Análisis de datos de Enfermedades Crónicas utilizando python y machine learning

Ailyn Camacho R.[1]

[1]Universidad del atlántico, Barranquilla-Atlántico

Resumen

En este artículo se analiza la prevalencia de enfermedades crónicas y su correlación con variables demográficas utilizando datos recopilados en Bucaramanga-Santander durante casi dos años. Se emplearon herramientas de análisis de datos y modelos de machine learning para comprender la distribución de enfermedades y desarrollar enfoques predictivos.

Introducción

Las enfermedades crónicas son condiciones de salud de larga duración que progresan lentamente y que generalmente no tienen cura completa. Estas enfermedades pueden persistir durante meses o años, y suelen requerir un manejo continuo y cuidados a largo plazo. Algunos ejemplos comunes de enfermedades crónicas incluyen la diabetes, la enfermedad cardíaca, la hipertensión arterial, el cáncer, la enfermedad pulmonar obstructiva crónica (EPOC) y la enfermedad renal crónica. Las enfermedades crónicas representan una carga significativa para los sistemas de salud y tienen un impacto negativo en la calidad de vida de las personas afectadas. Se considera que factores como el envejecimiento de la población, los estilos de vida poco saludables y la exposición a factores de riesgo contribuyen al aumento de la prevalencia de enfermedades crónicas a nivel mundial[1,2].

En este trabajo se utilizó un conjunto de datos recopilado en Bucaramanga-Santander durante un periodo de casi dos años (de junio 2020 a marzo 2022)[3] para analizar la presencia de enfermedades crónicas, como la artritis, la diabetes, la hipertensión y enfermedades huérfanas. El análisis se llevó a cabo con el objetivo de comprender mejor la distribución de estas enfermedades en la población y explorar posibles correlaciones con variables demográficas. Se utilizaron herramientas de análisis de datos proporcionadas por Python, como Pandas, Numpy, Matplotlib, Seaborn y Scikit-learn, para llevar a cabo el análisis. Estas bibliotecas permitieron realizar operaciones de manipulación y limpieza de datos, visualización de gráficos y modelos de aprendizaje automático. Adicionalmente se examinaron diversas variables demográficas presentes en el dataset como la edad, el sexo y la ubicación geográfica para determinar si tenían alguna influencia en la prevalencia de las enfermedades estudiadas. Además, se calcularon estadísticas descriptivas para obtener información sobre la distribución de las enfermedades y se exploraron posibles correlaciones entre ellas.

Es importante destacar que este estudio se basa en datos recopilados en un período y ubicación específicos, por lo que los resultados y conclusiones obtenidos se limitan a ese contexto. Sin embargo, los hallazgos pueden proporcionar información valiosa para comprender la prevalencia de enfermedades crónicas en la población estudiada y pueden servir como punto de partida para futuras investigaciones.

Metodología

El presente proyecto se realizó siguiendo la siguiente metodología:

1 Exploración de los datos: Se seleccionaron las carácterísticas de interés sobre las cuáles se basa el análisis realizado y se realizaron procesos correspondientes a la limpieza de los datos. Se calcularon algunos estadísticos descriptivos de las variables numéricas, como la edad para entender su distribución. También se realizarón gráficos para visualizar los patrones y tendencias.

2 Análisis de asociación: Se exploraron las posibles correlaciones o asociaciones entre las diferentes condiciones médicas y variables demográficas como la edad, el sexo y la ubicación.

3 Modelo de machine learning: Se desarrolló un modelo de machine learning con el objetivo de predecir una variable objetivo basada en las características seleccionadas. Se dividió el conjunto de datos en conjuntos de entrenamiento y prueba utilizando una proporción de 80:20, respectivamente.

4 Validación del modelo: Se realizó el entrenamiento utilizando el conjunto de entrenamiento y se evaluó su desempeño utilizando el conjunto de prueba. Se calcularon métricas de evaluación adicionales, como la matriz de confusión, y el área bajo la curva ROC para obtener una comprensión más completa de la calidad del modelo.

