



Innovación y Software

ISSN: 2708-0927

ISSN: 2708-0935

facin.innosoft@ulasalle.edu.pe

Universidad La Salle

Perú

Aquino Arcata, Rene; Cuevas Machaca, Ronald; Godoy Montoya, Luis; Rodríguez Puma, Heber
Aplicación de regresión logística para la predicción de demanda
por especialidad médica en consulta externa hospitalaria
Innovación y Software, vol. 2, núm. 2, 2021, Septiembre-Febrero, pp. 44-59
Universidad La Salle
Perú

Disponible en: <https://www.redalyc.org/articulo.oa?id=673870839004>

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

redalyc.org

Sistema de Información Científica Redalyc

Red de Revistas Científicas de América Latina y el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso
abierto

Tipo de artículo: Artículos originales

Temática: Inteligencia artificial

Recibido: 25/05/2021 | Aceptado: 02/07/2021 | Publicado: 30/09/2021

Identificadores persistentes:

ARK: [ark:/42411/s6/a45](https://nbn-resolving.org/urn:nbn:org:ark:42411/s6/a45)

PURL: [42411/s6/a45](https://nbn-resolving.org/urn:nbn:org:ark:42411/s6/a45)

Aplicación de regresión logística para la predicción de demanda por especialidad médica en consulta externa hospitalaria

Application of logistic regression for the prediction of demand by medical specialty in hospital outpatient consultation

Rene Aquino Arcata ¹[\[0000-0002-5041-7344\]](#)*, Ronald Cuevas Machaca ²[\[0000-0002-3887-6396\]](#), Luis Godoy Montoya ³[\[0000-0001-8860-8843\]](#), Heber Rodríguez Puma ⁴[\[0000-0003-1779-5738\]](#)

¹ Universidad Jorge Basadre Grohman, Tacna, Perú. reneaquino@gmail.com

² Hospital Regional de Moquegua, Moquegua, Perú. rzcuevasm@gmail.com

³ Universidad Jorge Basadre Grohman, Tacna, Perú. luis.godoy.montoya@gmail.com

⁴ Universidad Jorge Basadre Grohman, Tacna, Perú. wisapuma@gmail.com

* Autor para correspondencia: reneaquino@gmail.com

Resumen

En este trabajo se realizó el análisis de la información producto de la atención de pacientes en el servicio de consulta externa. Se han revisado trabajos que guardan relación con las metodologías posibles de utilizar, antes de la elección de una en particular. Posteriormente, se ha justificado y aplicado la metodología de regresión logística para evaluar, clasificar y pronosticar los resultados esperados conforme al objetivo trazado. En el Hospital Regional de Moquegua, desde el inicio de la emergencia sanitaria por el Covid-19, se suspendió la atención en el servicio de consulta externa, vale decir desde Marzo del 2020 a Junio 2021 no se tiene información de cuánto hubiese sido la demanda por especialidad en dicho servicio. El objetivo del trabajo es predecir, en base a variables de edad y sexo, la cantidad de pacientes de sexo femenino que solicitarán una cita para las especialidades de consulta externa, en un período de tiempo. Para la resolución del objetivo planteado, se aplicó el modelo de regresión logística de scikit-learn que, en un inicio ha permitido clasificar y determinar el grupo de importancia en base al cual está orientado nuestro objetivo, tomando como variables independientes y relevantes: el sexo y la edad. Los resultados iniciales obtenidos del procedimiento del modelo no mostraron correspondencia real a la predicción esperada. Las conclusiones determinan que el modelo propuesto requiere la inclusión de otras variables de entrada.

Palabras clave: atenciones médicas, covid-19, predicción, regresión logística

Abstract

In this work, the analysis of the information produced by the care of patients in the outpatient service was carried out. Studies have been reviewed that are related to the possible methodologies to be used, before choosing one in particular.

At the Regional Hospital of Moquegua, since the beginning of the health emergency due to Covid-19, care in the outpatient service was suspended, that is, from March 2020 to June 2021 there is no information on how much the demand would have been by specialty in said service. The objective of the work is to predict, based on age and sex variables, the number of female patients who will request an appointment for outpatient specialties, in a period of time. To solve the problem, the logistic regression technique was used, which initially allowed us to classify and determine the importance group on the basis of which our objective is oriented, taking sex and age as relevant variables. The results obtained from the initial procedure of the model did not show real correspondence to the expected prediction. The conclusions determine that the proposed model requires the inclusion of other input variables

Keywords: *medical care, covid-19, prediction, logistic regression*

Introducción

En los hospitales del país, el servicio de consulta externa, es el segundo nivel de atención después de los puestos y centros de salud. La diferencia en la prestación de servicios está dada por la atención especializada, por profesionales de la salud que han profundizado en el estudio especializado relativo a un área específica del cuerpo humano, a técnicas quirúrgicas específicas o a un método diagnóstico determinado.

