ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.248

Predictive Model Based on Machine Learning to Decrease Patient Attrition in Health Care Institutions in Lima Using Python.

Christian Ovalle[1]
[1]Universidad Tecnológica del Perú, Perú, dovalle@utp.edu.pe

Abstract– ***The present research addresses one of the main problems that prevent health problems in the country from being combated, identifying it as a significant challenge in health management. For this reason, it is necessary to generate a detection and/or prevention tool for these cases, so a predictive model is proposed to anticipate patients prone to drop out of the services of a health center. The research focuses on the Sanna El Golf clinic, where, by means of a predictive analysis, a 67% of assertiveness is obtained as a result, this approach shows substantial benefits for the clinic and highlights its contribution to meet the objectives set. In addition, the proposed model is positioned as a key tool in the prevention of medical attrition, identifying it as a significant challenge in health management.***

Keywords- Predictive Model, Machine Learning, Python, Logistic Regression.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Modelo predictivo basado en aprendizaje automático para disminuir la deserción de pacientes en instituciones de salud de Lima utilizando Python

Christian Ovalle[1]
[1]Universidad Tecnológica del Perú, Perú, dovalle@utp.edu.pe

Resumen– ***La presente investigación aborda uno de los principales problemas por los que no se pueden combatir los problemas de salud en el país, identificándolo como un desafío significativo en la gestión de la salud. Es por ello que, se hace necesario generar una herramienta de detección y/o prevención para estos casos, por lo que se propone un modelo de predicción para anticipar pacientes propensos a desertar de los servicios de un centro de salud. La investigación se centra en la clínica Sanna El Golf, donde, mediante un análisis predictivo, se obtiene como resultado un 67% de asertividad, esté enfoque muestra beneficios sustanciales para la clínica y destaca su contribución para cumplir con los objetivos planteados. Además, el modelo propuesto se posiciona como una herramienta clave en la prevención de deserciones médicas, identificándolo como un desafío significativo en la gestión de la salud.***

Palabras clave- Modelo Predictivo, Machine Learning, Python, Regresión Logística.

I. Introducción

Con la llegada de la pandemia, muchos establecimientos de salud se vieron afectados con relación a las atenciones que se realizaban, con ello las citas tuvieron que ser reprogramadas y buscar los canales necesarios para poder brindar las respectivas atenciones. Con ello, el MINSA indicó que se podían reaperturar las atenciones médicas que no sean COVID, llevando los implementos necesarios para las atenciones.

La deserción por parte de los pacientes que ya cuentan con una cita médica es uno de los principales problemas por los que no se puede combatir con los problemas de salud dentro del país. Es por ello, que es necesario generar una herramienta de detección y/o prevención para estos casos, se propone un modelo de predicción para identificar los pacientes potenciales a desertar los servicios de un centro de salud.

Según un estudio realizado en el 2019 en el Hospital Nacional del Perú [1], el 20.2% de consultas externas de 592 como total no fueron atendidas debido a que no se presentaron los pacientes, es decir, por pacientes desertores. Por lo que, esta situación no solo afecta la calidad de la atención médica, sino también la rentabilidad y eficiencia de los establecimientos de salud. En ese sentido, de no resolverse esta situación en las empresas del sector salud se incrementará la deserción de pacientes y las pérdidas potenciales en las empresas. Lo que nos trae una duda que viene aquejando a nuestro país desde ya hace unos años atrás: ¿Ha sido óptima la atención en primera instancia en nuestros centros de salud en estos últimos 10 años?

Es por ello por lo que se ha visto necesario hacer un análisis tanto en estudios teóricos como empíricos para descifrar cuál podría ser el causante de este problema con respecto al déficit en la gestión hospitalaria como en la atención oportuna de sus pacientes en general.

A. Machine Learning

Es una disciplina científica la cual haciendo uso de algoritmos, es capaz de aprender de si misma mediante la identificación de patrones que existen en un conjunto de datos para así poder predecir comportamientos futuros basándose en aquello aprendido previamente [2].

B. Regresión

Tiene como propósito establecer una relación entre un cierto número de características y una respuesta objetivo-continua, en la Fig. 1 se puede observar un ejemplo de Regresión Logística, lo cual se puede utilizar para predecir la probabilidad de que algún hecho ocurra [2].

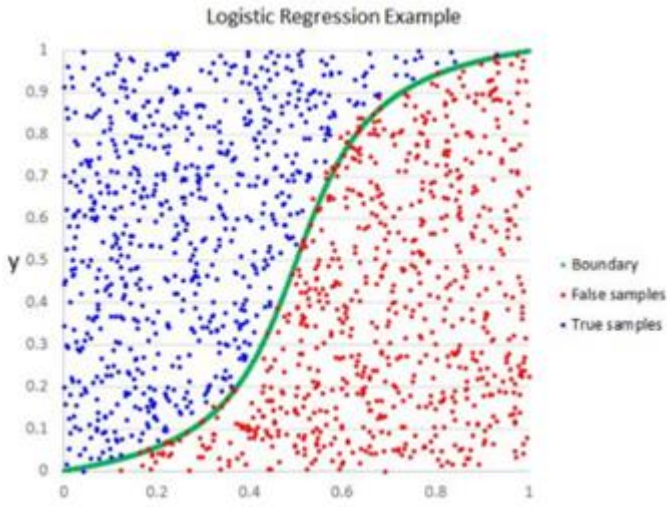

Fig. 1 Regresión continua.

