INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Evidencia 1: Proyecto de aprendizaje supervisado

Por

Daniel Eduardo Arana Bodart, A01741202

Jose Manuel Guerrero Arellano, A01747623

Valeria García Hernández, A01742811

Modelación del aprendizaje con inteligencia artificial (Grupo 201)

Dra. Maria Valentina Navárez Terán

24 de abril del 2024

INTRODUCCIÓN A LA PROBLEMÁTICA

El aprendizaje supervisado es uno de los tipos de **aprendizaje automático**, o machine learning como se conoce mayormente. Dentro del aprendizaje supervisado, los modelos de inteligencia artificial se entrenan con ayuda de un conjunto de datos de entrenamiento etiquetado. Estos conjuntos contienen datos de entrada y de salida en donde los datos de entrada son aquellos conjuntos de características que el modelo debe de aprender y reconocer su comportamiento con respecto a los datos de salida. Estos son normalmente conocidos como etiquetas, las cuales indican bajo el contexto de entrada, cual es la salida esperada, ya sea categórica o numérica.

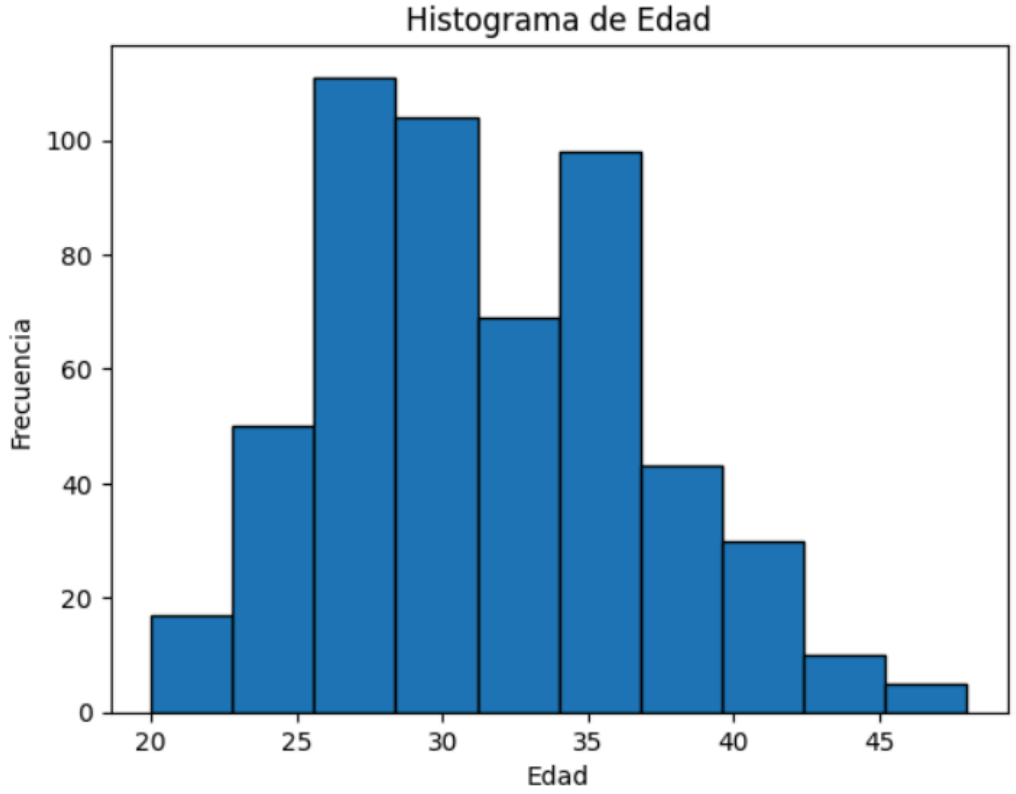
Existen diferentes tipos de algoritmos, o modelos, de aprendizaje supervisado, como la regresión, la clasificación, los árboles de decisión, K-Nearest Neighbor, redes neuronales artificiales, etc. Sus aplicaciones pueden ir desde el reconocimiento de voz, la visión por computadora, los sistemas de recomendación, la detección de fraudes, el análisis de sentimientos, etc. Una de las implementaciones radica en el **área de medicina y salud.** Bajo este contexto, se puede tomar como datos de entrada a aquellos registros e información recopilada de pacientes previos ingresada en conjuntos de datos, con el diagnóstico generado como etiqueta de salida. En este caso, los modelos pueden ser aplicados al diagnóstico de enfermedades, la predicción de riesgos, la recomendación de tratamientos, el desarrollo de fármacos e incluso la mejora de atención al paciente.

