Predicción del riesgo de hipertensión

Emmanuel Medina Espinosa
Unidad Profesional Interdisciplinaria de Ingeniería, Campus Zacatecas
Calle Circuito Cerro del Gato No. 202, Col. Ciudad Administrativa, Zacatecas, Zac. C.P. 98160
emedinae1700@alumno.ipn.mx

***Abstract*— En este trabajo se aplicó la técnica de regresión logística con descenso de gradiente estocástico dentro del marco teórico del modelo CRISP-DM para realizar predicciones del riesgo de hipertensión en personas en base a características de mediciones biomédicas, la actividad física, medidas antropomórficas y otros padecimientos médicos y mentales como la diabetes y la depresión, respectivamente.**

Se utilizaron sets de datos con información recolectada de encuestas médicas de diferentes pacientes, se desarrolló la metodología CRISP-DM para el preprocesamiento de datos, el análisis exploratorio de datos, el desarrollo del modelo predictivo y la evaluación del modelo con diversas métricas.

Palabras clave.- Minería de datos, Hipertensión, CRISP-DM, Regresión logística

I. Introducción

La hipertensión arterial, una condición médica caracterizada por niveles elevados y sostenidos de presión arterial, es una preocupación de salud pública global que afecta a millones de personas en todo el mundo. Con sus graves implicaciones para la salud cardiovascular y la calidad de vida, la identificación temprana de factores de riesgo y la predicción precisa del riesgo de desarrollar hipertensión se han convertido en áreas cruciales de investigación. En este contexto, la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) emerge como una herramienta robusta y estructurada para el desarrollo de modelos predictivos basados en datos.

La presente investigación se centra en la aplicación de la metodología CRISP-DM para construir un modelo de predicción del riesgo de hipertensión, utilizando datos clínicos. Este enfoque metodológico proporciona un marco sistemático para la gestión del ciclo de vida completo del proyecto, desde la comprensión del negocio y la comprensión de los datos hasta la implementación y evaluación del modelo final.

A través de la aplicación de las fases clave de la metodología CRISP-DM, este estudio tiene como objetivo no solo desarrollar un modelo de predicción preciso, sino también identificar los factores de riesgo más influyentes asociados con la hipertensión.

Este artículo proporcionará una descripción detallada de cada fase de la metodología CRISP-DM aplicada, destacando los desafíos específicos y las decisiones metodológicas tomadas durante el desarrollo del modelo. Además, se presentarán los resultados obtenidos y se discutirán las implicaciones clínicas y de salud pública derivadas de este enfoque predictivo. En última instancia, este trabajo busca contribuir al avance de la investigación en el campo de la salud preventiva, ofreciendo un enfoque estructurado y reproducible para la predicción del riesgo de hipertensión.

II. Marco teórico

1) CRISP-DM: La Ciencia de Datos ha emergido como un campo interdisciplinario que utiliza métodos y técnicas de estadística, aprendizaje automático y minería de datos para extraer conocimientos y patrones a partir de grandes conjuntos de datos. En este contexto, la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) se ha consolidado como un marco robusto y estandarizado para el desarrollo de proyectos de minería de datos y modelado predictivo.

El modelo CRISP-DM (Cross Industry Standard Process for Data Mining) es un proceso estándar utilizado en minería de datos para guiar el desarrollo de proyectos de minería de datos. El modelo se divide en seis fases principales, cada una de las cuales tiene subprocesos específicos. A continuación se presenta cada una de las fases:

1) Comprensión del negocio
2) Comprensión de los datos
3) Preparación de los datos
4) Modelado
5) Evaluación
6) Despliegue

La aplicación de CRISP-DM en el desarrollo de un modelo de predicción de riesgo de hipertensión busca aprovechar la estructura y la coherencia proporcionadas por esta metodología.

2) Gradiente Descendente Estocástico: El Gradiente Descendente Estocástico (SGD) es un algoritmo de optimización ampliamente utilizado en el entrenamiento de modelos de aprendizaje automático, especialmente en problemas de regresión y clasificación. Su eficacia radica en su capacidad para converger hacia el mínimo global de la función de costo al ajustar iterativamente los parámetros del modelo.

Consideremos un modelo de regresión lineal con m variables de entrada $(x_1, x_2, ..., x_m)$ y un objetivo de predicción y. El objetivo del SGD es minimizar la función de costo J(θ), donde θ representa el vector de parámetros del modelo.

