UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DESARROLLO DE UN MODELO PREDICTIVO DE LA ASISTENCIA A CITAS MÉDICAS DE TERAPIA FÍSICA EN LA RED DE CENTROS ASISTENCIALES DE ACHS.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

IGNACIO DANILO ORTIZ SÁEZ

PROFESOR GUÍA:

CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN:

ALEJANDRA PUENTE CHANDÍA

LUIS FERNANDO SOLARI DÍAZ

SANTIAGO DE CHILE

2022

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial.

POR: Ignacio Danilo Ortiz Sáez

FECHA: 2022

PROFESORA GUÍA: Carolina Segovia Riquelme

DESARROLLO DE UN MODELO PREDICTIVO DE LA ASISTENCIA A CITAS MÉDICAS DE TERAPIA FÍSICA EN LA RED DE CENTROS ASISTENCIALES DE ACHS.

La Asociación Chilena de Seguridad (ACHS) es la mutualidad más grande del país, la cual se dedica a administrar el seguro social contra accidentes del trabajo y enfermedades profesionales, teniendo distintas labores en torno a esta función.

Actualmente, hay una tasa de inasistencia a las sesiones kinesiológicas cercana al 10%, lo que se ve reflejado en la pérdida de 34.000 citas médicas anualmente, por este motivo en este proyecto se busca predecir cuáles son los pacientes que tienen una mayor tendencia a ausentarse.

Para llevar a cabo este proyecto se utiliza la metodología CRISP-DM de análisis de datos, en donde se aplican modelos de *machine learning* para lograr concretar el objetivo propuesto.

Se realiza un modelo de Redes Neuronales entrenado con datos balanceados con SMOTE, que contempla variables claves como la asignación de transporte para los pacientes, hora de la cita médica, tasa de inasistencia histórica de los pacientes y el porcentaje de citas médicas restantes para cada paciente. Con lo anterior, se tiene que el modelo predice con un 70% de precisión los casos inasistentes de las citas médicas de terapia física, obteniendo una AUC de 76% respecto de la curva ROC, lo cual indica que es un modelo robusto para realizar predicciones de citas médicas.

En base a los resultados obtenidos, se plantea un sistema de sobreagendamiento en base a las citas agendadas. Los resultados de este sobreagendamiento se dividen en dos: Hospital del Trabajador y otras sedes ACHS. Para el primer caso se obtiene que el sobreagendamiento es preciso, teniendo una falla de aproximadamente 1 citación por bloque horario, lo cual se considera positivo teniendo en cuenta que este centro tiene una gran capacidad y no sería problema que un kinesiólogo atienda un paciente extra (por el tipo de atención que se realiza). Sin embargo, para el segundo caso se observa que la precisión en el sobreagendamiento dista de la realidad de manera importante (5 agendamientos demás en promedio por hora) lo cual se atribuye a que no se consideró la variable de la sede en la construcción del modelo.

En conclusión, se recomienda utilizar el modelo únicamente para el Hospital del trabajador y fortalecer el modelo para las demás sedes agregando variables y entrenando con una mayor cantidad de datos.

AGRADECIMIENTOS

Quiero agradecer especialmente a mis padres Eric y Lorni por ser un pilar fundamental en mi vida, por su incondicionalidad y apoyo en todo momento. Gracias por sus enseñanzas, por su confianza infinita, por alegrarse en mis buenos momentos y abrazarme en los malos, esto no podría haberlo logrado sin ustedes. Las palabras se hacen escasas frente al amor que siento por ustedes.

A mi Nona y a mi Tata por alegrarme siempre con un pastelito, una conversación o simplemente una sonrisa, esos momentos los llevo siempre en mi corazón y son mi motivación para levantarme en el día a día.

Agradecer especialmente a mis compañeros y amigos de la Universidad quienes me han acompañado a lo largo de carrera, esta experiencia hubiera sido diametralmente distinta si no los hubiera conocido a Imprito, Lazos, CuarentenaFat, entre otros. Gracias por llenar mis días y noches de alegría.

A mis amigos del colegio quienes siempre estuvieron (y estarán) ahí para apoyarnos mutuamente, Angelo, Jorge y Cinthia son lo máximo.

Finalmente agradecer a cada uno de los que me ha apoyado en este largo proceso, estoy muy feliz por las personas que me rodean y que me rodearon en algún momento de la vida.

Seguramente se me queda mucha gente en el tintero, si usted no encontró su nombre arriba, estas líneas son suyas querido lector.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1. CONTEXTO ORGANIZACIONAL

1.1. Identificación y descripción del sector industrial

La Asociación Chilena de Seguridad pertenece al sector industrial de las mutualidades, las cuales según la definición de la Superintendencia de Seguridad Social son "corporaciones de derecho privado, sin fines de lucro, que administran el seguro de la Ley 16.744 (Ver Anexo 2), (…) otorgando las prestaciones preventivas, médicas y económicas que dicha Ley y sus reglamentos, establecen". [1]

En Chile sólo existen 4 instituciones que se dedican a este rubro: el Instituto de Seguridad del Trabajo (IST), la Asociación Chilena de Seguridad (ACHS), el Instituto de Seguridad Laboral (ISL) y finalmente la Mutual de Seguridad CChC (MUSEG). Entre las instituciones anteriormente mencionadas se protege a la totalidad de trabajadores del país.

En términos de posicionamiento, la ACHS tiene la mayor participación de mercado con un 38.9%, el cuál es logrado a través de sus de 84.310 empresas afiliadas. Si se analiza la accidentabilidad laboral de los afiliados, se evidencia una baja sostenida a través del último tiempo, siendo la menor tasa de accidentabilidad promedio del país. (Ver anexos 3, 4 y 5).

Dado que esta organización es una mutual, el principal marco regulatorio es la Ley 16.744 o Ley de Protección al Trabajador, en ella se establecen todas las normativas con las que deben cumplir este tipo de instituciones, la aplicación de esta ley es supervisada por el organismo estatal SUSESO (Superintendencia de Seguridad Social).

1.2. Tipo de Organización

Como se mencionó anteriormente, la ACHS es una corporación de derecho privado sin fines de lucro. En términos de gobierno corporativo es presidida por Paul Schiodtz Obilinovich, junto a otros 7 miembros del directorio. A principios de 2019, la mutual poseía más de 4.700 trabajadores propios, siendo de estos un 59% mujeres. [1]

La institución, en promedio, tiene 700.000 atenciones médicas anualmente, a través de sus 282 centros ambulatorios, 97 policlínicos y 421 ambulancias a lo largo de Chile, incluyendo el Hospital del Trabajador. Actualmente posee más de 2 millones de trabajadores afiliados, los cuales pertenecen a las 84.310 empresas en convenio.

1.3. Servicios y procesos

La organización se destaca por tener 3 principales funciones o servicios a la disposición de los trabajadores:

1.3.1 Prevención de Riesgos Laborales

Uno de los servicios que ofrece la mutual es la de prevención de riesgos laborales a través de talleres diseñados especialmente para las empresas y sus colaboradores. Creando planes de trabajo con actividades de detección de riesgos, entregando medidas concretas orientadas a la prevención, mejorando así la seguridad de los trabajadores y trabajadoras.

1.3.2 Prestaciones de Salud

Como segundo pilar dentro de los servicios ACHS, está la prestación de servicios de salud a los trabajadores. En caso de que los colaboradores de las empresas hayan sufrido un accidente de trabajo, de trayecto hacia el trabajo o alguna enfermedad profesional; es la ACHS la que se encarga de otorgar gratuitamente una cobertura médica integral hasta que el paciente se haya recuperado completamente o mientras aún presente síntomas producto del accidente o enfermedad. Se les entrega además una serie de facilidades de traslado hacia los centros de la institución.

1.3.3 Prestaciones Económicas

Como último punto, esta mutual también hace entregas de subsidios, indemnizaciones y pensiones a los colaboradores en caso de que existan incapacidades temporales, permanentes o bien el fallecimiento del trabajador; esto es aplicable únicamente en circunstancias relacionadas con el trabajo del mismo.

De los servicios anteriores, el que tendrá más relevancia respecto del trabajo a realizar en la memoria, es el segundo.

1.4. Clientes y usuarios

En base a lo anterior, se considera necesario diferenciar entre los clientes y los usuarios de los servicios de la mutual. Los clientes con los que trata la organización son principalmente las empresas empleadoras, dado que son ellas las que realizan las transacciones comerciales y financian la operación de esta institución en base a las cotizaciones. Estas últimas fluctúan entre un 0.93% a un 7% del sueldo total imponible por cada trabajador, este porcentaje varía según la tasa de siniestralidad de cada empresa (para más detalles revisar el anexo 9.6). Según las

definiciones mencionadas, la cantidad total de clientes de la institución es de 84.000 empresas aproximadamente, en base a las cifras más recientes.

Por otro lado, los usuarios son los trabajadores de las instituciones asociadas a la ACHS, quienes al momento de sufrir algún tipo de accidente o una enfermedad profesional, acceden a los servicios proporcionados por la mutual, recibiendo prestaciones médicas y/o subvenciones según corresponda.

Según los datos expuestos en la tabla 1, se observa que la mayor cantidad de trabajadores se encuentra en el sector de comercio y retail, seguido por el sector industrial y de educación. Además, el rubro con menor tasa de accidentabilidad es el de educación.

Industria	Accidentes	Masa	Tasa accidentabilidad
Acuícola	1.033	30.334	3,43
Agrícola	6.071	172.583	3,52
Comercio y retail	11.018	630.975	1,75
Construcción	5.285	171.908	3,07
Educación	1.566	318.868	0,49
Energía y telecom	1.381	66.171	2,09
Forestal maderero	1.583	59.410	2,66
Gubernamentales	1.545	218.972	0,71
Industrial	9.920	417.373	2,38
Minería	322	31.990	1,01
Pesca	284	6.956	4,07
Servicios de salud	1.734	122.239	1,42
Servicios financieros	1.467	174.127	0,85
Transporte	3.622	122.480	2,95

Tabla 1: Tasa accidentabilidad por rubro. ACHS Febrero 2020

2. DESCRIPCIÓN DEL PROBLEMA Y JUSTIFICACIÓN

Cuando un trabajador sufre un accidente laboral, debe acudir a un centro asistencial ACHS en donde es atendido y tratado de tratado de forma gratuita, porque está protegido por el seguro de accidentes del trabajo ligado a la Ley 16.744. En muchas ocasiones, el paciente no queda completamente recuperado con una sola atención en el centro y necesita, entre otras cosas, controles futuros con el médico, curaciones y terapias kinesiológicas, por lo que es necesario agendar citas médicas para que vuelva a acudir a un centro de la mutual.

Específicamente, este trabajo de título contemplará las terapias físicas realizadas dentro de la institución, esto corresponde a las sesiones kinesiológicas, las cuales son recetadas por los médicos de la mutual, generalmente traumatólogos en busca de que el paciente tenga una adecuada recuperación posterior a algún tipo de accidente o intervención quirúrgica. Según lo analizado en las bases de datos, en promedio la cantidad de citas de terapia física agendada por siniestro son 3, con un mínimo de 1 y un máximo de 30.

El principal problema que se desea abordar a través de este proyecto de memoria es el ausentismo a las citas médicas de Terapia Física, ya que según datos obtenidos de la base de datos de ACHS en 2020: hay una inasistencia de cercana al 10% del total de las citas concertadas de terapia física, lo que equivale a 34.000 agendamientos anualmente, lo cual es un costo para la mutual dado que se traduce en que la utilización promedio de los kinesiólogos es de un 82%, por lo que estos son recursos que se están desaprovechando. Es clave señalar que actualmente no se tiene ninguna noción de qué tipo de paciente es el que no asiste, ni tampoco se han establecido posibles causalidades.

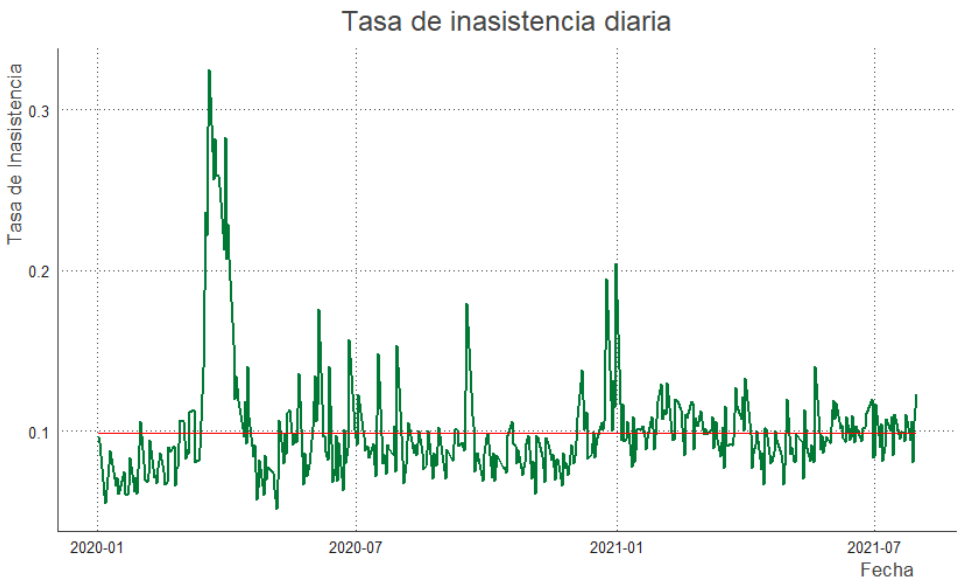


Figura 1: Tasa de inasistencia diaria

Bajo esa premisa, desde el área de *Analytics* surge la iniciativa de averiguar cuál es el tipo de paciente que se ausenta, ver la capacidad de predicción que podría tener un modelo y qué hacer para poder mitigar esta problemática. El área de *Analytics*, está enmarcada dentro de la Gerencia de Transformación Digital de la organización y se dedica a manejar la data producida por la misma, vale decir, la información de los pacientes, accidentes, datos de transporte, etc. Se encarga además de transformar los datos en información útil para los distintos proyectos que se generan dentro de la ACHS, como por ejemplo generar reportería, tableros y otras.

En ese contexto se visualiza la oportunidad de predecir a través de un modelo de clasificación, la asistencia o inasistencia de un paciente a su cita médica. Dado lo anterior, se analizará el aumento del nivel de utilización de los recursos a través de overbooking y pasos futuros.

Se plantea que la predicción de la asistencia a las citas por pacientes es una buena alternativa por tres motivos principalmente. El primero es que permite identificar de forma más precisa el perfil de persona que no está asistiendo a sus citas, lo que podría permitir que la institución focalice sus esfuerzos en este segmento, puede ser el caso que estos pacientes no vean valor en el trabajo realizado por la institución, busquen alternativas por no estar satisfechos con la

atención, lo que provoca que desistan de atenderse en la ACHS, habiendo espacio de mejora para la mutual. El segundo es que teniendo una mayor certeza de cuáles son los pacientes que no, se pueden establecer modelos de recordatorio o de confirmación de cita, los cuales además de aumentar la probabilidad de que el paciente asista, sin embargo se encuentra fuera del alcance de esta memoria y se plantea como trabajo futuro. Como tercer y último motivo es que en base a las predicciones se podrían ejecutar un sobrecupo con bajo margen de error buscando así aumentar la utilización de los recursos dispuestos para estas labores.

2.1. Estimativa del valor monetario

Con el fin de valorar económicamente el impacto de este trabajo para la institución, se realiza un cálculo en base a los siguientes supuestos:

Sueldo promedio de un Kinesiólogo: $900.000 (aproximado) [2]

- La ACHS cuenta con 62 kinesiólogos actualmente [3]
- La utilización de kinesiólogos es del 84% [3]
- Kinesiólogo trabaja en promedio 40 hrs semanales [3]
- Se agendan 240.000 citas anualmente [3]
- Se reporta un 10% inasistencia a las citas médicas de Terapia física anualmente (24.000) [3]
- Precio del transporte de paciente 22.000 ida y vuelta [3]
- 50% de las citas de kinesiología tienen transporte asignado [3]

En base a lo anterior se plantean 3 casos representativos: inasistencia de 34000 citas anualmente, 17000 inasistencias anuales (reducción a la mitad) y 0 inasistencias anuales (todos los pacientes asisten):

	Actualidad	Mejora 50%	Mejora 100%
Asistencia	306.000	323.000	340.000
Inasistencias	34.000	17.000	0
Utilización	0,84	0,89	0,93

Tabla 2: Casos posibles de mejora

Como se puede observar en la tabla 2 la utilización máxima alcanzada en el caso ideal de que asistan todos los pacientes a sus citas médicas es de un 93%, esto porque a pesar de la inasistencia, existen otros factores involucrados que afectan esta utilización.

En base al cálculo anterior de la utilización se procede a cuantificar las horas y transportes perdidos anualmente por la ausencia de los pacientes. Es clave notar que la cantidad de horas pérdidas se calcula en base a la utilización del 93%, ya que se evalúa solamente el impacto del ausentismo:

	Actualidad	Mejora 50%	Mejora 100%
Inasistencias	34.000	17.000	0
Utilización	0,84	0,89	0,93
Horas perdidas al año	11.160	4.960	0
Transportes perdidos	17.000	8.500	0

Tabla 3: Horas y transportes perdidos en base a inasistencia

Como se puede observar si se redujera el ausentismo en un 50%, se podrían aprovechar alrededor de 5.000 horas que actualmente están siendo perdidas. Ahora bien, si esto se traduce en impacto monetario, se alcanzan las siguientes cifras:

	Actualidad	Mejora 50%	Mejora 100%
Valor horas perdidas	62.775.000	27.900.000	0
Valor transporte perdido	374.000.000	187.000.000	0
Total costos	**436.775.000**	**214.900.000**	**0**

Tabla 4: Valorización de horas y transportes perdidos

Es evidente que, según los cálculos realizados, si se disminuye el ausentismo en un 50%, se obtiene un ahorro de aproximadamente 218 millones de pesos anualmente. En base a lo anterior se puede calcular cuánto equivale en promedio el valor de una inasistencia de un paciente: $13.000 aproximadamente.

3. OBJETIVOS

3.1. Objetivo general

El objetivo general de este proyecto es el **desarrollo de un modelo predictivo de ausentismo de citas médicas de terapia física para optimizar la utilización de recursos de la red de centros ACHS.**

3.2. Objetivos específicos:

- Determinar cuáles son las posibles variables claves que determinan la asistencia o no de un paciente a la cita médica.
- Generar modelos de predicción de asistencias a las citas médicas, que indiquen si el paciente asistirá a un centro médico. Seleccionar uno de los modelos anteriores en base a métricas de comparación.
- Crear planes de acción que aumenten la utilización de los recursos en base a los resultados del modelo de predicción.

4. MARCO CONCEPTUAL

El marco conceptual de este trabajo de memoria se dividirá en dos secciones, una hace referencia a la ciencia de datos y los modelos tentativos a utilizar y mientras que la otra se centra en las recomendaciones para aumentar la utilización de los recursos.

4.1. Ciencia de datos

Existen múltiples definiciones para esta temática, sin embargo una de las más apropiadas sería la de Oracle, que propone lo siguiente: "La ciencia de datos combina múltiples campos que incluyen estadísticas, métodos científicos y análisis de datos para extraer el valor de los datos" [4]. Por lo que este tipo de ciencia, de acuerdo a lo indicado en la cita anterior es multidisciplinaria y requiere de conocimientos en distintos tipos de temáticas, como lo viene siendo el reconocimiento de patrones, estadística, visualizaciones, manejo de datos, minería de datos; como se puede observar en la figura 5.