II. Estado del arte

La disputa entre las citas no asistidas y el aforo de citas ha obligado a realizar una búsqueda para la satisfacción del paciente, motivo por el cual se ha empezado a considerar la implementación de nuevas herramientas para obtener mejores resultados de la rentabilidad en la administración e ingresos de los establecimientos de salud, una de las herramientas es el

modelo predictivo basado en Machine Learning (ML), este se precisa en un grupo de procedimientos que mediante el aprendizaje automático, el recolectar datos históricos y la inspección de los patrones procura brindar una predicción de futuros resultados, cuyo objetivo principal sería definir la toma de decisiones, al indicar que ese encuentra basado en ML, definido como un tipo de Inteligencia Artificial (IA) que proporciona a un sistema (hardware y/o software) la capacidad de aprender, sin ser explícitamente programadas [3]. Con respecto a lo anterior, esta herramienta es importante, ya que permite realizar diversas tareas de manera autónoma, es decir es accesible a que los ordenadores puedan aprender por sí mismo.

El modelo predictivo contiene una serie de procesos sumarizados en un algoritmo para el respectivo análisis a realizar. El aprendizaje automatizado o ML es una disciplina del entorno de la IA, la cual tiene como principal característica que un sistema aprenda sin intervención humana, es decir en base a la identificación de patrones complejos dentro de una gran cantidad de datos predefinidos en el transcurso del tiempo [4]. Para realizar una buena manipulación de datos, en la presente investigación se debe considerar el historial de cada paciente en cada cita agendada y/o asistida. Asimismo, también revisar el historial de especialidades que se frecuentan de manera mensual para poder revisar a detalle si existe alguna reprogramación de las citas que se agendó en su momento y el motivo por el cual no se haya asistido.

Para mejorar el proceso de la predicción de las citas desertadas por los pacientes se hacen uso de diversos métodos. Se considera como deserción de una prestación de salud cuando un paciente deja de asistir a una atención médica sin notificación previa, no se suelen comunicar los motivos de tal decisión, los cuales suelen ser la calidad de servicio recibido o la satisfacción del paciente, entre otros [5], muchas clínicas cuentan con el manejo de satisfacción al paciente lo cual cada uno lleva una calificación de atención brindada. Estas estadísticas se tienen que controlar y de preferencia deben ser óptimas, este método es de gran utilidad, ya que se puede obtener datos en caso algo no sea lo esperado se pueda corregir y brindar una mejor atención al paciente con una mala experiencia. Según [4] menciona que, hoy en día, ningún centro prestador de servicios de salud puede dejar de evaluar la calidad en la atención de forma permanente, en este caso las estadísticas que nos indiquen los pacientes nos ayudarán a determinar porque los pacientes desertan las citas que son agendadas previas. Los establecimientos de salud como las clínicas en Lima tienen una gran demanda, puesto que originan un gran efecto a la sociedad.

Según un estudio, se ha observado que el modelo predictivo basado en ML tiene un bajo costo y alto impacto en las empresas como valor agregado a estas es muy recomendado ya que en conjunto a estrategias del negocio se pueden reducir pérdidas y generar valor a la información que se recolecta durante el transcurso del tiempo. Con ello las empresas tendrá una mejor planeación de sus atenciones a brindar a sus pacientes sin generar pérdidas [5].

El estudio [6], tuvo como objetivo fue reducir la cantidad de citas perdidas en clínicas ambulatorias de radiología, lo cual los investigadores emplearon un modelo de ML y utilizaron 32,957 registros de citas para predecir los casos en los que los pacientes no asistirían, para ello implementó el algoritmo XGBoost, logrando una precisión del 65%, una exactitud del 61%, y un AUC del 75%. En otra investigacipin [7], sobre la aplicación de un algoritmo de ML XGBoost para desarrollar y validar un modelo de predicción de no llegadas ambulatorios, se obtuvo que el valor AUC fue de 0.768 (0.767–0.770), para ello, identificaron las características más influyentes del modelo, que incluyen citas reprogramadas, tiempo de espera (medido en días desde la fecha programada hasta la fecha de la cita), proveedor de citas, tiempo transcurrido desde la última cita en el mismo departamento. Se observa que programar citas cercanas a la fecha de una cita previa reduce la probabilidad de que un paciente no asista. En términos generales, el modelo de predicción mostró una buena calibración para cada departamento, especialmente en el rango de probabilidad operativamente relevante del 0 al 40%. Por otro lado, otro estudio [8], utilizó modelos de ML, para el análisis de una gran cantidad de datos de citas médicas para predecir el comportamiento y las características de los pacientes, empleando diez algoritmos diferentes., incluyendo clasificadores como el de bosque aleatorio, árbol de decisión, regresión logística, XG Boost, entre otros lo cual obtuvo un alto nivel de precisión en la predicción, con un recall del 94%, exactitud del 86%, puntaje F1 del 87%, área bajo la curva del 92% y un error cuadrático medio mínimo de 0.106, lo cual estos modelos mejoran la precisión de las predicciones sobre el estado de las citas médicas de los pacientes.