El servicio de consulta externa del Hospital Regional de Moquegua es la unidad orgánica encargada de sistematizar la atención integral de la salud y la referencia y contrarreferencia de los pacientes nuevos y/o continuadores, a los cuales el Hospital venía atendiendo de forma continua hasta la declaratoria de la emergencia sanitaria en el país por la pandemia del Covid-19. Al servicio de consulta externa concurren pacientes con diferentes características, como tipos de seguro, sexo, etnia, de forma continua o por primera vez, tanto nacionales como extranjeros. De la información publicada en el Boletín Estadístico del Hospital – 2019 [1], se registra un total de 50,659 atenciones que corresponde al servicio de consulta externa, de las cuales 19,958 fueron de sexo masculino y 30,701 de sexo femenino.

Desde el 16 de marzo de 2020, fecha que entra en vigencia la cuarentena a consecuencia de la emergencia sanitaria por el Covid-19 en el Perú, se suspendió la atención total en el servicio de consulta externa del Hospital Regional de Moquegua. Los pacientes asegurados y no asegurados dejaron de recibir atención médica en un espacio de tiempo particular, dado que de enero a marzo abarca la estación de verano y el período vacacional para la mayoría de trabajadores y estudiantes del país.

De los pacientes atendidos en consulta externa, un porcentaje asiste por primera vez al hospital o regresan después de períodos prolongados de tiempo, que pueden ser meses o años. Otro conjunto de pacientes asisten continuamente al hospital ya sea para seguir un tratamiento prolongado o por nuevas dolencias derivadas o no de su enfermedad principal. En conjunto, todos los pacientes que se atendían por consulta externa han dejado de recibir atención por un lapso aproximado de 16 meses, lo que posiblemente habría provocado un mayor deterioro en la salud de la población. Otro hecho derivado de esta situación, luego de iniciada la reactivación económica, es la probable concurrencia de los pacientes a Instituciones Prestadoras de Servicios de Salud (IPRESS) privadas, lo que habría involucrado una afectación mayor a su economía personal o familiar, considerando los costos diferenciados con respecto a IPRESS públicas como el Hospital Regional de Moquegua. Estos hechos deben estar siendo evaluados por el máximo ente rector del sector salud en Perú, el Ministerio de Salud, así como por las demás autoridades en los distintos niveles de gobierno y el propio hospital, entonces es importante conocer cuál será la probable demanda futura en el servicio de consulta externa para las distintas especialidades ofertadas.

Con el avance y aplicación de la Inteligencia Artificial como herramienta para predecir, agrupar o clasificar grandes cantidades de datos, es la tecnología elegida para la resolución del problema planteado en el presente trabajo. De sus técnicas desarrolladas, es la regresión logística, por su aplicación al aprendizaje automático para clasificación, considerada como una red neuronal en miniatura y dado que ampliamente ha demostrado que funciona muy bien cuando hay muchísimos datos y las interrelaciones entre ellos no son muy complejas, sirve como sustento para priorizar su uso y aplicación.

De similar importancia son las Redes Neuronales, que más allá de imitar el funcionamiento de las redes neuronales de los organismos vivos. Se basan en una idea sencilla: dados unos parámetros hay una forma de combinarlos para predecir un cierto resultado. En suma, son un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo. En el lenguaje propio, encontrar la combinación que mejor se ajuste es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" la combinación.

En consecuencia, el objetivo del trabajo es predecir, en base a las variables de edad y sexo de los pacientes, cuántas mujeres, niñas o adultas, se atenderán en una especialidad de consulta externa, considerando un periodo de tiempo específico. Es necesario justificar nuestro enfoque en el grupo de interés señalado, principalmente porque representa el

60.60 % de la totalidad de pacientes atendidos en el último año (2019) del período considerado para nuestro análisis (2015-2019), vale decir dos tercios de la población atendida en el año previo a la pandemia del Covid-19. Para éste propósito, se ha se hecho un análisis de otros trabajos relacionados y su utilidad respecto a sus objetivos, éstos abarcan a áreas como la educación, enfocado en el rendimiento académico; también, al área de los negocios y el riesgo de quiebra por variables financieras; así como, el comportamiento de los afiliados a un organización profesional y el cumplimiento con sus obligaciones económicas. En conjunto, todos estos trabajos tienen una variable común, el tiempo.