La actualización de los parámetros θ en cada iteración t del SGD se realiza mediante la fórmula:

$$\theta_{t+1} = \theta_t - \alpha J(\theta_t; x_i, y_i)$$

Donde:

- αes la tasa de aprendizaje, un hiperparámetro que controla la magnitud de la actualización de los parámetros.
- $J(\theta_t; x_i, y_i)$ es el gradiente de la función de costo con respecto a los parámetros, evaluado en el punto (x_i, y_i) del conjunto de datos.

3) Extensión a Múltiples Variables de Entrada: Cuando el modelo posee múltiples variables de entrada, la formulación se extiende considerando el vector de características X y el vector de parámetros :

$$J() = \frac{1}{2m}\sum_{i=1}^{m}(h_\vartheta(x_i) - y_i)^2$$

Donde:

- $h_\vartheta(x_i)$ es la predicción del modelo para la observación i.
- m es el número total de observaciones.

La actualización de los parámetros para múltiples variables de entrada se realiza considerando la derivada parcial de J con respecto a cada parámetro:

$$\theta_j^{t+1} = \theta_j^t - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_(x_i) - y_i)x_{ij}$$

Donde:

- j representa la j-ésima variable de entrada.
- x_{ij}es el valor de la variable de entrada j en la observación i.

4) Funcionamiento del SGD:

1) nicialización: Se eligen valores iniciales para los parámetros del modelo (θ).
2) Iteración: Para cada observación en el conjunto de datos, se calcula la diferencia entre la predicción del modelo y la verdad conocida. Estas diferencias ponderadas por las variables de entrada se utilizan para actualizar los parámetros del modelo.
3) Convergencia: El proceso de iteración se repite hasta que se alcanza un criterio de convergencia, como un número fijo de iteraciones o la convergencia del error.

5) Regresión Logística Multivariable: La regresión logística es una técnica estadística ampliamente utilizada en el ámbito de la minería de datos y el aprendizaje automático, especialmente en problemas de clasificación binaria. La regresión logística es una técnica de regresión utilizada para predecir la probabilidad de que una instancia pertenezca a una clase particular. A diferencia de la regresión lineal, la regresión logística emplea la función logística, también conocida como función sigmoide, para transformar la salida del modelo a un rango entre 0 y 1. La función sigmoide tiene la propiedad de asignar valores cercanos a 0 o 1 a extremos opuestos, permitiendo así la interpretación probabilística. La regresión logística es una extensión de la regresión lineal que emplea la función logística para modelar la probabilidad de pertenencia a una clase. Dada una combinación lineal de las variables de entrada ponderadas por coeficientes, la función logística transforma este resultado en una probabilidad, que se interpreta como la probabilidad de pertenencia a la clase positiva. La función logística está definida como:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

donde Y es la variable dependiente binaria, $X_1, X_2, ...X_n$ son las variables independientes, y $\beta_1, \beta_2, ..., \beta_n$ son los coeficientes a estimar.

6) Funcionamiento de la Regresión Logística Multivariable:

1) Inicialización de Parámetros: Se seleccionan valores iniciales para los coeficientes del modelo (β).
2) Cálculo de Probabilidades: Se calculan las probabilidades predichas utilizando la función logística.
3) Función de Costo: Se define una función de costo que mide la discrepancia entre las predicciones y las observaciones reales, como la Entropía Cruzada.
4) Optimización de Parámetros: Se utiliza un algoritmo de optimización para ajustar los coeficientes del modelo y minimizar la función de costo.
5) Predicción: Una vez que los coeficientes están optimizados, se pueden realizar predicciones para nuevas observaciones calculando las probabilidades y aplicando un umbral de decisión.

7) Hipertensión: La hipertensión arterial, comúnmente conocida como presión arterial alta, es una condición médica crónica caracterizada por la elevación sostenida de la presión arterial en las arterias. La presión arterial se expresa en dos valores: la presión sistólica (la presión cuando el corazón late) y la presión diastólica (la presión cuando el corazón está en reposo entre latidos). Se considera que una presión arterial normal es alrededor de 120/80 mmHg, mientras que la hipertensión se define como una presión arterial igual o superior a 130/80 mmHg.