Respecto a la Clínica Sanna El Golf, se enfoca en brindar atención médica a los pacientes ya sea ambulatorio, emergencia y hospitalario, con la finalidad de ofrecer que su salud sea óptima y fuera de peligro. Uno de los principales problemas que no se puede combatir dentro del país es la deserción por parte de los pacientes que ya cuentan con una cita médica. Es por ello que, con la implementación de este modelo predictivo se podrá generar una herramienta de detección y/o prevención para estos casos, se propone este modelo de predicción para identificar los pacientes potenciales a desertar los servicios de un centro de salud, para optimizar de esta manera los horarios y tiempos de atención de los médicos, llevar un control sobre las citas atendidas y evitar pérdidas en las ganancias de la empresa.

Teniendo en cuenta la problemática, la finalidad de la presente investigación es diferenciar las particularidades que se presentan en la variable de ejecución del modelo predictivo en los establecimientos de salud, en relación con los estudios de investigación publicados en los últimos 5 años. Se justifica el trabajo de investigación, porque al realizar un análisis tanto en

estudios teóricos como empíricos se puede proporcionar datos que nos indiquen el motivo causante por el cual los pacientes no asisten a sus citas programadas en la Clínica Sanna El Golf, con esta información la clínica logrará alcanzar resultados precisos cuya finalidad será la mejora de la satisfacción del cliente y los ingresos a la clínica.

III. METODOLOGÍA

La investigación aplicada tiene como objetivo de estudio un problema destinado a la acción, donde puede aportar hechos nuevos, de modo que se pueda confiar en los hechos puestos al descubierto, por otra parte, "concentra su atención en las posibilidades concretas de llevar a la práctica las teorías generales, y destina sus esfuerzos a resolver las necesidades que se plantean la sociedad y los hombres" [6]. Por ende, la presente investigación, según su naturaleza es de tipo aplicada debido a que podemos obtener respuestas alentadoras por medio de conceptos teóricos y que al aplicarlos debidamente se concluye, en este caso, en soluciones innovadoras a problemas de gestión utilizando herramientas tecnológicas que se encuentran vigentes en el día de hoy.

Desde el punto de vista científico, la predicción se puede definir como una inferencia con respecto a un futuro acontecimiento para lo cual habría que hacer una traslación a través del tiempo de la explicación establecida desde un intervalo de tiempo pasado y conocido hasta otro intervalo futuro y por conocer [7]. Es por ello que el nivel de la presente investigación es de tipo predictive. ya que se utilizarán valores para ver el comportamiento y ratio de deserción en el pasado y en base a ellos poder realizar una predicción que nos ayude a anticipar estos sucesos y poder darle una solución óptima.

La presente investigación es de tipo experimental ya que se utilizará para establecer el efecto de una causa y se presentará la manipulación de una variable experimental no comprobada. Para esto, existen 2 acepciones para el término experimento. Nos centraremos en la primera que hace referencia a realizar una acción y después observa las consecuencias [8]. Las investigaciones con enfoque cuantitativo son las que por lo general siguen un patrón predecible y estructurado a lo largo del proceso de este [9]. De tal manera, esta investigación según sus características es de enfoque cuantitativo ya que sigue un proceso ordenado, detallado y sistemático desde el planteamiento de la pregunta de investigación hasta el análisis de los resultados obtenidos.

Para el proceso de la minería de datos vimos necesario utilizar la metodología propuesta por Fayyad [10] en 1996, denominada KDD. Esta metodología está dividida en 5 fases: Selección, preprocesamiento, transformación, minería de datos y evaluación e implementación.

Fig. 2 Fases del proceso de la minería de datos

A. *Selección de datos*

Para la etapa de selección se utilizaron en total 20 variables que se muestran en la tabla 1, de las cuales 19 son independientes y están asociadas a características demográficas y socioculturales de los pacientes, su historia clínica, grado de satisfacción y diferentes tiempos de espera para recibir los servicios brindados por el establecimiento de salud. También contamos con una variable dependiente que es el hecho de si un paciente ha desertado o no una atención médica en el centro de salud.