Trabajos relacionados

En la investigación propuesta en [2] El objetivo de este estudio es evaluar la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito/fracaso académico, partiendo de variables, como la asistencia y la participación en clase, cuya relevancia ya ha sido puesta de manifiesto en anteriores trabajos de nuestro equipo (Alvarado y García Jiménez, 1997). La muestra la constituyeron 175 universitarios de primero de psicología, tomándose los datos en la asignatura de «Métodos y Diseños de Investigación en Psicología I», del área de Metodología. Las conclusiones de este estudio son que (a) el rendimiento previo es un buen predictor del rendimiento futuro y (b) la asistencia y sobre todo la participación son variables con un peso importante en la predicción del rendimiento. Sin embargo, esta investigación tiene el inconveniente que el *dataset* está constituido por un reducido conjunto de datos a comparación del *dataset* que se utiliza en el presente trabajo.

En [3] se explora un método para realizar la predicción de la tendencia de cierre del indicador S&P 500 con un horizonte de pronóstico de 1 día, adecuando el problema de interés a un problema de clasificación binaria; asignando 1 si la predicción del indicador es creciente y 0 si es decreciente. Al final, se evalúan algunas pruebas de hipótesis para establecer conclusiones sobre la estacionalidad de la serie temporal, analizando los resultados más relevantes obtenidas en las implementaciones, de los cuales destacan niveles de exactitud de 52.51% y 64.04% para los modelos LSTM y Regresión Logística respectivamente.

En la investigación [4] se puso como objeto el desarrollo de un modelo de predicción de riesgo de quiebra con base en la metodología de regresión logística, para las micro y pequeñas empresas del sector comercial del Ecuador, identificando como factores influyentes las razones financieras de liquidez, solvencia, actividad, endeudamiento y

rentabilidad, así también, las variables de edad y tamaño de las empresas. Se identifica cuál de estos factores, son los que mayor impacto generan en la estabilidad.

Se concluye que el modelo propuesto en la investigación permite medir de una forma aceptable el nivel de probabilidad de riesgo de quiebra al que se expone una empresa comercial del Ecuador, logrando un 69.76% y 100% a tres y un año antes de que el fracaso ocurra.

En el trabajo [5] se presenta una comparación de indicadores de rendimiento del modelo de deserción actual de la Universidad del Bío-Bío (UBB), el cual está basado en la técnica de regresión logística y se compara con un nuevo modelo basado en árboles de decisión. El nuevo modelo se obtiene a través de metodologías de minería de datos y fue implementado a través de la herramienta SAP Predictive Analytics. Para entrenar, validar y aplicar el modelo se dispone de información real de las bases de datos de la Universidad de Bio Bio -UBB.

El modelo propuesto obtiene una exactitud del 86%, una precisión del 97% con un porcentaje de error del 14% en la predicción de la deserción estudiantil, muy superior al valor que entrega el modelo basado en regresión logística. Posteriormente, el modelo de predicción obtenido es optimizado considerando para ello otras variables logrando de este modo mejoras en los indicadores de predicción.

En el trabajo [6] se comparó una técnica de aprendizaje automático y una técnica clásica, con el objetivo de determinar qué técnica es más eficiente en la predicción de la morosidad de cuotas sociales en el Colegio de Ingenieros del Perú Consejo Departamental de Lambayeque. Las técnicas a comparar seleccionadas fueron máquina de soporte vectorial y regresión logística. Para efectuar la comparación de las técnicas se dispone de datos históricos de los colegiados cuya información se obtuvo de fuentes internas y externas a la organización. Posteriormente los datos recopilados por medio del proceso de extracción, transformación y carga (ETL) se limpió y estandarizó obteniéndose datos concisos y relevantes. Finalmente se aplicó las técnicas predictivas cuyos resultados son favorables para la máquina de soporte vectorial en comparación con la regresión logística.

Concluyendo que la técnica máquina de soporte vectorial es más eficiente para predecir la morosidad de cuotas sociales en el Colegio de Ingenieros del Perú Consejo Departamental de Lambayeque.

En la investigación [7] se estableció como objetivo el comparar el Modelo de regresión lineal Múltiple frente al Árbol de regresión, para ello se utilizó las variables Precio máximo de las acciones de Intel en función al Precio de apertura y

Volumen de ventas, de acciones por día. El *dataset* está conformado por todas las acciones de la empresa Intel desde su creación y a través del tiempo; se empleó el muestreo no probabilístico por conveniencia, se consideró desde mayo del 2018 hasta octubre del 2019 siendo un total de 410 registros recopilados a partir de la revisión documentaria. Las pruebas estadísticas usadas fueron el Análisis de regresión lineal múltiple y los Árboles de regresión. Los resultados obtenidos fueron; el Modelo de Regresión lineal múltiple con la técnica de eliminación de datos atípicos queda definida por la siguiente ecuación $Y=0.02856+1.003X_1+0.00000009405X_2$. Alcanzando una prueba F significativa y la bondad de ajuste es bastante alta $R^2=0.9979$, y un Error Estándar Residual de 0.2257 dólares, El Árbol de regresión establece que la variable para explicar el Precio máximo de acciones es el Precio de apertura, eliminando la variable volumen, el Error Medio Cuadrático es de 1.4480 dólares. Finalmente se concluye que el mejor modelo para predecir el precio máximo de acciones de Intel es el modelo de Regresión Lineal Múltiple con eliminación de puntos Outliers.

