

Predicción del ausentismo en citas médicas mediante Machine Learning

Catalina Isabel Valenzuela-Núñez
<https://orcid.org/0000-0002-3536-7051>
catalina.valenzuela1701@alumnos.ubiobio.cl
Departamento de Ingeniería Industrial
Facultad de Ingeniería
Universidad del Bío-Bío
Concepción, Chile

Fredy Humberto Troncoso Espinosa
<https://orcid.org/0000-0002-9972-3123>
ftroncos@ubiobio.cl
Departamento de Ingeniería Industrial
Facultad de Ingeniería
Universidad del Bío-Bío
Concepción, Chile

Guillermo Octavio Latorre-Núñez
<https://orcid.org/0000-0003-1655-2611>
glatorre@ubiobio.cl
Departamento de Ingeniería Industrial
Facultad de Ingeniería
Universidad del Bío-Bío
Concepción, Chile

Recibido (27/03/2023), Aceptado (11/07/2023)

Resumen: La programación de citas médicas es una actividad de gran importancia en un hospital, ya que se deben utilizar de forma eficiente diferentes capitales, tanto humanos como materiales. Uno de los problemas de este trabajo es la inasistencia de un paciente, lo que disminuye la eficiencia del uso de estos recursos. Para hacer frente a esto, diversos estudios han propuesto considerar el "ausentismo" para programar las citas médicas. Sin embargo, predecirlo es una tarea compleja. Esta investigación propone la predicción de la no asistencia a la citación para tres áreas médicas del Hospital Clínico Regional Dr. Guillermo Grant Benavente en la ciudad de Concepción, Chile. Para esto se entrenan y evalúan cinco algoritmos de Machine Learning. El mejor modelo entrenado logró ser una herramienta predictiva del nivel de ausentismo de un paciente para su próxima consulta y caracterizar a aquellos pacientes con mayores niveles de ausentismo.

Palabras clave: Machine learning, citas médicas, ausentismo.

Prediction of absenteeism in medical appointments using Machine Learning

Abstract.- The scheduling of medical appointments is an activity of great importance in a hospital since different capitals, both human and material, must be used efficiently. One of the problems of this work is the non-attendance of a patient, which decreases the efficiency of the use of resources. Several studies have proposed considering "absenteeism" for scheduling medical appointments to address this. However, predicting it is a complex task. This research proposes the prediction of absenteeism to medical appointments for three medical areas of the Hospital Clínico Regional Dr. Guillermo Grant Benavente in the city of Concepción, Chile. For this purpose, five Machine Learning algorithms are trained and evaluated. The best-trained model managed to be a predictive tool of a patient's absenteeism level for his next appointment and to characterize those patients with higher levels of no-show.

Keywords: Machine learning, medical appointments, absenteeism.



I. INTRODUCCIÓN

Todos los centros de salud buscan utilizar sus recursos de la forma más eficiente posible. Para lograr esto, se requiere de herramientas de apoyo para la toma de decisiones que les ayuden en la programación de sus recursos.

La programación de citas de pacientes (PAS, por sus siglas en inglés) es una actividad de vital importancia para todos los centros de salud, ya que un paciente requiere de recursos tanto humanos como materiales (médicos, enfermeras, tecnólogos, equipos, box, etc.) para su atención. Entre las decisiones del problema PAS se incluyen la asignación de un médico al paciente, un box de atención, además de la búsqueda del mejor tiempo de atención para todos los pacientes [1]. Existen diversas fuentes de incertidumbre que afectan la PAS como impuntualidad de los pacientes, el retraso de los médicos, las visitas sin cita previa y/o las urgencias, pero el mayor impacto lo producen los pacientes que, por diversas razones, no asiste a su cita médica, afectando los ingresos del centro de salud, y la eficiencia en el uso de los recursos [2]. Además, la salud de los propios pacientes se puede ver afectada por el aumento en los tiempos de espera indirectos producto del ausentismo. Las citas médicas a las cuales no asiste el paciente se pueden clasificar como "perdidas" para la institución de salud en cuestión, puesto que conducen a un control deficiente de la utilización de los recursos destinados a esta.

Algunos de los motivos más frecuentes de inasistencia son el olvido de la cita programada y la poca comunicación con el establecimiento. Aquellos pacientes con un mayor nivel de inasistencia son aquellos con actividades relacionadas al servicio de urgencias y de atención postoperatoria.

Para reducir el ausentismo, los centros de salud alrededor del mundo utilizan diversas estrategias, tales como los recordatorios (vía telefónica, mensajes y correos) lo que es insuficiente para la prestación del servicio y la gestión de la demanda de estas citas médicas. Además, cabe mencionar que en China es un problema persistente [3]. En el estudio de [2], realizado en una clínica de Estados Unidos, mencionan la mala comunicación que existe entre distintas áreas de esta, haciendo más lento el proceso de atención del paciente y no atendidos en sus horarios correspondientes.