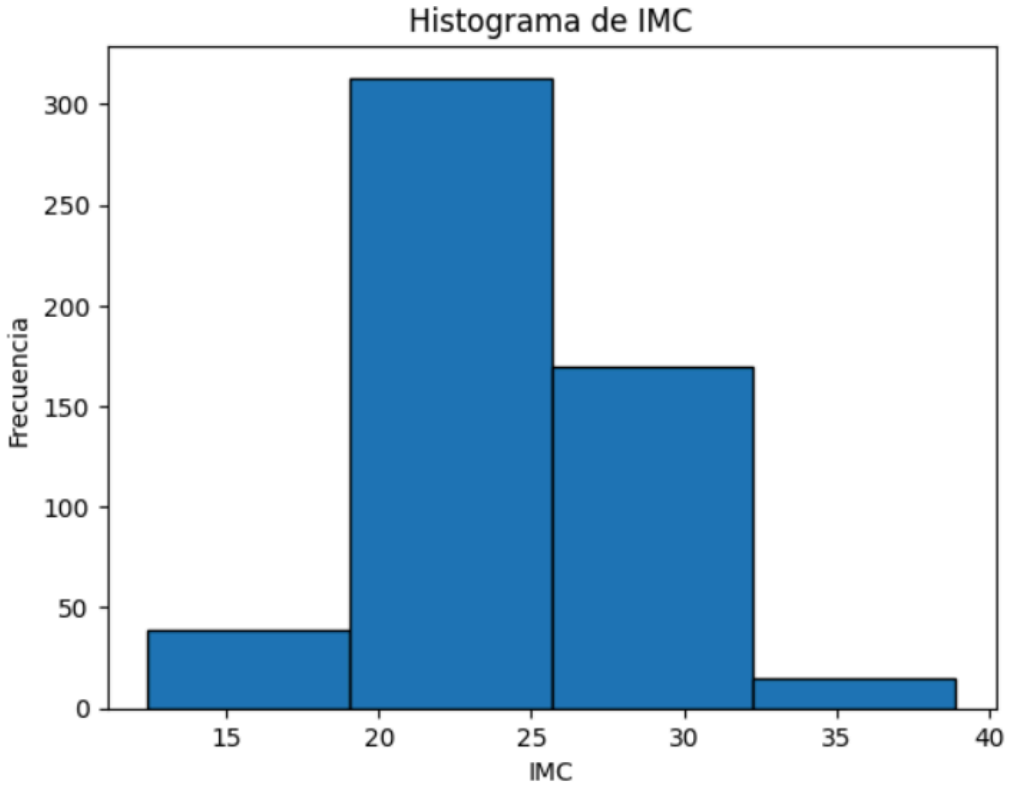
Un ejemplo de aplicación del aprendizaje supervisado de datos en esta industria es la investigación realizada por Google Health y el grupo Northwestern Medicine para lograr que el teléfono celular logre eventualmente ayudar a la detección de problemas dermatológicos. En esta publicación, Anna Martí (2021), menciona brevemente el proceso que se está realizando y cómo es que se está entrenando el modelo de inteligencia artificial con por lo menos 65,000 registros en ese entonces.

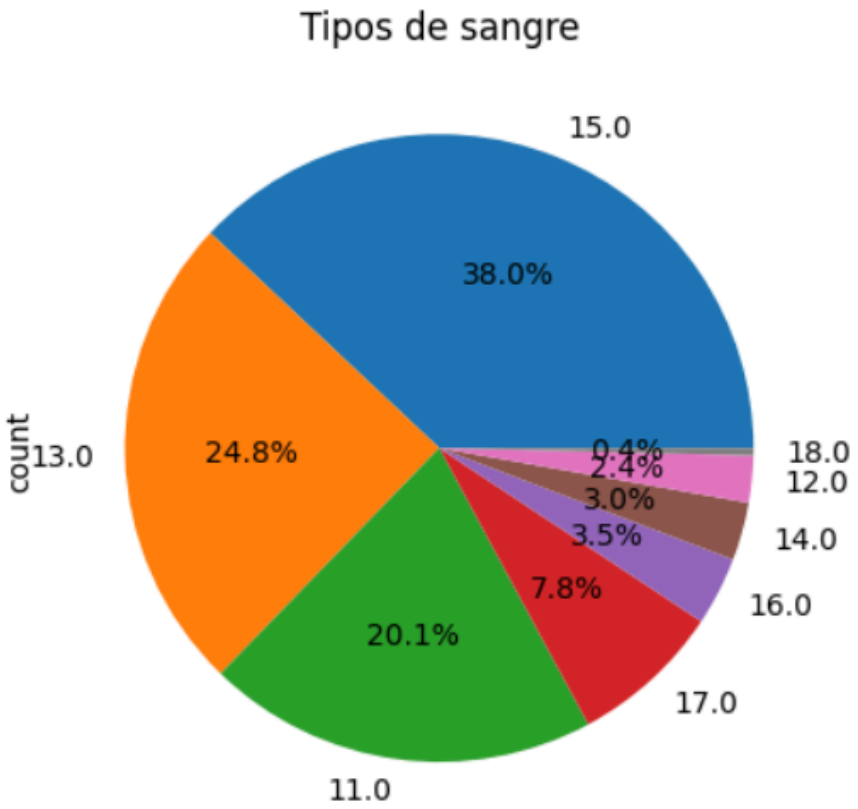
El blog de Datos Maestros menciona cómo la calidad de los datos en el sector salud se refiere a la precisión, integridad, confiabilidad y oportunidad que poseen en la industria. Posteriormente, mencionan los cinco tipos de datos que existen en este sector, los cuales son datos clínicos, administrativos, demográficos de los pacientes, farmacéuticos y de salud pública, (2023). En este caso, se trabajará con el set de datos para la detección del **Síndrome de Ovario Poliquístico**, o PCOS por sus siglas en inglés. Este dataset contiene los parámetros clínicos y físicos para la detección del síndrome, recopilados por 10 hospitales diferentes en la región de Kerala, en la India, este dataset tiene una forma inicial de 999 registros con 45 columnas, que, luego de dar un vistazo a la composición del set, se encontró que existía una columna, así como múltiples registros completos con valores nulos, por lo que se decidió quitar ambos casos del set de datos. Posteriormente, se encontró con 4 columnas que tenían problemas para poder visualizar sus medias de distribución, por lo que se decidió implementar una transformación logarítmica para poder tener una mejor visualización de sus distribuciones.