Materiales y métodos o Metodología computacional

En la actualidad nos vemos inmersos en la denominada era de la información, en la que el conocimiento es un gran activo para las compañías, sobre todo cuando se trata de conocer más al público objetivo únicamente basándose en información que se genera como resultado del usos de servicios que de alguna forma son almacenados en sistemas de información como datos históricos, datos que con las herramientas y estrategias adecuadas pueden servir para identificar hábitos de consumo, preferencias y costumbres ya sea en la gran red Internet como en el uso de servicios como supermercados, tiendas de consumo, grifos, entidades financieras, etc.

Toda la información que se produce diariamente y en cada instante influye en gran medida para orientar campañas de publicidad, campañas de lanzamiento de nuevos productos, habiendo predicho de antemano la respuesta del público. Este universo de conocimiento es el insumo del cual se nutre la Ciencia de Datos mediante la técnica del Machine Learning, logrando con gran aproximación identificar comportamientos, tendencias, patrones a lo largo de periodos de tiempo que con mucha probabilidad se volverían a presentar en un futuro próximo. En resumen, es así como se genera la predicción sobre el conjunto de datos.

En el presente trabajo de investigación se empleó como herramienta principal el entorno web GOOGLE COLABORATORY, adicionalmente se emplearon librerías para la implementación de algoritmos de machine learning (para el presente trabajo, librerías que permitan implementar algoritmos de árbol de decisiones: pandas, numphy, keras,

matplotlib) en el lenguaje de programación Python sobre la plataforma de cuaderno de notas (Notebook) de Jupiter (Jupyter Note Book).

Google Colaboratory. Colaboratory, también llamado "Colab", permite ejecutar y programar en Python en nuestro navegador o como se conoce actualmente en la "nube" y tiene las siguientes ventajas:

- No requiere configuración
- Da acceso gratuito a GPUs
- Permite compartir contenido fácilmente

Machine learning. Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos.

Regresión Logística. Es una técnica de aprendizaje automático que proviene del campo de la estadística.

Es uno de los algoritmos más utilizados actualmente en aprendizaje automático. Este algoritmo tiene como principal aplicación los problemas de clasificación binaria. Dada su simplicidad, en el que se pueden interpretar fácilmente los resultados obtenidos e identificar por qué se obtiene un resultado u otro. A pesar de su simplicidad funciona realmente bien en muchas aplicaciones y se utiliza como referencia para pruebas de rendimiento respecto a otros algoritmos.

Python. En la actualidad como lenguaje de programación interpretado goza de gran preferencia por los programadores, entusiastas y todo aquel que se ve atraído por temas como inteligencia artificial, robótica, procesamiento de imágenes, etc. Este lenguaje da mucha importancia a la legibilidad de su código. Lenguaje de programación multiparadigma, soporta parcialmente la orientación a objetos, programación imperativa y, en menor grado, la programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma, pudiéndose ejecutar en sistemas operativos como mac OS, windows y linux.

Pandas. Es una librería de las muchas que han sido desarrolladas por la gran legión de programadores de Python que la mantienen, está orientada al manejo especializado y análisis de estructuras de datos. Soporta archivos en formatos CSV, Excel y base de datos SQL.

Keras. Es una biblioteca de Redes Neuronales de Código Abierto escrita en Python. Es capaz de ejecutarse sobre TensorFlow, Microsoft Cognitive Toolkit o Theano, siempre de manera transparente hacia el usuario, sin que este tenga que preocuparse de nada.

Ha sido diseñada para hacer posible la experimentación en muy poco tiempo con redes de Aprendizaje Profundo.

Matplotlib. Es una librería de Python especializada en la creación de gráficos en dos dimensiones.

Permite crear y personalizar los tipos de gráficos más comunes, entre ellos:

Diagramas de barras

- Histograma
- Diagramas de sectores
- Diagramas de caja y bigotes
- Diagramas de violín
- Diagramas de dispersión o puntos
- Diagramas de líneas
- Diagramas de áreas
- Diagramas de contorno
- Mapas de color

y combinaciones de ellos.

Seaborn. Es una librería desarrollada para Python, dotando a este lenguaje de la posibilidad de generar fácilmente elegantes gráficos. Seaborn está basada en matplotlib y proporciona una interfaz de alto nivel con una curva rápida de aprendizaje. Dada su gran popularidad se encuentra instalada por defecto en la distribución Anaconda, y también puede ser importada en la plataforma Google COLAB.