Resultados

Al hacer una agrupación de los datos por sexo (Figura 1) se encontró que hay una proporción ligeramente mayor de mujeres en el estudio, representando el 57.8 % de los pacientes, mientras que los hombres representan el 42.4 %.

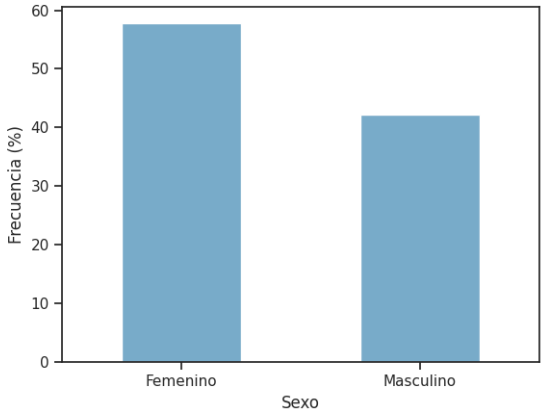

Figura 1. Distribución de géneros dentro del conjunto de datos

Una de las variables más importantes en este caso es la edad de los pacientes en el conjunto de datos. La Figura 2 muestra que las edades en el conjunto de datos estudiado siguen una distribución aproximadamente normal. Esto implica que la mayoría de las personas tienen edades cercanas a la media (63 años) y que las edades en los extremos son menos comunes. El rango intercuartil contiene el 50 % de los datos centrales, es decir, el 50 % de las observaciones están comprendidas entre el primer cuartil (Q1) y el tercer cuartil (Q3), que corresponden a personas con edades entre 54 y 74 años. Por otro lado, la Figura 3 muestra que la distribución de las edades es similar tanto en los pacientes de sexo masculino como en los pacientes de sexo femenino.

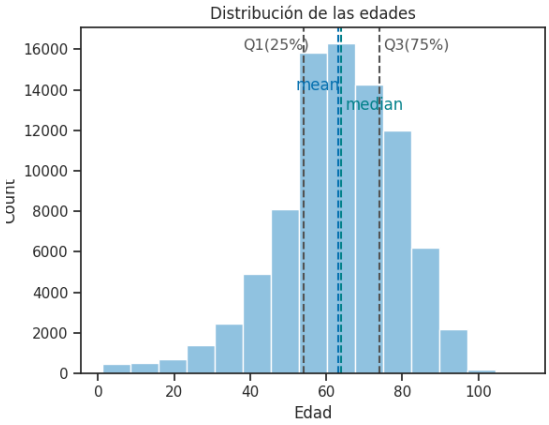

Figura 2. Distribución de las edades de los pacientes en el estudio. Se encuentran señaladas la media, moda y el rango intercuantílico

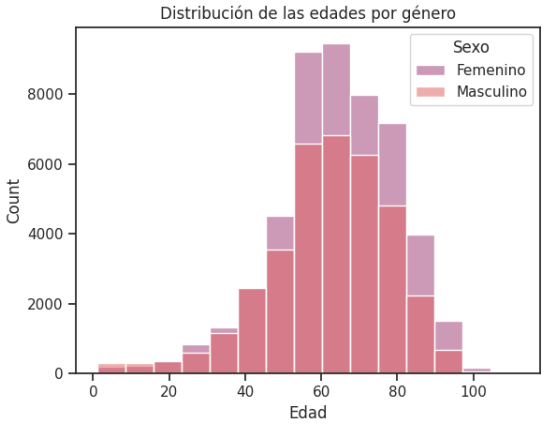

Figura 3. Distribución de las edades por sexo

Como se muestra en la Figura 4, la patología con más frecuencia en el estudio es Huerfanas- Hemofilias y otras Coag (99.3 %), seguida de la Insuficiencia cardiaca (99 %) y Atritis (98.4 %). Por otro lado, la enfermedad menos frecuente es la Insuficiencia renal crónica (IRC_Si), con una frecuencia del 3.2 %. La Figura 5 muestra además que la EAPB más frecuente en

el estudio es Nueva EPS, con un porcentaje del 24.9 %. Esto es consistente con el hecho de que Nueva EPS es la EPS con mayor número de afiliados en Colombia[4]. Esta información puede ser útil para identificar las EAPB con mayor carga de enfermedad crónicas y tomar medidas para mejorar la atención y los servicios de salud.