8) Factores de Riesgo Asociados con la Hipertensión:

a) Edad: La incidencia de hipertensión aumenta con la edad. A medida que las arterias envejecen, tienden a volverse menos flexibles, lo que contribuye al aumento de la presión arterial.

b) Genética: Existe una clara predisposición genética a la hipertensión. Si hay antecedentes familiares de hipertensión, el riesgo de desarrollarla puede aumentar.

c) Estilo de Vida:

- Dieta no saludable: El consumo elevado de sodio y una dieta baja en potasio pueden contribuir al desarrollo de la hipertensión.
- Inactividad física: La falta de actividad física regular puede contribuir al aumento de peso y a la hipertensión.

d) Obesidad: El exceso de peso corporal, especialmente en la zona abdominal, está asociado con un mayor riesgo de hipertensión.

e) Consumo de Alcohol y Tabaco:

- Consumo excesivo de alcohol: El abuso de alcohol puede elevar la presión arterial.
- Tabaquismo: Los productos químicos del tabaco pueden dañar las arterias, contribuyendo al aumento de la presión arterial.

f) Estrés: El estrés crónico puede tener efectos negativos en el sistema cardiovascular, contribuyendo a la hipertensión.

g) Enfermedades Subyacentes:

- Enfermedades renales: Problemas en los riñones pueden afectar la regulación de los fluidos y las sales, influenciando la presión arterial.
- Diabetes: La diabetes tipo 2 está estrechamente vinculada a la hipertensión.

III. Materiales y Métodos

A. Materiales

- Set de datos Cuestionario de antropometría y tensión arterial
- Set de datos Determinaciones para enfermedades crónicas y deficiencias
- Set de datos Actividad física - Adolescente y adultos
- Set de datos Cuestionario de salud de adultos (20 años o más)
- Sets de datos disponibles en: Data Sets
- Python 3.11.5
- Jupyter notebook

B. Metodología enfocada al modelo CRISP-DM

1) Comprensión del negocio: Esta fase tiene como objetivo entender el problema de negocio y los objetivos que se desean alcanzar con el proyecto de minería de datos. Se identifican los requerimientos, se define el problema, se establecen los objetivos y se elabora un plan de proyecto.

2) Comprensión de los datos: En esta fase se lleva a cabo la recopilación y exploración de los datos necesarios para el proyecto de minería de datos. Se identifican los conjuntos de datos relevantes, se obtienen y se exploran, se verifica su calidad y se documenta el proceso de exploración.

3) Preparación de los datos: En esta fase se lleva a cabo la preparación de los datos para su uso en el modelado. Se seleccionan los datos relevantes, se limpian, se transforman y se integran para crear un conjunto de datos apto para el modelado.

4) Modelado: En esta fase se construyen los modelos de minería de datos que se utilizarán para resolver el problema de negocio. Se seleccionan las técnicas de modelado adecuadas, se construyen los modelos y se validan.

5) Evaluación: En esta fase se evalúan los modelos construidos en la fase anterior. Se determina si los modelos son lo suficientemente buenos para resolver el problema de negocio y si satisfacen los criterios de éxito definidos en la fase de comprensión del negocio.

6) Despliegue: En esta fase se implementan los modelos seleccionados en la fase anterior y se entregan al cliente para su uso en el negocio. Se elabora un plan de implementación, se lleva a cabo la implementación y se monitorea el desempeño del modelo en producción.

IV. Experimentación y Resultados

A. Comprensión del negocio

La problemática a resolver se centra en proponer un modelo de predicción del riesgo de hipertensión en función a variables de datos clínicos obtenidos de encuestas y mediciones médicas a pacientes, se hace un enfoque en datos relacionados a actividades físicas, mediciones antropomórficas, métricas biomédicas y otros padecimientos, los sets de datos recaban la información de los pacientes en los tópicos previamente descritos, contienen de forma desglosada la información que se debe de utilizar para desarrollar el modelo de predicción, se propone realizar la limpieza de los datos, la selección de variables para el desarrollo del modelo, la limpieza de impuresas de los sets de datos, el preprocesamiento de daots, generar un análisis exploratorio de las variables de estudio, desarrollar agrupamientos de los datos con métodos de machine learning para el análisis de patrones, la implementación de un algoritmo de predicción utilizando la regresión logística y finalmente una evaluación del modelo utilizando diversas métricas de evaluación.

