MODELAMIENTO PREDICTIVO DE “PACIENTE NO SE PRESENTA” EN CONSULTAS AMBULATORIAS MEDIANTE ALGORITMOS DE MACHINE LEARNING

Carolina Lindsay Brain. BQ, MSc, PhD.
Departamento de Investigación
Hospital Dr. Franco Ravera Zunino

INTRODUCCIÓN

La inasistencia a consultas médicas programadas representa un desafío crítico para los sistemas de salud pública, e impactan directamente no sólo en la eficiencia y producción hospitalaria, sino que también en la salud de los pacientes (1). En hospitales de alta complejidad, donde los recursos son limitados y la demanda es elevada, un índice elevado de ausencias de pacientes a sus consultas médicas agendadas retrasa la resolución de las listas de espera, debido a la subutilización de la capacidad clínica (1,2). Si bien las causas del ausentismo de los pacientes son diversas, el desarrollar herramientas predictivas que permitan identificar a los pacientes con mayor riesgo de ausentarse a su cita agendada, facilitaría la implementación de intervenciones específicas como mecanismos de contactabilidad diferentes, recordatorios personalizados, reprogramaciones condicionales o derivaciones a apoyo social (3,4).
El objetivo de este proyecto es generar un modelo de Machine Learning capaz de predecir la probabilidad de ausentismo de los pacientes a sus citas médicas agendadas en base a factores sociodemográficos y clínicos.

METODOLOGÍA

Recolección de datos: La fase de recolección de datos consistió en la extracción anonimizada de registros históricos de consultas ambulatorias programadas durante el año 2025 del Hospital Dr. Franco Ravera Zunino, hospital público de alta complejidad y referencia regional. Los datos se obtuvieron a partir del sistema de agendamiento FONENDO y las bases de datos institucionales de la Unidad de Estadística de la Subdirección de Gestión Estratégica y Eficiencia Hospitalaria, asegurándose la integridad y confidencialidad de la información sensible bajo los protocolos de ética y protección de datos del establecimiento (SGC-PR-MDYC y Solicitud de Manejo de datos clínicos con fines no asistenciales). Los datos fueron pre-procesados para homogenizar los registros, y las variables con alta cardinalidad fueron agrupadas por afinidad. Las variables recolectadas corresponden a factores sociodemográficos y clínicos, incluyendo edad, género, comuna de residencia, tipo de previsión de salud, especialidad médica de la consulta, tipo de profesional que atiende, tipo de consulta/procedimiento, fecha de cita, hora de cita. La variable objetivo/target corresponde a “Estado”, con las categorias de "Paciente No Se Presenta" y "Paciente Asiste/Atendido".

Preparación de datos: En una primera etapa, se realiza la carga del conjunto de datos y un análisis exploratorio inicial, con el objetivo de comprender su estructura, distribución y calidad. Posteriormente, si bien no existen valores nulos en ninguna columna, se efectúa la limpieza de los datos abordando valores faltantes indicados como “Sin información” en la variable “Previsión” mediante imputación con la moda de la columna respectiva. La variable

"Estado" se codifica como 1 "Paciente No Se Presenta" y 0 "Paciente Asiste/Atendido". El conjunto de datos es dividido en 2 subconjuntos; un subconjunto corresponde a todo el set de datos originales sin la columna (etiqueta/target) "Estado" (variable X), y el otro subjunto corresponde solo a la columna "Estado" (variable y). La data de cada subconjunto fue separada en data de entrenamiento (train, 80%) y validación (test, 20%). El set de datos posee solo variables categóricas, las cuales fueron transformadas mediante técnica de One hot enncoding (OHE) o Target Encoding (TE) para su uso en los modelos ML. Asimismo, al detectarse un desbalance en la distribución de las los registros en las categorías de la variable "Estado", se procede a aplicar SMOTE como método de balanceo.