Scikit-learn. Es una biblioteca desarrollada para Python orientada al aprendizaje automático de software libre. Esta librería incluye varios algoritmos de clasificación, regresión y análisis de grupos como ser: máquinas de vectores de soporte, bosques aleatorios, Gradient boosting, K-means y DBSCAN. Así también está diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy.

Numpy. Es una librería de Python especializada en el cálculo numérico y el análisis de datos, sobre todo grandes volúmenes de datos.

Esta librería, incorpora una clase de objetos llamados arrays que permite representar colecciones de datos de un mismo tipo en varias dimensiones, y funciones muy eficientes para su manipulación.

SciPy. Es una biblioteca libre y de código abierto para Python. Se compone de herramientas y algoritmos matemáticos. SciPy contiene módulos para optimización, álgebra lineal, integración, interpolación, funciones especiales, FFT, procesamiento de señales y de imagen, resolución de ODEs y otras tareas para la ciencia e ingeniería.

Procedimientos

Se dispone de un Dataset en formato csv (BD_SALUD_ESPECIALIDADES.csv), correspondiente a las citas para atención en un establecimiento de salud tipo Hospital ubicado en la Región de Moquegua, citas que corresponden a las atenciones generadas durante un periodo de 05 años.

A continuación, se resumen los procedimientos seguidos en el presente trabajo:

- Primeramente, estando en un nuevo proyecto de Jupyter notebook en Google Colab, se importará el data set arriba mencionado.
- A continuación se verifica los nombres de las columnas (features) el contenido (filas) del data set importado.
- Se realizan tareas de tratamiento de datos: conversión de dato del tipo object a int64 (sx_valor y edad_menor)
- Se generan totalizados mediante comandos de agrupamiento a fin de verificar la cantidad de registros vinculados con las variables de nuestro interés (sexo y especialidad).
- Se procede a presentar la descripción del dataset, lo cual nos da una idea de valores como total de registros, valor promedio, desviación estándar, percentil 25%, percentil 50%, percentil 75%, valor mínimo, valor máximo, para cada columna que forma parte del dataset.
- Se procede a agrupar el conjunto dataset por servicio y sexo mostrando el valor promedio de las columnas que conforman el dataframe.
- Se procede a agrupar el conjunto dataset por la columna numérica calculada menor_edad.
- Se genera un gráfico de barra mostrando el número de hombres y mujeres, agrupados en menor_edad y mayor de edad.

- Se genera un gráfico del tipo histograma mostrándonos la distribución de las citas por edades en el rango de 0 a 100 años. Este gráfico permite apreciar los bloques de edades en los que se concentran la mayor cantidad de datos del dataframe.

Preparación del modelo:

- Se crean los conjuntos de datos a X que corresponden a las entradas e Y que corresponden a las salidas esperadas; para lo cual se eliminan las columnas que a nuestro criterio no son significativas para el modelo de regresión lineal.
- Se crea el modelo de regresión lineal, se entrena con un conjunto de datos que corresponde al 80% del dataframe, y posteriormente se prueba o valida con el restante 20% del conjunto de dataframe.
- Luego de la validación, se realizan las predicciones que se consideren necesarias.
- Finalmente, con los resultados obtenidos se procede al análisis y discusión de los mismos.

Resultados y discusión

El modelo se basa en el análisis de las tuplas generadas por las atenciones realizadas en el servicio de consulta externa del Hospital Regional de Moquegua, se han analizado un total de 171,797 tuplas, que abarcan desde el año 2015 al 2019. Cada tupla contiene 11 columnas con datos relacionados a la atención, de entre las cuales se definió las variables de entrada del modelo propuesto. La variable Sexo es relevante dado que permitió filtrar a los pacientes atendidos en dos grupos diferenciados y determinar la prioridad de uno de ellos en particular: el femenino. La variable Edad es importante porque asociada a la variable anterior nos mostró con mayor precisión el enfoque hacia donde debíamos orientar el modelo propuesto. La Especialidad Médica se considera en el modelo por ser la variable que contiene el propósito de concurrencia de un paciente al servicio de consulta externa.

En la figura 1, se muestra un cuadro de barras que representan los totales de pacientes atendidos registrados, para nuestro caso podemos observar la cantidad de pacientes femeninos menores de edad.