B. Comprensión de los datos

En función al dataset de determinaciones bioquímicas que originalmente contiene una recopilación de un procedimiento médico donde se sigue la estructura de un documento para cerciorarse de que el procedimiento se cumple y recabar la información, se revisó el dataset y se seleccionaron las variables características de mediciones biomédicas de los pacientes donde las variables reprecentan diferentes mediciones de diferentes componentes químicos del organismo que se relacionan con el padecimiento de la hipertensión como se muestra en la Fig. 1. que muestra el nombre de todas las variables seleccionadas y su descripción.

Fig. 1. Determinaciones bioquímicas dataset

En el dataset de medidas antropomórficas de manera similar al dataset anterior es un procedimiento que recaba información de las medidas antropomórficas de los pacientes, para la selección de las variables se tomaron las variables

escalares enfocadas a medidas del cuerpo y de tensión arterial, así como una ponderación final, como se muestra en la Fig. 2. se observa el nombre de la variable y su descripción.

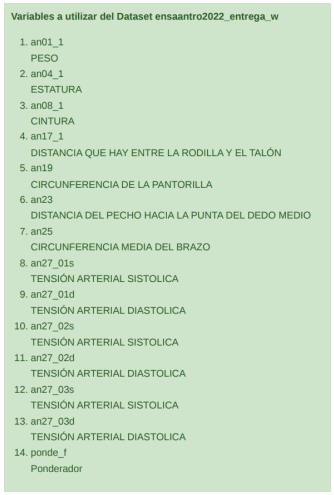


Fig. 2. Medidas antropomórficas dataset

En el dataset de actividad física se muestra un cuestionario que registra las respuestas de los pacientes sobre el tipo de actividad física que realizan, la frecuencua con que realizan esa actividad física, además de otros aspectos como el tiempo que toman para dormir, el tiempo que realizan de traslado a sus trabajos o escuelas y el tiempo que pasan sentados o frente a un dispositivo electrónico, como se muestra en la Fig. 3. se tomaron las variables más características referente a las categorías de tipos de actividades físicas, vigorosa, moderada o baja, se tomaron las variables que toman el tiempo en horas y se toman en cuenta solamente actividades entre semana, así como el tiempo de reposo o que pasa frente a un dispositivo electrónico.

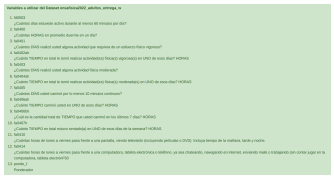


Fig. 3. Actividad física dataset

Uno de los factores que contribuyen o derivan de la hipertensión son otros padecimientos como la diabetes, fallas renales, depresión, entre otras, como se menciona en el marco teórico, por esta razón se toman de este dataset los padecimientos más comunes y característicos relacionados con la hipertensión, para esto como se muestra en la Fig. 4. se observan los padecimientos seleccionados además de que en este dataset se encuentra la variable objetivo (a0401) donde se señala si el paciente ha sido diagnosticado con hipertensión por algún doctor, la mayoría de las variables de este dataset son variablles categoricas.

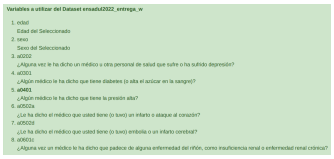


Fig. 4. Otros padecimientos datset

C. Preparación de los datos

Para la preparación de los datos se optó por sólo tomar las columnas distintivas de cada dataset previamente descritas para cada uno y mezclarlos en función a las coincidencias que se encuentren en el folio del paciente, esto con la intención de reducir la cantidad de variables y de complejidad del dataset como se muestra en la Fig. 5. Además de mezclar los datasets se realizaron varias operaciones a los mismos necesarias para el correcto funcionamiento del preprocesamiento de los datos, se cambiaron las comas por puntos y se cambiaron los espacios por NaN para poder conocer la cantidad correcta de datos faltantes, de forma similar se establecieron todas las columnas con variables numéricas.

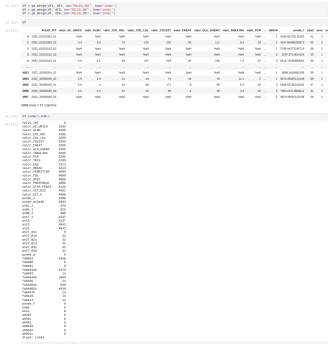


Fig. 5. Coincidencias en el dataset

Como se observa en la siguiente figura se puede observar la relación que tiene cada una de las variables de datos faltantes en relación con el total de datos, se tomó el criterio de que las columnas que cuenten con un porcentaje mayor al 60% sean eliminadas ya que no proporcionan la cantidad suficiente para tomarlas en cuenta y evitar sesgos utilizando algún método para llenar los campos vacíos de las columnas.