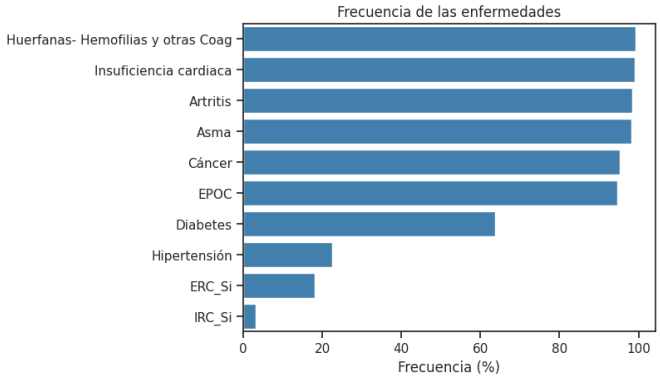

Figura 4. Frecuencia de las patologías registradas

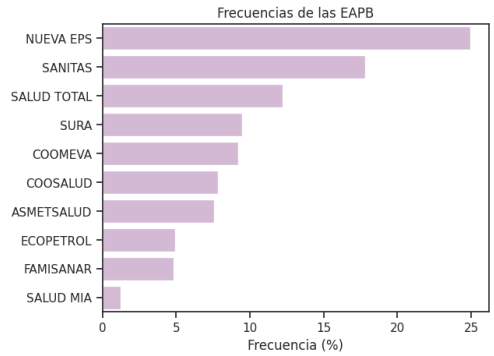

Figura 5. Frecuencia de las EAPB (Entidad Administradora de Planes de Beneficios de Salud)

El análisis de la distribución de las enfermedades por comuna (gráfico de barras apiladas) en la Figura 6 muestra que la comuna 13 tiene la mayor presencia de enfermedades crónicas, seguida de San Francisco y Centro. Esto puede indicar que ciertas áreas geográficas pueden estar más afectadas por las enfermedades crónicas, lo cual puede ser relevante para la planificación de servicios de salud y recursos. Además, se puede observar que las enfermedades con mayor frecuencia son las mismas en cada comuna. Las enfermedades parecen tener la misma proporción en todas las comunas, siendo las más polulares Artritis, EPOC, Asma, Insuficiencia cardiaca, Cáncer y Huérfanas-Hemofilias y otras Coag.