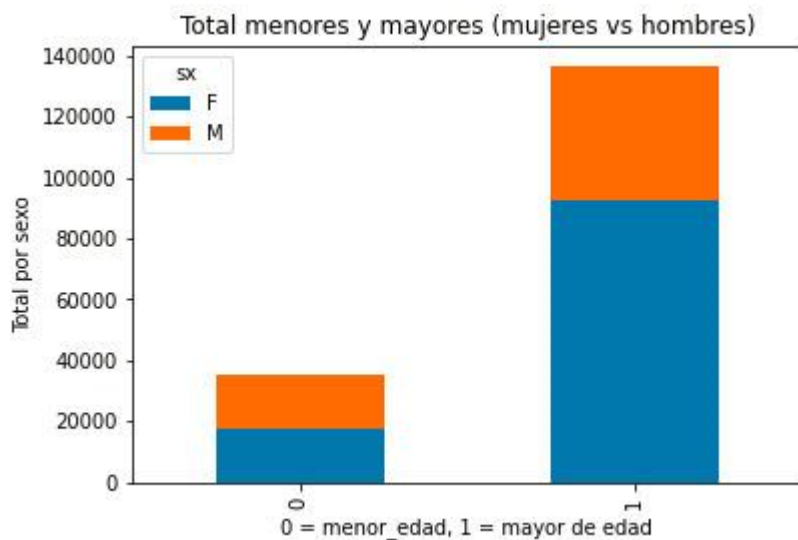


Figura 1. Cantidad de pacientes de ambos sexos y su agrupamiento por edad.

En la figura 2, visualizamos en formato de historial los tres Features o características de entrada con nombres “edad”, “menor_edad” y “sx_valor” podemos ver gráficamente entre qué valores se comprenden sus mínimos y máximos y en qué intervalos concentran la mayor densidad de registros.

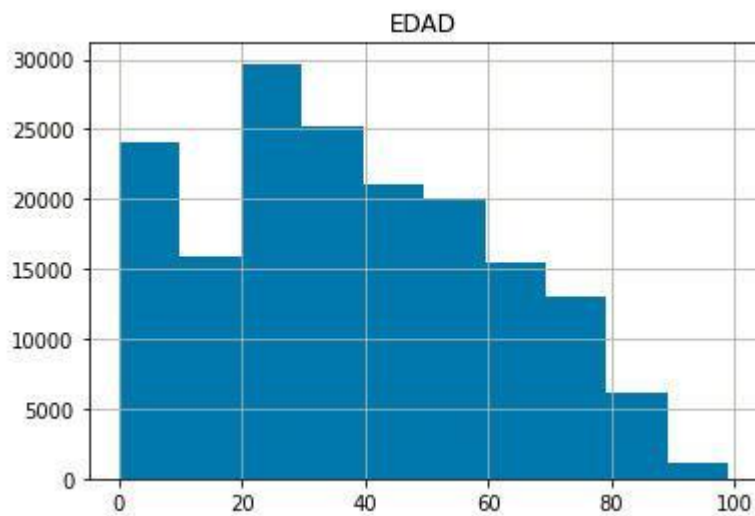
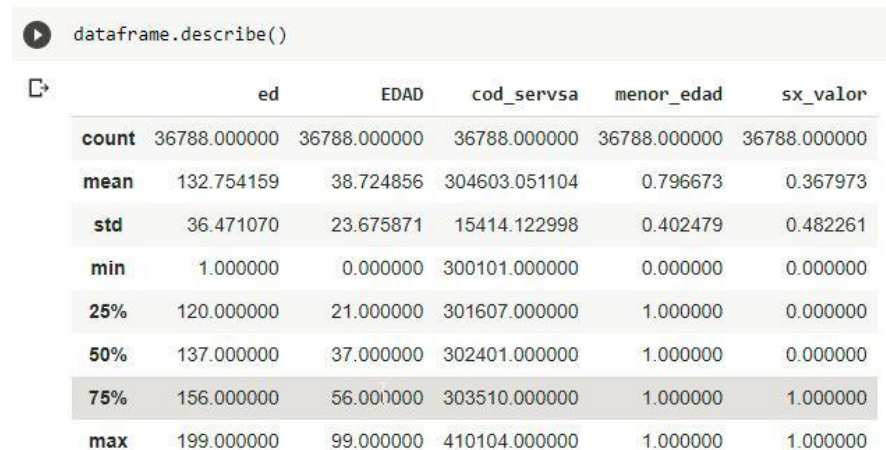


Figura 2. Agrupamiento del total de pacientes de grupos etarios por décadas.