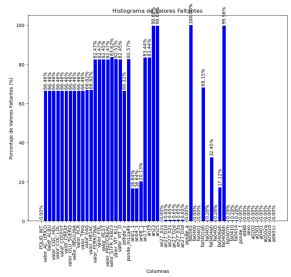


Fig. 6. Relación de datos faltantes

Resultante de la eliminación de las columnas con un gran déficit de información, las variables con una tasa de datos

faltantes menor al 50%, se optó por utilizar un método para completar los campos faltantes, para las variables con menor tasa de déficit de información se utilizó la media para llenar las celdas con falta de información como se muestra a continuación.

Fig. 7. Completado de variables con baja tasa de défict de datos

Para las variables con un mayor déficit de datos se optó por utilizar un método más robusto para el completado de los datos, se realizó una comparación entre dos métodos, Random Forest y KNN, se utilizó una métrica de evaluación para revisar la calidad de la predicción del modelo, la métrica del MSE señala que el método más óptimo para el completado de los datos es el método KNN con 5 vecinos más cercanos como se muestra en la siguiente imagen, el método de KNN cuenta con un MSE de cero menor al método de Random Forest, métrica que indica la eficiencia de las predicciones del modelo, por lo que se aplicó este modelo para el completado de las variables restantes del dataset.

Posterior al preprocesamiento de los datos se realizó un análisis a la correlación de los datos del dataset para ver la relación que tienen las diferentes variables, así como de ver el comportamiento de las variables en la variable objetivo para seleccionar las variables que tengan la mayor relación para el desarrollo del modelo de predicción.

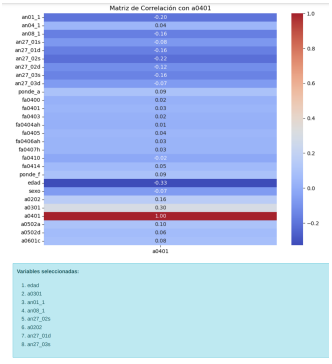


Fig. 8. Correlación de variables

En la figura anterior se muestra la relación que tienen las variables sobre la función objetivo, así como las variables que se seleccionaron para el desarrollo del modelo de predicción.

Siguiendo con el proceso del modelo CRISP-DM toma parte el análisis exploratorio de los datos, para esta etapa teniendo ya las variables con las cuales se desarrollará el modelo se renombraron las columnas para un mayor entendimiento, de manera similar para el EDA se busca obsrvar el comportamiento y los patrones que se puedan encontrar en los datos así como una limpieza a los datos que no correspondan a una distribución normal.

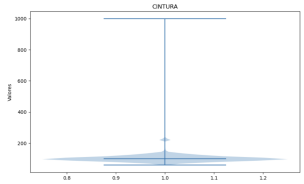


Fig. 9. Diagrama de violín de la variable peso

Primero se realizó una gráfica de violín para cada una de las variables para observar la distribución de los datos y cómo estan agrupados, así como valores que no respetan la distribución normal, para las variables que se observaron valores atípicos se realizó una limpieza de esos valores atípicos.

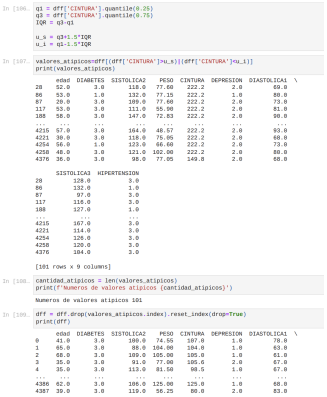


Fig. 10. Eliminación de datos atípicos

Tras el análisis exploratorio de los datos se puede observar el resultado de la distribución y diversas características de las variables en la siguiente gráfica, donde se muestran características estadísticas principales.

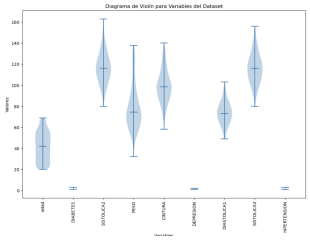


Fig. 11. Enter Caption

D. Modelado

Iniciando con la etapa de modelado del modelo de CRIPS-DM en función a la etapa de comprensión del negocio uno de los objetivos es la implementación de un modelo KNN

para el agrupamiento de las diferentes variables para observar patrones de los datos, para conocer cual es el valor óptimo de clusters a definir se utilizó el método del codo, como se muestra en la siguiente figura, la cual señala que por sus características que donde se observa la mayor disminución de pendiente en el diagrama es el número que se toma como óptimo.