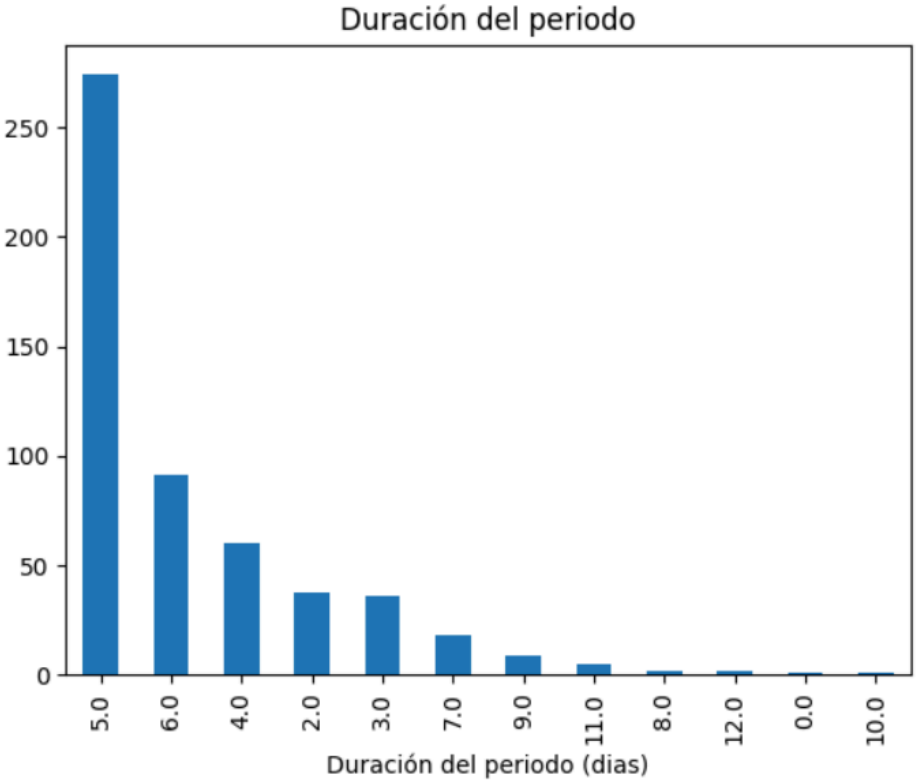
Una vez con el dataset limpio, seleccionado las variables de interés y haciendo la transformación logarítmica a aquellas variables que lo necesitaban procedimos a realizar un análisis exploratorio de aquellas variables que consideramos más importantes y que, de acuerdo con nuestro criterio, tendrían más correlación con que alguna mujer pudiera padecer de PCOS. Estas variables decidimos graficarlas de diferentes maneras de acuerdo con la distribución de los datos para cada variable. También dividimos la base de datos en variables numéricas y categóricas (booleanas) para que sus análisis fueran más sencillos.

Para las variables numéricas algunos de los resultados más interesantes se muestran en las siguientes gráficas. (Para ver todas las gráficas realizadas acceder al notebook anexado donde se comparte todo nuestro proceso).

