A. Flocke - Time Spent in Face-to-Face Patient Care and Work Outside the Examination Room - Annals Of Family Medicine – 2005; Vol. 3, No. 6

[25] Alice Migongo, Richard Charnigo, Margaret Love, Richard Kryscio, Steven Fleming, Kevin Pearce – Factors Relating to Patient Visit Time With a Physician – Medical Decision Making – 2012;32: 93-104

[26] Boletín de Vigilancia – Enfermedades no transmisibles y factores de riesgo – Ministerio de Salud de la República Argentina – 2013

[27] Dan-Avi Landau, Yaacov G. Bachner, Keren Elishkewitz, Liav Goldstein, Erez Barneboim - Patients' Views on Optimal Visit Length in Primary Care – Medical Practice Management – Julio/Agosto 2007

[28] Delia Outomuro, Andrea Mariel Actis – Estimación del tiempo de consulta ambulatoria en clínica médica – Rev Med de Chile – 2013;141:361-366

[29] Michele Samorani, Linda LaGanga – Improving appointment scheduling with data mining – Production and Operations Managment Society – Vancouver – May 2010

[30] Linda LaGanga – Lean Service Operations: Reflections and New Directions for Capacity Expansion in Outpatient Clinics – Journal of Operations Managment – July 2011

[31] Linda LaGanga & Stephen Lawrence – Increasing Access to Healthcare Services through Service Time Process Improvements – POMS 20th Annual Conference – May 2009

[32] Linda LaGanga & Stephen Lawrence – Clinic Overbooking to Improve Patient Access and Increase Provider Productivity – Decision Sciences – May 2007

[33] Juan José Cubillas & M. Isabel Ramos & Francisco R. Feito & Tomás Ureña - An Improvement in the Appointment Scheduling in Primary Health Care Centers Using Data Mining - J Med Syst (2014) 38:89

[34] Laura P. Sands, Joanne K Daggy, Purdue University, Mark Lawley, Deanna Willis, Debra Thayer - Using No-Show Modeling to Improve Clinic Performance - School of Nursing Faculty Publications - 12-2010

[35] KJ Glowacka, RM Henry and JH May - A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling - Journal of the Operational Research Society - 2009 (60, 1056 –1068)

[36] Badr HSSINA, Abdelkarim MERBOUHA,Hanane EZZIKOURI,Mohammed ERRITALI - A comparative study of decision tree ID3 and C4.5 - (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications - 2014

[37] Sonia Singh, Priyanka Gupta - COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY - International Journal of Advanced Information Science and Technology (IJAIST) - Vol.27, No.27, July 2014

[38] Terry Therneau, Beth Atkinson, Brian Ripley - Package 'rpart' Recursive Partitioning and Regression Trees - April 21, 2017

[39] Ministerio de Salud de la Nación - http://www.salud.gob.ar/index.php/component/content/article/46-ministerio/194-vuelta-a-clases

[40] Fawcett, Tom - ROC graphs: Notes and practical considerations for researchers - Machine learning, 31(1), 1-38 (2004).

[41] Emiliana Valderrama Gama , Fernando Rodríguez Artalejo , Antonia Palacios Díaz , Pilar Gabarre Orús y Jesús Pérez del Molino Martín - Consumo De Medicamentos En Los Ancianos: Resultados De Un Estudio Poblacional - Rev. Esp. Salud Pública vol.72 no.3 Madrid may. 1998

[42] Thomas J. Moore, Donald R. Mattison – How Many Adults in the United States Are Taking Psychiatric Drugs? - Jama Internal Medicine (2016)

[43] Daniel Ferrante, Mario Virgolini - Encuesta Nacional de Factores de Riesgo 2005: resultados principales. Prevalencia de factores de riesgo de enfermedades cardiovasculares en la Argentina - Rev. Argent. Cardiol. v.75 n.1 Buenos Aires ene./feb. 2007

[44] Maria-Luiza Antonie, Osmar R. Zaıane, Alexandru Coman - Application of Data Mining Techniques for Medical Image Classification - Second International Workshop on Multimedia Data Mining (2001)

[45] Health Level Seven International - http://www.hl7.org/index.cfm

[46] Instituto Nacional de Estadística y Censos (INDEC) - http://www.indec.gov.ar/censos_provinciales.asp?id_tema_1=2&id_tema_2=41&id_tema_3=135&p=06&d=000&t=3&s=0&c=2010

[47] MOHAMMED M MAZID, A B M SHAWKAT ALI, KEVIN S TICKLE - Improved C4.5 Algorithm for Rule Based Classification - School of Computing Science - Central Queensland University (2010)

[48] Leo Breiman – Random Forest - Machine Learning, 45, 5–32, 2001 - Kluwer Academic Publishers