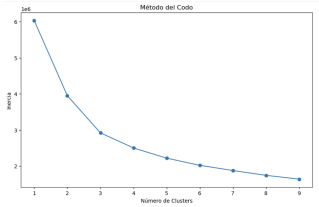


Fig. 12. Método del codo

Como se puede observar se muestran tres grupos característicos diferentes.

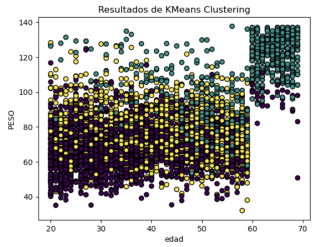


Fig. 13. Cluster peso y edad

En la figura anterior se puede observar la distinción de tres grupos diferentes en relación al peso y a la edad, donde un grupo de azul más oscuro se muestra en personas con menor peso y menos edad, el segundo grupo de color amarillo para valores medios de edad y de peso, finalmente se observa un último grupo para las personas de mayor edad y de mayor peso.

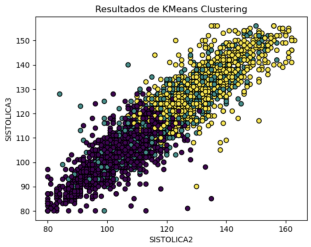


Fig. 14. Sistolicas

En la imagen anterior se muestra la relación de la sistolicas que son una medida de hipertensión donde se observan dos grupos característicos, entre menor graduación de ambas medidas se encuentra el grupo azul y entre mayor graduación se encuentra el grupo amarillo, relacionado con el diagrama anterior se puede relacionar que las personas que precentan una menor graduación de las sistolicac son personas jóvenes de menor peso mientras que el grupo amarillo representa personas de mediana edad con peso medio.

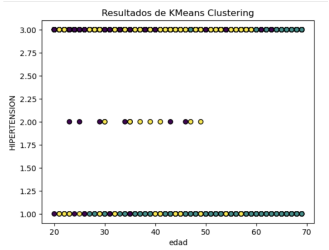


Fig. 15. Hipertension y edad

Otras de las relaciones que se pueden observar es que las personas que presentan en tendencia la mayoría de los casos de hipertensión son personas con mayor medida de sistolicas que a su vez son las personas con un peso y edad medias, mientras que las personas de mayor edad que presentan hipertensión son personas mayores y que casi no se encuentran personas de bajo peso y jóvenes con problemas de hipertensión.

Siguiendo con el proceso del modelo de CRISP-DM se continua con el desarrollo de modelo de predicción para el riesgo de hipertensión que para este caso se utilizó el método de regresión logística para varias variables de entrada, que la regresión logística actúa como un clasificador binario haciendo uso de la función logit, también se utiliza el método de descenso de gradiente estocástico para la optimización del método de predicción, para el entrenamiento se utilizaron todas las variables seleccionadas en el análisis exploratorio de datos con el objetivo de hacer predicciones de la variable objetivo.

Fig. 16. Regresión logística con algoritmo de descenso de gradiente estocástico

El código anterior muestra la implementación del método de predicción, las librerías necesarias y la división de variables independientes y la variable objetivo, además de la separación entre casos de entrenamiento y casos de prueba, el ajuste de los datos, el entrenamiento de los datos y el apartado de las predicciones así como el uso de casos de probabilidad para su evaluación de funcionamiento.

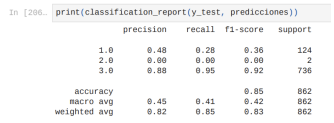
E. Evaluación

Para la evaluación del modelo se utilizó la métrica de precisión que resultó en un 85% de presición con una tasa de error del 15%, también en la siguiente figura se muestra la matriz de confusión que muestra los errores y aciertos de clasificación donde se observa que la mayoría de los errores que tiene el modelo es en predecir cuando alguien no tiene hipertensión en función a las variables seleccionadas, debido

a la distribución de la variable de hipertensión que tiene más variables que representan cuando si se tiene hipertensión que cuando no.