Modelos de Machine Learning implementados: Se implementan distintos algoritmos de clasificación/predicción supervisada:

a) Regresión Logística.
b) Random Forest con codificación OHE y con codificación TE
c) Gradient Boosting

Identificación de importancia de variables: La importancia de las variables fue evaluada utilizando la reducción media de impureza (feature_importances_). Este análisis permitió identificar las variables con mayor contribución relativa a la capacidad predictiva del modelo de Random Forest.

Ajuste de Hiperparámetros: Los modelos fueron entrenados utilizando configuraciones de hiperparámetros definidas a priori según buenas prácticas, sin optimización exhaustiva mediante GridSearchCV, debido a restricciones computacionales. Para Regresión Logística, se entrenó el modelo en primera instancia con los parámetros predefinidos (solver='liblinear', C=1), luego con C=0.1 y C=10 con el objetivo de modificar la regularización (~ penalización). Para Random Forest, se entrenó el modelo con hiperparámetros predefinidos (n_estimators=100, max_depth=None, min_samples_leaf=1), y luego con n_estimators=200, max_depth=20 y min_samples_split=5 (Más árboles se refleja en menor varianza y mayor estabilidad, menor profundidad refleja mejor generalización y menor ruido)

Evaluación de desempeño de Modelos de Machine Learning implementados: La evaluación de los modelos se realiza mediante métricas estándar de clasificación, tales como exactitud (accuracy), precisión, recall, F1-score y matrices de confusión. Asimismo, se realiza una comparación de estas métricas entre los distintos modelos y los distintos ajustes de hiperparámetros.

Enlace a repositorio: https://github.com/clindsayhfrz/Diplomado-IA/blob/main/codigo_Modelo_ML_Diplomado.ipynb

RESULTADOS

Se presenta una base de datos con 717162 registros (filas) y 12 variables categóricas (columnas): ID, Estado, Sexo, Previsión, Rango Edad, Estival Fecha Cita, Rango Hora Cita, Agenda Biomédica, Profesional, Tipo de Atención, Distancia Comuna y Tipo Cita. El análisis descriptivo de cada variable (recuento, número de categorías, moda y frecuencia de la moda) y la distribución de los datos de la base de datos se indica en Tablas 1 y 2. Para

visualizar la distribución de las variables se realizaron gráficos de barra horizontal (categoría versus recuento) los cuales se muestran en la Figura 1.

Tabla 1.

Variable	Frecuencia	Variable	Frecuencia
Estado		**Agenda Biomédica**	
Atendido	637104	Otros	196023
No se presentó	80058	Medicina Clínica (Adultos)	136363
Sexo		Especialidades Quirúrgicas	104130
Femenino	397090	Oncología y Enfermedades Complejas	69481
Masculino	320072	Órganos de los Sentidos	67146
Distancia Comuna		Apoyo Terapéutico y Diagnóstico	60521
Cercana (hasta 40 km)	544313	Pediatría y Neonatología	45772
Intermedia (41-80 km)	154145	Medicina Familiar y Rehabilitación	30397
Lejana (+80 km)	18704	Salud Mental	7329
Tipo de Cita		**Profesional**	
Presencial	707776	Medicina	353622
Telefónica	9386	Enfermería	164331
Rango de Edad		Kinesiología	99339
Personas Mayores	309755	Odontología	25483
Adultos	293480	Fonoaudiología	19612
Adolescencia	50041	Matronería	19021
Primera Infancia	40629	Terapia Ocupacional	16906
Niñez	23257	Psicología	7851
Rango hora Cita		Nutrición	6034
Mañana	239929	Asistente Social	3572
Media Mañana	196382	Farmacia	1391
Tarde	164582	**Tipo de Atención**	
Media tarde	109802	Consultas Médicas	301383
Noche	6467	Procedimientos y Cirugías	163747
Estival Fecha Cita		Consultas No Médicas	142668
Invierno	208169	Gestión Clínica y Administrativa	44721
Primavera	204846	Otros	40722
Otoño	189111	Hospitalización y Cuidados Críticos	11890
Verano	115036	Atención Domiciliaria	7389
Previsión		Rehabilitación y Educación en Salud	4642
FONASA Tipo B	454016		
FONASA Tipo D	99998		
FONASA Tipo C	86535		
FONASA Tipo A	73638		
Isapre	1443		
Sin Información	1414		
Otro	118		

Tabla 2.