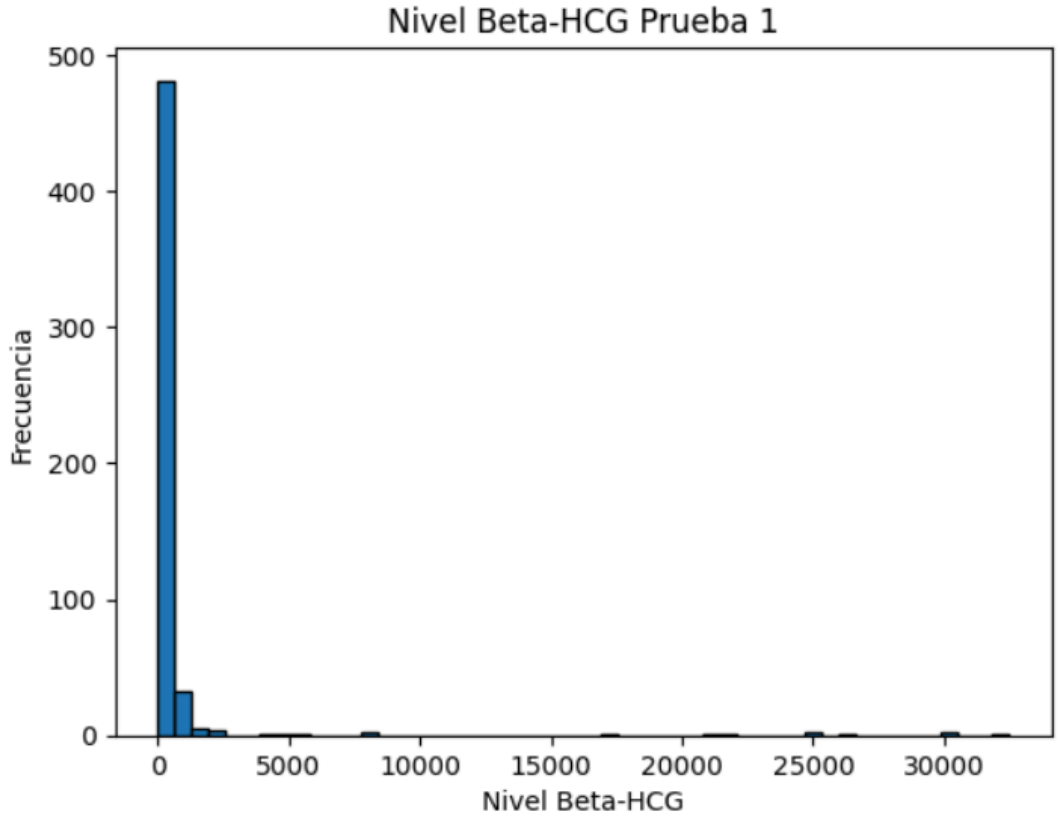


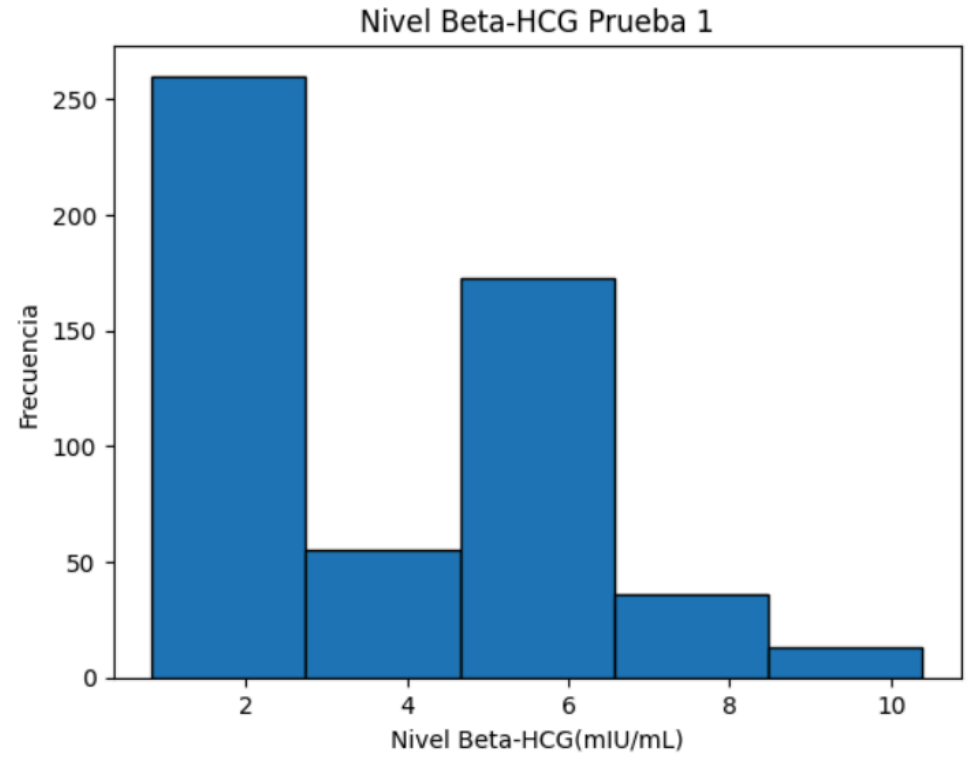



Tipos de sangre
count
15.0
38.0%
13.0
24.8%
11.0
20.1%
17.0
7.8%
16.0
3.5%
14.0
3.0%
12.0
2.4%
18.0
0.4%

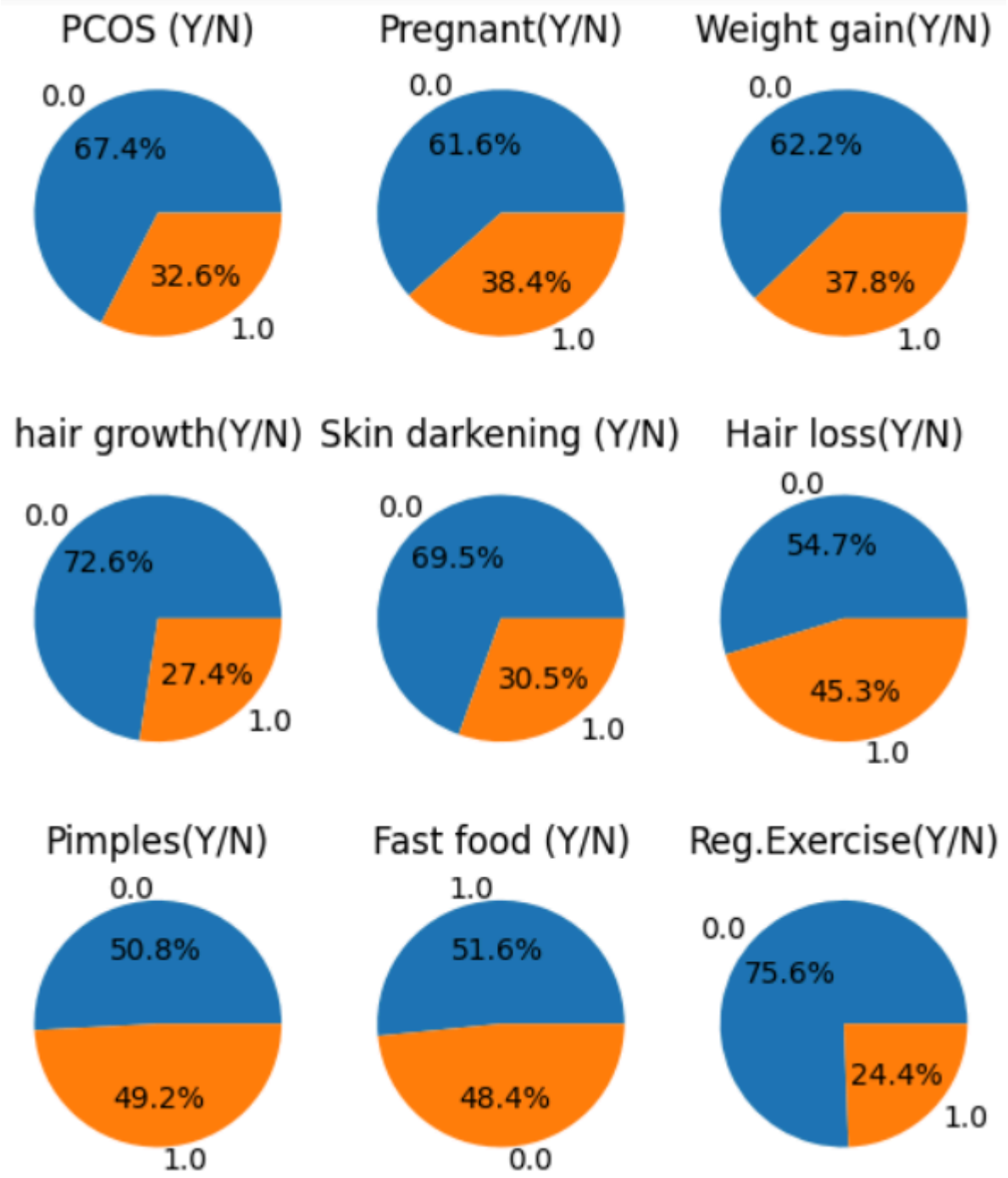

Duración del periodo
250
200
150
100
50
0
5.0
6.0
4.0
2.0
3.0
7.0
9.0
11.0
8.0
12.0
0.0
10.0
Duración del periodo (dias)

Ejemplo de la transformación logarítmica para visualización de los datos:





Las variables categóricas se muestran a continuación:



METODOLOGÍA

En total se aplicaron dos métodos de aprendizaje supervisado con el objetivo de hacer una comparativa entre modelos, como método visto en clase se optó por realizar un modelo de **redes neuronales**, esto debido a que las redes neuronales son especialmente útiles cuando se tienen muchas columnas el cual es nuestro caso. Originalmente teníamos pensado aplicar árboles de decisión, pero a la hora de mostrar el árbol, para conseguir un modelo adecuado era necesario tener un número de ramas(max_depth) muy alto, por lo que a la hora de observar el árbol quedaba un resultado ilegible, además que un gran número de ramas necesita de una alta cantidad de recursos para funcionar. Como método no visto en clase se optó por realizar el método **random forest**; aunque el método sean árboles de decisión y ya se explicó porque no se usó ese método antes. Consideramos que el hecho de realizar una gran cantidad de árboles de decisión excluye la necesidad de crear un solo árbol sumamente complejo y se puede dividir en 100 árboles más sencillos donde se pueden tomar en cuenta todas las variables.