TABLA I
MATRIZ DE MODELO DE INFORMACIÓN

Variable	Descripción
Nacionalidad	País de nacimiento
estado_civ	Estado civil del paciente
edad	Grupo etario del paciente
	0-10: 1
	11-17: 2
	18-30: 3
	31 >: 4
sexo	género de nacimiento de la persona
brevete	Si el paciente posee o no licencia de conducir
t_espera_avg_cita	Tiempo de espera promedio desde la reserva hasta la atención de una cita
Satisfaccion	Grado de satisfacción promedio entre 1 - 10
t_espera_cola	tiempo de espera promedio en la cola
cant_aten_amb	Cantidad de atenciones ambulatorias atendidas en los últimos 30 días
cant_aten_eme	Cantidad de atenciones de emergencia atendidas en los últimos 30 días
cant_aten_hos	Cantidad de atenciones hospitalarias atendidas en los últimos 30 días
grupo_aseg	Agrupador de Aseguradoras
sit_laboral	Situación laboral (Dependiente/No dependiente)
cn_res_web	Cantidad citas reservadas por el canal de página web
cn_res_app	Cantidad citas reservadas por el canal de app móvil
cn_res_call	Cantidad citas reservadas por el canal de call center
cn_res_pre	Cantidad citas reservadas por el canal presencial
desercion	Indicador de deserción
tipo_pac_pps	Indicador de tipo de paciente pago por servicio
tipo_pac_cpm	Indicador de tipo de paciente costo paciente mes

- **Análisis Univariado**

Se realizó un análisis univariado, donde se analizaron datos del año 2022 en comparativa con el año anterior 2021 y la cantidad de deserciones por mes, esto nos sirvió para determinar patrones que existen cerca a los meses de diciembre – enero y entre junio – julio. También se observa una tendencia positiva sobre deserciones, concorde a la cantidad de atenciones que van en aumento en el transcurso del tiempo, para mayor detalle véase la Fig. 3.

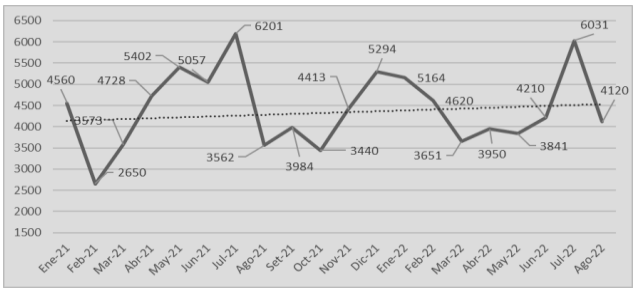

Fig. 3 Deserciones mensuales 2021-2022

Con respecto a las variables sociodemográficas, en la Fig. 4 se puede observar que el 42% son del género masculino y el 58% son del género femenino, para los valores inválidos se clasificó como "Otros" para su correcto procesamiento.

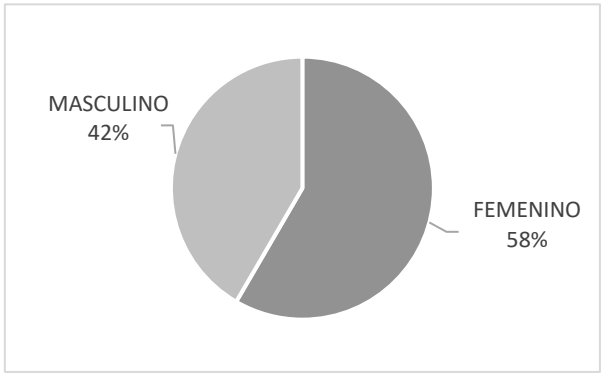

Fig. 4 Distribución por género

- **Análisis Bivariado**

Teniendo en relación el grado de satisfacción de los pacientes el cual se ha medido a través de encuestas enviadas vía correo electrónico, estas se obtuvieron mediante la plataforma web Qualtrics. Se transformaron los datos y se cruzaron en conjunto a la base de datos de los pacientes para así poder determinar el grado de satisfacción de cada uno de los pacientes en sus últimas atenciones. Se tiene como el concepto de NPS (Net Promoter Score) una puntuación de 1 a 10, a los pacientes que puntúan los servicios recibidos entre 1 y 6 se les categoriza como "detractores", los que puntúan entre 7 y 8 como "pasivos" o "neutros" y a los que dan una puntuación entre 9 y 10 como "promotores". Al momento de analizar los grados de satisfacción con el sexo de los pacientes nos indica que, si bien hay una cantidad considerable de pacientes promotores, la gran mayoría tiene votos detractores, mientras que los pacientes masculinos tienen mayor cantidad de votos neutros y promotores, para mayor detalle véase la Fig. 5.