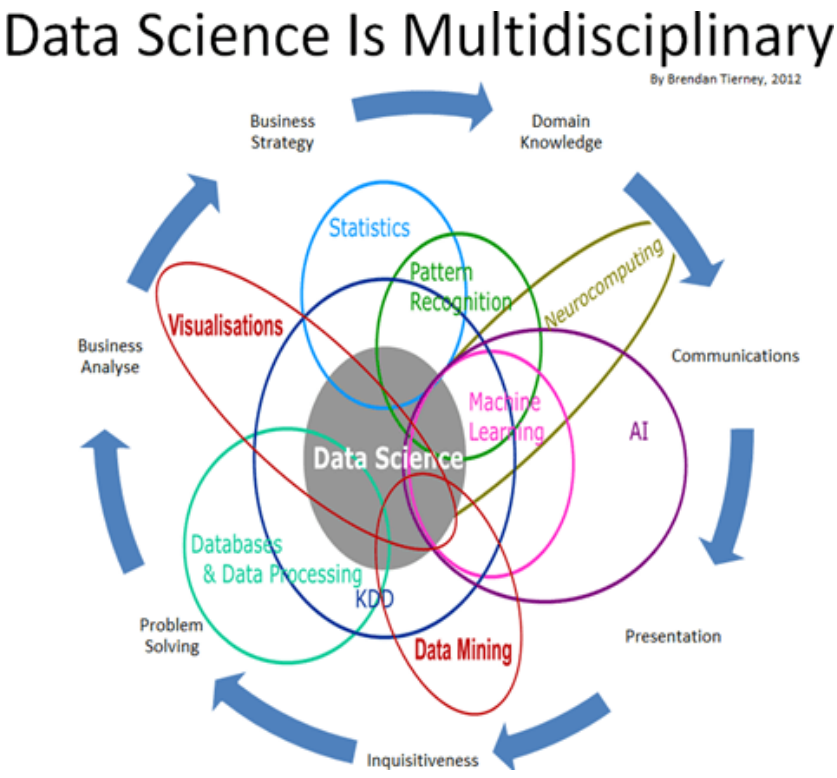


Figura 2: Data Science is multidisciplinary. Brendan Tierney (2012)

En base al breve contexto explicado anteriormente, en el trabajo de título se utilizarán distintas disciplinas, como por ejemplo el procesamiento de datos, visualizaciones,

estadísticas, reconocimiento de patrones y *machine learning*. Respecto a este último campo, se procede a especificar en el siguiente ítem.

4.1.1 Modelos

Algoritmos no lineales

Con el objetivo de tener resultados más precisos y robustos, se utilizan métodos de clasificación supervisados no lineales como Redes neuronales y XGBoost, los cuales tienen como característica aprender patrones.

Redes neuronales

Las redes neuronales son algoritmos inspirados en el funcionamiento del cerebro humano para que un computador los imite. Funcionan a través de una unidad básica llamada perceptrón, que tiene distintas entradas de información, el conjunto de estos se llaman capas, que terminan formando la red. Existen varios tipos de redes como por ejemplo multicapas y convolucionales, a mayor número de capas, más complejas serán las funciones a realizar. Presentan 2 hiperparámetros: *size* y *decay*. El primero hace referencia a el número de nodos en la capa oculta y *decay* es un parámetro de regularización de los pesos.

Se utilizan habitualmente en problemas de clasificación, predicción, reconocimiento de tendencias, entre muchas otros.

XGBoost

Es un algoritmo de aprendizaje supervisado basado en arboles de decisión, se caracteriza por entregar buenos resultados con poco esfuerzo. Usa modelos débiles (árboles de decisión) secuencialmente para generar un modelo robusto, a partir de un algoritmo de optimización. Los parámetros de los modelos débiles se ajustan en cada iteración, con el fin de encontrar el mínimo de la función objetivo y así encontrar un óptimo. Sus 2 parámetros principales *Max Depth*, que indica la cantidad de profundidad o número de nodos de bifurcación de los árboles de decisión usados en el entrenamiento y también el ETA, que es la tasa de aprendizaje del modelo.

4.1.2 Evaluación de los modelos

Dado que se va a trabajar con diferentes tipos de modelos estadísticos, resulta necesario tener métricas para poder compararlos entre ellos. Por consiguiente, en esta sección se mostrarán los principales.

Accuracy

Se calcula en base a la matriz de confusión y representa el porcentaje de casos que fueron predichos de manera correcta por el modelo en comparación a todos los casos predichos. De forma matemática se expresa de la siguiente forma:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

En donde *TP* y *TN* son los casos *True Positives* y *True Negatives* respectivamente, los cuales hacen referencia a la frecuencia de predicciones correctas. En cambio, los *FP* y *FN* son los casos en los que no se predijo correctamente. Notar que, en este caso para la memoria, se considerará como *TP* a la asistencia predicha correctamente y *como TN* a la inasistencia predicha correctamente.

Recall

El Recall es el ratio que representará los casos verdaderos negativos (inasistencias) sobre los casos que fueron marcados como inasistentes por el modelo, ya sea correcta o incorrectamente. Esto indica el porcentaje de las predicciones que son correctas entre las predicciones totales del caso de interés.

$$Precisión = \frac{TN}{TN + FN} \quad (2)$$

Curva ROC

Gráfico que muestra en un eje la tasa de falsos positivos y en el otro la tasa de falsos negativos. Permite comparar los modelos estableciendo cuál es el que tiene mejor rendimiento como clasificador, siendo el área bajo la curva (AUC) uno de los principales indicadores, mayor AUC implica mejor clasificador.

4.2. Overbooking

Como se ha mencionado, este trabajo planteará la posibilidad de realizar *overbooking* en base a los resultados obtenidos por los algoritmos de predicción, con la finalidad de que se pueda aumentar la productividad de los recursos, cuidando la calidad de la atención. Bajo esa propuesta, se analiza documentación científica que trata del tema con casos de éxito [10, 11, 12], en donde se evidencia que es posible sobreagendar de manera precisa para tener una mayor tasa de utilización de recursos. Es relevante acotar que para poder realizar los cálculos de sobreagendamientos, se recurrirá al texto de gestión de operaciones *Pricing and revenue optimization* [13], específicamente el capítulo 9 que trata de esta temática.

4.3. Trabajo de título anterior

Se considera además como marco teórico el trabajo de memoria realizado por Bastián Elgueta [14], titulado “Desarrollo de un modelo predictivo para apoyar la gestión de la agenda en un centro médico”, en el cual se realizan modelos predictivos similares al que se desarrolla en el presente informe, con la diferencia que la institución en la cual se realiza es con fines de lucro (se le cobra a los pacientes), en donde no realizan transporte de pacientes, los médicos pueden establecer horarios donde no atienden y se realizan confirmaciones previas a las citas médicas.

Dadas las condiciones anteriores en aquella memoria se desarrolla un modelo que prediga los bloqueos de los médicos, con una granularidad diaria que se realiza en conjunto a las citas contiguas de cada profesional. Se asume que la probabilidad de bloqueo tiene mayor variación debido a la constante alteración de la agenda del profesional.

En cambio, el modelo de predicción de inasistencias considera dos instantes de citas: la creación y confirmación. Se balancean las bases de datos de entrenamiento del modelo con el algoritmo SMOTE con distintos métodos, *oversampling* y *undersampling*. Además, en los algoritmos se calcula el peso relativo de las clases. El algoritmo utilizado son redes neuronales y redes neuronales recurrentes, lo interesante de estas últimas es que permite entregar secuencias de eventos como input al modelo.

5. METODOLOGÍA

La metodología a seguir es la CRISP-DM [15] ya que es la principal técnica que se utiliza dentro de la industria del manejo de datos y también es la que se utiliza dentro de la institución. Esta contempla un ciclo de vida del proyecto que se divide en 6 fases principales, las cuales no son rígidas y pueden ser iterativas dependiendo de la necesidad de proyecto.

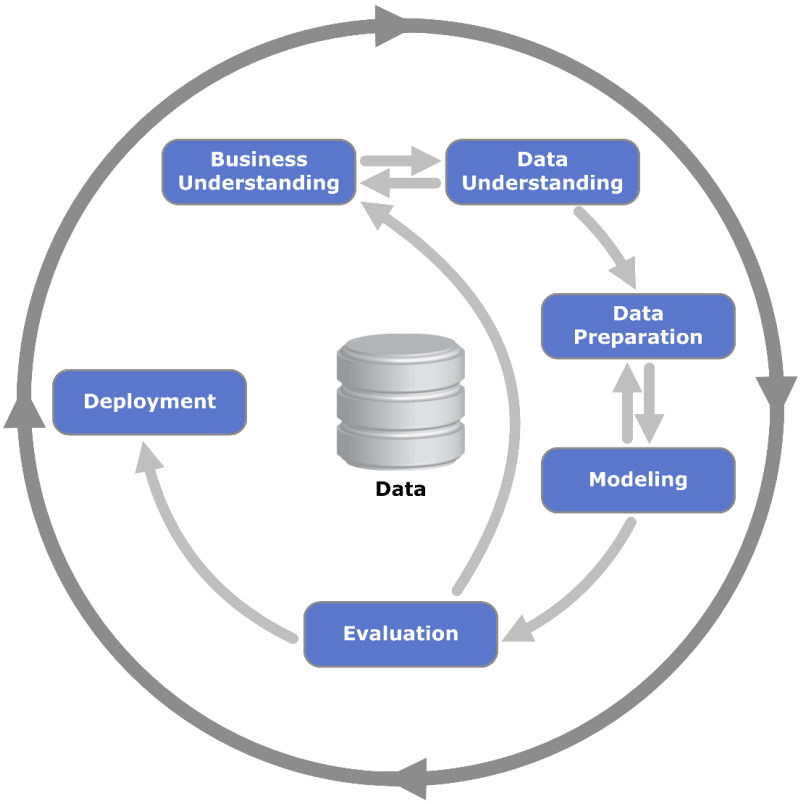


Figura 3: Metodología CRISP-DM.

En la fase inicial se buscó conocer del negocio al cuál se dedica la institución y de las necesidades de los pacientes, en este caso se analizaron las casuísticas correspondientes a las atenciones en los centros, respondiendo preguntas del tipo: cuáles son las sedes con mayor afluencia de público, cómo se realizan las atenciones a los pacientes, por qué los pacientes asisten a las mutualidades, etc. Lo anterior con el fin de interiorizarse y definir de mejor forma el problema.

Se realizaron reuniones con las personas encargadas de la rehabilitación en ACHS, para profundizar el conocimiento específico de las Terapias físicas, entendiendo cuáles son las

casuísticas por las que un paciente es derivado a kinesiología, cómo se receta tal tratamiento, la perspectiva de la institución respecto a las inasistencias, etc.

Además de tener reuniones con el área de consultoría interna para poder conocer el levantamiento de información realizado en años anteriores, dado que se reestructuró completamente el área de rehabilitación, con el fin de tener una mayor eficiencia organizacional.

La siguiente fase fue analizar e interiorizarse en las bases de datos que posee la institución, con la intención de identificar variables interesantes para el darle solución al problema, evaluar problemas de calidad, limpieza de datos, y generar posibles cruces. Para lo anterior se solicita en primera instancia un set de datos que contengan variables tanto del paciente (sexo, edad, tipo de trabajo, residencia), del siniestro (tipo de accidente, días perdidos, fecha de declaración, diagnóstico) y de las citas médicas (fecha de atención, asistencia o inasistencia, lugar de la cita, etc), estas se visualizaron con el fin de poder obtener claridad sobre la calidad de los datos.

La tercera fase corresponde a solucionar los problemas identificados en la fase anterior, eliminando variables de poco interés como por ejemplo los diagnósticos médicos, los cuales eran campos de texto en donde los doctores escribían las características de las lesiones, se eliminaron registros de pacientes externos a la institución por falta de datos, se agruparon las categorías de isapres, cambio en el nombre de comunas, entre otras. Todo lo anterior con el fin de obtener una base de datos final y consolidada, considerando las variables creadas u nuevas variables solicitadas con las que se trabajarían los modelos finales.

En la cuarta fase se han generado distintos tipos de modelos en base al análisis exploratorio de los datos, seleccionando y aplicando técnicas que permitan alcanzar los objetivos propuestos en los ítems anteriores. Se tienen actualmente 8 modelos, que se diferencian principalmente en los algoritmos utilizados (Redes neuronales y XGBoost) y en las bases de entrenamiento utilizadas (desbalanceadas y balanceadas).

Finalmente, la quinta fase es la evaluación de los resultados obtenidos por los modelos en donde se identifica la efectividad y precisión de ellos, buscando tener un buen modelo que prediga los inasistentes a las citas médicas, siendo de utilidad para la institución.

La sexta fase de la metodología CRISP-DM no se aplicará ya que hace referencia al despliegue del modelo, en su reemplazo se realizarán recomendaciones y pasos futuros con los resultados obtenidos, dentro de los que destacan la posibilidad de sobreagendar.

6. DESARROLLO METODOLÓGICO

6.1. Entendimiento del negocio

La kinesiología es la ciencia que estudia los movimientos del cuerpo humano, ya sean biomecánicos, fisiológicos o psicodinámicos. Siendo su enfoque principal la adquisición de habilidades, rehabilitación y fisiología del deporte.

En este caso, la especialidad de la ACHS es la terapia física, una rama de la kinesiología enfocada a la rehabilitación de los pacientes que han sufrido algún tipo de accidente laboral o de trayecto. El principal centro de trauma de la institución es el Hospital del trabajador, el cual es uno de los más importantes de Latinoamérica.

Esta área actúa en forma coordinada con los demás profesionales, tales como traumatólogos, terapeutas ocupacionales y psicólogos; esto con el fin de que el paciente tenga una rehabilitación integral, apoyando su inserción laboral, familiar y social.

Para poder ser atendido, inicialmente el paciente comienza realizando una declaración de accidente laboral o accidente de trayecto en el centro ACHS más cercano. Al declarar lo anterior, se le agenda una hora de atención médica para ser evaluado y ser calificado ya sea positiva o negativamente. Esta calificación consiste en determinar si el accidente debe ser cubierto por el seguro laboral, o bien se encuentra fuera de cobertura. En el caso de que el médico determine que debe ser atendido por la mutual, puede ser derivado a traumatología o bien directamente derivado a la atención kinesiológica. Es importante destacar que el médico no es escogido por el paciente y es asignado según la disponibilidad existente, sin embargo si se respeta que los siguientes controles sean agendados con el mismo profesional tratante.

Las citas de terapia física son agendadas en “paquetes”, es decir el médico receta una cierta cantidad de sesiones a las cual el paciente debe asistir y luego estas son agendadas por el personal de recepción del centro médico, en horarios de Lunes a Viernes de 08:00 a 17:00, según la disponibilidad de los kinesiólogos. Cabe destacar que todas estas citaciones son agendadas desde el comienzo y no se envía ningún tipo de recordatorio a los pacientes ni tampoco se realiza alguna confirmación de cita médica. En ocasiones, cuando los pacientes tienen problemas físicos de movilidad se les asigna un transporte, los cuales generalmente son ambulancias, furgones o taxis, los cuales van a buscar al paciente a su hogar para

trasladarlo al centro médico y luego lo trasladan a su casa. De los anteriores, los únicos transportes que podrían ser comunes (de varias personas) son los furgones, sin embargo raramente operan de aquella forma, generalmente son para el paciente en particular.

Como la labor de la mutualidad es la recuperación de los trabajadores del país, a pesar de que las personas estén en tratamiento (y por tanto no asisten al trabajo) su salario no se ve modificado, ya que el seguro lo cubre y es la institución quien le paga a la empresa los días perdidos de trabajo del paciente. Además de lo anterior, cuando los pacientes son atendidos por la ACHS no deben pagar dinero en ninguna de las consultas.

La mutualidad tiene distintos tipos de especialidades de terapia física:

- Extremidad Superior
- Extremidad Inferior
- Columna y Lesionados Medulares
- Respiratorio
- Quemados y Cirugía Plástica
- Neurología
- Neurorrehabilitación
- Pacientes amputados
- Unidad del Dolor
- Biomecánica

Aunque lamentablemente estos no tienen representación en las bases de datos de la organización.

6.2. Descripción de los pacientes atendidos

Con la finalidad de comprender las características de los pacientes atendidos en el área de terapia física, se expone lo siguiente.

6.2.1 Edad

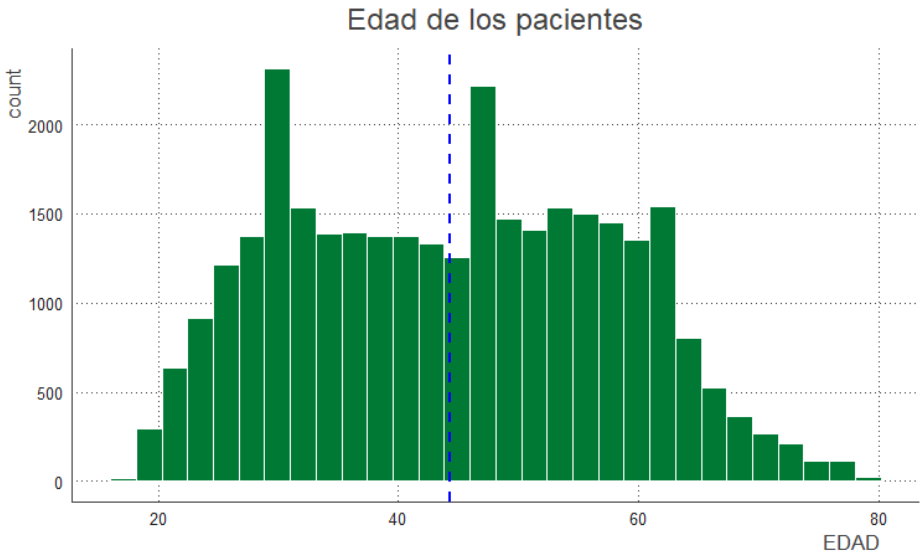


Figura 4: Histograma edad de los pacientes

Como se observa en la figura 4, la edad de los pacientes se distribuye de manera bastante homogénea entre los 30 y 60 años. Se evidencia una menor cantidad de pacientes menores de 30 años y mayores de 65 años. Lo anterior se presume por la poca accidentabilidad de trabajadores jóvenes y que la edad de jubilación es de 65 años para hombres y 63 para mujeres.

6.2.2 Género

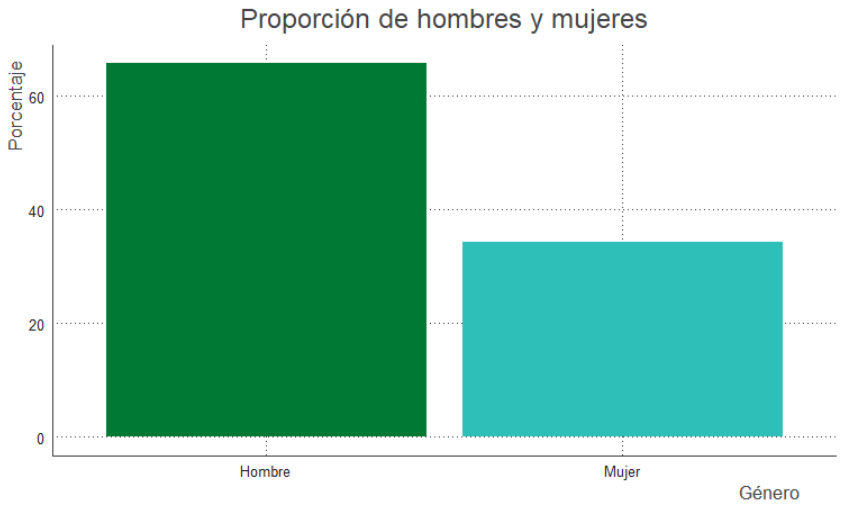


Figura 5: Porcentaje de hombres y mujeres atendidos

Se observa que el género predominante de accidentados son hombres bordeando un 60%, mientras que el porcentaje restante corresponde a mujeres.

6.2.3 Previsión

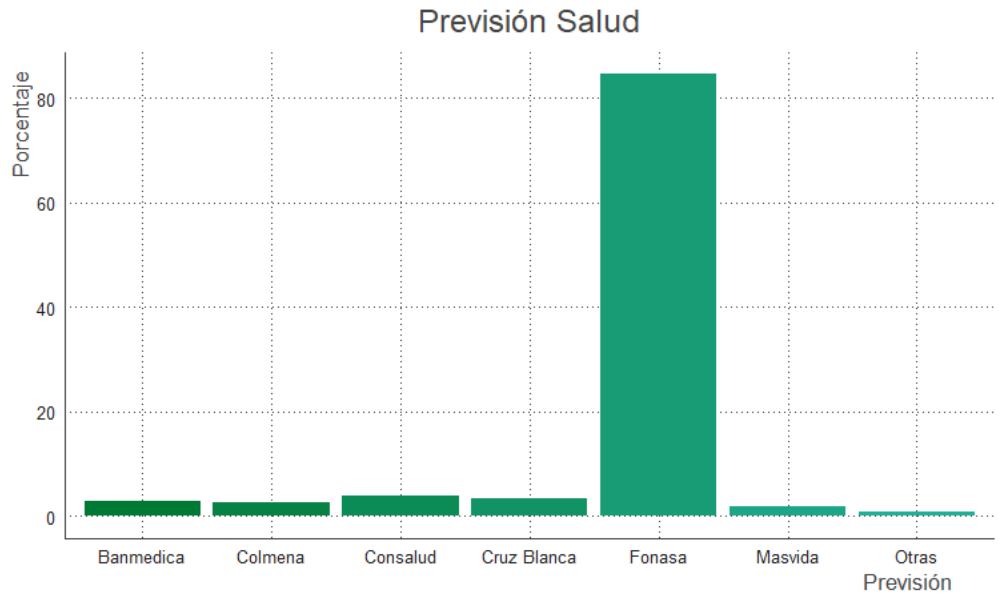


Figura 6: Previsión de salud de los pacientes

Respecto a la previsión de salud, hay una predominancia importante de FONASA (84%) dentro de los pacientes accidentados, seguido por Consalud con un 4%, Cruz blanca, Banmédica y Colmena. Es importante notar que el porcentaje de residentes en Chile que tienen cobertura en FONASA es del 77% [19], por lo que era esperable que este porcentaje sea alto.