Para hacer frente al a este problema, diversos estudios proponen considerar el "ausentismo" al momento de realizar la PAS. Entre algunas estrategias que se han aplicado a la PAS se encuentra el overbooking, que permite mejorar la ocupación del bloque de atención [1], [3]. Lamentablemente, no es fácil para los servicios hospitalarios poder predecir el nivel de ausentismo de un paciente. Por lo tanto, uno de los principales desafíos para crear PAS más eficientes, consiste en poder predecir de forma correcta la probabilidad de asistencia o inasistencia.

Dada la información disponible en los centros médicos en la actualidad, una herramienta que permite la predicción del ausentismo o inasistencia es Machine Learning. Algunos autores [4] utilizan algoritmos de Machine Learning para clasificar los perfiles de los usuarios y determinar las probabilidades de asistencia de los pacientes a su cita médica. Otras investigaciones [5] comparan algoritmos de Machine Learning menos frecuentes como técnicas de Bagging, Random Forest y Boosting, las cuales muestran un alto rendimiento predictivo. Trabajos como el de AlMuhaideb et al [6] utiliza datos históricos de una clínica y emplea algoritmos de Machine Learning para predecir la inasistencia de los pacientes a las intervenciones, permitiendo reconocer aquellos con alto riesgo de no presentarse, reduciendo así los costos asociados a la tasa de inasistencia. Otros autores [7] comparan distintos modelos de predicción como AdaBoost, Logistic Regression, Random Forest, Naive Bayes y multicapa perceptrón, junto a un modelo de optimización, para realizar la PAS basándose en probabilidades de asistencia que dependen de la etnia del paciente (blanco, afroamericano, asiático y otros). Sin embargo, Mohammadi et al [8] y Devasahay et al [9] también prueban distintos modelos para predecir la inasistencia de diferentes clínicas o centros de atención, determinando que los modelos de Naive Bayes, Decision Tree y Logistic Regression logran las mejores estadísticas de predicción.

Las variables más utilizadas en la investigación del problema PAS son: la edad, el género, el día de la semana, el tiempo de espera entre la fecha de la consulta y la cita, las ausencias anteriores, la hora de la cita y la distancia con respecto al establecimiento, fecha de creación de la cita, la fecha de registro, la hora de llegada del paciente, el ingreso estimado y la zona donde vive el paciente [5], [10].

El Hospital Clínico Regional Dr. Guillermo Grant Benavente en la ciudad de Concepción, Chile, es un hospital de alta complejidad que atiende a gran parte de la población de la provincia de Concepción y que posee un alto nivel de ausentismo a citas médicas. Este hospital actualmente no incorpora en sus sistemas de agendamiento la probabilidad de que un paciente no asista, lo que no permite generar una planificación eficiente. Dado los buenos resultados de la aplicación de algoritmos de Machine Learning para predecir la asistencia o inasistencia a citas médicas, este trabajo propone el uso de estos algoritmos para predecir el nivel de ausentismo de un paciente a citas médicas en las tres especialidades con mayor inasistencia de pacientes que son Ginecología, Neurología y Otorrinolaringología. Para esto se considera una base de datos de cerca de 7.000 registros de citas médicas.

II. DESARROLLO

A. Selección y preprocesamiento

En la primera etapa de selección se escogen los atributos e instancias que se consideran relevantes para el estudio, y mediante las técnicas de filtrado de registros y atributos, se eliminan las observaciones irrelevantes para el análisis. Se crea un conjunto objetivo utilizando toda la base de datos disponible. En la limpieza, se eliminan aquellas variables que pertenecen a la identificación personal de salud o del paciente. Además, se identifican algunas columnas como la "Hora de Recepción del Paciente" que tiene un 97,63% de celdas vacías, que también se quitan del conjunto de datos.

B. Transformación

Consiste en transformar los registros de forma apropiada para la extracción de información. Se decide añadir una nueva variable que contabilice el número de veces que el paciente ha asistido y faltado a sus citas médicas, como se observa en la Fig. 1.



Fig. 1. Nuevo atributo creado a partir del historial del paciente

Antes de comenzar con la determinación de factores relevantes para la aplicación de los algoritmos, es importante conocer a priori algunas características presentes en la base de datos. Como se observa en la Fig. 3, las edades de los pacientes que más se ausentan a su cita médica comprenden entre 1 y 15 años y luego tienen un alza entre los 18 y 40 años aproximadamente, con un total de 15.333 registros (Fig. 2).