En la siguiente tabla 1, complementando con información consolidada se tiene que el promedio de edad de los pacientes es de 39 años, que representa un 56% de la población en estudio y así mismo considerar que la edad mínima del paciente es de 0 años y la máxima de 99 años

Tabla 1. Resumen del reporte consolidado.



	ed	EDAD	cod_servsa	menor_edad	sx_valor
count	36788.000000	36788.000000	36788.000000	36788.000000	36788.000000
mean	132.754159	38.724856	304603.051104	0.796673	0.367973
std	36.471070	23.675871	15414.122998	0.402479	0.482261
min	1.000000	0.000000	300101.000000	0.000000	0.000000
25%	120.000000	21.000000	301607.000000	1.000000	0.000000
50%	137.000000	37.000000	302401.000000	1.000000	0.000000
75%	156.000000	56.000000	303510.000000	1.000000	1.000000
max	199.000000	99.000000	410104.000000	1.000000	1.000000

Para un mejor procesamiento de los datos, se ha añadido al *dataset* inicial dos columnas: “menor_edad”, la cual tomará como valor 0 si es menor a 18 años o 1 si es mayor de edad; así mismo, se agregó la columna “sx_valor” en la cual se transforma numéricamente los valores de sexo femenino y masculino en 0 y 1 respectivamente. Con la estructura de la data actualizada se procedió a hacer una evaluación por sexo y edad de la información, dando como resultado que un total de 17,796 son de sexo femenino y menores de edad, y 92,798 también de sexo femenino pero mayores de edad, como datos de importancia para el modelo.

En un primer resultado, luego de evaluar la interrelación de las entradas de a pares, para ver cómo se concentran linealmente las salidas de especialidades médicas por colores, es decir las variables: edad, “menor_edad” que determina si es mayor o menor de edad un paciente, “sx_valor” que agrupa a los mismos en femenino o masculino, se ha podido evidenciar su concentración en base a la clasificación por especialidad médica, representada en la variable “desc_servs”.