6.2.4 Nacionalidad

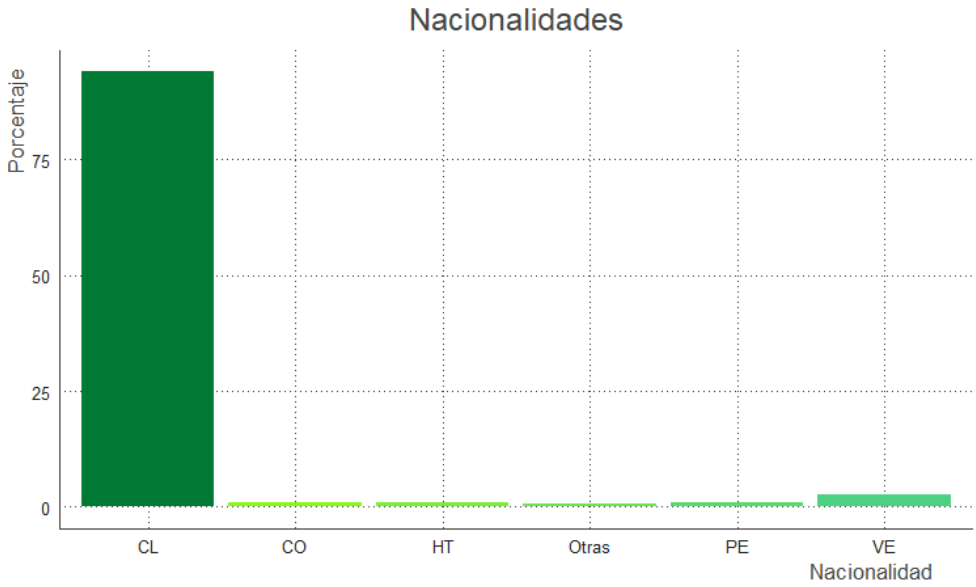


Figura 7: Nacionalidad de los pacientes

Como es esperado, la principal nacionalidad de los pacientes es chilena (94%), estando en segundo lugar la nacionalidad venezolana con apenas un 2%.

6.2.5 Región de residencia

La principal región de residencia de los pacientes accidentados es la RM de Santiago bordeando el 50%, seguida por la región del Bio Bio, Maule y Valparaiso. Es clave notar que estas son las regiones de residencia de los pacientes y no necesariamente es la región en donde se les atiende.

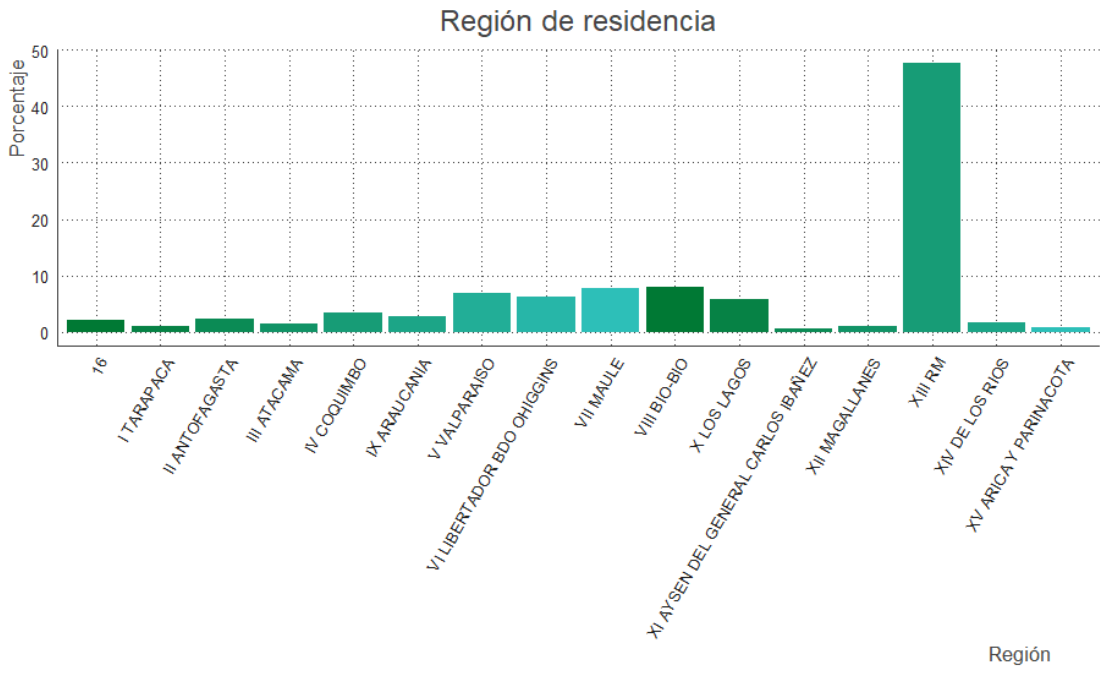


Figura 8: Región de residencia pacientes

6.2.6 Sector económico de las empresas de los trabajadores

Como se observa en la figura 9, el principal sector económico en donde trabajando los accidentados, es el del comercio y retail, seguido por el sector industrial, agrícola y la construcción.

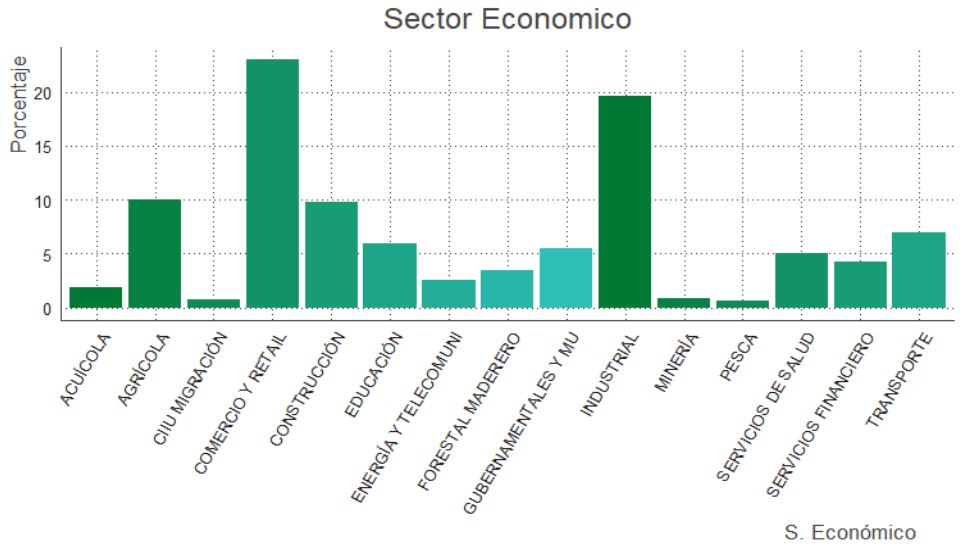


Figura 9: Sector económico de las empresas de los trabajadores

6.3. Entendimiento de la data

Como se explica en la sección anterior, se realiza una limpieza de los datos con el fin de poder trabajar adecuadamente con ellos y además se realiza un análisis exploratorio con el objetivo.

La base de datos con la que se está trabajando cuenta con 782.000 observaciones, correspondientes a citaciones médicas de pacientes exclusivamente de Terapia Física desde Enero de 2020 a Julio de 2021. Dentro de esta base de datos, se poseen las siguientes:

6.3.1 Variables

Variable	Tipo	Descripción	Comentarios
Episodio	Numérica	ID del episodio de la cita. Es único para cada citación	
Clase_episodio	Categórica	Indica si el episodio es ambulatorio u hospitalario	
Fecha_cita	Fecha	Fecha de la cita médica	Contiene fechas entre Enero de 2020 y Julio de 2021, solamente de Lunes a Viernes.
Hora_cita	Hora	Hora de la cita médica	
Unidad_organizativa	Categórica	ID de la unidad de terapia física del centro ACHS	
BP_paciente	Numérica	ID del paciente	
Fecha_nacimiento	Fecha	Fecha de nacimiento del paciente	Se opta por descartar esta variable y ocupar la edad del paciente, por términos prácticos.
Edad	Numérica	Edad del paciente	
Desc_genero	Categórica	Género del paciente (hombre/mujer)	
Comuna	Categórica	Comuna de residencia del paciente	
Región	Categórica	Región de residencia del paciente	
Des_estado_civil	Categórica	Estadio civil del paciente	
Nacionalidad	Categórica	Nacionalidad del paciente	
Sede_cita	Categórica	Sede de ACHS donde se realiza la atención	
Territorio_cita	Categórica	Agrupación de sedes ACHS según zonas geográficas	

Tabla 5: Variables de la base de datos

ID_siniestro	Numérica	ID del accidente	
Tipo_siniestro	Categórica	Indica si el tipo de siniestro es de Trabajo, Trayecto o Enfermedad Profesional	
Fecha_presentación	Fecha	Fecha de declaración del accidente en el centro ACHS	1301 NAs
Fecha_inicio_reposo	Fecha	Fecha en que se inicia el reposo del paciente	623274 NAs
Fecha_alta	Fecha	Fecha de alta médica del paciente	623407 NAs
Sector_económico	Categórica	Sector económico al que pertenece el paciente	2804 NAs
Cod_diagnostico_1	Categórica	Código del diagnóstico principal realizado por el médico	
Desc_diagnostico_1	Texto	Descripción del diagnóstico principal realizado por el médico	
Cod_diagnostico_2	Categórica	Código del diagnóstico secundario realizado por el médico	64423 NAs
Desc_diagnostico_2	Texto	Descripción del diagnóstico secundario realizado por el médico	64423 NAs
Cod_diagnostico_3	Categórica	Código del diagnóstico terciario realizado por el médico	135724 NAs
Desc_diagnostico_3	Texto	Descripción del diagnóstico terciario realizado por el médico	135724 NAs
Nombre_aseguradora	Categórica	Nombre de la previsión del paciente	
Inicio_transporte	Categórica	Indica si el paciente posee transporte asignado	
Marca	Categórica	Indica si el paciente asiste a la cita o no	

Tabla 6: Variables de la base de datos

6.3.2 Descripción del agendamiento de las citas médicas

Como se está trabajando con las citas médicas, se considera relevante hacer un análisis de estas y su distribución temporal.

Respecto a los agendamientos de citas según el día de la semana, se observa que se agendan una mayor cantidad de citas médicas los días Lunes y Miércoles, siendo el día con menor cantidad de citas agendada el día martes. Es clave mencionar que los centros solo atienden de lunes a viernes, por lo que los fines de semana no hay citas agendadas.

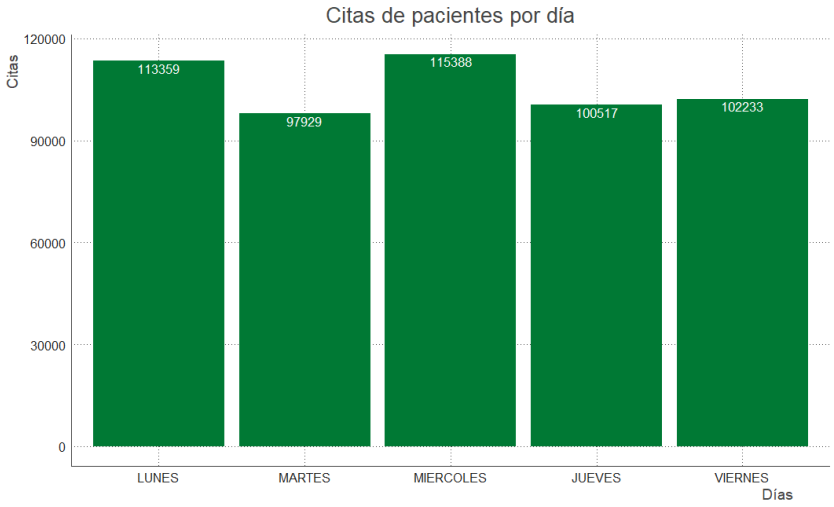


Figura 10: Cantidad total de citas agrupadas por día

Respecto a la cantidad de citas agendadas según la hora del día, se observa que claramente hay una preferencia al agendamiento de citas en el horario de la mañana, siendo el más agendado a las 10 de la mañana. Se evidencia también una clara disminución en el agendamiento en torno a las 13 hrs, probablemente porque este horario esté destinado al almuerzo de los funcionarios. Durante la tarde, la mayor cantidad de citas agendadas se produce en el bloque de las 15 hrs.

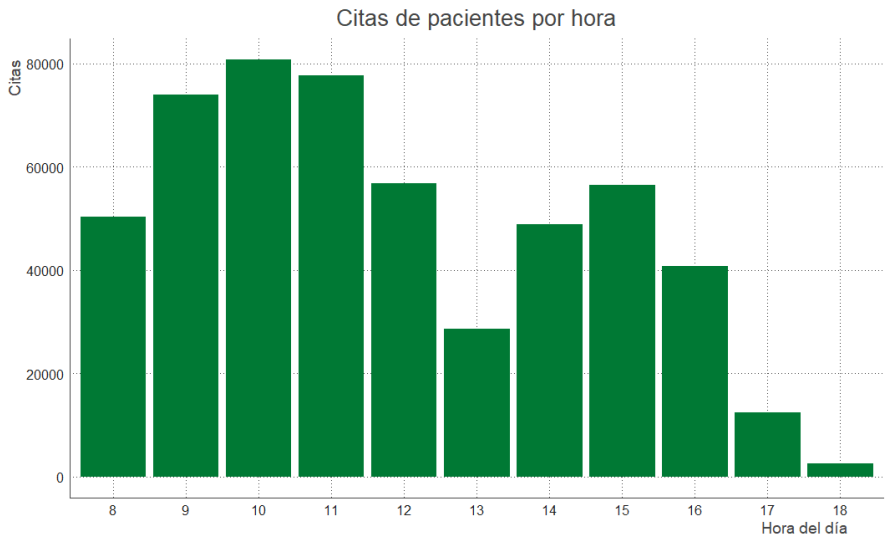


Figura 11: Cantidad total de citas agrupadas por horas

Es clave mencionar que al realizar el mismo análisis anterior pero esta vez incluyendo los diferentes días de la semana, no se evidencian diferencias sustanciales entre los días, el gráfico se encuentra en anexos.

6.3.3 Descripción de las variables más relevantes

Clase episodio

Variable categórica que indica si el episodio es ambulatorio, es decir, el paciente asiste al centro a su sesión u hospitalario, lo que implica que el paciente está dentro del recinto y no debe desplazarse para asistir a la sesión. Se tiene que el 99.5% de los episodios de la BD son ambulatorios y los restantes son hospitalarios. Por las características del problema que se está abordando, solo se trabajará con los casos de la categoría “ambulatorio”.

Hora cita

Indica la hora de la cita médica. Los horarios de las citas médicas se discretizan dejando solamente las horas. Se observa que se distribuyen entre 8:00 a 18:00, observándose 432 outliers a las 19 hrs y 1 outlier a las 23 hrs que fueron eliminados. Se observa además que hay una mayor cantidad de citas agendadas por la mañana, con una disminución importante en el horario de almuerzo (13 hrs).

Es importante mencionar que según la distribución probabilística de asistencia (figura 12) según la hora de la cita, pareciera existir una diferencia importante en las horas agendadas en las tardes, teniendo una mayor probabilidad de inasistencia.

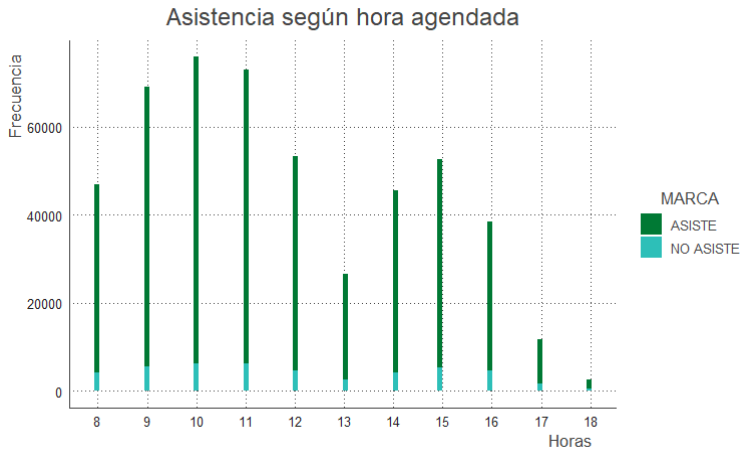


Figura 12: Asistencia según hora agendada

Edad

Variable que indica la edad de los pacientes. Para poder observar la distribución de edades de los pacientes se procede a realizar una segunda base de datos en donde se tiene solo una cita por paciente. Los datos indican que la edad promedio de los pacientes es de 44 años, teniendo un mínimo de 17 años y un máximo de 92 años, no se observan NAs. Se observa además que, si se compara esta variable con la asistencia a los centros, la gente de mayor edad podría tener una leve tendencia hacia asistir a su cita médica, respecto de la gente de menor edad. Para corroborar esto se realiza un análisis estadístico en donde se evidencia que la diferencia entre estos grupos es significativa, sin embargo al aplicar esta variable en los modelos, no resulta determinante para realizar una predicción.

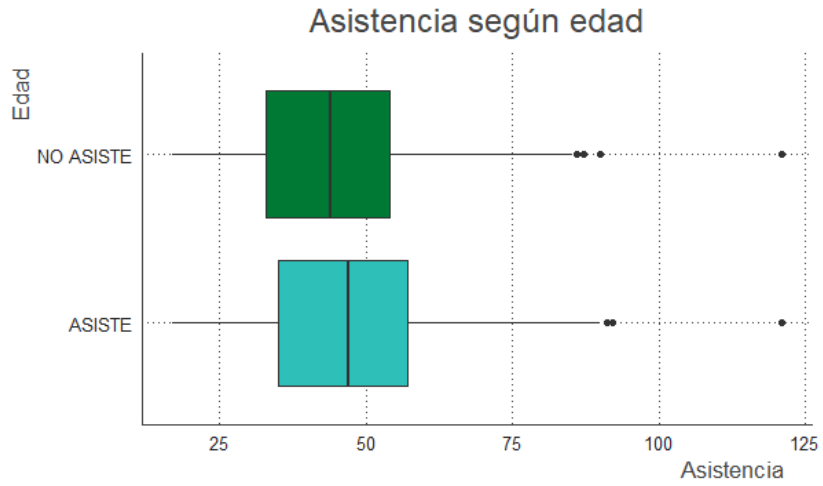


Figura 13: Asistencia según la edad del paciente

Desc genero

Variable que indica el género del paciente (hombre/mujer). Se evidencia que el porcentaje de hombres es de un 65,7%, mientras que las mujeres corresponden a un 34.2%. Respecto a la asistencia por género, en términos porcentuales se ven bastante similares por lo que no pareciera ser una variable de interés para el estudio.