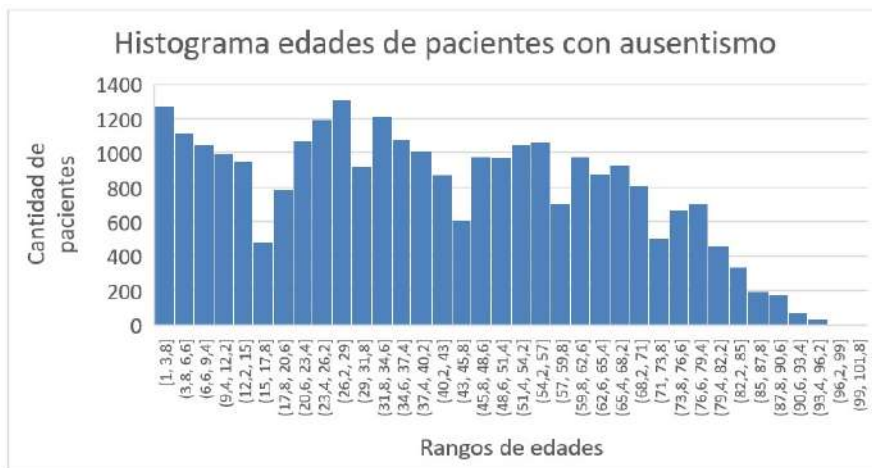


Fig. 2. Nuevo atributo creado a partir del historial del paciente

C. Selección de variables

La selección de variables requiere evaluar criterios que dividan los datos de forma adecuada. Para lograr esto se utilizan medidas de relevancia como la ganancia de información que examinan la precisión de la clasificación, comprobando la idoneidad de los modelos para realizar buenas predicciones [14]. La ganancia de información se basa en la entropía, midiéndose la utilidad de una cierta característica en la clasificación, es decir, se utiliza para medir la relevancia del atributo en la clase. Cuanto mayor sea este valor, mayor será la relevancia del atributo. Se seleccionaron las variables con una ganancia de información mayor o igual al promedio. Para su aplicación se transformaron las variables numéricas o categóricas. De un total de 27 atributos, se consideraron 22 para la confección del modelo.

D. Minería de datos

En la etapa de minería de datos se determinan los algoritmos de Machine Learning. Como se señaló anteriormente, se probaron cinco algoritmos: Árbol de Decisión, Red Neuronal, Support Vector Machine, Regresión Logística y Naive Bayes.

Los algoritmos se aplicaron utilizando el software RapidMiner. Este permite el desarrollo de procesos mediante el encadenamiento de operadores a través de un entorno gráfico, utilizado en investigación y aplicaciones empresariales.

El proceso de entrenamiento, validación y prueba de los modelos se muestra en la Fig. 3. La primera fila muestra el conjunto de datos de enero a mayo de 2021. Los primeros cuatro meses son para el entrenamiento del modelo, y el quinto mes es para el ajuste de parámetros y validación. Luego del ajuste de parámetros y validación, se entrenan los algoritmos con la información hasta mayo y se prueba su desempeño con el mes de junio de 2021. Debido a la existencia de un desbalance de clases (asistente y ausente), se utiliza la técnica "SMOTE" para generar un conjunto de datos balanceados para el proceso de entrenamiento, validación y prueba. Para ajustar los parámetros de los algoritmos y probar si el desempeño se utilizó validación cruzada.



Fig. 3. Conjunto de entrenamiento y validación para el modelo.

La Tabla 2 muestra, que, a partir de las cuatro medidas de evaluación utilizadas, los algoritmos de mejor desempeño son Support Vector Machine, Red Neuronal y Árbol de Decisión.

Tabla 2. Resultados de desempeño de modelos aplicados.

	Exactitud	Precisión	Exhaustividad	ABC
Árbol de decisión	0,7453 ± 0,0219	0,7582	0,7202	0,800 ± 0,027
Red neuronal	0,7586 ± 0,0227	0,7723	0,7334	0,828 ± 0,020
SVM	0,7267 ± 0,0173	0,6595	0,9372	0,851 ± 0,018
Regresión logística	0,6747 ± 0,0256	0,6984	0,6149	0,734 ± 0,026
Naive Bayes	0,6371 ± 0,0168	0,6659	0,5500	0,701 ± 0,018

III. METODOLOGÍA

Este trabajo consistió en la medición de la predicción de ausencias médicas en las especialidades de mayor ausentismo, Ginecología, Neurología y Otorrinolaringología del Hospital Clínico Regional Dr. Guillermo Grant Benavente, por esas tres, las de mayor concurrencia de pacientes y, por ende, las de mayor falta de los mismos.

Para preprocesar los cerca de 7.000 registros de citas médicas, entrenar los algoritmos de Machine Learning y evaluar sus resultados, se aplicó el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD).