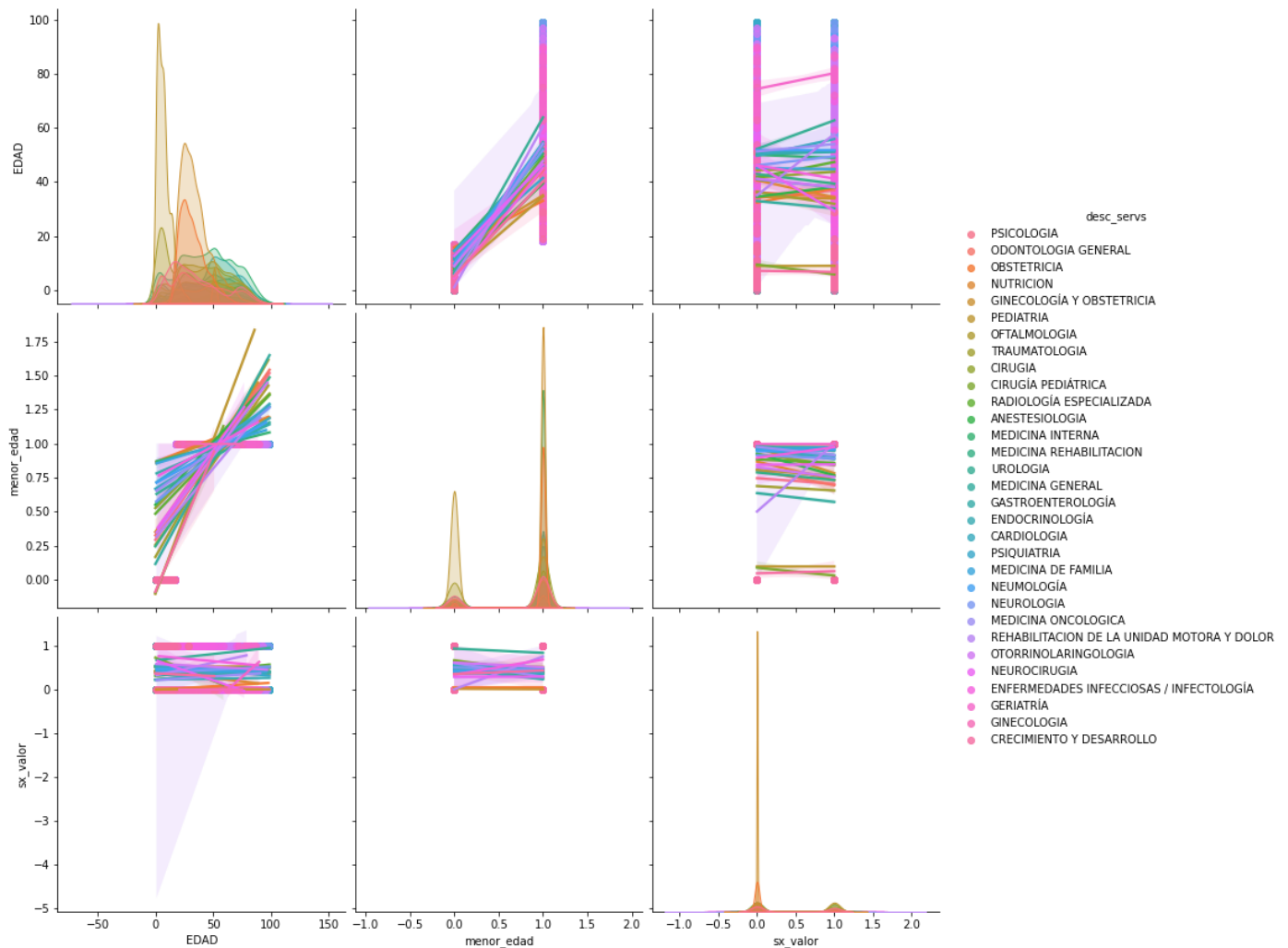


Figura 3. Clasificación por especialidad médica con la interrelación de variables de entrada por sexo y edad.

Luego de la aplicación del método, como se puede apreciar en la figura 4, los resultados del entrenamiento del modelo han obtenido el 23.67% de proximidad o confiabilidad en el modelo de la predicción inicial. Esto debido a las variables intervinientes edad, sx_valor (sexo de tipo numérico), menor_edad. Lo que indica que nuestro modelo aún es perfectible. Para una mayor precisión del modelo, se podrían considerar otras variables adicionales que tengan relación con la variable determinante o predictora.

```
[ ] predictions = model.predict(X)
    print(predictions)

['PEDIATRIA' 'PEDIATRIA' 'TRAUMATOLOGIA' ... 'GINECOLOGÍA Y OBSTETRICIA'
 'GINECOLOGÍA Y OBSTETRICIA' 'GINECOLOGÍA Y OBSTETRICIA']

▶ model.score(X,y)

📄 0.2367076220506687
```

Figura 4. Resultado de la predicción inicial.

La matriz de confusión es una herramienta muy útil para valorar cómo de bueno es un modelo de clasificación basado en aprendizaje automático. En particular, sirve para mostrar de forma explícita cuándo una clase es confundida con otra, lo cual nos permite trabajar de forma separada con distintos tipos de error.

Para la evaluación del modelo, como se muestra en la figura 5, vamos a echar mano de la matriz de confusión. Para ello, dividimos el *dataset* en dos partes. Dejamos un 80% de los datos como datos de entrenamiento (train), y reservamos el 20% restante como datos de prueba (test). A continuación, entrenamos el modelo de nuevo, pero ahora sólo con los datos de entrenamiento.

En consecuencia, como se observa en la tabla 2, el resultado de la regresión logística arroja un 23.58% de confiabilidad que indica que el porcentaje de precisión del modelo es muy bajo o poco confiable.

```
[ ] validation_size = 0.20
    seed = 7
    X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, y, test_size=validation_size, random_state=seed)

[ ] name='Logistic Regression'
    kfold = model_selection.KFold(n_splits=10, random_state=seed, shuffle=True)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

Logistic Regression: 0.235848 (0.008124)

[ ] predictions = model.predict(X_validation)
    print(accuracy_score(Y_validation, predictions))

0.2386518075564012
```

Figura 5. Resultado de la validación del modelo de regresión logística.

Las variables de entrada, seleccionadas para el modelo, hasta el momento, no muestran correlación respecto a la variable de salida: especialidad médica. En ese sentido, se sugiere la incorporación de otras variables de entrada que nos permitan mejorar el porcentaje de predicción del modelo propuesto.

Referencias

- [1] “Estadísticas” Boletín Estadístico 2019, [online document], 2019. Disponible: Web Hospital Regional de Moquegua, <http://www.hospitalmoquegua.gob.pe>.
- [2] Jiménez, M. V. G., Izquierdo, J. M. A., & Blanco, A. J. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema*, 12(Su2), 248-525.
- [3] Guzmán Aristizábal, S. M., & Hurtado Franco, J. C. (2021). Predicción de la tendencia del indicador S&P 500.
- [4]. Erazo Garzón, J. F. (2019). Desarrollo de un modelo de predicción de riesgo de quiebra empresarial para el sector comercial del Ecuador: un enfoque de regresión logística (Doctoral dissertation, Universidad Autónoma de Nuevo León).
- [5] Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2013). Análisis Comparativo de Técnicas de Predicción para Determinar la Deserción Estudiantil: Regresión Logística vs Árboles de Decisión. *Arquitectura*, 2014, 2015.
- [6] Huamán Bernilla, J. N. (2020). Comparación de máquina de soporte vectorial y regresión logística en la predicción de morosidad de cuotas sociales del colegio de ingenieros del Perú consejo departamental Lambayeque.
- [7] Maydana Huanca, A. R. (2021). Elección del mejor modelo entre regresión lineal múltiple y árboles de regresión para predecir el precio máximo de las acciones de Intel en función al precio de apertura y volumen de ventas de acciones por día-2019.