Antes de hacer los modelos fue necesario primero preparar la base de datos como ya se explicó anteriormente, para ello se empleó la librería pandas con el dataset que se puede observar en el anexo y se remueven todos los datos NaN de la base de datos (no se buscó normalizar los datos o emplear medias porque solamente había dos registros en un dataset de más de 500 filas, por lo que podemos desprendernos de ellos sin mayores consecuencias). Después de eliminar los NaN hicimos un análisis de las diferentes variables para decidir cuáles no aportan valor al dataset y fueron eliminadas, variables tales como IDs o razones entre dos variables.

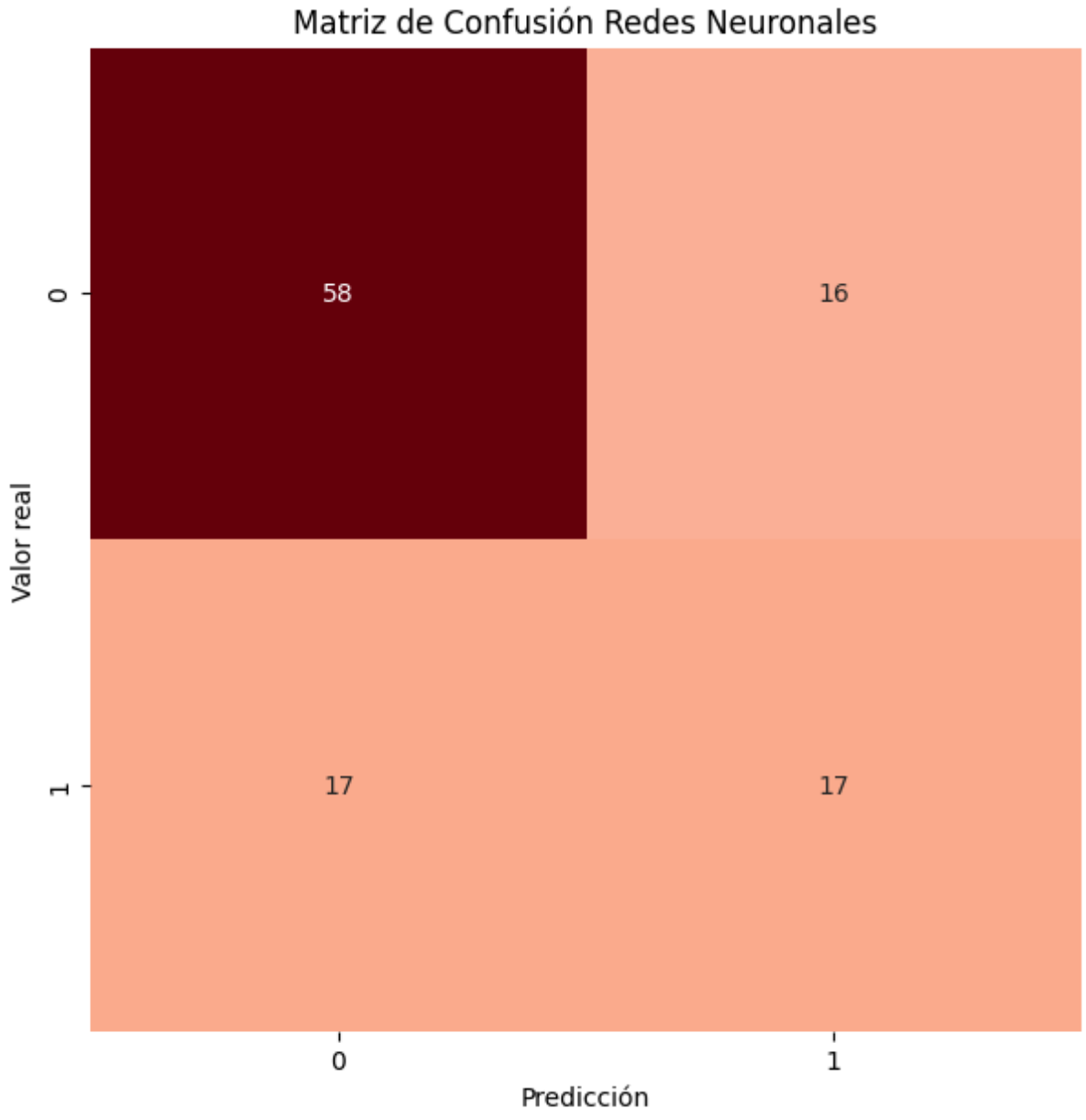
Una vez con la base de datos ya limpia se realizó el análisis exploratorio de los datos con las gráficas que ya se mostraron previamente y procedimos a realizar el modelo de redes

neuronales. Para este modelo fue necesario emplear la librería sklearn y dividir la base de datos en un 80-20 para entrenamiento y pruebas respectivamente, con una semilla de aleatoriedad designada (42) para asegurar reproducibilidad en nuestro modelo. Posteriormente, con ayuda de la función GridSearchCV dimos una serie de dimensiones para la red neuronal, así como una lista con el número máximo de iteraciones. También empleamos Cross Validation para descartar que el desempeño del modelo fuera dado exclusivamente por suerte. Estuvimos probando hasta que la variabilidad entre los modelos obtenidos fuera muy poca. A prueba y error fuimos eliminando números máximos de iteraciones (2, 5, 20, 50, 100) y dimensiones de la red neuronal ((10,10),(50,50),(100,50),(100,100),(300,300)) hasta quedar con opciones que tenían un buen rendimiento. Es importante mencionar que la función GridSearchCV realiza modelos diferentes cada vez que se emplea, por lo que es necesario tener un poco de paciencia hasta encontrar un modelo con un desempeño deseable. Al final terminamos con una red neuronal con capa oculta de dimensión (50,25) y un número máximo de 2500 iteraciones como nuestro modelo con mejor desempeño general después de la Cross Validation.