El proceso KDD involucra cinco fases que se pueden observar en la Fig. 4 y que se describen a continuación [11]

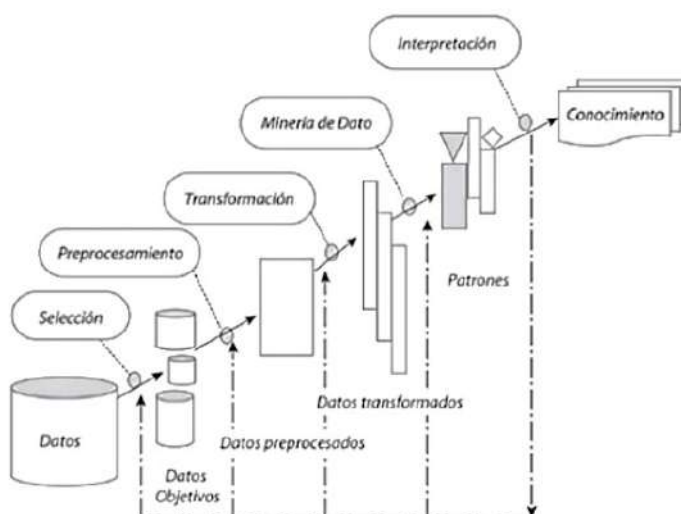


Fig. 4. Slump del concreto normal y con incorporación de la fibra de acero.

- Selección: se crea un conjunto de datos objetivo, seleccionando todos los datos o una muestra representativa de estos, sobre el cual se realiza el proceso de descubrimiento.
- Preprocesamiento: en esta etapa se analiza la calidad de los datos, se aplican operaciones como la remoción de registros que están significativamente fuera del rango de valores esperados, se seleccionan estrategias para el manejo de información desconocida, filas duplicadas y técnicas estadísticas para su reemplazo.
- Transformación: en la tercera etapa se buscan características útiles para representar los datos. Las técnicas de reducción de estos están orientadas a obtener una representación reducida de los originales, manteniendo en lo posible, la integridad y la información existente de los datos. No se considera estrictamente obligatoria, ya que usualmente se utiliza cuando los tiempos de ejecución o el tamaño son mayores de lo que se espera. Las técnicas más relevantes son la selección de atributos, selección de instancias o la discretización.
- Minería de datos: el objetivo de esta etapa es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando algoritmos de Machine Learning.
- Interpretación: aquí se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Por otra parte, se consolida el conocimiento descubierto para verificar y resolver conflictos potenciales.

Los algoritmos de Machine Learning supervisado seleccionados según su aplicación en la literatura son:

- Árbol de Decisión: es uno de los algoritmos más útiles y potentes en la minería de datos, su principal ventaja consiste en facilitar la comprensión del conocimiento, explicando el comportamiento de una determinada decisión.
- Red Neuronal: están inspiradas en las redes neuronales biológicas, pues consisten en un gran número de elementos simples de procesamiento llamados nodos (neuronas) que están organizados en capas. Son algoritmos adaptativos que aprenden de la experiencia.
- Support Vector Machines (SVM): la tarea de este algoritmo es encontrar la mejor separación posible entre clases de elementos. Debido a que los problemas de aprendizaje automático tienen un gran número de dimensiones, se encuentra el hiperplano que maximiza el margen de separación entre clases.
- Regresión Logística: es un modelo estadístico en el cual se desea conocer la relación entre una variable dependiente cualitativa dicotómica con una o más variables explicativas independientes.
- Naive Bayes: es un clasificador probabilístico que calcula un conjunto de probabilidades contando la frecuencia y combinaciones de valores en un conjunto de datos utilizando el teorema de Bayes.

Para la evaluación de los algoritmos de Machine Learning, se utilizarán cuatro medidas de desempeño, accuracy (Exactitud), precisión, recall (Exhaustividad) y AUC (Área bajo la curva). Las primeras tres medidas se basan en datos resumidos en la matriz de desempeño de cada modelo llamada matriz de confusión, donde el accuracy representa la proporción total de predicciones correctamente clasificadas por el modelo, el precision corresponde al porcentaje de elementos correctamente clasificados como clase 1 y el recall indica el porcentaje de observaciones que pertenecen a la clase 1 y que fueron correctamente clasificadas por el modelo. La matriz de confusión, mostrada en la Tabla 3, es una herramienta que muestra, explícita y efectivamente, el comportamiento de los algoritmos teniendo como ventaja la fácil interpretación de los resultados obtenidos.

Las filas de la matriz representan los valores de las clases predictivas y las columnas, los valores de las clases actuales. El valor de "True Positives (TP)" representa los elementos de la clase 1 que fueron correctamente predichos por el modelo y "Falses Negatives (FN)" representan los elementos de la clase 1 que fueron incorrectamente predichos, es decir los errores tipo I. Asimismo, "True Negatives (TN)" son los elementos de la clase 0 que fueron correctamente predichos por el modelo y "Falses Positives (FP)" los elementos que fueron incorrectamente predichos, es decir los errores tipo II [12] (Tabla 3).