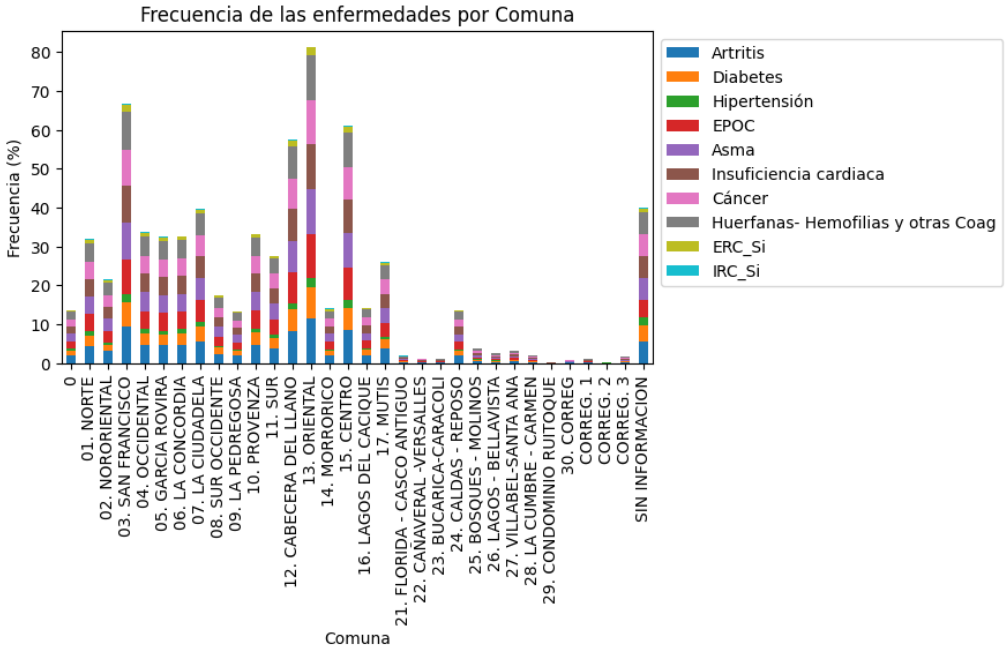

Figura 6. Distribución de las enfermedades por comuna

Al analizar la frecuencia de las patologías entre hombres y mujeres (Figura 7) se encontró que Huerfanas- Hemofilias y otras Coag (H-H y otras) y la Insuficiencia cardiaca (IC) son las enfermedades más comúnes en ambos sexos (41.9 %, 53.4 % y 41.6 %, 57.3 % respectivamente). A estas le siguen la Artritis (41.8 %) en hombres y Asma (56.6 %) en mujeres. Aunque la gráfica muestra una mayor frecuencia de enfermedades en las mujeres, se debe tener en cuenta que la mayoría de los pacientes

registrados son mujeres, por lo que no se puede concluir que el sexo es necesariamente un factor de riesgo para la presencia de alguna patología.