Variable	Registros	N° categorias	Moda	Frecuencia moda
Estado	717162	2	Atendido	637104
Sexo	717162	2	Femenino	397090
Previsión	717162	7	FONASA Tipo B	454016
Rango Edad	717162	5	Personas Mayores	309755
Estival Fecha Cita	717162	4	Invierno	208169
Rango Hora Cita	717162	5	Mañana	239929
Agenda Biomédica	717162	9	Otros	196023
Profesional	717162	11	Medicina	353622
Tipo atención	717162	8	Consultas Médicas	301383
Distancia comuna	717162	3	Cercana (hasta 40 km)	544313
Tipo Cita	717162	2	Presencial	707776

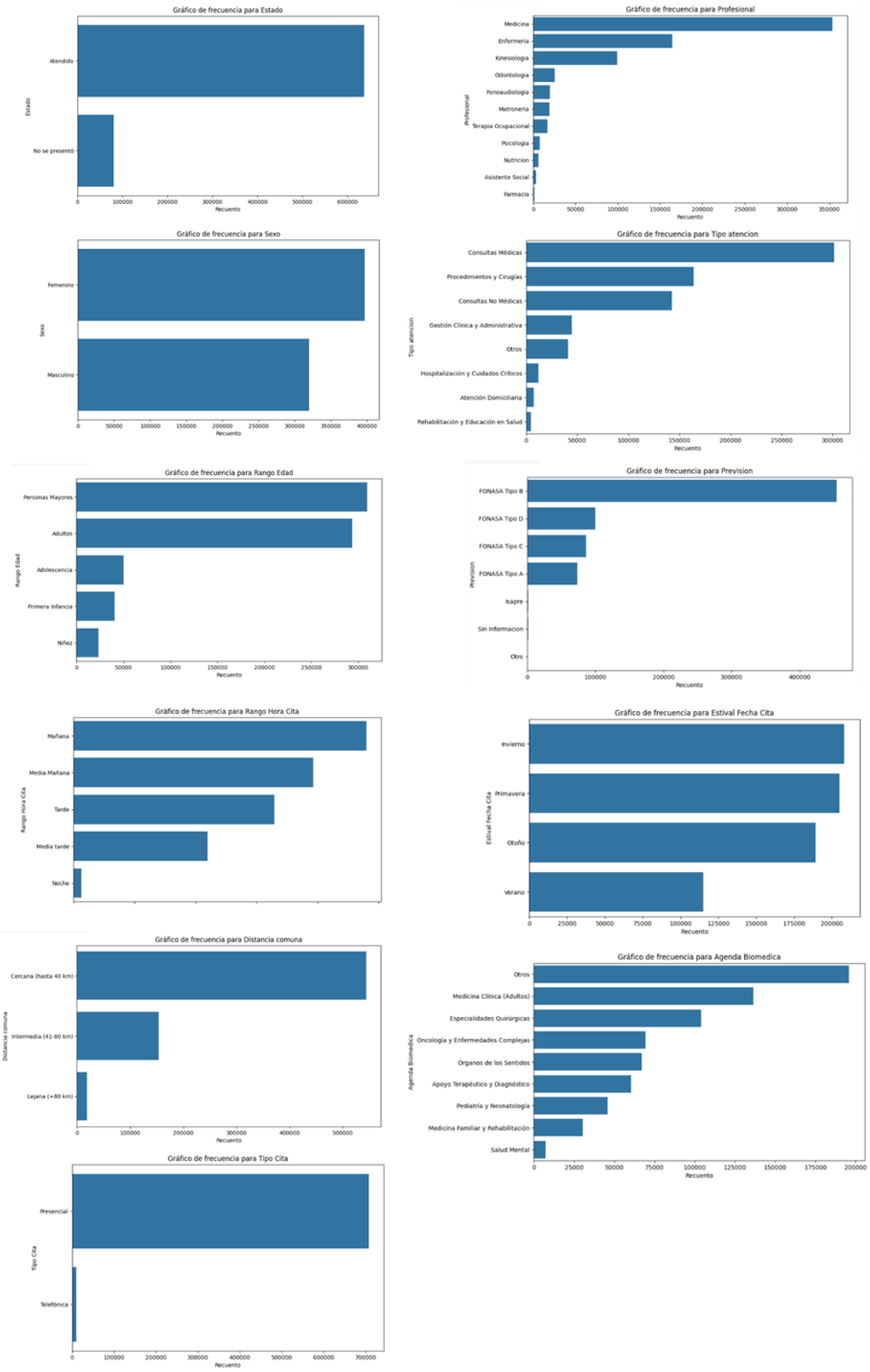
Gráfico de frecuencia para Estado
Estado
Atendido
No se presentó
Recuento
Gráfico de frecuencia para Profesional
Profesional
Medicina
Enfermeria
Kinesiologia
Odontologia
Fonoaudiologia
Matroneria
Terapia Ocupacional
Psicologia
Nutricion
Asistente Social
Farmacia
Recuento
Gráfico de frecuencia para Sexo
Sexo
Femenino
Masculino
Recuento
Gráfico de frecuencia para Tipo atencion
Tipo atencion
Consultas Médicas
Procedimientos y Cirugías
Consultas No Médicas
Gestión Clínica y Administrativa
Otros
Hospitalización y Cuidados Críticos
Atención Domiciliaria
Rehabilitación y Educación en Salud
Recuento
Gráfico de frecuencia para Rango Edad
Rango Edad
Personas Mayores
Adultos
Adolescencia
Primera Infancia
Niñez
Recuento
Gráfico de frecuencia para Prevision
Prevision
FONASA Tipo B
FONASA Tipo D
FONASA Tipo C
FONASA Tipo A