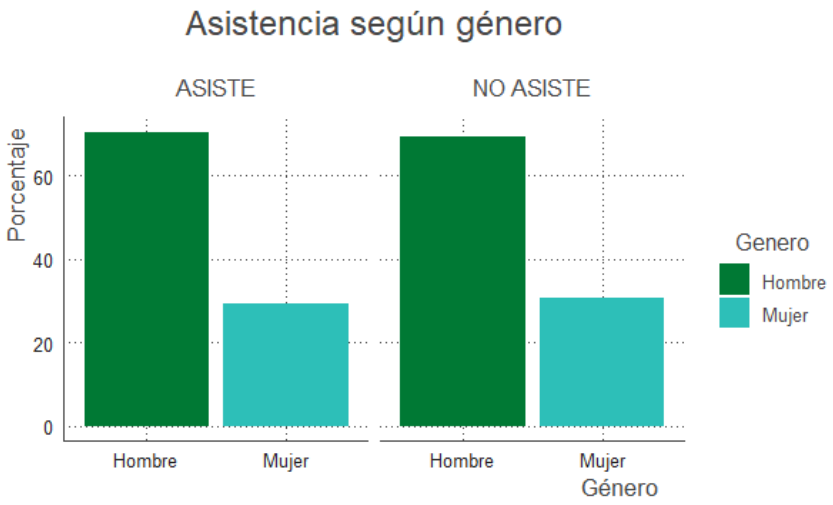


Figura 14: Asistencia según el género del paciente

Sede cita

Variable que indica la sede donde se realiza la cita. Se identifica variabilidad en la cantidad de citas médicas según la sede, lo cual es esperable ya que sus tamaños difieren, así como la cantidad de profesionales. La mayor cantidad de citas médicas la posee el Hospital del trabajador con 181718. Mientras que el promedio considera 9867 citas y una mediana de 4831 citas.

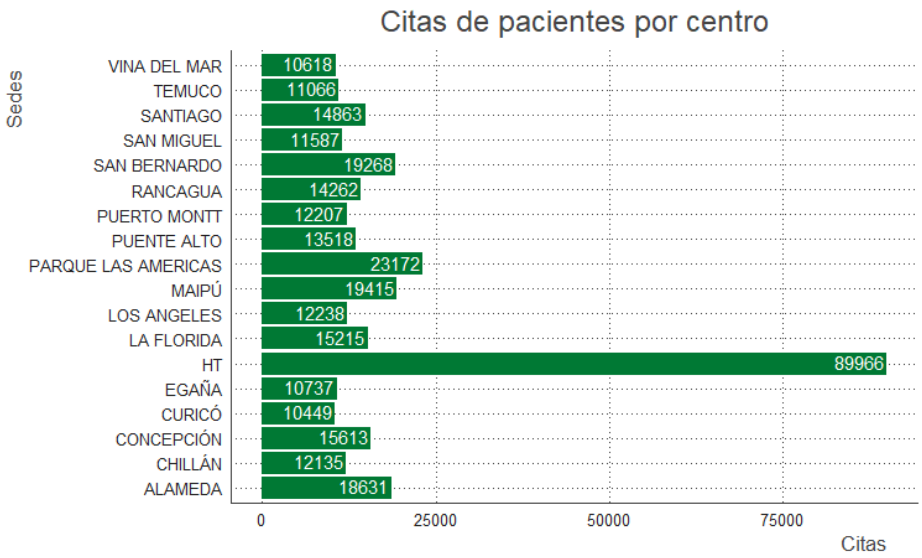


Figura 15: Citas totales agrupadas por centros médicos (centros con +6.000 citas)

Inicio Transporte

Esta variable indica si el paciente tuvo un traslado asignado, es decir si debía ser llevado desde su casa hacia el centro de atención y viceversa. Se observa que el 52.2% de las citas médicas tienen indicación de transporte, mientras que el resto no lo tiene. Como se observa en la figura, es clave notar que dentro de los pacientes que no asisten a sus citas médicas, un porcentaje mayoritario no tiene transporte asignado, por lo que se considera variable de interés.

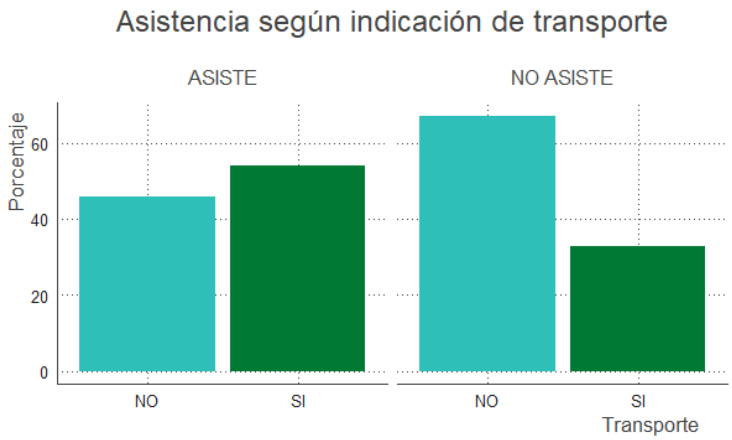


Figura 16: Asistencia a los centros según la indicación de transporte

Marca (Asistencia)

Variable que indica la asistencia de un paciente a su cita concertada de terapia física. No se presentan NAs y al ser una variable categórica no presenta outliers. Se tiene que se ha asistido al 91.4% de las citas mientras las restantes son inasistencias.

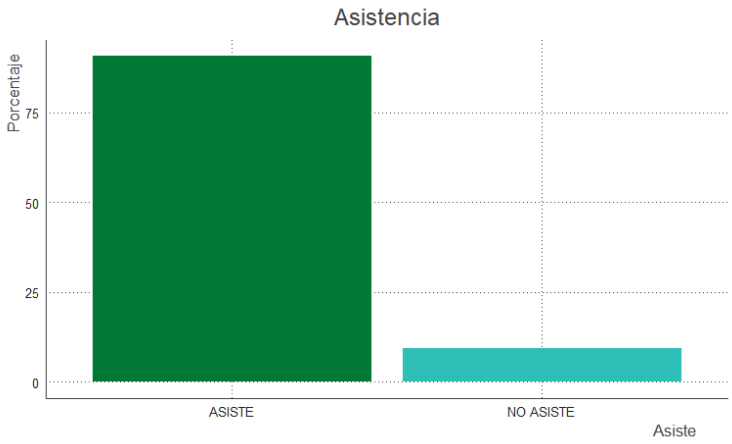


Figura 17: Asistencia a las citas médicas

6.4. Preparación de la data

6.4.1 Limpieza

Respecto a la limpieza de los datos, la principal eliminación corresponde a los pacientes de tipo ISL (Instituto de Seguridad Laboral), estos son pacientes externos a la institución, por lo que por este motivo la ACHS no posee datos ni de los pacientes, ni del tipo de accidente que

tuvo ya que solamente actúa como un prestador de servicios. Es por ello que poseen una gran cantidad de datos faltantes y se procede a eliminar de la base de datos.

La cantidad de registros eliminados corresponden al 16.4% del total de citas que existen en la base de datos, de todas formas es importante notar que antes de borrar los registros, se evaluó la proporción de asistencias que tenían estos pacientes en comparación con los pacientes ACHS y no se evidenciaron diferencias sustanciales (ambos tipos tenían una inasistencia cercana al 10%).

Por otro lado, también se eliminaron registros de citas médicas duplicadas por irregularidades en los tratamientos en el CRM SAP que ocupa la institución, en ocasiones hay citas médicas duplicadas, esto según lo investigado con el personal del negocio de la institución, no es correcto y por tanto se procede a eliminar una de las dos citas, sin privilegiar una por sobre la otra ya que son idénticas.

6.4.2 Creación de variables

En base a conversaciones realizadas con el área de negocios, se identifican diversos factores claves, que son diferenciadores del resto de las especialidades médicas. El primero es que las citas médicas de terapia física se paquetizan, es decir, se genera un plan de trabajo kinesiológico y no son citas individuales, lo que pudiera provocar que para los pacientes con menos citas restantes, tengan una menor probabilidad de asistencia, ya sea por cansancio o sentimiento de recuperación total; en otras palabras la ganancia marginal de cada cita de terapia de decreciente con el tiempo en la mayoría de los casos.

Con el *insight* anterior se generan las siguientes variables:

Días desde declarado el siniestro

Variable numérica que indica la cantidad de días que han pasado desde que se declaró el siniestro (primera vez que asistió a ACHS) hasta el día de la cita médica. Se realiza esta variable suponiendo que a mayor cantidad de días pasados desde que se declara el siniestro, habría una mayor probabilidad de inasistencia a la cita, sin embargo según lo evidenciado en el gráfico, esta intuición pareciera ser incorrecta, por lo que la variable queda descartada.

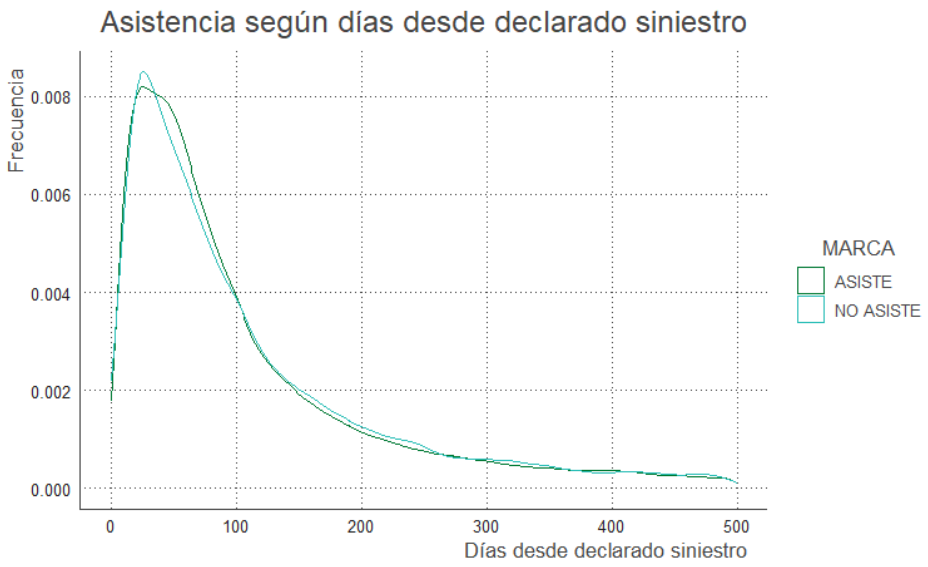


Figura 18:Asistencia según los días desde la declaración del siniestro en ACHS

Porcentaje de citas restantes para el paciente

Variable numérica que indica el porcentaje de citas restantes a las cuales el paciente debería asistir al centro médico respecto del total de citas agendadas para el siniestro (tratamiento). Se puede notar que los pacientes que no asisten a sus citaciones tienen en general un promedio de citas restantes menor, es decir, mientras menos citas les queden (respecto de su tratamiento) es probable que el paciente se sienta recuperado, por lo que tendría motivos para no asistir a su citación. Con el argumento anterior, se podría suponer que esta variable tiene una incidencia importante en la asistencia a las citas.

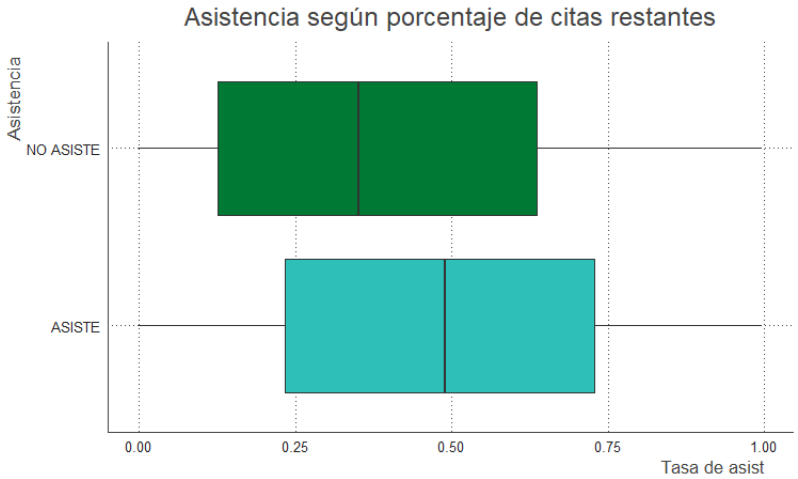


Figura 19: Asistencia según el porcentaje de citas restantes agendadas

Días desde la última cita

Variable numérica que indica la cantidad de días que han transcurrido desde que tuvo la cita de terapia física anterior hasta el día de la cita actual. Se plantea que esta característica puede ser relevante, ya que a medida que más pasa el tiempo desde la cita anterior, el paciente podría olvidar su atención agendada, haciendo que sea mucho más probable una inasistencia.

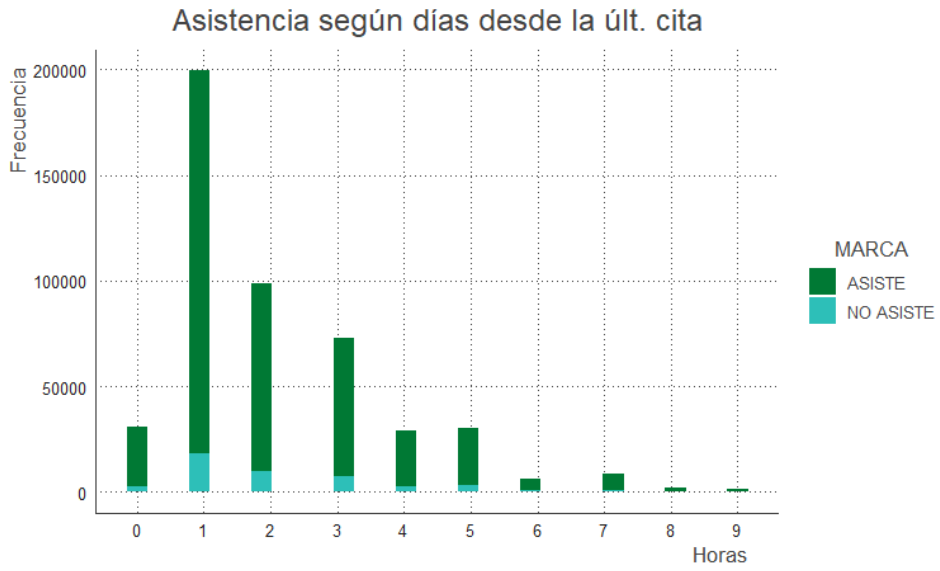


Figura 20: Distribución de la asistencia en base a la cantidad de días transcurridos desde la última cita médica

Casos Covid-19

Variable numérica que indica la cantidad de casos Covid-19 que fueron reportados el día de la cita médica. Se realiza porque se presumía una correlación positiva entre el porcentaje de inasistencias y la cantidad de los casos COVID, sin embargo no se observa tal relación a lo largo del tiempo. Es importante mencionar que si se evidencia un gran porcentaje de inasistencia en el comienzo de la pandemia por covid-19, en marzo/abril de 2020, que luego se normaliza. Estos datos al ser considerados "anormales" u *outliers* son eliminados y no considerados para el proceso de entrenamiento de los modelos.

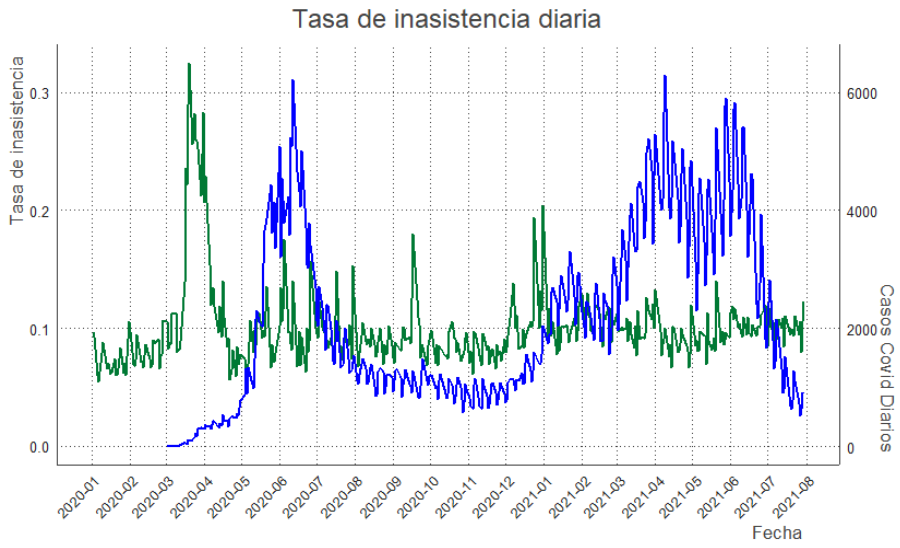


Figura 21: Tasa de inasistencia diaria y casos COVID reportados diariamente.

Temperaturas del día y precipitación

Con el fin de evaluar si la temperatura del día o precipitación del mismo pueda afectar la asistencia de los pacientes, se genera un subset de datos con información exclusiva de los centros ubicados en la Región metropolitana de Santiago. Los datos meteorológicos son obtenidos directamente de la estación de Pudahuel.

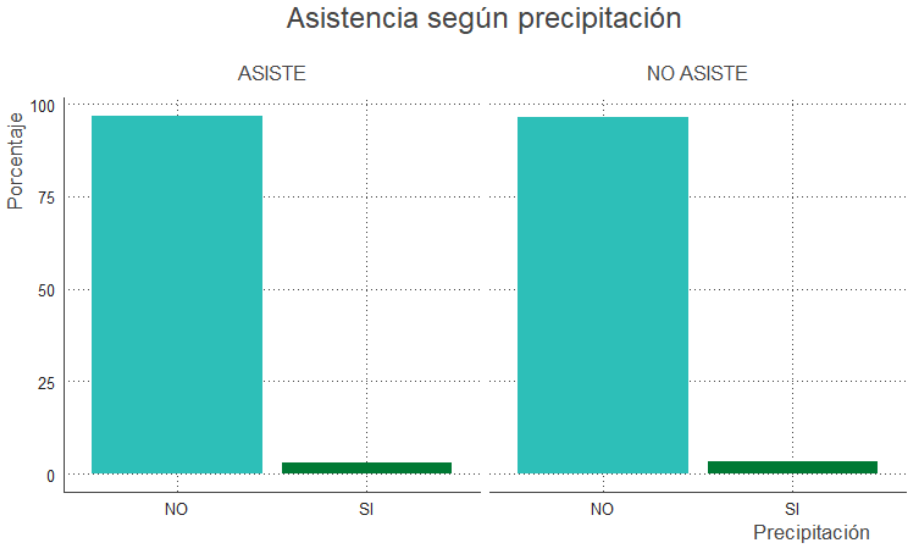


Figura 22: Asistencia a los centros en base a la variable precipitación

Según lo observado en los datos, al parecer la precipitación no sería un factor clave en las asistencias a las citas de terapia física, al menos en Santiago, ya que si bien existe una diferencia es bastante pequeña, cercana al 0.02%

Paciente vive en misma comuna del centro

Se genera una variable binaria que indica si el paciente vive en la misma comuna del centro de ACHS, con el fin de visualizar la cercanía del paciente al centro. Dado que solamente se poseen las comunas de residencia y también se tienen las coordenadas geográficas de los centros médicos, se opta poder discretizar, de modo que se ubicará un punto de referencia de la comuna del paciente en el medio de la misma, entonces si la sede de la cita médica está a una distancia menor de 5 km (radio promedio de una comuna), se etiqueta como que el paciente vive en la misma comuna del centro médico.

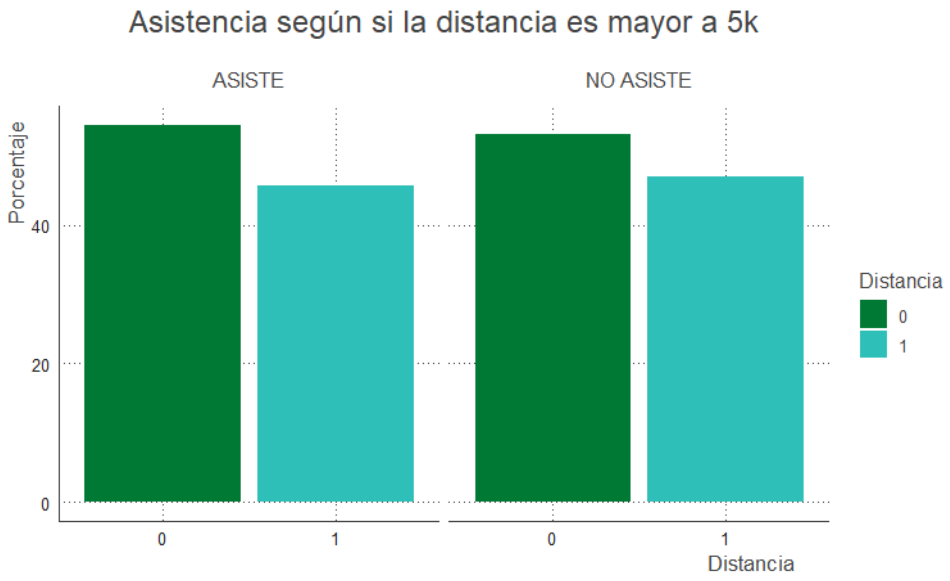


Figura 23: Asistencia según la distancia de la comuna del paciente a la sede ACHS

Sin embargo, a pesar de la intuición que se tenía en un comienzo, los datos no reflejan este pensamiento y como se observa en el gráfico, esta variable parecería ser no relevante para el análisis de asistencia de los pacientes.