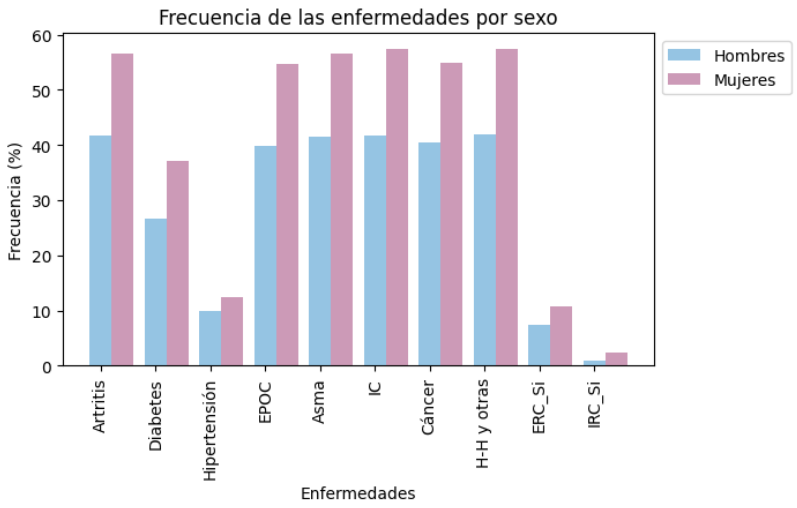

Figura 7. Distribución de las enfermedades por sexo

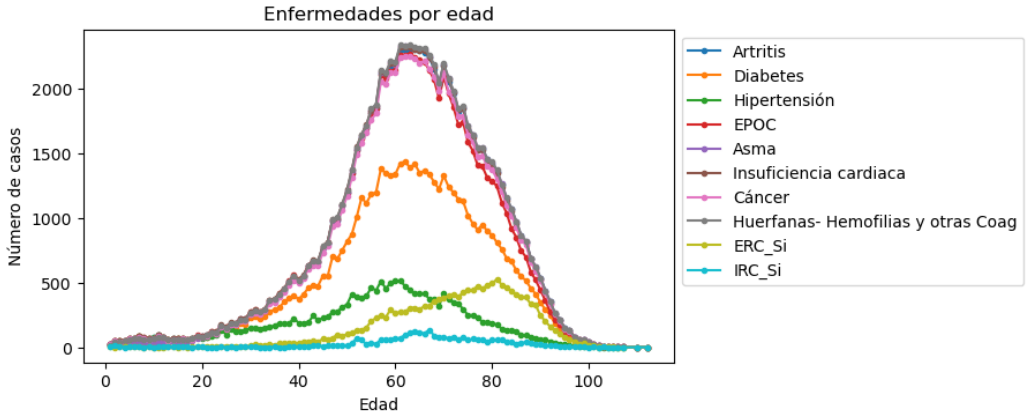

Figura 8. Distribución de las enfermedades por edad

La distribución de los casos según la edad (Figura 8) muestra que el mayor número de casos para cada patología ocurre en un rango de edades entre 50 y 70 años, que corresponde al grupo de edad de la mayoría de los pacientes en el estudio. Esto indica que las enfermedades crónicas presentes en el estudio afectan principalmente a personas en esta etapa de la vida. Sin embargo, para la enfermedad renal crónica (ERC), se observa un pico en el número de casos cerca de los 80 años. Estos hallazgos resaltan la importancia de la detección temprana y el manejo adecuado de las enfermedades crónicas en estos grupos de edad.

En esta última parte del análisis se construyó un modelo de regresión logística para predecir la presencia de una enfermedad teniendo en cuenta la edad, sexo y presencia de otras patologías. Este algoritmo es útil en este caso ya que se trata de un problema de clasificación: presencia o no de una enfermedad. Un primer modelo se construyó para predecir la presencia de la enfermedad Insuficiencia cardíaca, utilizando como características del modelo todas las patologías restantes (además de la edad y el sexo). Este modelo sin embargo, presentaba algunas inconsistencias. Hay que tener en cuenta que la mayoría de patologías presentes en el conjunto de datos tienen mayor número de ocurrencia que de no ocurrencia. En el caso de la Insuficiencia cardiáca, 84667 de los pacientes registrados tienen la enfermedad y 882 no la tienen. La diferencia entre el número de ocurrencias y no ocurrencias de una variable objetivo puede afectar la regresión logística debido a un desequilibrio en los datos. Esto se conoce como desequilibrio de clases y puede tener un impacto en el rendimiento del modelo.

Cuando existe un desequilibrio de clases, el modelo puede verse sesgado hacia la clase mayoritaria, lo que resulta en una menor capacidad para predecir la clase minoritaria. En el caso de la Insuficiencia cardíaca, donde hay un mayor número de ocurrencias que de no ocurrencias, el modelo predecía incorrectamente la ausencia de la enfermedad (clase 0).

El desequilibrio de clases puede conducir a varios problemas, como una baja sensibilidad (capacidad para detectar correctamente los casos positivos) y una alta especificidad (capacidad para detectar correctamente los casos negativos). Esto se debe a que el modelo puede inclinarse hacia la clase mayoritaria y tener dificultades para capturar los patrones y características de la clase minoritaria[5,6].

Se utilizó pues como variable dependiente del modelo la Diabetes, ya que esta presenta un mejor balance: 54554 pacientes registrados tienen la enfermedad y 30995 no la tienen. Se utilizaron como variables predictoras del modelo el resto de patologías, el sexo y la edad de los pacientes.