Tabla 3. Resultados de desempeño de modelos aplicados.

		Valor real	
		1	0
Valor predicho	1	True Positives (TP)	Falses Positives (FP) Error tipo II
	0	Falses Negatives (FN) Error tipo I	True Negatives (TN)

El resultado de la predicción de un algoritmo de Machine Learning supervisado de clasificación es un valor entre 0 y 1 y no un 0 o un 1. Por regla general si este valor es mayor o igual a 0,5 se determina que la predicción será de la clase 1 y si es menor a este valor de la clase 0. A este valor de decisión se le conoce como valor de umbral. La variación de este valor de umbral afecta los valores de la matriz de confusión y, por tanto, los valores de las medidas de desempeño. Una forma de medir el desempeño predictivo de un algoritmo considerando todos los valores de umbral posibles es mediante la curva ROC y el valor del área bajo su curva AUC [13]. La curva ROC en una gráfica que relaciona la tasa verdadero positivo con la tasa falso positivo, para distintos valores de umbral. El valor del AUC varía entre 0 y 1. A mayor valor, mejor es el desempeño de un algoritmo.

IV. RESULTADOS

En la Tabla 2 se comparan los resultados de los diferentes modelos. Se observa que todos tienen un desempeño superior al 60% en cada una de las métricas analizadas. Por lo tanto, todos los modelos pueden dar una estimación aceptable sobre la predicción de ausencia del paciente. La técnica que mejor se adapta al comportamiento de los datos fue la de Support Vector Machine (SVM), el cual fue seleccionado dado su alto valor de Recall y AUC.

Un resultado importante de la aplicación del algoritmo SVM es la obtención de los pesos de las variables consideradas. El algoritmo maximiza la separación de las clases y el hiperplano de separación permiten evidenciar la importancia de cada atributo en la separación de las clases. La importancia de cada atributo se muestra en la Tabla 4. Dado estos valores, es posible caracterizar a los pacientes con mayor nivel de ausentismo. Estas características son:

- Pacientes que tienen una cita médica mediante el sistema de interconsulta del hospital (pacientes que son derivados desde otros centros de asistencia para ser evaluados por un especialista).
- Pacientes que tienen una nueva consulta (pacientes que por primera vez solicitan atención en una especialidad).

- Las actividades que se encuentran dentro del ausentismo son la evaluación postoperatoria, lactancia materna y telemedicina de otorrinolaringología.
- Los establecimientos de donde son derivados los pacientes que pertenecen al grupo de ausentismo son centros de atención familiar ubicados alrededor de la comuna de Concepción.
- La diferencia de días entre la fecha de creación del registro y la fecha de la cita real se encuentran entre 15 y 24 días (mediano plazo) y más de 25 días (lejano plazo).
- Pacientes que tienen un alto índice de ausentismo (tienen un registro de más de cinco veces sin asistir a una cita médica).

En la Tabla 4 se observa el peso de cada atributo considerado para el estudio con su respectivo peso.

Tabla 4. Características de pacientes con ausentismo y sin ausentismo.

Características pacientes con ausentismo		Características pacientes sin ausentismo	
Atributo	Peso	Atributo	Peso
Tipo Solicitud = SISTEMA INTERCONSULTA	685,46	Sub Especialidad = CONSULTA NEUROLOGÍA	-690,22
Tipo Documento = INTERCONSULTA	674,29	Actividad = CONSULTA INFANTIL	-690,22
Tipo Atención = NUEVA	530,62	Categoría Atención = RECETA	-690,22
Actividad = EVALUACION PREOPERATORIA	402,33	Tipo Atención = RECETA	-690,22
Establecimiento Origen = CENTRO MÉDICO DE ATENCIÓN 1	388,48	Frec. Asistencias = CINCO	-688,15
Diferencia Creación Cita = ENTRE 5 Y 10 DÍAS	384,06	Citas Totales = CINCO	-500,06
Actividad = LACTANCIA MATERNA	374,86	Sobre Cupo = SI	-445,37
Establecimiento Origen = CENTRO MÉDICO DE ATENCIÓN 2	374,00	Frec. Asistencias = MAYOR A CINCO	-424,65
Categoría Atención = CONSULTA	365,59	Diferencia Creación Cita = MENOR A 5 DÍAS	-420,03
Mes Creación = MAYO	307,61	Actividad = EDUCACION INDIVIDUAL	-417,67
Frec. Ausencias = MAYOR A CINCO	301,91	Citas Totales = MAYOR A CINCO	-365,12
Diferencia Creación Cita = MAYOR A 15 DÍAS	286,55	Mes Creación = JUNIO	-363,36
Establecimiento Origen = CENTRO MÉDICO DE ATENCIÓN 3	285,74	Actividad = CABEZA Y CUELLO	-294,96
Establecimiento Origen = CENTRO MÉDICO DE ATENCIÓN 4	272,89	Tipo Solicitud = VENTANILLA	-288,08
Actividad = TELEMEDICINA OTORRINO	272,17	Tipo Documento = NO TIENE	-287,01