Tasa de asistencia por paciente

Dado que se entiende que la asistencia de los pacientes a sus citaciones pudiera depender de su comportamiento anterior, se plantea generar una tasa de asistencia por cada paciente, que

se calcula como la cantidad de citas asistidas sobre la cantidad de citas totales agendadas. Los resultados de esta métrica se encuentran expuestos en la figura 18, en donde se observa que los pacientes que asisten tienen una tasa de asistencia histórica mayor.

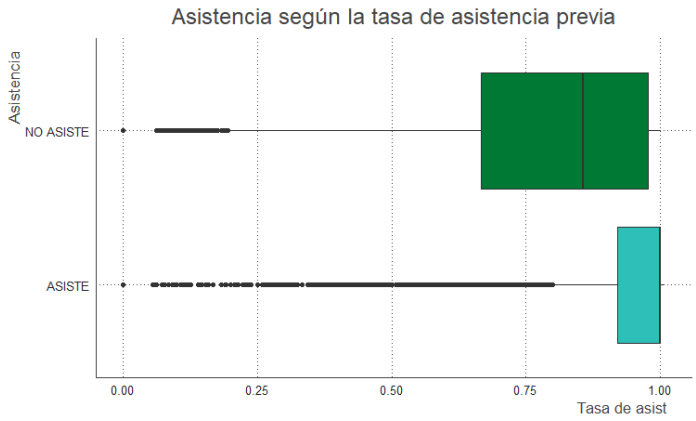


Figura 24: Asistencia de los pacientes según la tasa de asistencia previa

Si bien es cierto que se podría suponer que el comportamiento de los pacientes cambia a lo largo del tiempo (años) y que esta tasa de asistencia no considere el total histórico de las citas, en base a los datos que se están trabajando, no hace sentido hacerlo, ya que el 75% de los programas médicos agendados se cumplen en 60 días, siendo el promedio 19 días, lo anterior calculado como la diferencia entre la Fecha de la última cita del paciente y la primera cita del paciente.

6.4.3 Balanceo de datos

En base a la literatura analizada [18], gran parte de los trabajos de investigación de predicción de no shows que se han realizado hasta el momento demuestran que es conveniente tener una data balaceada (ya sea natural o sintéticamente) ya que fortalecen la capacidad de predicción de los modelos, sobre todo en los que tienen factores bajo el 15%. Por lo anterior se realiza un balanceo a través de distintos algoritmos, como ROSE (*Random Over Sampling Examples*) Y SMOTE (*Synthetic Minority Oversampling Technique*), que busca equilibrar la cantidad de inasistencia y asistencias de la base de datos, con la finalidad de que los algoritmos queden mejor entrenados y así tener una capacidad de predicción mayor sobre los casos inasistentes. Los modelos que fueron entrenados con balance de datos y los tipos de balanceo son especificados en la siguiente etapa.

6.5. Modelamiento

6.5.1 Variables de los modelos

Con la información obtenida del procesamiento de datos anterior, se establece que las variables que incluirán este primer modelo serán las siguientes:

Variable	Justificación
Marca	Variable a predecir
Inicio Transporte	Según lo evidenciado en la parte anterior, se observa que los pacientes que tienen indicación de transporte tienen una mayor tendencia a asistir a la terapia.
Porcentaje de citas restantes	Se visualizó que a menor porcentaje de citas restantes, mayor es la probabilidad de que un paciente no asista a su cita médica.
Hora cita	Dado que los pacientes tienden a tener una menor probabilidad de asistir durante las horas de la tarde, se considera una variable de interés.
Tasa asistencia histórica	Se observa que los pacientes que asisten tienen una tasa de asistencia histórica mayor que la de los pacientes que no asisten a sus citas médicas.

Tabla 7: Variables de los modelos

A pesar de que no se intuye una posible correlación entre las variables anteriormente mencionadas, se realiza un gráfico de correlación entre ellas. En él se observa que no están relacionadas y se previenen problemas de multicolinealidad.

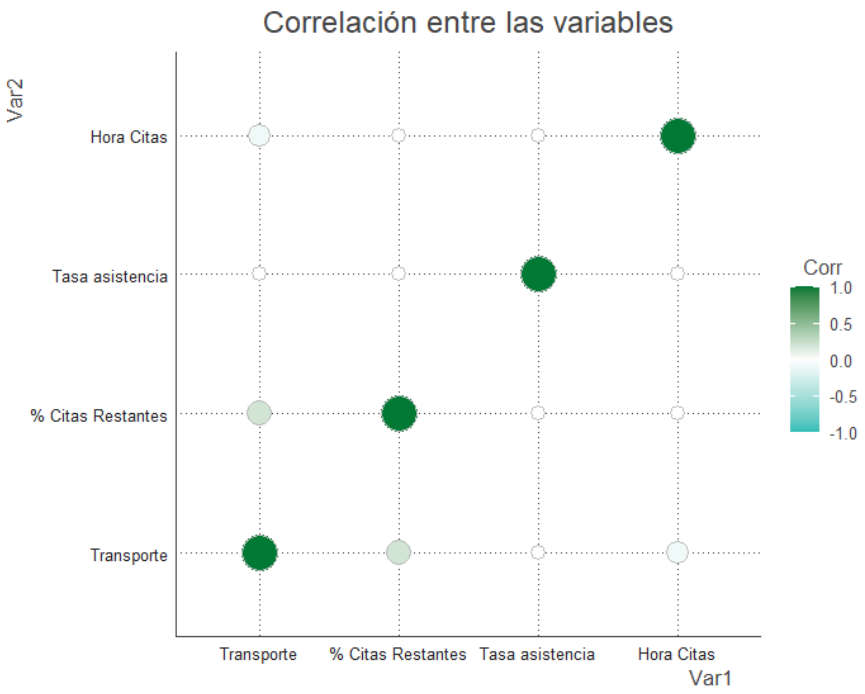


Figura 25: Correlación entre variables

6.5.2 Consideraciones del modelo planteado

Como las citas médicas de Terapia física son agendadas en una sola ocasión, no conviene realizar la predicción de la asistencia a estas citas de inmediato, si no que se plantea que el modelo se ejecute de manera semanal, de modo que pueda considerar las variables que van cambiando con el tiempo, como lo es la tasa de asistencia a las citaciones.

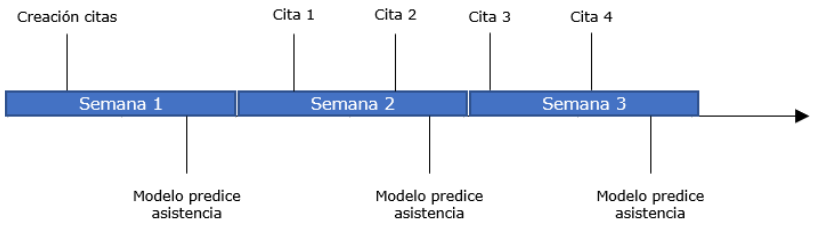


Figura 26: Diagrama funcionamiento modelo

Como se puede observar, está planificado para que se ejecute previo al final de la semana, con el objetivo de que se pueda planificar la semana entrante en base a la predicción del modelo, de modo que se pueda hacer sobreagendamiento cuando se considere necesario.

6.5.3 Tuning Parameters

Dado que se está trabajando con algoritmos, y los parámetros que se le indican a estos afectan directamente en los resultados de los modelos y dado que no se conoce de antemano cuales podrían ser los hiperparámetros más adecuados por modelo, se considera prudente probar con variaciones a modo de encontrar lo óptimo, esto está respaldado ampliamente literatura, considerado una práctica habitual.

Redes neuronales

En el caso de las redes neuronales se encuentran presentes 2 hiperparámetros: *size* y *decay*. El primero hace referencia a el número de nodos en la capa oculta y *decay* es un parámetro de regularización de los pesos. Los valores seleccionados para estas iteraciones corresponden a los siguientes:

Size	[1,10]
Decay	[0.1,0.1]

Tabla 8: Matriz hiperparámetros redes neuronales

XGBoost

Para el caso de XGboost se realiza el *tuning* con sus 2 parámetros principales *max Depth*, que indica la cantidad de profundidad o número de nodos de bifurcación de los árboles de decisión usados en el entrenamiento y también el ETA, que es la tasa de aprendizaje del modelo. Respecto al número de iteraciones que se realizarán antes de detener el proceso de ajusto del modelo, se escogió que fueran 1000.

En base a los siguientes valores se armó una matriz, para así escoger los modelos óptimos:

Max.depths	[2,8]
ETA	[0.1,0.8]

Tabla 9: Matriz hiperparámetros XGBoost

6.5.4 Particionamiento de la data

Para particionar los datos y así generar las bases de datos de entrenamiento, testeo y validación de los modelos se utilizó la lógica con la que trabajará el mismo, es decir se particionaron los datos en base a fecha, suponiendo que los datos desde el 1 de enero de 2020 hasta el 1 de abril de 2021 eran datos solamente para entrenar los modelos. Luego los datos

del 01 de abril de 2021 hasta el 01 de junio de 2021 fueron datos para validar los modelos entrenados con la base de datos de entrenamiento. Finalmente, la base de datos de validación corresponde al intervalo de tiempo que abarca desde 01 de junio de 2021 hasta el 31 de julio de 2021, siendo esta exclusiva para la validación del overbooking planteado en base a los modelos realizados.

Figura 27: Diagrama partición base de datos

Como se puede observar en la figura 27, la proporción entre las bases de datos planteadas, corresponde a 80%, 10% y 10% para entrenamiento, testeo y validación respectivamente.

6.5.5 Modelos

Dado que las variables parecieran ser lo suficientemente explicativas, las únicas variaciones entre los distintos modelos son el tipo de datos con los que se entrenan los mismos. Se tienen 3 tipos: balanceo con *undersampling* (ROSE), balanceo con *oversampling* (ROSE) y balanceo con *oversampling* (SMOTE).

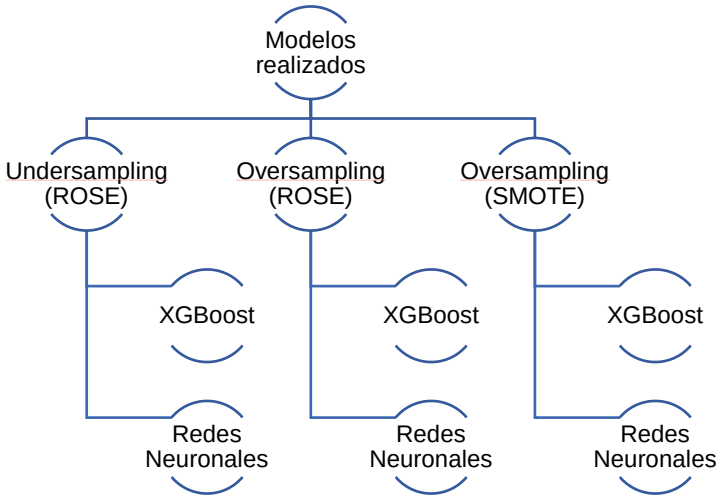


Figura 28: Modelos realizados

Modelo 1: Datos balanceados con *undersampling* (ROSE)

Para estos modelos se realiza un balance de la variable a predecir (asistencia), estableciendo que a proporción de asistencia e inasistencia sea de 50% cada una. Lo anterior se lleva a cabo

a través del algoritmo ROSE reduciendo la cantidad de datos de asistencia, la cual era la categoría con mayor proporción (90%). Lo anterior se lleva a cabo solamente con la base de entrenamiento, ya que la base de testeo no se modifica, esto porque así sería la realidad a la que se enfrentaría el modelo.

Redes neuronales

En este modelo, el número de neuronas en la capa oculta (*size*) óptimo es el de 7, con un *decay* de 0.4. Fue entrenado con un preprocesamiento de datos centrado y escalado.

Matriz de confusión:

Predicción		Real	
		No asiste	Asiste
	No asiste	2869	11153
	Asiste	1546	33252

Tabla 10: Matriz de confusión redes neuronales balanceadas con undersampling

XGBoost

Para este modelo los hiperparámetros óptimos en base a la matriz realizada, resultó ser una profundidad de los árboles de decisión (*max.depth*) equivalente a 3 y una tasa de aprendizaje del modelo (eta) correspondiente a 0,1.

Matriz de confusión:

Predicción		Real	
		No asiste	Asiste
	No asiste	2802	10788
	Asiste	1613	33617

Tabla 11: Matriz de confusión XGBoost balanceado con undersampling

Modelo 2: Datos balanceados con *oversampling* (ROSE)

Para estos modelos se realiza un balance de la variable a predecir (asistencia), estableciendo que la proporción de asistencia e inasistencia sea de 50% cada una, a través del algoritmo ROSE aumentando la cantidad de datos de inasistencia en base a los ya existentes, esto porque era la categoría con menor proporción (10%). Al igual que el caso anterior, se realiza solamente con los casos de entrenamiento, teniendo el set de testeo original.

Se utilizan los mismos algoritmos del caso anterior:

Redes neuronales

En este modelo, el número de neuronas en la capa oculta (*size*) óptimo es el de 3, con un *decay* de 0.5. Fue entrenado con un preprocesamiento de datos centrado y escalado.

Predicción		Real	
		No asiste	Asiste
	No asiste	2867	11094
	Asiste	1548	33311

Tabla 12: Matriz de confusión red neuronal balanceada con oversampling (ROSE)

XGBoost

Para este modelo los hiperparámetros óptimos en base a la matriz realizada, resultó ser una profundidad de los árboles de decisión (*max.depth*) equivalente a 5 y una tasa de aprendizaje del modelo (eta) correspondiente a 0,2.

Matriz de confusión:

Predicción		Real	
		No asiste	Asiste
	No asiste	2856	11341
	Asiste	1559	33064

Tabla 13: Matriz de confusión XGBoost balanceado con oversampling (ROSE)

Modelo 3: Datos balanceados con *oversampling* (SMOTE)

Se realiza un balanceo de datos a través del algoritmo SMOTE generando datos sintéticos de la clase minoritaria, en este caso de las inasistencias. Logrando generar una base de datos de entrenamiento con alrededor de 340.000 observaciones, equilibradas en proporción 50/50.

Se utilizan los mismos algoritmos del caso anterior, con los siguientes resultados:

Redes neuronales

En este modelo, el número de neuronas en la capa oculta (*size*) óptimo es el de 6, con un *decay* de 0.3. Fue entrenado con un preprocesamiento de datos centrado y escalado.

Predicción		Real	
		No asiste	Asiste
	No asiste	3094	14019
	Asiste	1321	30386

Tabla 14: Matriz de confusión red neuronal balanceada con oversampling (SMOTE)

XGBoost

Para este modelo los hiperparámetros óptimos en base a la matriz realizada, resultó ser una profundidad de los árboles de decisión (*max.depth*) equivalente a 2 y una tasa de aprendizaje del modelo (eta) correspondiente a 0,1.

		Real	
Predicción		No asiste	Asiste
	No asiste	2866	11199
	Asiste	1549	33206

Tabla 15: Matriz de confusión XGBoost balanceado con oversampling (SMOTE)

6.6. Evaluación y comparación de los modelos

En esta sección se evalúan y comparan los modelos, poniendo especial énfasis en su capacidad de predicción de inasistencias.

6.6.1 Métricas de precisión

Para poder evaluar y comparar los modelos/algoritmos, se calculan las métricas *precision*, *recall*, *accuracy*, F1 score y el posible dinero ahorrado. La tabla con los errores asociados a los modelos se encuentra adjunta en el anexo 10.10

El dinero ahorrado se entiende como el valor promedio del costo de la inasistencia ($13.000) multiplicado por la cantidad total de inasistencias predichas correctamente. Es importante notar que todas estas métricas están calculadas en base a las matrices de confusión expuestas en la sección anterior, en donde los modelos fueron testeados en una base de datos desbalanceada, con datos de citas médicas reales que fueron realizadas entre abril y junio de 2021.

Modelo	Algoritmo	Precision	Recall	Accuracy	F1	Posible dinero ahorrado
Undersampling (ROSE)	Redes Neuronales	0,205	0,650	0,740	0,311	$186.485.000
	XGBoost	0,206	0,635	0,746	0,311	$182.130.000
Oversampling (ROSE)	Redes Neuronales	0,205	0,649	0,741	0,312	$186.355.000
	XGBoost	0,201	0,647	0,736	0,307	$185.640.000
Oversampling(SMOTE)	Redes Neuronales	0,181	0,701	0,686	0,287	$201.110.000
	XGBoost	0,204	0,649	0,739	0,310	$186.290.000

Tabla 16: Métricas y comparación de modelos

Si se observan los resultados del primer tipo de modelo, que fue entrenado con una base balanceada con *undersampling* de datos, se observa que Redes neuronales tiene un *recall* superior respecto de XGBoost, bordeando el 65%, por lo que el porcentaje de casos inasistentes que clasifica correctamente el modelo es un poco mejor. Por consiguiente el posible dinero ahorrado de cara a la institución es mayor, siendo este uno de los principales

enfoques de la memoria. Respecto al *accuracy* de los modelos de *undersampling* es mejor el de XGBoost dada la mejor clasificación en los casos asistentes.

En los modelos balanceados con *oversampling* mediante el algoritmo ROSE, se evidencia que redes neuronales es el modelo que más dinero ahorraría a la institución, ya que es el que tiene una mayor certeza a la hora de predecir las inasistencias de los pacientes. Su capacidad de predicción de asistencias también es aceptable, logrando un *accuracy* del 74.1% siendo la segunda más alta. Es importante notar que en este tipo de entrenamiento los modelos obtienen una métrica *recall* mucho mayor que los modelos entrenados con *undersampling*.

Si se ocupa la técnica de *oversampling* con el algoritmo SMOTE, se evidencia un incremento en el Recall respecto de los demás modelos entrenados con otros tipos de balanceo. El caso de las redes neuronales es el que alcanza una mayor precisión a la hora de clasificar los casos inasistentes con un 70,1%, siendo el mejor de todos los modelos medidos por esta variable. El modelo entrenado con el algoritmo XGBoost mediante esta técnica presenta resultados similares a los modelos comentados anteriormente. Notar que en estos casos disminuye el *accuracy*, sin embargo no se considera de gran relevancia al ser un caso desbalanceado, poniendo énfasis en el caso de menor recurrencia (inasistencias).

6.6.2 Curva ROC y AUC

Es sabido que además de las métricas de precisión de los modelos existen otros métodos de compararlos, uno de aquellos métodos es la curva ROC o Característica Operativa del Receptor (por sus siglas en inglés) y su área debajo de la curva (AUC). Esta curva corresponde a representación gráfica de la entre la razón de los verdaderos positivos versus los falsos negativos. Respecto a este gráfico, como se observará más adelante se traza una diagonal desde la esquina inferior izquierda a la superior derecha, esta recta representa un clasificador aleatorio, es decir que tiene un 50% de probabilidades de clasificar cualquiera sea de los casos, para este trabajo vendría siendo asistencia o inasistencia. Se considera un mejor modelo a medida que la curva del clasificador se acerque a la parte superior del gráfico, generando una mayor área bajo la curva.

Dado que los resultados de las curvas, son en general muy similares (se puede observar en anexos) y están todos en torno al 76% del área bajo la curva, se hará mención únicamente al

modelo que destacó de la sección anterior, el que fue entrenado con redes neuronales balanceado mediante el método de *oversampling* SMOTE.

Como se observa en la figura 28, la curva ROC del modelo en cuestión, presenta un área del 76%, siendo la mayor en comparación a los demás algoritmos. Se evidencia que la curva está por sobre la recta, lo que indiciaria que las predicciones realizadas serían mejores que un clasificador completamente aleatorio.

Por el simple motivo de que estas curvas son similares en los demás modelos, no se considera como un factor diferenciador para poder identificar el mejor modelo.