Fig. 17. Métricas de evaluación

La clase 2.0 muestra un rendimiento deficiente, mientras que las clases 1.0 y 3.0 muestran resultados mixtos. La precisión global del modelo es razonablemente alta (85%), pero se deben considerar mejoras específicas para las clases individuales.



	precision	recall	f1-score	support
1.0	0.48	0.28	0.36	124
2.0	0.00	0.00	0.00	2
3.0	0.88	0.95	0.92	736
accuracy			0.85	862
macro avg	0.45	0.41	0.42	862
weighted avg	0.82	0.85	0.83	862

Fig. 18. Reporte de clasificación

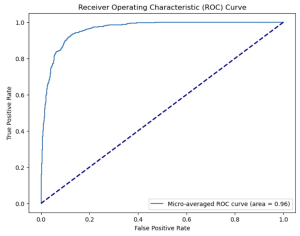


Fig. 19. Curva ROC

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada en estadísticas y aprendizaje automático para evaluar el rendimiento de un modelo de clasificación binaria. Representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) a medida que se varía el umbral de decisión del modelo.

La curva ROC es especialmente útil para comprender cómo se comporta un clasificador en diferentes puntos de corte de probabilidad, ayudando a encontrar un equilibrio óptimo entre la sensibilidad y la especificidad. Cuanto más se acerca la curva ROC al área bajo la curva (AUC) de 1, mejor es el rendimiento del modelo.

F. Despliegue

V. CONCLUSIONES

En este estudio, se aplicó la metodología CRISP-DM para desarrollar un modelo de predicción del riesgo de hipertensión, utilizando datos clínicos recopilados de encuestas médicas. Se realizaron análisis exploratorios de diferentes conjuntos de datos relacionados con mediciones biomédicas, actividad física, medidas antropomórficas y otros padecimientos médicos y mentales como la diabetes y la depresión.

La preparación de datos involucró la limpieza y selección de variables relevantes, así como la integración de conjuntos de datos para construir un conjunto de datos coherente. Se utilizó el método KNN para completar los datos faltantes, tras lo cual se realizó un análisis exploratorio de la correlación entre variables y la identificación de posibles patrones.

En la fase de modelado, se implementó un algoritmo de regresión logística con descenso de gradiente estocástico, utilizando múltiples variables de entrada seleccionadas durante el análisis exploratorio. La evaluación del modelo reveló una precisión del 85%, con un análisis detallado de métricas como la matriz de confusión y el reporte de clasificación.

Aunque la precisión general del modelo es aceptable, se observaron deficiencias específicas en la clasificación de la clase 2.0. Se sugiere una revisión y mejora específica para esta clase. Además, la curva ROC proporciona una visión detallada del rendimiento del modelo en diferentes umbrales de decisión.

VI. DISCUSIÓN

1) Manejo de Datos Desbalanceados: Dada la baja cantidad de instancias para la clase 2.0, considerar estrategias de manejo de datos desbalanceados, como el sobremuestreo o submuestreo de la clase minoritaria.
2) Optimización de Hiperparámetros: Ajustar los hiperparámetros del modelo, como la tasa de aprendizaje y el número de iteraciones, puede ayudar a mejorar el rendimiento del modelo.
3) Exploración de Otras Técnicas: Evaluar la aplicación de otras técnicas de aprendizaje automático y minería de datos, como árboles de decisión, bosques aleatorios o máquinas de soporte vectorial, para comparar y mejorar el rendimiento.

REFERENCES

[1] F. Nelli, Python Data Analytics. US: Apress, 2015.
[2] P. Bruce y A. Bruce, Practical Statistics for Data Scientists. United States of America: O'Reilly, 2016.
[3] P. Bhatia, Data Mining and Data Warehousing Principles and Practical Techniques. United Kingdom: Cambridge University Press, 2019.
[4] C. Bishop, Pattern Recognition and Machine Learning. US: Springer, 2006.
[5] OMS. "Hipertensión". World Health Organization (WHO). Accedido el 1 de diciembre de 2023. [En línea]. Disponible: Sitio
[6] "Sklearn.linear_model.SGDClassifier — documentación de scikit-learn - 0.24.1". Site not found · GitHub Pages. Accedido el 1 de diciembre de 2023. [En línea]. Disponible: Sitio
[7] "Hipertensión arterial". IMSS. Accedido el 2 de diciembre de 2023. [En línea]. Disponible: Sitio