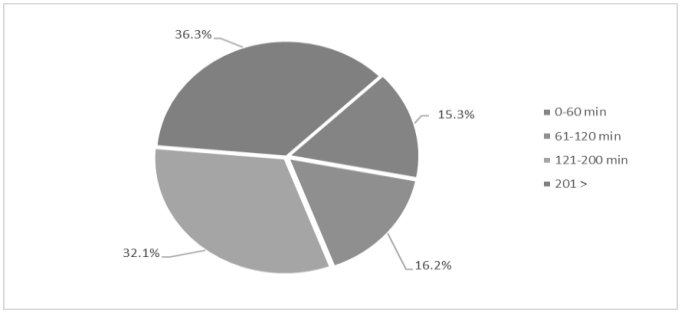

Fig. 5 Grado de satisfacción por género

Con esta información podemos sugerir que tanto el género de un paciente como su grado de satisfacción están directamente relacionados y por consiguiente a la deserción de una atención médica.

Se manejó como concepto cuantitativo y medible de una deserción médica el siguiente enunciado: abandono de una cita médica para cualquier atención ambulatoria, se consideran solo consultas externas y no procedimientos de apoyo al diagnóstico, este abandono tiene que haberse realizado en el transcurso del último mes. El paciente pudo o no, haber tenido atenciones médicas con anterioridad, todos los pacientes al momento de su registro en el aplicativo móvil o plataforma web proporcionan su fecha de nacimiento, en caso de reservas de manera presencial o por call center ocurre lo mismo. Teniendo la variable de edad se analizó con respecto al indicador de deserción. En la Fig. 6, se puede observar que los pacientes con mayor tendencia de deserción son los que tienen atenciones pediátricas entre 0 y 10 años y también los que tienen más de 30 años. Los pacientes que no se encuentran en esos rangos tienen índices de deserción menores al 22,4% con respecto al total de atenciones realizadas. Lo que sugiere que el grupo etario sea un factor que influya en la deserción.

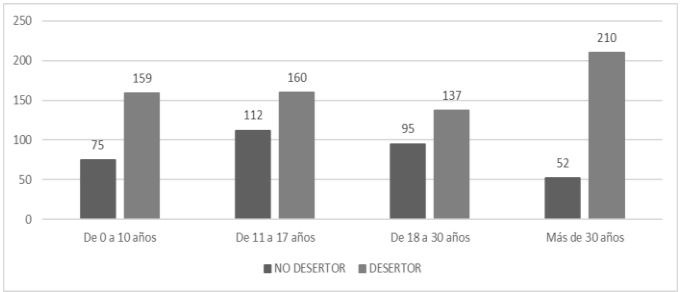

Fig. 6 Deserciones por grupo etario

Al momento de analizar la variable de tiempos de espera se pudo observar que en el macroproceso de agendamiento y preatención de una consulta médica existen dos tiempos de espera primordiales. El primero es el tiempo de espera hasta la siguiente consulta disponible, es decir el tiempo de espera entre el agendamiento hasta la atención el cual puede durar desde tiempos cortos menores a un día hasta varias semanas. El segundo se define como el tiempo de espera en el sector de admisión en el cual se efectúa el pago por la atención antes de ser recibida, al tener un tiempo de espera mayor en cualquiera de ambos casos, podemos observar que el porcentaje de deserción aumenta cuando el tiempo de espera es mayor, para mayor detalle véase la Fig. 7.

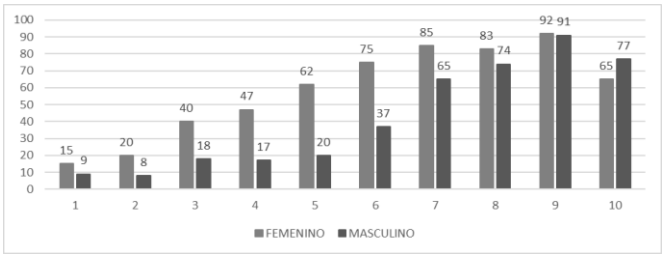

Fig. 7 Deserciones por tiempo en cola

B. Tratamiento y limpieza de datos

Para este punto se realizaron los siguientes procesos y técnicas de limpieza y tratamiento, teniendo en cuenta que se cuentan con variables tanto cuantitativas como cualitativas. De igual manera para datos inválidos, nulos y atípicos se realizaron las correspondientes correcciones teniendo en cuenta las mejores prácticas de ciencia de datos, como se puede observar en la tabla 2.

TABLA II
CLASIFICACIÓN DE TRATAMIENTO

Problemática en datos	Tratamiento realizado	
Datos nulos o perdidos	Imputación paramétrica	
		Variable cualitativa: Mediana
		Variable cuantitativa: Mediana
Datos categóricos o etiquetas	Codificación numérica	
		Label encoder
Datos atípicos o anómalos	Recodificación	
		Percentiles

C. Modelo de Machine Learning

Para la siguiente investigación se seleccionó el modelo Extreme gradient boosting (XGBoost), este modelo se fundamenta en el concepto de árboles de decisión y utiliza un enfoque de entrenamiento en serie, donde conjuntos de datos se utilizan para mejorar gradualmente la capacidad predictiva, por lo que implica combinar predictores más débiles para formar predictores más robustos, y se lleva a cabo mediante entrenamientos sucesivos para optimizar la función objetivo y su proceso de entrenamiento se detiene una vez que la función objetivo alcanza su valor mínimo [11]. Lo cual, resulta esencial, para identificar los principales drivers o variables explicativas que motivan a un paciente a desertar una atención médica, es decir su relación a la variable dependiente. Cabe recalcar, que la selección de este modelo se basó en investigaciones previas, para dar una mayor validez y respaldo por varias investigaciones [6], [7], [8]. Asimismo, se hizo la partición del dataset en un set de entrenamiento y uno de validación, sus proporciones fueron 66% y 33% respectivamente, como se observa en la Fig. 8.