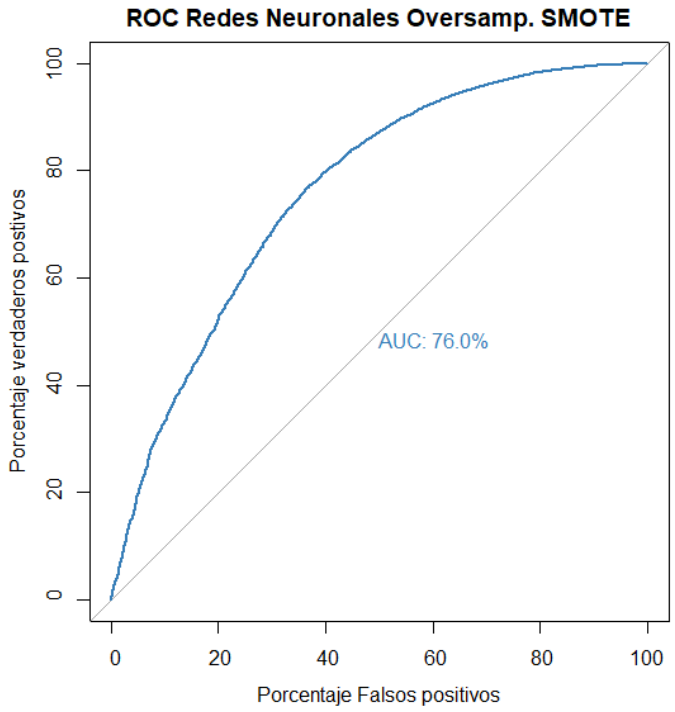


Figura 29: Curva ROC Redes neuronales Oversampling SMOTE

6.6.3 Importancia de las variables

Luego de haber comparado las métricas y curvas ROC, se plantea la necesidad de conocer la importancia o contribución de cada variable agregada a los modelos, con el fin de saber qué factor es el más relevante a la hora de predecir la inasistencia de un paciente.

En base a lo anterior se utiliza el algoritmo de Garson aplicado a las redes neuronales, que permite utilizar los "pesos" que conectan las variables en la red neuronal para poder hacer un análogo a los coeficientes utilizados en las regresiones lineales. Lo que realiza el algoritmo

es calcular la importancia relativa entre las variables de entrada en base a la salida del modelo, esta es determinada en base a las conexiones entre los nodos de la red.

Este proceso se realiza solamente sobre el modelo que fue entrenado con la base de datos balanceada con SMOTE, ya que es el de mejor desempeño y los resultados se observan en la figura 29.

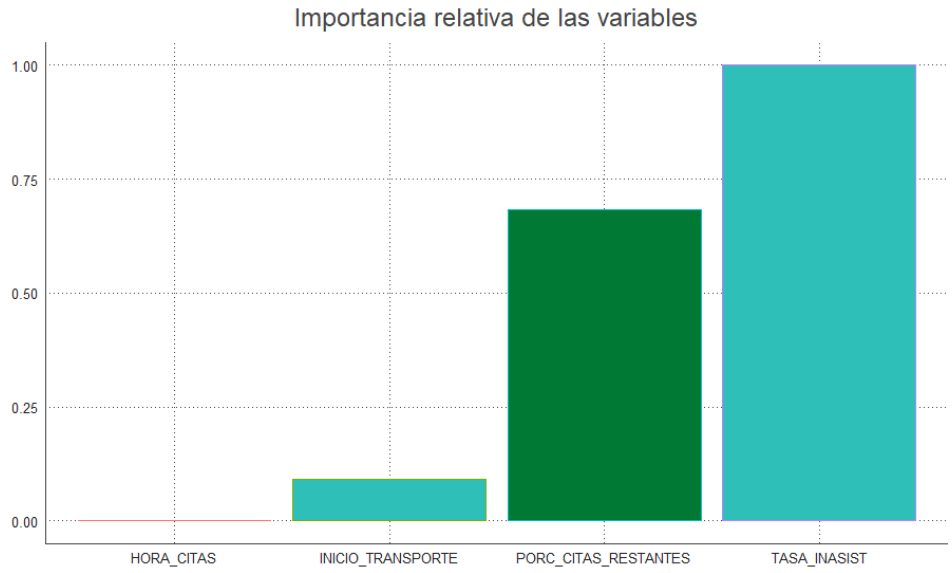


Figura 30: Importancia relativa de las variables modelo redes neuronales oversampling SMOTE.

Como se evidencia en el gráfico, el eje vertical tiene un rango entre 0 y 1, en donde a medida que se tiene un mayor valor, mayor es la importancia de la variable en el momento de la predicción. Dado lo anterior, es notable la importancia que toma la tasa de inasistencia histórica del paciente, lo cual es muy esperable ya que se entiende que la asistencia a una cita médica depende principalmente de la decisión de la persona en asistir y la importancia que esta le da a su tratamiento. Es por esto que hace sentido que las otras variables tengan una importancia relativa menor.

Respecto al porcentaje de citas restantes, también hace sentido que tenga una importancia relativa mayor a la asignación de transporte y la hora de la cita médica, porque se podría interpretar como un tipo de indicador de progreso en el tratamiento del paciente, de modo que se espera que a medida que el tratamiento está más avanzado, la ganancia marginal para el paciente por cada cita médica disminuye, porque no observa/siente un cambio sustancial.

La asignación de transporte que aparece en 3er lugar de importancia, tal vez sería esperable que tuviera un poco más de relevancia, ya que se pensaba que si se ofrecen más comodidades para que el paciente se pueda desplazar al centro médico, es posible que tenga una mayor asistencia, de igual modo que si no se ofrece podría esperarse que tenga una mayor probabilidad a no asistir. Sin embargo lo que indican los datos dista de esta suposición y se interpreta como que si bien es importante, depende mucho más del comportamiento del paciente y de su tratamiento la asistencia a la cita médica.

Finalmente, respecto a la hora de la cita médica que es la variable que menos relevancia tiene según el algoritmo, se le da una interpretación similar a la que se había mencionado anteriormente, los factores externos a los pacientes tienen una menor importancia al momento de decidir la asistencia a su cita médica, tomando en cuenta que en general estas horas se agendan de común acuerdo entre la persona y el centro médico.

6.6.4 Elección del modelo

En base a los resultados expuestos en las secciones anteriores, el modelo elegido es el de Redes Neuronales con balanceo de datos con *Oversampling* SMOTE. Se elige este porque es el que tiene una mejor precisión en la predicción de las inasistencias (Recall), siendo esta un 70,1% lo cual permite utilizarlo como *input* para realizar overbooking, lo cual se expondrá en la siguiente sección.

6.7. Modelo Overbooking

Lo que se ha planteado desde el comienzo de la memoria es la posibilidad de hacer un sobreagendamiento en base a los resultados obtenidos a través del modelo, es precisamente esto lo que se quiere lograr en los siguientes párrafos.

Es importante recalcar nuevamente que se está analizando las citas de terapias físicas (kinesiológicas) de los pacientes las cuales son agendadas en el momento posterior a la indicación del médico, siendo designada una hora en común acuerdo con el paciente. Su predicción sería realizada dos veces por semana con el fin de poder prevenir posibles ausentismos y actuar sobre estos resultados.

Inicialmente se plantea realizar un *overbooking* sobre la capacidad de atención de los centros de la institución, lo cual es lógico porque cada sede tiene distinta dotación de personal, así

como capacidad de salas. Lo anterior sumado a que actualmente se encuentran limitantes de aforo en las salas de atención kinesiológica producto de la pandemia dificulta esta labor, por el simple motivo de que no se tiene registro sobre los aforos permitidos en los distintos centros de manera centralizada. Esto se considera relevante, ya que si se pretende hacer un mayor agendamiento, se debe cuidar que no sobrepasen los límites permitidos por la autoridad y así el sistema no colapse.

En base a lo anterior se cambia la forma de realizar el sobreagendamiento, tomando como referencia únicamente los pacientes agendados, esto bajo el supuesto que la cantidad de agendamientos realizados para cierto día, en cierta sede y en cierto periodo cumplen con los estándares de seguridad.

Según lo explicado, el *overbooking* que se estudia implementar corresponde a cubrir los no shows predichos por el modelo, aumentando la utilización de los recursos disponibles. De este modo, no se pretende realizar un sobreagendamiento buscando el 100% de la utilización, porque se entiende que hay más factores que afectan esta tasa y no son solamente la cantidad de pacientes inasistentes.

Ahora tomando en consideración los resultados del modelo seleccionado, se tiene que predice de correctamente el 70% de los casos no show, teniendo un *accuracy* general de un 68% aproximadamente. Dado el funcionamiento del algoritmo permite determinar cuántos son los pacientes que pueden ser inasistentes en un bloque horario y así sobreagendar estos.

En base a que la precisión del modelo para predecir los no asistentes corresponde a un 70% en la base de testeo, se plantea que la cantidad de no asistencias que indique el modelo sea ponderada por esta misma tasa, así en promedio se tendría una precisión mayor de predicción. Lo anterior porque en particular se considera relevante el número total de inasistentes por bloque horario en el caso de hacer un sobreagendamiento, no así el paciente en específico que no asistirá a esta cita médica, porque según lo comunicado por la institución no existe actualmente los medios para confirmar o cancelar las horas de estos pacientes y tampoco se planea implementar en el corto/mediano plazo.

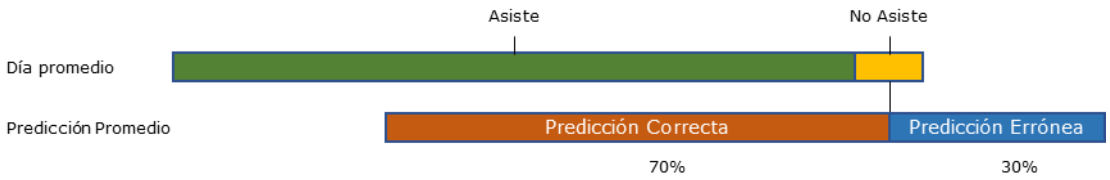


Figura 31: Diagrama de asistencia y predicciones.

6.7.1 Predicción Hospital del Trabajador

Tomando como base lo anterior, se realiza un testeo emulando el uso del modelo de predicción y su hipotético sobreagendamiento durante los meses de Junio y Julio de 2021. Para esto se extraen las citas de la base de datos de validación, específicamente del hospital del trabajador, esto porque es la sede que más citaciones tiene y si se quiere ver la validación del modelo hace sentido utilizarla. Por lo tanto se corre el modelo de redes neuronales para predecir el estado de asistencia de los pacientes y luego se corrige esta predicción multiplicándola por el ponderador que se mencionó en el párrafo anterior. Esto da como resultado lo siguiente

Fecha	Hora	Asiste Predicho	No asiste Predicho	Asiste Real	No Asiste Real	Diferencia No Asiste
03-06-2021	8	23	2	23	2	0
03-06-2021	9	26	4	28	2	2
03-06-2021	10	29	6	32	3	3
03-06-2021	11	31	5	32	4	1
03-06-2021	12	20	1	18	3	-2
03-06-2021	13	20	1	18	3	-2
03-06-2021	14	28	3	28	3	0
03-06-2021	15	23	6	26	3	3
03-06-2021	16	8	2	9	1	1
03-06-2021	17	3	0	2	1	-1
03-06-2021	18	0	0	0	0	0

Tabla 17: Resultados predichos por el modelo para el día 03/06/2021.

En la tabla 16 se observa un extracto de la predicción para el Hospital del trabajador, correspondiente al día 3 de junio, esto según el bloque horario al que pertenecen las citas médicas.

Descripción columnas de la tabla 16:

- Fecha: fecha de las citas médicas
- Hora: bloque horario de 1 hr de las atenciones. Por ejemplo “8” contempla desde las 8 am a 9 am.
- Asiste predicho: cantidad de asistencias predichas por el modelo sumado a las inasistencias multiplicados por 0,3.
- No asiste predicho: cantidad de asistencias predichas por el modelo multiplicadas por el ponderador 0,7.
- Asiste real: cantidad de asistencias reales
- No asiste real: cantidad de inasistencias reales
- Diferencia No asiste: resta entre las inasistencias predichas y las reales

De tal modo, la diferencia entre las inasistencias predichas y reales podría ser un buen indicador, para observar la precisión del *overbooking* que se está realizando, porque si el indicador es 0 esto quiere decir que el modelo acertó en cantidad de asistencias e inasistencias para esa hora en ese día. Si el resultado es negativo, indica que subestimó la cantidad de inasistencias, es decir se podrían haber agendado más personas. Por último si el resultado es positivo, esto quiere decir que se sobreestimó la cantidad de inasistencias por lo que llegarían más personas de las que estaba planificado atender.

Como es importante poder visualizar estos datos de manera clara, se genera el siguiente gráfico, en donde se observa en frecuencia relativa porcentual de la diferencia entre inasistencias predichas y reales. Es relevante aclarar que en este gráfico no se consideraron los casos en donde hay 0 atenciones para el bloque horario, porque esto podría sesgar los resultados.

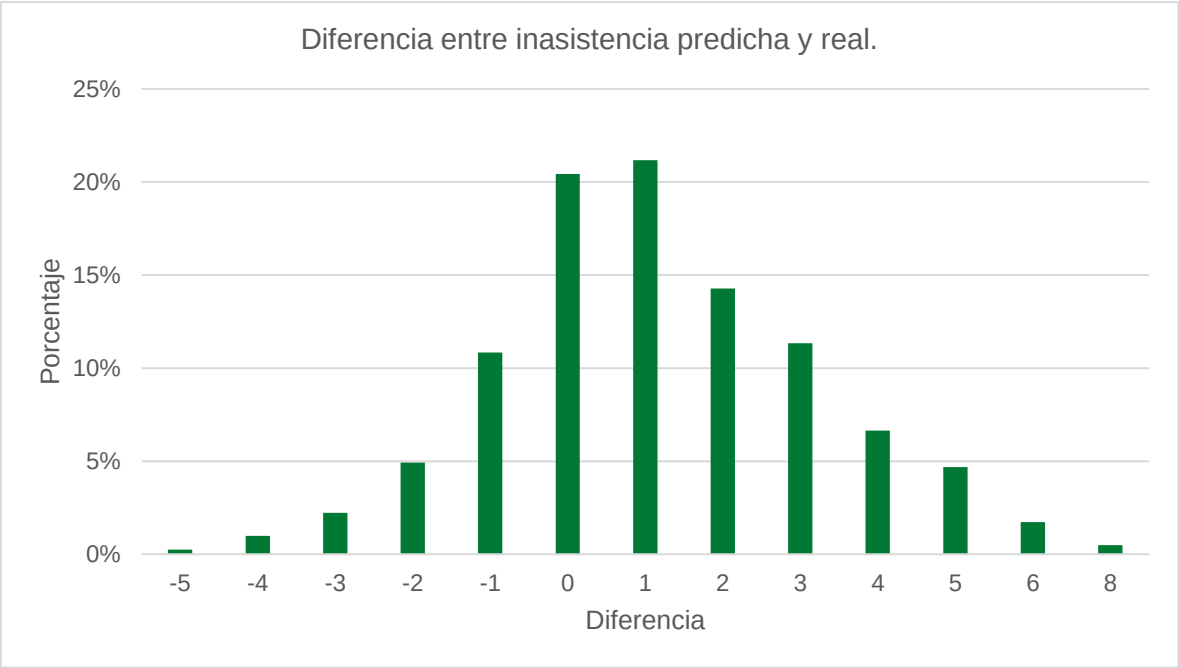


Figura 32: Diferencia entre predicciones inasistentes y realidad Hospital del Trabajador.

Se observa claramente que en porcentaje, la mayor frecuencia es el número 1 el 0 lo que indicaría que el modelo tiende a predecir de manera correcta en la mayor parte de los casos. También es importante notar que el 78% de los casos se encuentra concentrado entre el -1 y el 3 por lo que este seria unos de los principales rangos a considerar como factor de desajuste del modelo.

Al sobreagendar en demasía se corre el riesgo de que no existan suficientes profesionales para atender a los pacientes o que la capacidad del centro no sea lo suficientemente grande para albergar tal cantidad de pacientes. Es por esto que a pesar de que se cambió el método para realizar los sobreagendamientos, sería prudente tener la capacidad de cada centro para evaluar límites.

Al realizar un análisis de las medidas de tendencia central de este indicador se obtiene lo siguiente:

Min	1er quintil	Mediana	Media	3er quintil	Máx
-4	0	1	1.15	2	9

Tabla 18: Medidas de tendencia central de diferencia entre el predicción y realidad.

En la tabla 17 se observa que el mínimo observable es de -4, es decir en el caso extremo se subagendaron 4 citas médicas que podrían haberse utilizado para atender pacientes, por otro

lado en el máximo se evidencia que es 9, lo cual podría ser riesgoso ya que se podría colapsar el centro médico. La parte positiva de este análisis es que la mediana y la media tienden a ser 1, es decir en promedio el modelo se equivoca en predecir la cantidad de inasistentes en sólo 1 caso, lo que indica que es relativamente cercano a la realidad.

No está demás decir que a modo general el comportamiento del modelo en este caso tiende a sobreestimar las inasistencias, sin embargo también se entiende por parte del negocio que la única pérdida que tiene la institución respecto a los agendamientos es cuando se producen inasistencias, porque la ACHS al ser una mutual no gana por paciente atendido, si no que solamente puede perder utilización de recursos ante la falta de asistentes.

6.7.2 Predicción otras sedes

Dado que el caso anterior se realizó solamente para el Hospital del trabajador, se estudiarán las citas de las demás sedes agrupadas, ya que no se considera relevante estudiarlas una a una porque la cantidad de citaciones no es lo suficientemente grande. Es importante notar que esto se realiza solamente para evaluar el posible comportamiento de sobreagendamiento y no sería parte de un entregable para el uso de la institución.

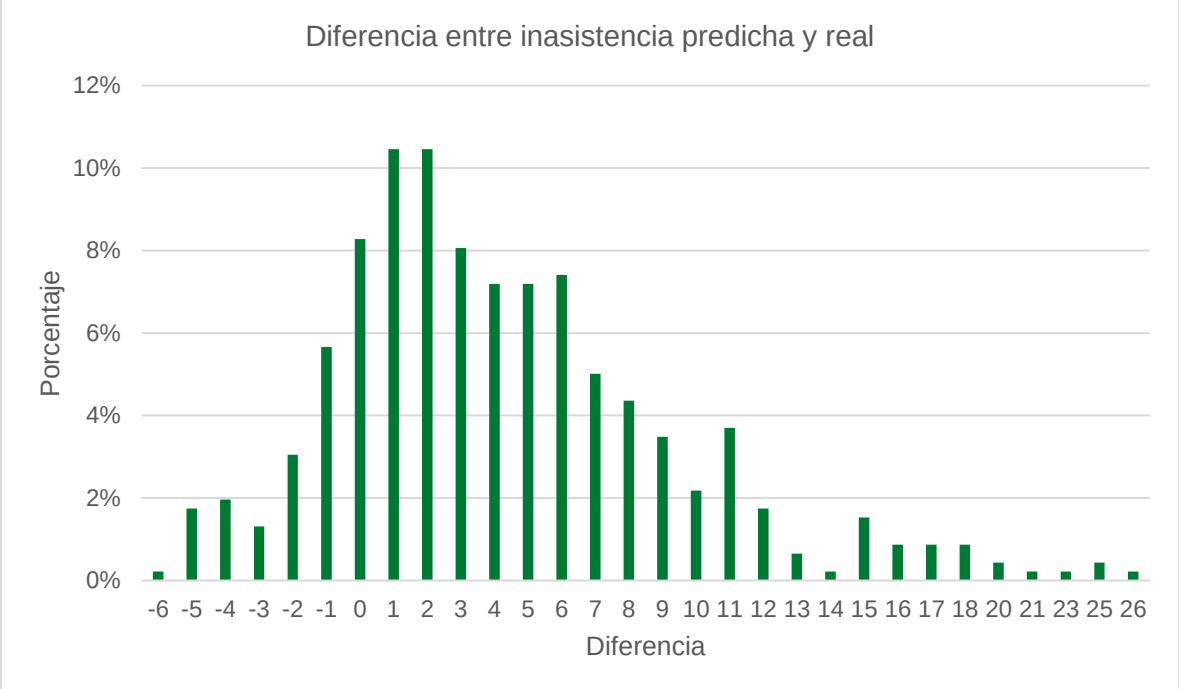


Figura 33: Diferencia entre predicción inasistencias y realidad otras sedes.