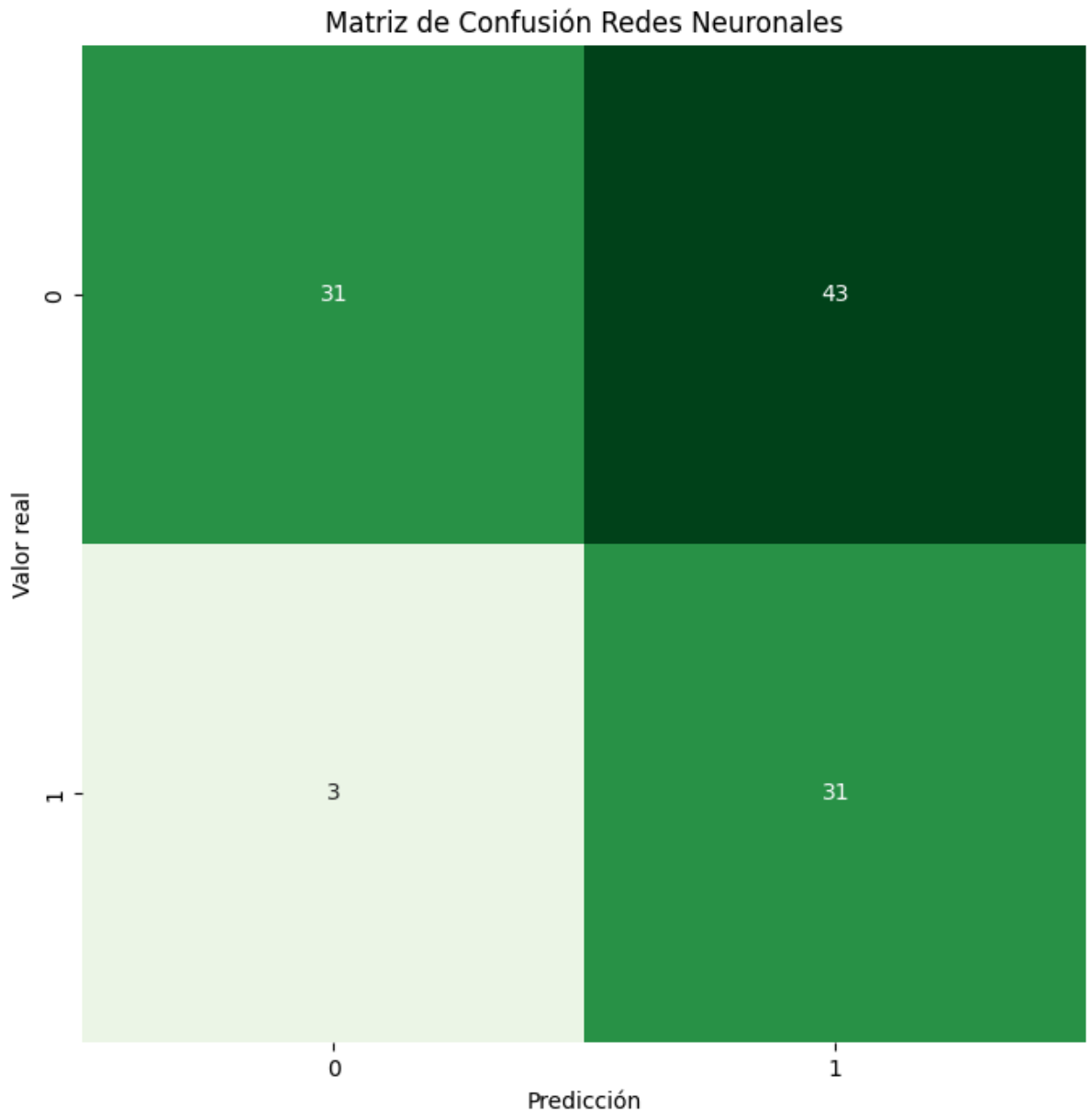
El modelo random forest fue un poco más directo, para este empleamos de igual manera la librería sklearn. Para este caso fue necesario dividir el dataset en uno que tuviera solamente la clase a predecir, y a otro que tuviera todo el resto del dataset sin la clase a predecir. Esto debido a que si no se separan los árboles de decisiones creados simplemente buscarán el valor de la clase a predecir asignado a cada registro y lo asignan a su predicción, mientras que buscamos todo lo contrario. Una vez con el dataset separado se dividió de igual manera en un 80-20 el dataset para el apartado de entrenamiento y el de prueba respectivamente y empleamos la semilla asignada de aleatoriedad para el modelo anterior (42), de esta manera nos aseguramos que los datos de entrenamiento y prueba en ambos modelos sean los mismos.

Con esto en mente se entrenó el modelo con un hiperparámetro de máximo 100 árboles aleatorios (n_estimators) y tenemos el modelo listo para realizar un análisis de desempeño así como una comparación entre ambos modelos.