Isapre
Sin Informacion
Otro
Recuento
Gráfico de frecuencia para Rango Hora Cita
Rango Hora Cita
Mañana
Media Mañana
Tarde
Media tarde
Noche
Gráfico de frecuencia para Estival Fecha Cita
Estival Fecha Cita
Invierno
Primavera
Otoño
Verano
Recuento
Gráfico de frecuencia para Distancia comuna
Distancia comuna
Cercana (hasta 40 km)
Intermedia (41-80 km)
Lejana (+80 km)
Recuento
Gráfico de frecuencia para Agenda Biomedica
Agenda Biomedica
Otros
Medicina Clínica (Adultos)
Especialidades Quirúrgicas
Oncología y Enfermedades Complejas
Órganos de los Sentidos
Apoyo Terapéutico y Diagnóstico
Pediatría y Neonatología
Medicina Familiar y Rehabilitación
Salud Mental
Recuento
Gráfico de frecuencia para Tipo Cita
Tipo Cita
Presencial
Telefónica
Recuento

Figura 1.

Durante el proceso de imputación de datos, en la variable Previsión, se reemplazaron 1414 registros “Sin información” por la moda de la variable “FONASA Tipo B”. Luego, durante el proceso de balanceo de registros mediante SMOTE se obtuvo un cambio de distribución de 637104 de Paciente Asiste y 80058 de Paciente no se presenta, a 509596 de Paciente Asiste y 509596 de Paciente no se presenta .

Finalmente, se calcularon las métricas de desempeño de cada modelo implementado con el objetivo de predecir la no presentación del paciente a su cita. Se presentan los resultados en la Tabla 3

Tabla 3.