Se realizó el mismo tratamiento que en el análisis anterior y se observa una mayor dispersión en la precisión de la predicción de inasistencias, esto podría tener sus causas principalmente en que como la mayor cantidad de citas de la base de datos de entrenamiento corresponden al hospital del trabajador, es probable que haya un sobreajuste con esa sede, lo que permite que el modelo clasifique de mejor forma estos casos por sobre las sedes pequeñas. Esto se podría mitigar entrenando con una mayor cantidad de datos de sedes de regiones o bien creando muestras sintéticas como se realizó con el desbalance. Este último punto no se consideró dentro del trabajo de memoria, por el hecho de que ya se está ocupando un modelo con muestras sintéticas y acotar esa generación de datos podría haber sido contraproducente.

De todas formas, el modelo continúa siendo consistente, sobreagendando con mayor prioridad que subagendar pacientes, sin embargo claramente en este caso agrupado se escapa de los límites recomendados.

6.7.3 Análisis monetario overbooking

Resulta interesante evaluar económicamente la reducción de costos para la institución producto de la realización de *overbooking* en base a los modelos predictivos, ya que esta es una de la principal finalidad de este proyecto. Para lo anterior se tomará como referencia la sección 2.1, de donde se obtiene que el costo promedio de una inasistencia para la institución es de $13.000.

Para realizar este análisis monetario se propone acotarlo específicamente a la sede del Hospital del Trabajador, ya que en base a los resultados mostrados en la sección anterior, no sería prudente aplicarlo a otras sedes, por la falta de precisión del modelo. En este caso se supondrá que el 100% de las citas propuestas por el modelo para sobreagendar, fueron efectivamente agendadas y realizadas.

Prosiguiendo con lo anterior, para los meses de Junio y Julio de 2021, el modelo se corrió sobre 9.957 citaciones, de las cuales 9.013 fueron asistencias y 916 fueron inasistencias. Según lo predicho por el modelo, las asistencias hubieran sido 8572 y las inasistencias 1385, es decir hay una diferencia del 50%, lo cual puede sonar fuerte sin embargo como se observa en el gráfico de la sección anterior, no habría grandes complicaciones en el corto plazo dado que las terapias kinesiológicas no necesitan de la completa atención del personal en todo momento, de hecho se atienden 4 pacientes en promedio por profesional y en teoría se podría

llevar a cabo un overbooking de ese calibre. Sin embargo al largo plazo podría existir algún tipo de complicación por un posible agotamiento del personal médico, por lo que esto tema tendría que ser probado y luego ver los resultados en torno a la calidad de atención y desgaste del personal.

Dado lo anterior, suponiendo que de esas 1.385 citas predichas se hayan sobreagendado correctamente, se tiene que en ahorros para la institución serían del orden de los 18 millones de pesos en 2 meses, por consiguiente se hablaría de alrededor de 108 millones de pesos anualmente por concepto de sobreagendar cupos de terapia física.

7. CONCLUSIONES Y TRABAJO FUTURO

Existen múltiples investigaciones acerca de la predicción de estados finales de citaciones médicas con el fin de reducir los impactos económicos que pueden causar las inasistencias o no shows a las instituciones de salud, sin embargo hasta el momento no se había observado ningún trabajo realizado en torno a una mutualidad. Lo anterior es particularmente interesante, dado que el modelo de negocios es distinto a lo que podría ser una organización con fines de lucro, esto dado que la ACHS no tiene ganancias por pacientes atendidos, si no que solamente por trabajadores afiliados, de tal forma el único costo que pudiese provocar una inasistencia es la baja en la utilización de los recursos.

Para poder realizar este modelo predictivo se utilizó la metodología CRISP-DM a través del cual se analiza la perspectiva del negocio de la institución, la perspectiva de los pacientes, posteriormente se analizan y limpian los datos de la base obtenida, se seleccionan y visualizan las variables claves para posterior generar modelos de clasificación supervisados y evaluarlos. Finalmente se realiza una propuesta de modelo de *overbooking* en base a los resultados obtenidos por el modelo elegido.

Respecto al primer objetivo especifico, se cumplió en la fase del análisis de la información contenida en la base de datos, en donde se identificó que las principales variables claves que determinan la asistencia de un paciente a la cita médica son la tasa de inasistencia propia del paciente, el porcentaje de sesiones restantes del paciente y si el paciente tiene traslado o no al centro de atención.

El segundo objetivo específico también fue cumplido a cabalidad, realizándose 6 modelos de predicción de asistencias, con algoritmos de redes neuronales y XGBoost con distintos balanceos de data.

Respecto al modelo de predicción de inasistencias de las citas médicas elegido, corresponde a un algoritmo de Redes Neuronales entrenado con datos balaceados con *oversampling* a través de SMOTE, se observa que en base a los datos de testeo se alcanzó un Recall del 70% que se interpreta como el porcentaje con el cual el modelo clasifica correctamente los casos inasistentes sobre los clasificados como inasistentes, lo cual se considera positivo. Por otro lado, si se grafica la curva ROC del modelo se obtiene un área bajo la curva del 76%, siendo

este un buen indicador para clasificadores como el que se trabaja en la memoria. En relación a la importancia de las variables que contiene el modelo, se observa que en orden de importancia se encuentra la tasa de inasistencia histórica de los pacientes, el porcentaje de citas restantes que tiene el paciente respecto de su tratamiento, la asignación de transporte para el paciente y finalmente el bloque horario de la cita médica.

En línea con el tercer objetivo específico, el modelo de predicción se aplicó para simular un sobreagendamiento sobre las citas médicas de Junio y Julio de 2021 desglosado por hora, día y centro médico. Respecto a lo anterior se evidenció que el modelo predice de forma efectiva para las citas médicas del Hospital del trabajador, teniendo una diferencia entre lo real y lo predicho bastante bajo, siendo el promedio 1 cita médica sobreagendada de más. Es importante mencionar que en general el modelo predice de favoreciendo el sobreagendar más que el subagendamiento, lo cual hay que tenerlo en consideración.

El comportamiento del modelo para las citas médicas de las demás sedes no tiene un comportamiento tan bueno como para el Hospital Del Trabajador, sin embargo sus predicciones se concentran en torno a una diferencia entre real y predicho de 5 citaciones. Esto puede ser peligroso ya que en agencias ACHS que son pequeñas, hacer un overbooking de 5 citaciones puede ser riesgoso ya que no se podría cumplir con todas las atenciones.

En base a lo mencionado se recomienda que la institución utilice los modelos como marcha blanca de manera gradual en el Hospital del Trabajador, probando el funcionamiento práctico de los sobreagendamiento, verificando que la teoría se relacione con la práctica. Para las demás sedes de ACHS se recomienda entrenar el modelo con una mayor cantidad de citaciones médicas de esas sedes.

7.1. Trabajo futuro

Como recomendación para un trabajo futuro se plantea la utilización de variables que indiquen el tiempo de desplazamiento de un paciente hacia el centro médico, lo que se estima influye directamente en las asistencias de los pacientes, esto según investigación realizada y también según intuición del negocio. Esto no se abordó en la memoria actual por la falta de datos facilitados por la institución.

Así mismo, también se recomienda fuertemente tener una mayor cantidad de datos por sedes de las instituciones a evaluar, esto porque permitiría segmentar modelos por sus características propias, ya sea de tamaño, capacidad, colaboradores, entre otros.

En el caso de que este tipo de modelos se pretenda utilizar en otro tipo de instituciones se debe considerar que este está diseñado y medido favoreciendo la predicción de inasistencias, por el tipo de modelo de negocios que tiene la ACHS, en el caso de que se quiera aplicar en alguna institución con fines de lucro, en donde se gana por paciente atendido se deberían considerar otras métricas de evaluación.

Por otro lado, si se quiere utilizar el modelo para una organización similar pero para una especialidad diferente, es necesario considerar que las citaciones de dicha especialidad sean del mismo tipo que las analizadas en este trabajo, es decir que sean sesiones que se agenden con anterioridad y que sean mayor a una.

8. BIBLIOGRAFÍA

[1] ACHS. (2019). Memoria Asociación Chilena de Seguridad 2019.

[2] ACHS. (2022). Entrevista con RRHH y Gestión de Terapia Física.

[3] Consultoría interna ACHS. (2020, diciembre). Nuevo modelo de rehabilitación integral ACHS. ACHS.

[4] Oracle. (s. f.). ¿Qué es la ciencia de datos? Oracle Chile. Recuperado 6 de junio de 2021, de https://www.oracle.com/cl/data-science/what-is-data-science/

[5] Wooldridge, J. (2008). Introductory Econometrics: A Modern Approach, 4th Edition (4.a ed.). South-Western Pub.

[6] Devasahay, S. R., Karpagam, S., & Ma, N. L. (2017). Predicting appointment misses in hospitals using data analytics. mHealth, 3, 12. https://doi.org/10.21037/mhealth.2017.03.03

[7] Innovation, A. (2021, 26 mayo). Qué son las redes neuronales y sus funciones. ATRIA Innovation. https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

[8] N. (2020, 15 julio). Algoritmo k-Nearest Neighbor. Aprende Machine Learning. https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

[9] Autor: Mg. Daniel Paredes Inilupu. (2020, 26 junio). Capítulo 10 Aprendizaje Supervisado | Data Science con R. Aprendizaje Supervisado. https://bookdown.org/dparedesi/data-science-con-r/aprendizaje-supervisado.html

[10] Parente, C. A., Salvatore, D., Gallo, G. M., & Cipollini, F. (2018). Using overbooking to manage no-shows in an Italian healthcare center. BMC Health Services Research, 18(1). https://doi.org/10.1186/s12913-018-2979-z

[11] Reid, M. W., May, F. P., Martinez, B., Cohen, S., Wang, H., Williams, D. L., & Spiegel, B. M. R. (2016). Preventing Endoscopy Clinic No-Shows: Prospective Validation of a Predictive Overbooking Model. American Journal of Gastroenterology, 111(9), 1267–1273. https://doi.org/10.1038/ajg.2016.269

[12] Reid MW, Cohen S, Wang H, Kaung A, Patel A, Tashjian V, Williams DL Jr, Martinez B, Spiegel BM. Preventing patient absenteeism: validation of a predictive overbooking model. Am J Manag Care.

[13] Phillips, R. L. (2005). Pricing and Revenue Optimization. Amsterdam University Press.

[14] Elgueta Gallardo, Bastián Ignacio;, B. I. E. G. (2021). Desarrollo de un modelo predictivo para apoyar la gestión de la agenda médica en un Centro Médico.

[15] Schröe C., Kruse F., & Marx J. (2020). A Systematic Literature Review on Applying CRISP-DM Process Model. https://doi.org/10.1016/j.procs.2021.01.199

[16] Davies, M. L. (2016, 16 febrero). Large-Scale No-Show Patterns and Distributions for Clinic Operational Research. PubMed. https://pubmed.ncbi.nlm.nih.gov/27417603/

[17] ACHS. (s. f.-b). Estadísticas de Gestión ACHS. Asociación Chilena de Seguridad. Recuperado 6 de junio de 2021, de https://www.achs.cl/portal/ACHS-Corporativo/EstadAchs/Paginas/estadisticas-ACHS.aspx

[18] Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., & Arribas-Gil, A. (2020). Patient No-Show Prediction: A Systematic Literature Review. Entropy, 22(6), 675. https://doi.org/10.3390/e22060675

[19] FONASA (2021). Cuenta pública participativa Fonasa 2021.

ANEXOS

ANEXO A: Definiciones

Días Perdidos

Se entiende por número de días perdidos aquellos en que el trabajador, conservando o no la calidad de tal, se encuentra temporalmente incapacitado debido a un accidente o enfermedad profesional, sujeto a pago de subsidio, sea que éste se pague o no.

Accidentes del trabajo

El total de lesiones, a causa o con ocasión del trabajo, y que le produzca muerte, incapacidad permanente o incapacidad con tiempo perdido (no incluye trayecto), ocurridos a los trabajadores protegidos, es decir, los trabajadores dependientes por quienes se declararon cotizaciones, se hayan pagado éstas o no, más los trabajadores independientes adheridos a una Mutualidad de Empleadores, siempre y cuando se encuentren al día en el pago de las cotizaciones previsionales

Enfermedad Profesional

Toda aquella enfermedad causada de una manera directa por el ejercicio de la profesión o el trabajo que realice una persona y que le produzca incapacidad temporal, permanente o muerte

Masa de trabajadores

Número de trabajadores protegidos por el seguro de la Ley N° 16.744

Tasa promedio de Siniestralidad Temporales

Es el promedio de las Tasas de Siniestralidad por Incapacidades Temporales de los Períodos de 12 meses considerados. Las Tasas de Siniestralidad por Incapacidades Temporales, como se muestra en el aviso, corresponden al cociente entre el total de Días Perdidos en un Período de 12 meses y el Promedio mensual de Trabajadores del mismo, multiplicado por cien y expresado con dos decimales

Tasa de Siniestralidad por Invalidez y Muerte

Es el valor que según la tabla del numeral ii en el punto 1, capítulo III, de la sección “B. Cotización adicional diferenciada” del “Compendio de Normas del Seguro Social de Accidentes del Trabajo y Enfermedades Profesionales”, corresponde al promedio de Factores

de Invalidez y Muerte considerados en el Período de Evaluación. Para el cálculo de la Tasa de Siniestralidad por Invalidez y Muerte, deberán considerarse las invalideces que sean iguales o superiores al 15%

Tasa de Siniestralidad Total

Corresponde a la suma de la Tasa Promedio de Siniestralidad Temporal y de la Tasa de Siniestralidad por Invalideces y Muertes

Tasa de cotización

La tasa de cotización pagada por cada uno de sus trabajadores de forma mensual para efectos del seguro de la Ley N° 16.744. Esta se compone por una tasa básica de 0,90%, una tasa adicional y una tasa de 0,03% por concepto del Seguro para el Acompañamiento de Niños y Niñas que padezcan enfermedades graves (Ley SANNA)

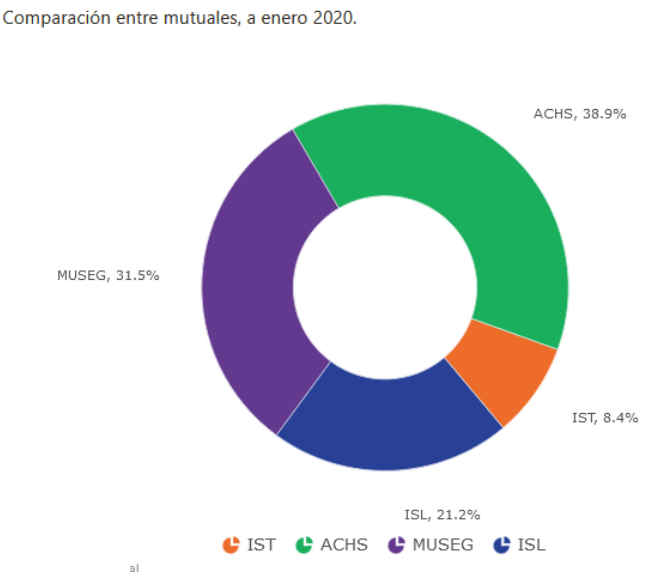
Tasa adicional

Se determina en función del riesgo presunto o efectivo que la empresa presenta según la actividad económica que desarrolla. La tasa de cotización por riesgo presunto (empresas nuevas, esto es, con menos de dos años consecutivos de actividad) se determina según lo establecido en el D.S. N° 110

Ley 16.744

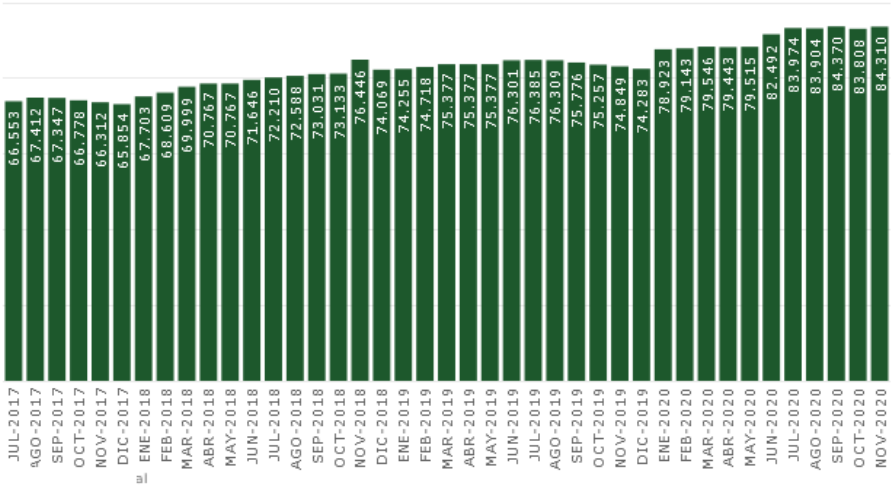
"La Ley N° 16.744 establece normas sobre Accidentes del Trabajo y Enfermedades Profesionales. Mediante esta ley se declara obligatorio el Seguro Social contra riesgos de Accidentes del Trabajo y Enfermedades Profesionales, y se establecen disposiciones para su aplicación." (SUSESO, Superintendencia de Seguridad Social, 2021)

ANEXO B: Participación de mercado



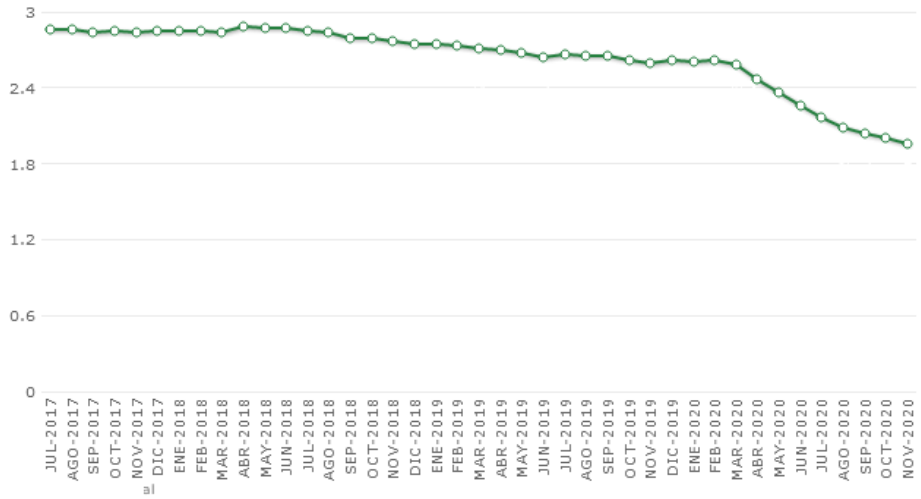
Fuente: Panorama Mensual Seguridad y Salud en el Trabajo, Superintendencia de Seguridad Social.

ANEXO C: Cantidad de empresas afiliadas



Fuente: ACHS [14]

ANEXO D: Tasa de Accidentabilidad con reposo por cada 100 trabajadores afiliados a la ACHS a lo largo del tiempo.



Fuente: ACHS [14]

ANEXO E: Cálculo de la tasa de cotización

Para una empresa “X”, su tasa de cotización se calcula de la siguiente manera:

Tasa de cotización total (X) = 0,93% + Tasa Adicional(x)

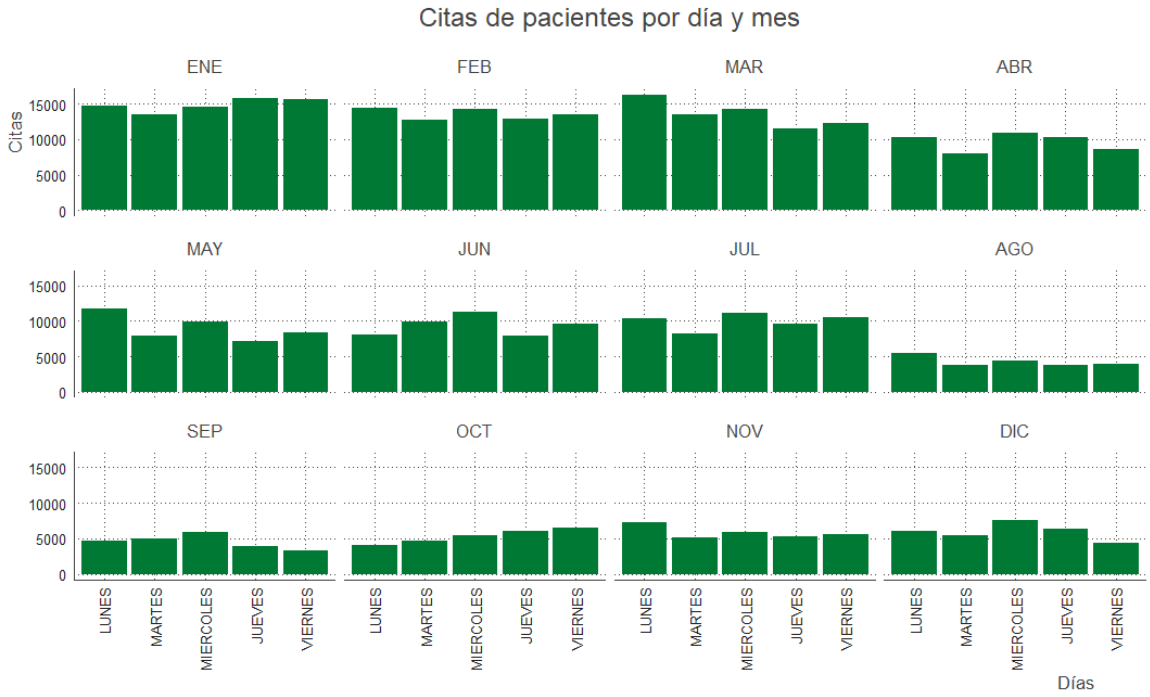
Donde la tasa adicional corresponde al valor asignado en la tabla abajo, disponible en el artículo N°5 del Decreto Supremo N°67 para un rango de tasa de Siniestralidad Total.