Fig. 8 Distribución de datos de entrenamiento y validación

D. Evaluación de modelo

Se definió la métrica de Exactitud (ACC) para evaluar el desempeño del algoritmo XGBooost, ya que esta métrica se refiere a la proporción de las predicciones correctas realizadas por el modelo en comparación con el total de predicciones realizadas [12]. Además, brinda una visión de la efectividad de este modelo, En este contexto, se consideran los verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN) [13], como se ve en (1), lo que contribuye a calcular la precisión del modelo.

$$ACC = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

Asimismo, en la presente investigación se utiliza la siguiente herramienta de evaluación del rendimiento del modelo, como la Receiver Operating Characterisctic o Curva ROC, ya que es una representación gráfica que muestra el desempeño de un modelo de clasificación binaria en un plano bidimensional, además muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos [14], por lo que nos será de gran ayuda para evaluar y comparar la capacidad del modelo.

IV. Resultados

Para poder mostrar los resultados de la presente investigación, se tuvo en consideración que el entendimiento del negocio ya es implícito en la propuesta de los objetivos planteados. Con la anterior premisa establecida, se hizo abordando el análisis realizado a las variables del modelo, lo cual se obtuvieron los siguientes resultados.

Como se mencionó anteriormente, se adoptó la proporción de 66% y 33% según la partición del dataset. Posteriormente en la Fig. 9 se observa la valoración de las variables explicativas con respecto a la variable objetivo según el modelo XGBoost, es decir el peso que tiene cada uno de estos aspectos con la decisión de un paciente por desertar o no una atención médica. Las variables con mayor puntuación son más valoradas como por ejemplo el tiempo de espera en la cola siendo la mayor valorada con una puntuación de 98, mientras que la menor significativa es si el paciente tiene o no un brevete con un puntaje de 20.

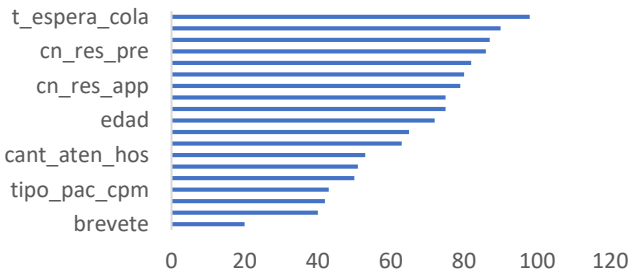

Fig. 9 Puntuación de relevancia de variables descriptivas

Una vez definidas las variables explicativas se procedió el entrenamiento del modelo de regresión logística mediante la librería sklearn. Posterior a ello, se entrenó el modelo, lo cual se ejecutó el algoritmo de regresión el cual según la configuración de cada iteración se evalúa y se obtiene el valor predicho.

La herramienta utilizada para poder mostrar los resultados fue la matriz de confusión, lo cual representan los aciertos y no aciertos como se puede apreciar en la tabla 3. Se obtuvieron un total de 672 aciertos y 328 no aciertos teniendo un total de 67.20% de precisión.

TABLA III
MATRIZ DE CONFUSIÓN

	deserción	
predicción	0	1
0	348	198
1	130	324

Asimismo, para medir el entrenamiento se empleó la Curva Roc, lo cual se generó una curva teniendo los FP y VP como ejes X e Y respectivamente. Se considero que el mejor método posible y más optimista se acerca a la esquina superior izquierda del gráfico mientras que un modelo no predictivo se encontraría en la diagonal de aleatoriedad. Los resultados del entrenamiento del modelo podemos observar que el área generada debajo de la curva o AUC es de 0.67, como se puede observar en la Fig. 10. Este valor indicaría que la predicción del modelo es regular-optimista.