	Modelos de Machine Learning			
	Regresión Logistica	Random Forest con OHE	Random Forest con TE	XGBoost
Métricas				
Accuracy	0,5438	0,6773	0,6771	0,6603
Precision	0,9456	0,9329	0,9329	0,936
Recall	0,5166	0,6864	0,6862	0,6632
F1-score	0,6682	0,7909	0,7907	0,7763
Matriz de Correlación				
True Negative	12136	9630	9627	10142
False Positives	3789	6295	6298	5783
False Negatives	61640	39986	40013	42941
True Positives	65868	87522	87495	84567

Luego, se calcularon las métricas de desempeño del modelo de Regresión Logística y Random Forest con los distintos hiperparámetros ajustados. Se presentan los resultados en la Tabla 4

Tabla 4.

	Métricas			
Modelo de ML e hiperparámetros	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Regresión Logística (Predeterm.)	0.5438	0.9456	0.5166	0.6682
Regresión Logística (C=0.1)	0.5439	0.9455	0.5167	0.6682
Regresión Logística (C=10)	0.5439	0.9456	0.5167	0.6682
Random Forest con OHE (Predeterm.)	0.6773	0.9329	0.6864	0.7909
Random Forest con OHE (Ajustado)	0.6362	0.9449	0.6273	0.7540
Random Forest con TE	0.6771	0.9329	0.6862	0.7907
XGBoost (Predeterm.)	0.6603	0.9360	0.6632	0.7763

Finalmente, se identificaron las características más importantes según el modelo de Random Forest en base a si la categoría de la variable es usada frecuentemente para dividir nodos, genera grandes reducciones de impureza y/o aparece en niveles altos del árbol (promedio de importancia de la variable sobre todos los árboles). Las 10 variables/categorías con mayor importancia se indican en la Tabla 5.

Tabla 5.

Variable_Categoría	Importancia
Profesional_Enfermeria	0,07178
Agenda Biomedica_Apoyo Terapéutico y Diagnóstico	0,044037
Tipo atencion_Gestión Clínica y Administrativa	0,038324
Agenda Biomedica_Otros	0,032112
Tipo atencion_Consultas Médicas	0,031243
Prevision_FONASA Tipo	0,030351
Estival Fecha Cita_Primavera	0,028579
Rango Edad_Personas Mayores	0,028411
Estival Fecha Cita_Invierno	0,028343
Profesional_Medicina	0,027855

Interpretación de Resultados

El objetivo del análisis es predecir la no asistencia del paciente a su cita, priorizando la capacidad de detección de ausentismo. Esto sugiere que la métrica de interés corresponde al Recall para la clase “Paciente no se presenta”, lo cual corresponde al porcentaje de pacientes que no se presentarán a su cita que el modelo identifica correctamente.
Respecto de los resultados, en primera instancia se implementó un modelo de Regresión Logística por sus características para tareas de predicción. Si bien el modelo muestra un buen desempeño global (accuracy), posee un Recall bajo, lo que implica que no logra identificar adecuadamente a los pacientes que no asisten. Es posible que el modelo tienda a favorecer la clase mayoritaria (asistentes), lo cual es esperable en un contexto de desbalance de clases, incluso tras técnicas de balanceo. En resumen, el desempeño de este modelo es insuficiente, ya que el costo principal se basa en no detectar a los pacientes ausentes.

En segunda instancia se implementó un modelo de Random Forest entrenado con las mismas variables codificadas mediante One-Hot Encoding (OHE). Este modelo muestra una clara mejoría del Recall con una disminución marginal de la precisión respecto a la regresión logística, probablemente en base a su capacidad de capturar relaciones no lineales e interacciones. Sin embargo, es observable que el uso de OHE aumenta considerablemente la dimensionalidad de los datos, lo cual probablemente esté limitando la generalización del modelo. Para abordar aquella dificultad, se procede a entrenar nuevamente un modelo de Random Forest utilizando Target Encoding (TE) para las variables categóricas, reduciendo así la alta dimensionalidad. Sin embargo, este modelo obtiene un Recall similar al modelo de Random Forest con OHE, al igual que en las demás métricas.