Tasa de Siniestralidad	Cotización Adicional (%)
0 a 32	0,00
33 a 64	0,34
65 a 96	0,68
97 a 128	1,02
129 a 160	1,36
161 a 192	1,70

193 a 224	2,04
225 a 272	2,38
273 a 320	2,72
321 a 368	3,06
369 a 416	3,40
417 a 464	3,74
465 a 512	4,08
513 a 560	4,42
561 a 630	4,76
631 a 700	5,10
701 a 770	5,44
771 a 840	5,78
841 a 910	6,12
911 a 980	6,46
980 y más	6,80

Fuente: ACHS

ANEXO F: Citaciones de pacientes según mes y día



ANEXO G: Validación diferencia estadística entre asistencia de hombres y mujeres.

Para realizar un test de diferencia estadístico, se proponen las siguientes hipótesis:

$$H_0: \mu_A = \mu_{NA}$$
$$H_a: \mu_A > \mu_{NA}$$

En donde se tiene que H_0 es la hipótesis nula que indica que no existe diferencia estadística entre la edad de los asistentes μ_A y de los no asistentes μ_{NA} y H_a es la hipótesis alternativa que indica que si hay diferencia significativa.

Se tiene que los asistentes son 450056 y los no asistentes son 46038. El promedio de edad de los asistentes es 46,06 años y de los no asistentes es 43,74 años. Con una desviación estándar de asistentes de 13,39 años y no asistentes de 13,22 años.

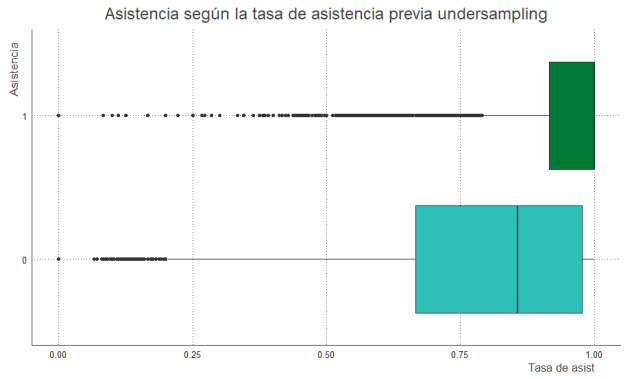
Entonces, dado lo anterior se tiene que usando el estadístico:

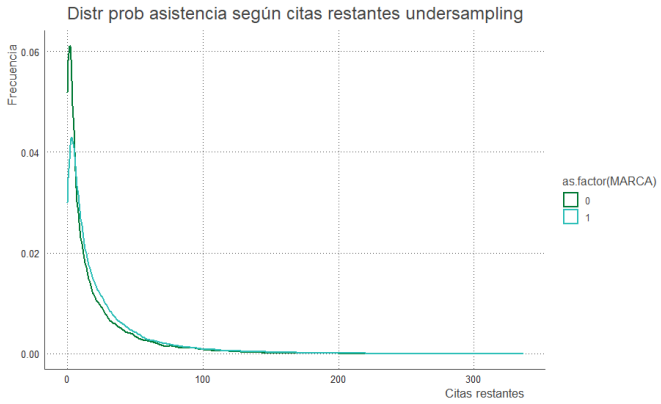
$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_{NA}^2}{n_{NA}}}}$$

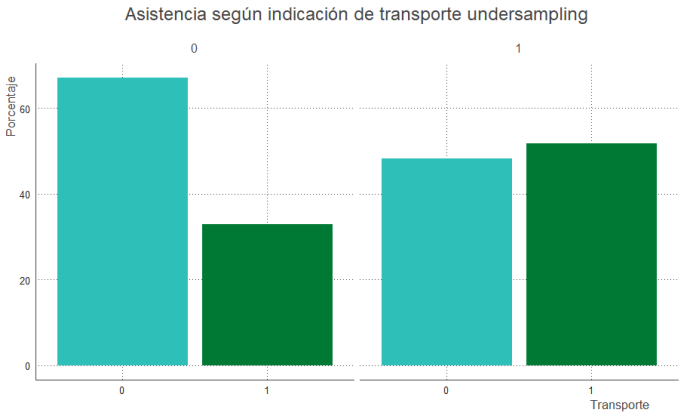
$$Z_p = \frac{46{,}06 - 43{,}74}{\sqrt{\frac{179{,}55}{450056} + \frac{174{,}88}{46038}}} = 35{,}80$$

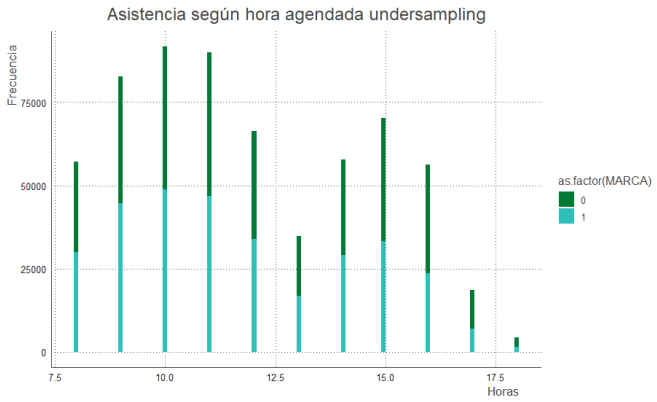
Se ocupa una significancia estadística al 5%, lo que implica que $Z_\alpha = 1{,}64$ por consiguiente como $Z_\alpha < Z_p$, se rechaza H_0lo que implica que si hay diferencias significativas entre los dos grupos y que el grupo que asiste a las sesiones tiene una mayor edad que el que no asiste.

ANEXO H: Validación de la base de datos undersampling.

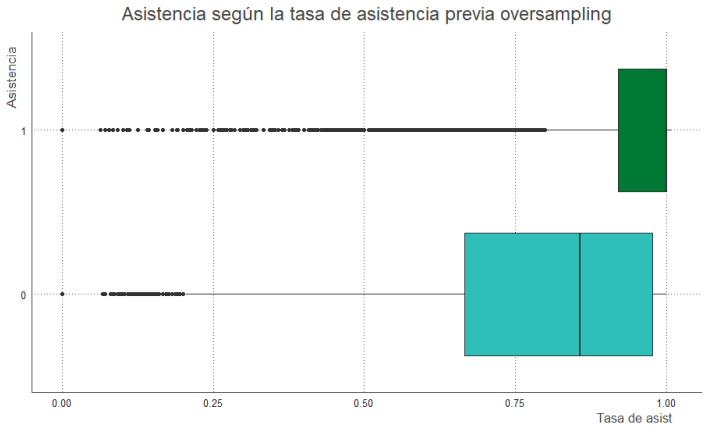


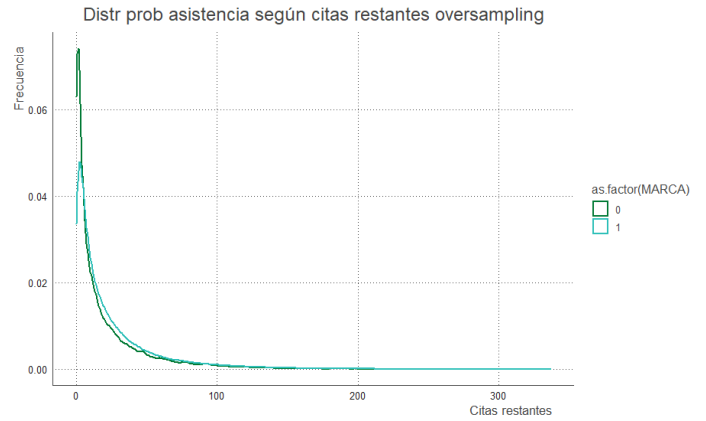


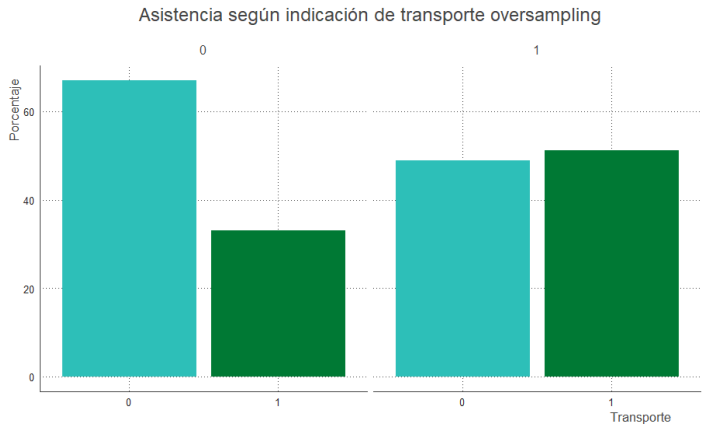

Asistencia según indicación de transporte undersampling
0
1
Porcentaje
60
40
20
0
0
1
0
1
Transporte

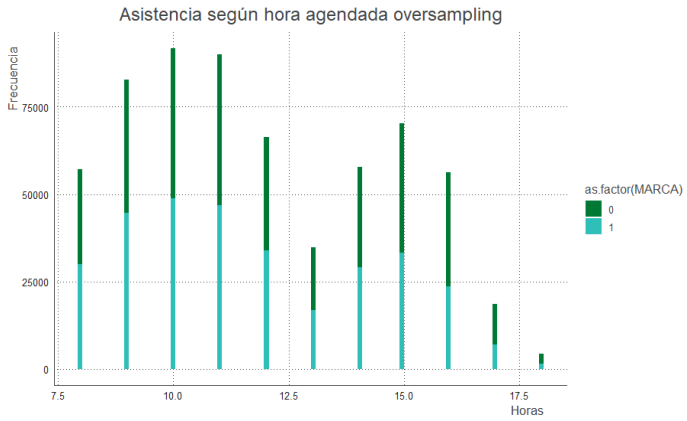

Asistencia según hora agendada undersampling
Frecuencia
75000
50000
25000
0
7.5
10.0
12.5
15.0
17.5
Horas
as.factor(MARCA)
0
1

ANEXO I: Validación base de datos oversampling





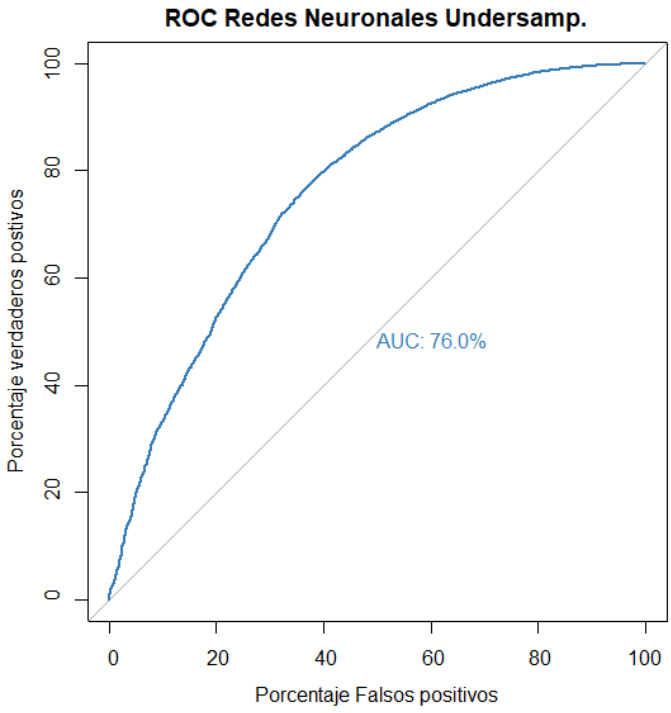




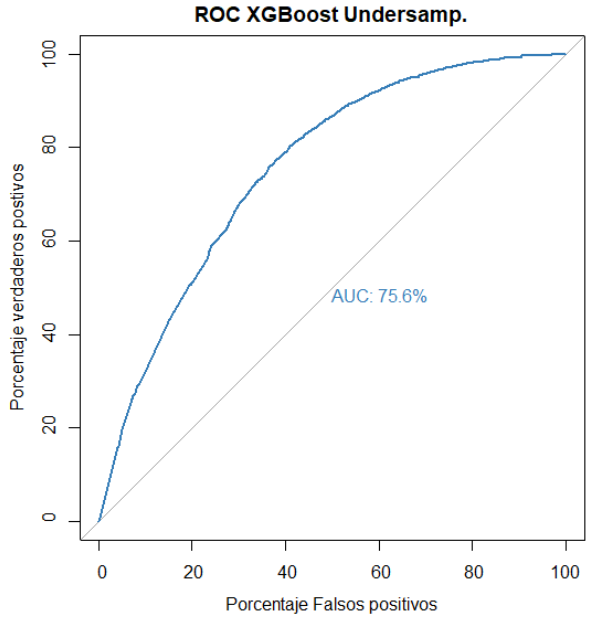
ANEXO J: Curvas ROC

Modelos Balanceados Undersampling

Redes neuronales

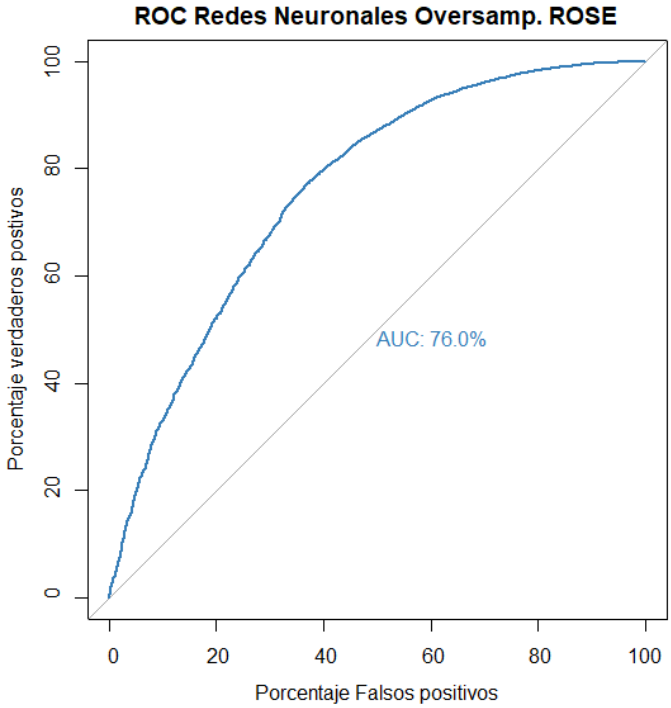


XGBoost

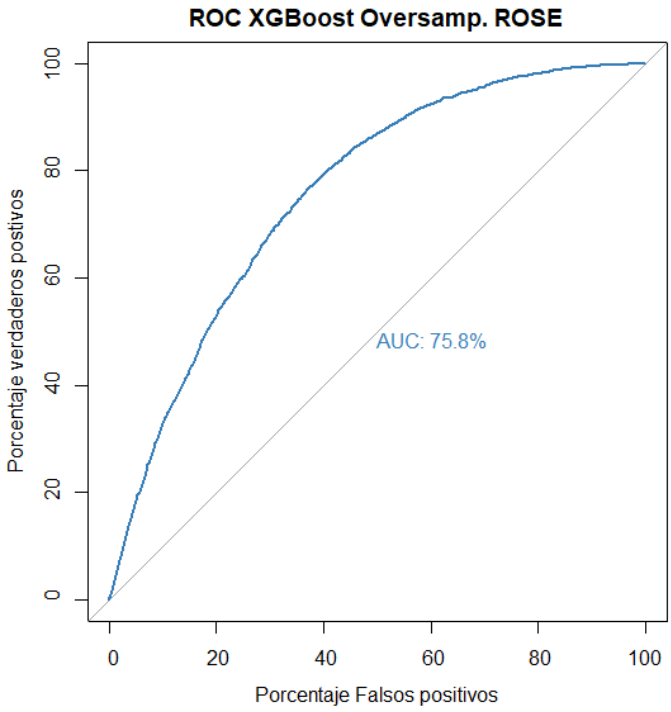


Modelos balanceados con oversampling ROSE

Redes neuronales

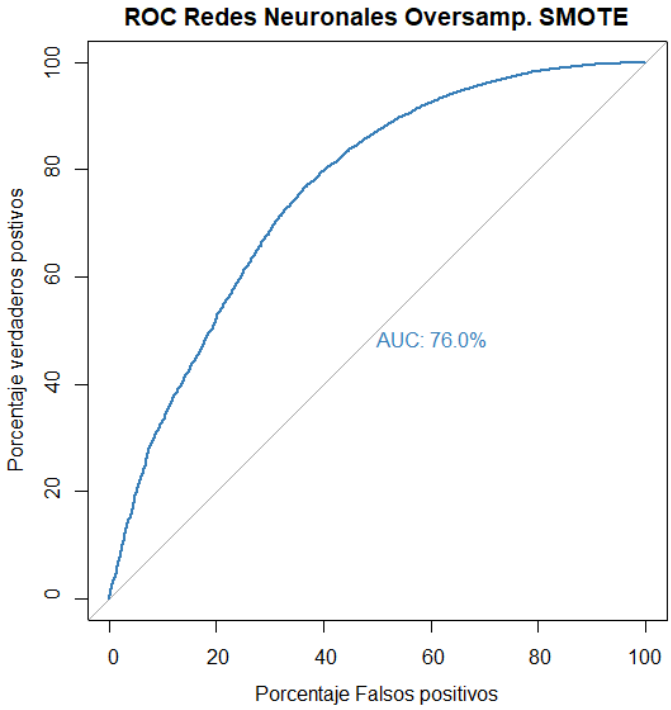


XGboost

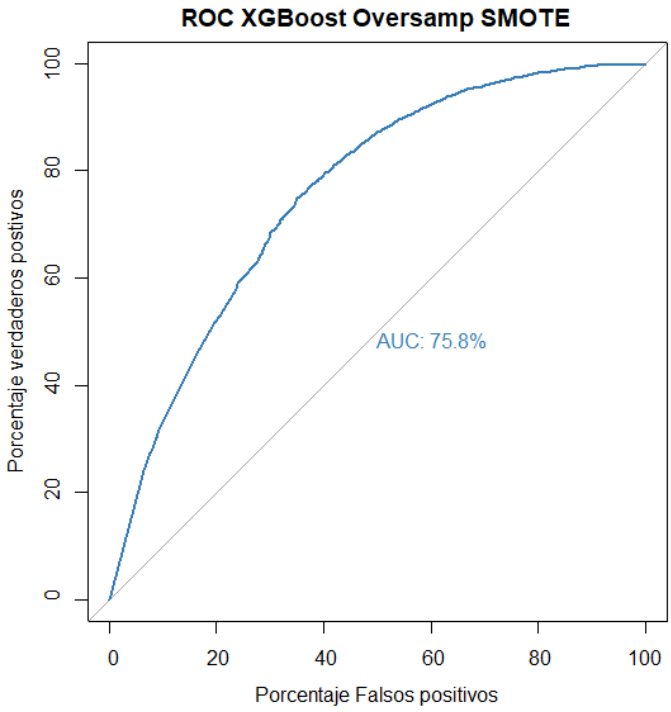


Modelos balanceados con oversampling SMOTE

Redes neuronales



XGBoost


ROC XGBoost Oversamp SMOTE
Porcentaje verdaderos postivos
0
20
40
60
80
100
AUC: 75.8%
Porcentaje Falsos positivos