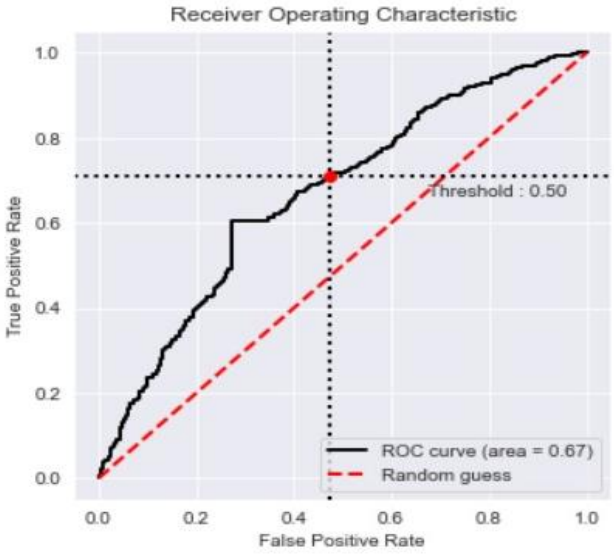

Fig. 10 Curva ROC del modelo predictivo

V. Conclusiones

A. *Discusión*

El objetivo general de la presente investigación realizada fue implementar un modelo de regresión logística basado en ML que permita predecir las deserciones en pacientes, para poder determinar la factibilidad y practicidad de una herramienta de bajo costo y alto impacto en el negocio, esto se comprueba al haber generado un modelo predictivo con sklearn. La hipótesis propuesta se comprueba al haber demostrado que hay una relación directa entre los datos de los pacientes con sus decisiones por desertar atenciones médicas como se ve en la Fig. 10. Se comprueba que tiene más impacto los tiempos de espera, el grado de satisfacción y la cantidad de atenciones previas, las variables antes mencionadas tienen relación directa a las deserciones en el establecimiento de salud, teniendo como instrumento el árbol de toma de decisiones XGBoost.

Esta solución favorece a la empresa del caso de estudio y a las demás empresas del sector salud al proponerse como una herramienta de bajo costo, y que disminuye las pérdidas de ingresos por producción al optimizar los horarios de programaciones médicas El modelo de ML sklearn alcanzó

indicadores aceptables a las métricas de validación de modelos predictivos obteniendo un índice de sensibilidad del 0.67 en la curva ROC. Lo cual indica que para atenciones médicas desertadas se tuvo un acierto de predicción del 67%. Los factores más relevantes para la identificación de pacientes más propensos a desertar una atención médica son los tiempos de espera, el grado de satisfacción en atenciones previas y el canal de reserva de las citas generadas. Para concluir, este modelo de Machine Learning tiene una correcta identificación o especificidad del 67% con respecto a las deserciones de atenciones médicas, comprobando así que un modelo predictivo de ML está directamente vinculado a las deserciones de atenciones médicas.

Asimismo, el estudio [6] obtuvo una exactitud del 61% empleando el algoritmo XGBoost, lo cual se centró en reducir las ausencias a citas en clínicas ambulatorias de radiología. En otra investigación [7], también aplicó el algoritmo XGBoost para desarrollar y validar un modelo de predicción de ausencias en citas ambulatorias, alcanzando un valor AUC de 0.768, lo cual en esa investigación dentificaron características influyentes como citas reprogramadas, tiempo de espera, proveedor de citas y tiempo transcurrido desde la última cita en el mismo departamento, por lo que el modelo mostró una buena calibración. Por otro lado, el estudio [8] utilizó diversos modelos de ML para analizar una gran cantidad de datos de citas médicas y predecir el comportamiento de los pacientes, empleando diez algoritmos diferentes, uno de ellos fue el XGBoost logrando unaexactitud del 86% y AUC del 92%, lo cual estos resultados sugieren que este modelo mejoran la precisión de las predicciones sobre las deserciones de los pacientes.

B. Limitaciones

En la presente investigación se obtuvo como limitación la disponibilidad de datos respecto a los horarios de los médicos dentro de la clínica, puesto que los horarios al ser rotativos o de emergencia, así como la falta de un sistema de control para rastrear el tiempo de espera del paciente en situaciones de emergencia. Además, la disponibilidad de los médicos puede variar según la especialidad y el tipo de atención requerida en un momento dato, lo cual requirió un mayor tiempo para recopilar información de manera precisa. Lo cual, esta falta de datos detallados sobre los horarios de los médicos pudo haber influido en la capacidad del modelo para predecir con precisión la asistencia de los pacientes a sus citas médicas. Para ello, se implementó un enfoque proactivo que incluyó la colaboración del personal médico y administrativo de la clínica, asimismo se establecieron canales de comunicación para recopilar información actualizada sobre los horarios médicos y cualquier cambio relevante en la disponibilidad de atención.

Si bien la predicción de la deserción de los pacientes se basó en datos de citas agendadas, lo cual implica ciertas limitaciones en este enfoque. Para ello, se implementó medidas adicionales para poder mitigar estas limitaciones, como la validación cruzada de los datos, lo cual se realizó para evaluar la estabilidad y la precisión del modelo, dividiendo repetidamente los datos en conjuntos de entrenamiento y prueba. Además, se analizaron las tendencias históricas de deserción para identificar patrones y factores influyentes adicionales permitiendo obtener un porcentaje de deserciones con la finalidad de no generar pérdidas por las atenciones no asistidas y poder brindar atención a otros pacientes.