En tercera instancia, se implementó un modelo XGBoost como alternativa de ensamble boosting. Este modelo muestra un Recall muy similar al Random Forest, sin una mejora en el resto de las métricas de desempeño.

El análisis posterior de identificación de importancia de las variables según el modelo de Random Forest identificó que las categorias Profesional_Enfermeria, Agenda Biomedica_Apoyo Terapéutico y Diagnóstico, y Tipo atencion_Gestión Clínica y

Administrativa corresponden a las de mayor importancia, sin corresponder a las de mayor frecuencia en los registros, lo que sugiere que su relevancia predictiva no está determinada por su frecuencia de ocurrencia, sino por su capacidad para discriminar entre pacientes que asisten y no asisten a sus citas, destacando el valor del modelo para identificar patrones no evidentes en el análisis descriptivo.

Respecto del ajuste manual de hiperparámetros en comparación con la configuración por defecto, se observan diferencias leves en el modelo de Random Forest, pero no así en el modelo de Regresión Logística. El uso de un mayor número de árboles junto con restricciones en la profundidad y el tamaño mínimo de los nodos permite controlar la complejidad del modelo de Random Forest y favorece la capacidad de generalización. Sin embargo, esto no se reflejó en mejores métricas, probablemente por la ausencia de tuning/optimización de hiperparámetros mediante GridSearchCV.

CONCLUSIÓN

Para el análisis predictivo de pacientes que no se presentan a su cita ambulatoria agendada en el Hospital Dr. Franco Ravera Zunino, los modelos de Random Forest mostraron un desempeño consistentemente superior a los de Regresión Logística, particularmente en términos de Recall, métrica prioritaria dada la relevancia de identificar pacientes ausentes. Si bien tanto Random Forest como XGBoost alcanzaron valores comparables de desempeño, el modelo de Random Forest entrenado con Target Encoding posee resultados de mejor desempeño, al ofrecer el mejor equilibrio entre capacidad predictiva, estabilidad y eficiencia computacional sin pérdidas sustantivas en precisión. No obstante, el ajuste de hiperparámetros podría mejorar las métricas de desempeño del modelo. En conjunto, estos resultados respaldan la selección del Random Forest con Target Encoding como el modelo más apropiado para una eventual implementación en el contexto clínico.

REFERENCIAS

1. Marbouh, D., Khaleel, I., Al Shanqiti, K., Al Tamimi, M., Simsekler, M. C. E., Ellahham, S., Alibazoglu, D., & Alibazoglu, H. (2020). Evaluating the Impact of Patient No-Shows on Service Quality. Risk management and healthcare policy, 13, 509–517. https://doi.org/10.2147/RMHP.S232114

2. Carreras-García D, Delgado-Gómez D, Llorente-Fernández F, Arribas-Gil A. Patient No-Show Prediction: A Systematic Literature Review. Entropy (Basel). 2020;22(6):675. Published 2020 Jun 17. doi:10.3390/e22060675

3. Russotto, A., Ragusa, P., Catozzi, D., De Angelis, A., Durbano, A., Siliquini, R., & Orecchia, S. (2025). Understanding No-Show Patterns in Healthcare: A Retrospective Study from Northern Italy. Healthcare (Basel, Switzerland), 13(15), 1869. https://doi.org/10.3390/healthcare13151869

4. Toker, K., Ataş, K., Mayadağlı, A., Görmezoğlu, Z., Tuncay, I., & Kazancıoğlu, R. (2024). A Solution to Reduce the Impact of Patients' No-Show Behavior on Hospital Operating Costs: Artificial Intelligence-Based Appointment System. Healthcare (Basel, Switzerland), 12(21), 2161. https://doi.org/10.3390/healthcare12212161