RESULTADOS:

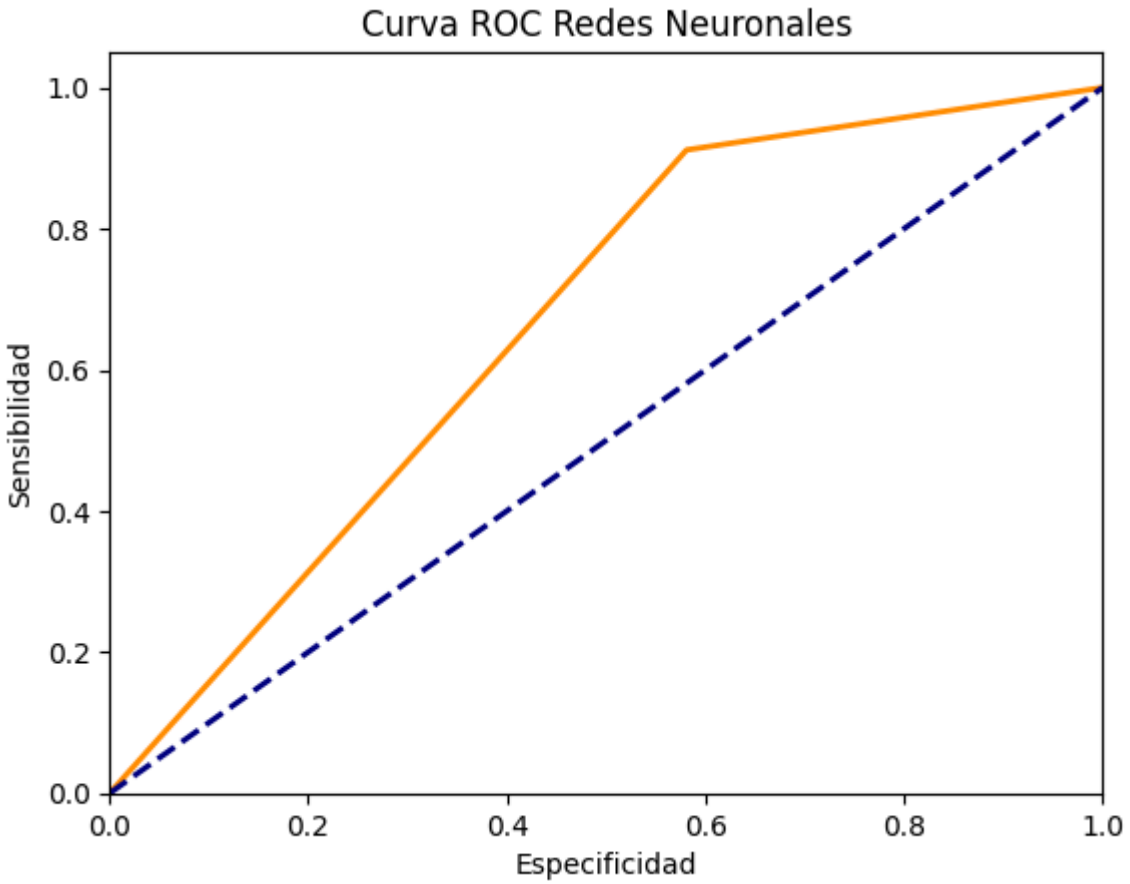


Matriz de confusión para red neuronal con parámetros (50,25) con max_iter = 100, cabe destacar que este no es el modelo final implementado. Esta matriz demuestra cómo el mal entrenamiento de una red neuronal con hiper parámetros diferentes puede llegar a un desempeño mucho peor, en este caso con precisión de 0.52 y recall de 0.50 para las etiquetas 1 y 0.77 y 0.78 para las etiquetas 0, respectivamente.