E. Costos de clasificación e implementación del modelo

Para un análisis más profundo para la implantación de este modelo, se analizan los costos de error de clasificación. Como se observa en la matriz de confusión de la Tabla 3, existen dos tipos de errores el tipo I y tipo II. En esta investigación el error tipo I ocurre cuando un algoritmo predice que un paciente asistirá a su cita médica, pero en realidad no lo hará. El error tipo II ocurre cuando el algoritmo predice que un paciente se ausentará cuando realmente asistirá a su cita médica. Ambos errores están asociados a un costo económico para el hospital al utilizar un algoritmo de Machine Learning. Llamaremos a CEI como el costo del error tipo I y a CEII el costo de error tipo II. Llamaremos TFN a la tasa de ocurrencia del error tipo I, es decir los falsos negativos dividido los fasos negativos más los verdaderos positivos y llamaremos TFP a la tasa de ocurrencia de los falsos positivos, es decir, los falsos positivos dividido por los falsos positivos más los verdaderos negativos.

De esta forma, el costo esperado por error de clasificación queda expresado de la siguiente manera:

$$E(\text{Costo}) = TFN * C_{EI} + TFP * C_{EII} \tag{1}$$

Si expresamos esta curva de costos en términos de los ejes de la curva ROC, es decir reemplazando en la ecuación (1) $E(\text{Costo}) = 0$ y (2) $TFN = 1 - TPP$, la curva de costos queda representada por la siguiente ecuación:

$$TPP = \left(\frac{C_{EI}}{C_{EII}}\right) * Tasa FP + 1 \tag{3}$$

Al intersecar esta curva de costos con la curva ROC podemos obtener el valor de umbral de clasificación que minimiza el costo de error de clasificación $E(\text{Costo})$.

De acuerdo a la información del hospital el costo del error tipo I implica tiempo ocioso del box y equipos, acumulación del tiempo de espera por el servicio, horas laborales adicionales del médico y personal asociado. Por otro lado, el costo del error tipo II está asociado a tomar acciones preventivas como llamada recordatoria para la cita médica del paciente y el costo de oportunidad de la utilización del box de atención. Se estima para la relación de costos $\left(\frac{C_{EI}}{C_{EII}}\right)$ un valor de 3/7.

El procedimiento se llevó a cabo mediante el software RapidMiner con el operador "Find Threshold", el que permite encontrar el umbral óptimo para el modelo. En la Fig. 5 se presenta la curva ROC y su respectiva función de costos asociados al problema de clasificación de ausentismo en los pacientes mediante el algoritmo de Support Vector Machine. La línea tangente de color azul representa la función de costos, la curva hacia debajo de color verde es el umbral, mientras que la curva hacia arriba de color rosado corresponde a la curva ROC. La intersección entre estas dos curvas corresponde al valor del umbral óptimo.

Para esta instancia de clasificación binomial (dos valores), la regla de probabilidad implica que cada registro tiende a pertenecer a una clase u otra. Por defecto, los pacientes con probabilidad de ausentismo tienen un valor menor a 0,5 y aquellos pacientes que si asistirán a su cita médica tienen un valor mayor o igual a 0,5. Para cada modelo se establecen distintos valores de umbrales, los cuales permiten determinar una probabilidad por la cual los registros serán clasificados. El valor del umbral que minimiza el costo de error de clasificación se observa en la intersección antes mencionada, resultando un valor de 0,4380. Esto quiere decir que al clasificar a un paciente que no asistirá a la cita médica, debe tener una probabilidad mayor a 0,4380.

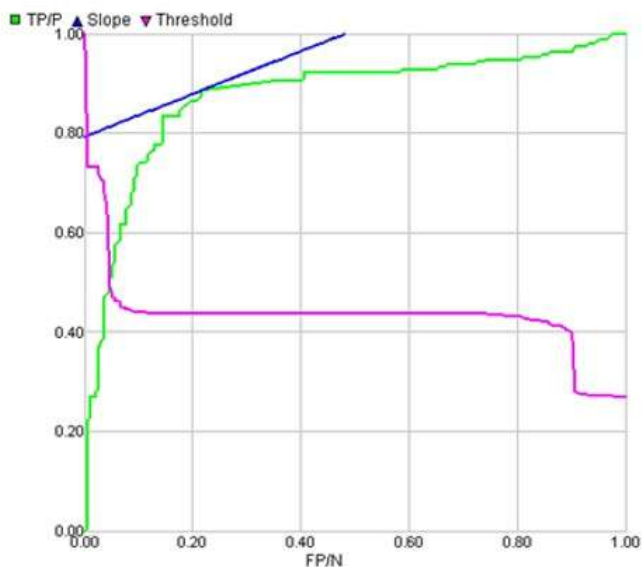


Fig. 5. Curva ROC, función de costos y umbral óptimo.