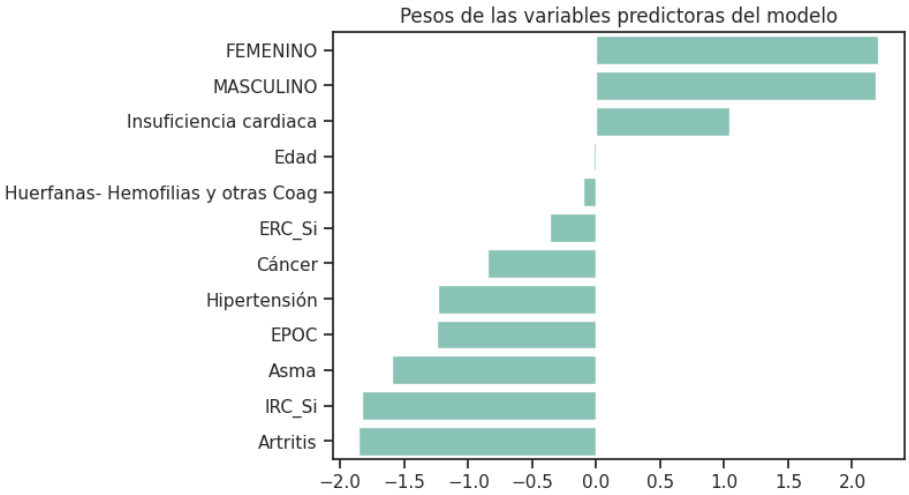

Figura 9. Coeficientes del modelo asociados a cada variable independiente

De la Figura 9 se observa que la mayoría de las variables predictoras del modelo aportan a la clase negativa (ausencia de la enfermedad). El sexo y la insuficiencia cardiaca son las únicas variables que aportan a la clase positiva (presencia de la enfermedad).

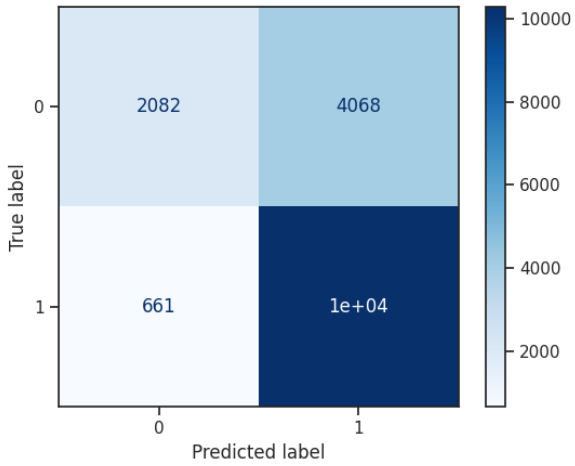

Figura 10. Matriz de confusión del modelo

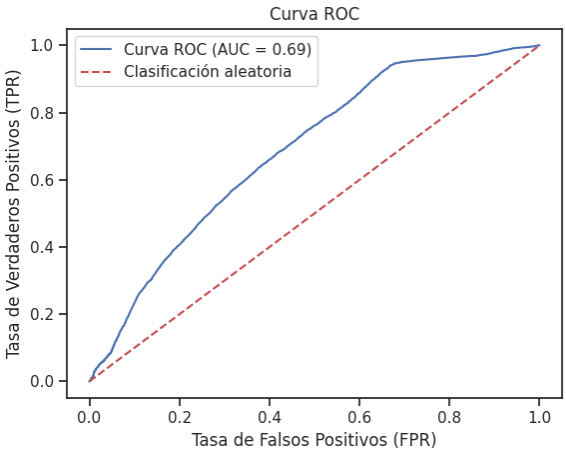

Figura 11. Curva ROC

La Figura 10 muestra la matriz de confusión obtenida. Podemos decir que el modelo tiene un buen desempeño al predecir los casos negativos, ya que acertó en la mayoría de los casos (2082 verdaderos negativos). Sin embargo, muestra ciertas dificultades para predecir los casos positivos, con un número considerable de falsos positivos (4068) y falsos negativos (661). Esto sugiere

que el modelo tiene una tendencia a sobreestimar la presencia de la enfermedad, clasificando erróneamente a algunas personas como enfermas cuando en realidad no lo están (falsos positivos). Además, también puede pasar por alto algunos casos de enfermedad, clasificando erróneamente a algunas personas como sanas cuando en realidad tienen la enfermedad (falsos negativos).