C. Conclusiones

Con el análisis realizado actualmente en la clínica Sanna El Golf, mediante un modelo predictivo (XGBoost) se obtiene como resultado un 67% de asertividad, el cual beneficia a la clínica poder tomar acción sobre la deserción de pacientes, lo cual representa un paso significativo hacia la optimización de los procesos de atención en salud, puesto que al obtener el conocimiento sobre la deserción de sus pacientes a sus citas le permite poder agendar a otros pacientes en espera y evitar pérdidas por las citas separadas.

Cabe recalcar que, si bien este modelo beneficia a la Clínica Sanna El Golf, su aplicación se puede considerar en diferentes contextos de salud, ya que el modelo basado en XGBoost es ampliamente utilizado en el campo de ML, demostrando su eficacia en gran variedad de aplicaciones y entornos, lo cual podría adaptarse fácilmente en diferentes contextos de la salud. Pero es necesario reconocer que existen oportunidades para futuras investigaciones más profundas permitiendo un visón más completa y detallada de la deserción de pacientes en los centros de salud, contribuyendo así a la mejora continua de los servicios de atención médica.

El modelo predictivo basado en ML implementado en la clínica Sanna El Golf presenta un avance significativo, lo cual demostró su contribución y su finalidad de los objetivos propuestos, además se sugierte considerar la aplicación de este modelo en diferentes contextos de salud, pero se reconoce la necesidad de seguir investigando y mejorando este enfoque para abordar de manera más afectiva el problema de deserciones de pacientes en los centros de salud.

REFERENCIAS

[1] C. Díaz, V. Benites, E. Peña, M. Apolaya, and D. Urrunaga, "Factores asociados a deserción en consulta externa en hospital del Seguro Social del Perú," *Revista Médica del Instituto Mexicano Seguro Social*, vol. 57, no.5, pp. 307-313, 2019. Available: http//www.monografias.com/trabajos15/investcientífica/investcientífica.shtm

[2] D. Rolnik, *et al*. "Tackling Climate Change with Machine Learning," *ACM Computing Journals*, vol. 55, no. 2, Mar. 2023, doi: 10.48550/arXiv.1906.05433

[3] A. Walid, M. Ahmed, M. Zeyad, S. Galib, and M. Nesa, "Analysis of machine learning strategies for prediction of passing undergraduate admission test," *International Journal of Information Management Data Insights*, vol. 2, pp. 22-53, Nov. 2022, doi: 10.1016/j.jjimei.2022.100111

[4] E. Stenwig, G. Salvi, P. Rossi, and N. Skjærvold, *"*Comparative analysis of explainable machine learning prediction models for hospital mortality,*"* *BMC Medical Research Methodology*, Feb. 2022, doi: 10.1186/s12874-022-01540-w
[5] O. Álvarez, *"*La inteligencia artificial en la gestión de proyectos de inversión pública del Ministerio de Vivienda, Construcción y Saneamiento,*"* Perú, Jun. 2021
[6] L.R. Chong, K.T. Tsai, L.L. Lee, S.G. Foo, P.C. Chang, "Artificial intelligence predictive analytics in the management of outpatient MRI appointment No-shows," *Am J Roentgenol*, vol.215, no. 5, pp. 1155-1162, 2020, doi: 10.2214/AJR.19.22594
[7] K. Coppa, EJ, Kim, MI. Oppenheim, *et al*. "Application of a Machine Learning Algorithm to Develop and Validate a Prediction Model for Ambulatory Non-Arrivals," *J GEN INTERN MED*, vol. 38, pp. 2298–2307, 2023, doi: 10.1007/s11606-023-08065-y
[8] Z. Qureshi, A. Maqbool, A. Mirza, M. Z. Iqbal, *et al*., "Efficient Prediction of Missed Clinical Appointment Using Machine Learning," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021, doi: 10.1155/2021/2376391
[9] G. Baena, *"*Metodología de la investigación,*"* 3a. ed., México: Patria, pp. 18, 2017.
[10] E. De Gotari, *"*La metodología, una discusión y otros ensayos sobre el método,*"* México, D.F. Grijalbo, pp. 69, 1980
[11] R. Hernández, *"*Metodología de la investigación,*"* 6a ed., pp. 151, *2014*.
[12] R. Hernández *"Metodología de la investigación,"* 6a ed., pp. 7, *2014*.
[13] U. Fayyad, G. Piatetsky-Shapiro, *"*From Data Mining to Knowledge Discovery in Databases,*"* American Association for Artifiical Intelligence, United States, 1996
[14] S. Roopashree, J. Anitha, S. Challa *et al*, "Mapping of soil suitability for medicinal plants using machine learning methods," *Sci Rep*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-54465-3.
[15] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-54807-1.
[16] K. Rahman, A. Ghani, S. Misra, and A. U. Rahman, "A deep learning framework for non-functional requirement classification," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-52802-0.
[17] C. Gui, "Link prediction based on spectral analysis," *PLoS One*, vol. 19, no. 1, Jan. 2024, doi: 10.1371/journal.pone.0287385.