Mediante este valor, en la Fig. 6 se observa la propuesta para la implementación de la metodología de aplicación del algoritmo SVM. Se comienza por establecer la base de datos de cierto período en específico y los pacientes de estudio. Luego se requiere la obtención de variables relevantes para el modelo, seguido de la aplicación del algoritmo de predicción. Finalmente, se realiza la aplicación de las estrategias de ocupación médica como lo son el overbooking, la cancelación de citas médicas, los llamados recordatorios, entre otros.

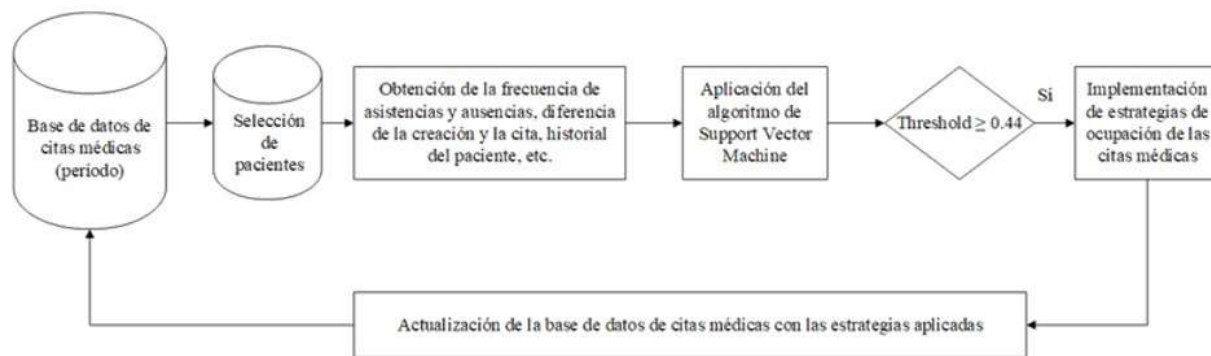


Fig. 6. Diagrama del proceso de identificación y contramedidas para pacientes con ausentismo.

La predicción del ausentismo de los pacientes ha sido estudiada desde los años 80 y, de acuerdo a la literatura estudiada, diferentes investigaciones han aplicado algoritmos de la minería de datos para predecir este comportamiento [6]. Las principales técnicas utilizadas en la literatura son con el método de Regresión Logística, Random Forest y Árbol de Decisión, pues obtienen un desempeño superior al 60% de AUC. Las variables más frecuentes utilizadas en la literatura se encontraban disponibles en la base de datos del Hospital Regional de Concepción. Sin embargo, en el proceso de KDD se eliminan algunas variables como el día de la cita médica y la comuna de residencia del paciente. Aun así, el algoritmo entrenado muestra un valor de AUC = 85,1% que supera el valor 79,2% presentado en [15].

Kaplan-Lewis y Percac-Lima[3], mencionan que la principal estrategia implementada para el agendamiento de citas son los programas basados en tecnologías de la información para generar contacto con los usuarios mediante diferentes medios, ya sea por mensajería instantánea, mensajes de confirmación o e-mails. Como el Hospital Regional de Concepción es un recinto de gran complejidad, el sistema de programación de citas debería incluir estas tecnologías u otras tecnologías y registrar el motivo específico de inasistencia del paciente, de manera de contar con información adicional para el proceso de predicción.

CONCLUSIONES

La aplicación de técnicas de Machine Learning, permiten predecir el nivel de inasistencia de un paciente a su cita médica. De los modelos entrenados, el de mejor desempeño resultó ser Support Vector Machine, SVM. Mediante su aplicación se pudo determinar las variables más influyentes y mediante su análisis se determinó que los pacientes con mayor nivel de inasistencia son quienes poseen alguna de estas características: Pacientes con cita médica mediante el sistema de interconsulta del hospital, pacientes que tienen una nueva consulta, pacientes con cita postoperatoria, pacientes de lactancia materna, pacientes de telemedicina, pacientes derivados desde centros de atención familiar de otras comunas.