Precisión	0.7236119228521333
Precisión positiva (Sensibilidad)	0.7168511171434537
Sensibilidad (Recall)	0.9396897810218978
Puntuación F1	0.8132822679354048

Cuadro 1. Métricas de evaluación del modelo

Fold 1	Score:	0.6389246054938632
Fold 2	Score:	0.6512565751022794
Fold 3	Score:	0.7646405610753945
Fold 4	Score:	0.7618936294564582
Fold 5	Score:	0.7418902332105909

Cuadro 2. Resultados de la validación cruzada

De la Tabla 1 se observa que el modelo tiene una precisión del 72.36 %, lo que significa que alrededor del 72.36 % de las predicciones realizadas por el modelo son correctas. La sensibilidad del modelo es de 71.69 %, lo que significa que alrededor del 71.69 % de los casos positivos fueron identificados correctamente por el modelo. El modelo tiene un recall del 93.97 %, lo que indica que es capaz de identificar correctamente alrededor del 93.97 % de todos casos positivos. Por último, la puntuación F1 del modelo es de 0.8133, lo que indica un equilibrio razonable entre la precisión y la sensibilidad.El área bajo la curva ROC (0.680) en la Figura 11 indica la capacidad del modelo para distinguir entre las clases positiva y negativa. Un valor de 0.68 sugiere un rendimiento moderado, donde el modelo tiene una ventaja sobre una clasificación aleatoria (línea punteada roja) pero aún puede haber margen de mejora. En general, estos valores indican que el modelo tiene un rendimiento razonable en la clasificación de las clases positiva y negativa. Para verificar que no el modelo no está experimentando overfitting se utilizó el método de validación cruzada k-fold, que es comúnmente utilizado para evaluar el rendimiento del modelo en múltiples divisiones de datos (5 en este caso). Ta Tabla X muestra que el modelo tiene un rendimiento consistente en la mayoría de las divisiones. El rendimiento promedio es de aproximadamente 70 % lo que indica que, en promedio, el modelo es capaz de clasificar correctamente alrededor del 70 % de las muestras.

Conclusiones

En este proyecto, se aplicaron técnicas de análisis de datos para estudiar la presencia de enfermedades crónicas en una población determinada. Estas técnicas demostraron ser útiles para modelar y analizar el comportamiento de los datos, proporcionando información valiosa para evaluar la presencia de enfermedades en función de variables como la edad, el sexo y la presencia de otras patologías. Se implementó también un modelo de machine learning para precedir la presencia de una enfermedad en función de variables como la edad, el sexo y la presencia de otras patologías y se obtuvo que la presenciia de diabetes podría estar relacionada con la Insuficiencia cardiaca[7]. El modelo mostró una precisión promedio del 72 % en la clasificación de las muestras. Sin embargo, es importante tener en cuenta que el modelo puede presentar algunas inconsistencias y limitaciones, como la desigualdad en el número de ocurrencias y no ocurrencias de enfermedades en los datos. Es importante considerar las limitaciones de este estudio y realizar análisis más exhaustivos para obtener conclusiones más sólidas y precisas. El análisis de datos en el campo de la medicina es crucial para comprender, prevenir y tratar las enfermedades crónicas. Proporciona información precisa y basada en evidencia que respalda la toma de decisiones clínicas, personaliza el tratamiento, impulsa la prevensión, etc.

Referencias

1. WHO, W. H. O. Enfermedades no transmisibles. *World Heal. Organ. WHO* (2022). [Online; accessed 2023-06-02].

2. CDC. Coronavirus disease 2019. https://www.cdc.gov/chronicdisease/index.htm (2023). [Online; accessed 2023-06-02].

3. Enfermedades crónicas junio 2020 a marzo 2022. https://www.datos.gov.co/Salud-y-Protecci-n-Social/39-Enfermedades-cr-nicas-junio-2020-a-marzo-2022/4iz7-suhz (2022). [Online; accessed 2023-06-02].

4. Conozca las EPS con más usuarios en el país. *El Tiempo* (2022). [Online; accessed 2023-06-02].

5. Haibo He & Garcia, E. Learning from imbalanced data. *IEEE Transactions on Knowl. Data Eng.* **21**, 1263–1284 (2009). [Online; accessed 2023-06-02].

6. Mazumder, S. What is Imbalanced Data. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/ (2021). [Online; accessed 2023-06-02].

7. Cardiovascular disease and diabetes. https://www.heart.org/en/health-topics/diabetes/diabetes-complications-and-risks/cardiovascular-disease–diabetes (2018). [Online; accessed 2023-06-02].