El algoritmo entrenado, Support Vector Machine, muestra un valor de AUC = 85,1% que supera el valor 79,2% presentado en la literatura. La utilización este algoritmo de machine learning, permitirá al Hospital Clínico Regional Dr. Guillermo Grant Benavente, identificar aquellos pacientes con mayor probabilidad de inasistencia, priorizar las acciones para evitar su inasistencia. El nivel de ausentismo de cada paciente puede ser utilizado como dato de entrada para herramientas de apoyo para la toma de decisiones que puedan complementarse

con estrategias de overbooking, recordatorios o números de emergencia.

Como trabajo futuro se considera desarrollar algoritmos exactos y aproximados para la programación de citas médicas, los que consideren la inasistencia del paciente, estrategias de overbooking y que sean alimentados con la información proporcionada por la metodología propuesta en esta investigación.

REFERENCES

- [1] G. Lamé, O. Jouini y J. Stal-Le Cardinal, «Outpatient chemotherapy planning: A literature review with insights from a case study,» *IIE Transactions on Healthcare Systems Engineering*, vol. 6, p. 127–139, 2016.
- [2] M. C. Rossi y H. Balasubramanian, «Panel size, office visits, and care coordination events: a new workload estimation methodology based on patient longitudinal event histories,» *MDM Policy & Practice*, vol. 3, p. 2381468318787188, 2018.
- [3] E. Kaplan-Lewis y S. Percac-Lima, «No-show to primary care appointments: why patients do not come,» *Journal of primary care & community health*, vol. 4, p. 251–255, 2013.
- [4] M. Samorani, S. L. Harris, L. G. Blount, H. Lu y M. A. Santoro, «Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling,» *Manufacturing & Service Operations Management*, 2021.
- [5] G. Fan, Z. Deng, Q. Ye y B. Wang, «Machine learning-based prediction models for patients no-show in online outpatient appointments,» *Data Science and Management*, vol. 2, p. 45–52, 2021.
- [6] S. AlMuhaideb, O. Alswailem, N. Alsubaie, I. Ferwana y A. Alnajem, «Prediction of hospital no-show appointments through artificial intelligence algorithms,» *Annals of Saudi medicine*, vol. 39, p. 373–381, 2019.
- [7] S. L. Harris y M. Samorani, «On selecting a probabilistic classifier for appointment no-show prediction,» *Decision Support Systems*, vol. 142, p. 113472, 2021.
- [8] I. Mohammadi, H. Wu, A. Turkcan, T. Toscos y B. N. Doebbeling, «Data analytics and modeling for appointment no-show in community health centers,» *Journal of primary care & community health*, vol. 9, p. 2150132718811692, 2018.
- [9] S. R. Devasahay, S. Karpagam y N. L. Ma, «Predicting appointment misses in hospitals using data analytics,» *Mhealth*, vol. 3, 2017.
- [10] D. B. Ferro, S. Brailsford, C. Bravo y H. Smith, «Improving healthcare access management by predicting patient no-show behavior,» *Decision Support Systems*, vol. 138, p. 113398, 2020.
- [11] S. R. Timarán-Pereira, I. Hernández-Arteaga, S. J. Caicedo-Zambrano, A. Hidalgo-Troya y J. C. Alvarado-Pérez, «El proceso de descubrimiento de conocimiento en bases de datos,» *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*, p. 63–86, 2016.
- [12] F. H. T. Espinosa, «Prediction of recidivism in thefts and burglaries using machine learning,» *Indian Journal of Science and Technology*, vol. 13, p. 696–711, 2020.
- [13] M. N. Sulaiman y R. W. Rahmat, «Improving accuracy metric with precision and recall metrics for optimizing stochastic classifier,» 2011.
- [14] S. Tangirala, «Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm,» *International Journal of Advanced Computer Science and Applications*, vol. 11, p. 612–619, 2020.
- [15] A. Nelson, D. Herron, G. Rees y P. Nachev, «Predicting scheduled hospital attendance with artificial intelligence,» *NPJ digital medicine*, vol. 2, p. 1–7, 2019.

LOS AUTORES



Fredy Troncoso Espinosa, Doctor en Sistemas de Ingeniería, Universidad de Chile, Ingeniero Civil Industrial Universidad del Bío-Bío, Chile. Académico e Investigador Departamento de Ingeniería Industrial, Universidad del Bío-Bío. Concepción, Chile.



Guillermo Latorre-Núñez, Doctor en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile. Magíster en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile. Magíster en Ingeniería Industrial, Universidad del Bío-Bío. Ingeniero Civil Industrial, Universidad del Bío-Bío. Licenciado en Ciencias de la Ingeniería, Universidad del Bío-Bío.



Catalina Valenzuela-Núñez, Magíster en Ingeniería Industrial, Universidad del Bío-Bío. Ingeniera Civil Industrial Universidad del Bío-Bío, Chile. Licenciada en Ciencias de la Ingeniería, Universidad del Bío-Bío.