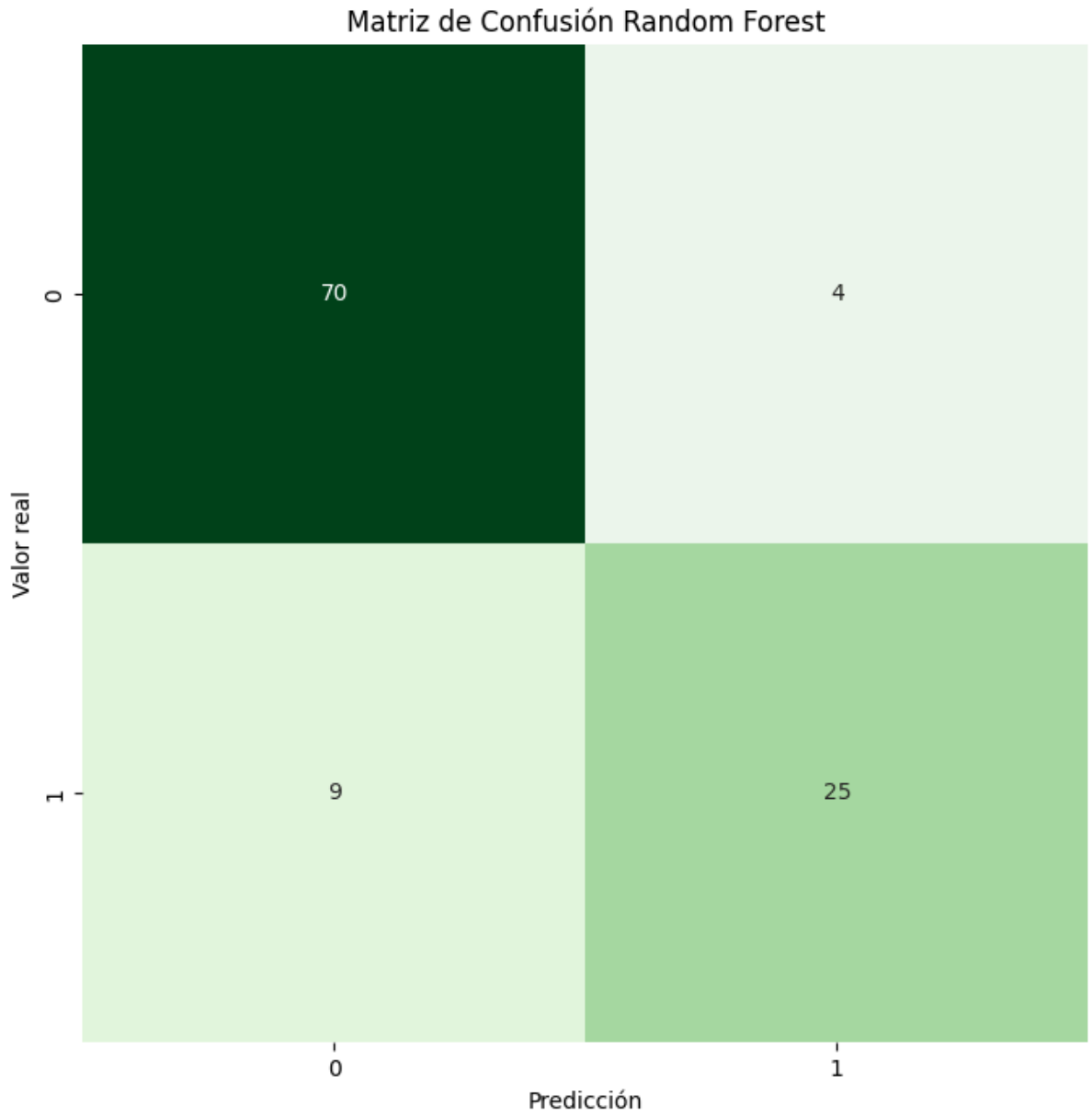


Esta es la matriz de confusión para el modelo de red neuronal que se entrenó a manera que se procuró mejorar los resultados ante la métrica recall. La red neuronal se entrenó con mallas

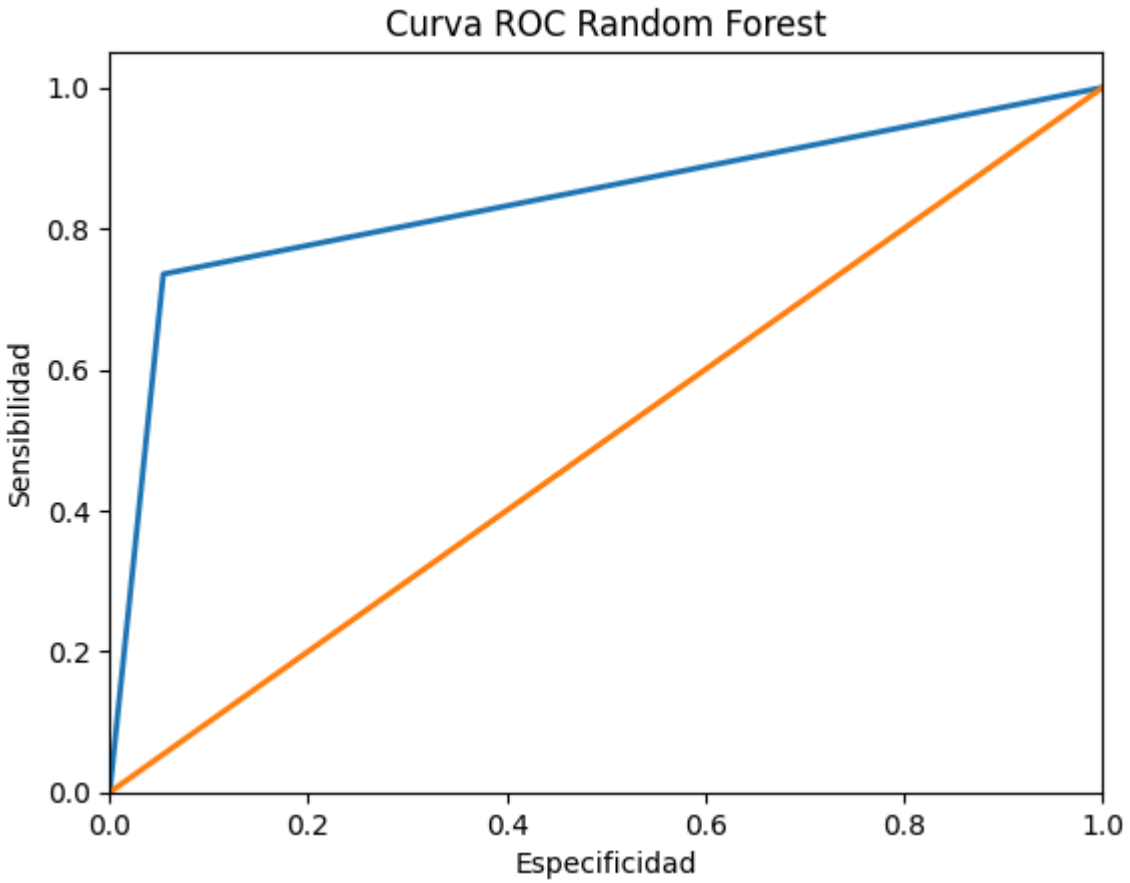
ocultas de tamaño (50,25) y una iteración máxima de 2500. Esto permitió que el modelo pudiera obtener las siguientes medidas, 0.91 en precision y 0.42 en recall para la etiqueta 0, y 0.42 y 0.91 respectivamente para la etiqueta 1. A través de estos valores y de analizar la matriz, se puede ver que el modelo es muy bueno para predecir a la gran mayoría de gente que padece PCOS al costo de dar muchas predicciones falsas, es decir, que no tenían PCOS, sin embargo el modelo arrojó que sí tenían.



En esta gráfica de ia curva ROC, se puede ver cómo es que el modelo de red neuronal tiene una mejor predicción que el azar, por lo que sigue siendo mejor alternativa aplicar este modelo en los pacientes que simplemente ir determinando aleatoriamente si algún paciente padece de PCOS.



Por otro lado, está la matriz de confusión para el modelo random forest, en donde se obtuvieron las medidas de precision de 0.89 y recall de 0.95 para la etiqueta 0, y 0.86 y 0.74 respectivamente para la etiqueta 1. Interpretando esta información, se puede ver cómo el modelo es muy bueno para descartar aquellos casos en donde no existe el síndrome, sin embargo, es sustancialmente peor en comparación al otro modelo, en lograr predecir el síndrome a la mayor cantidad de personas posible, debido a una proporción mucho menor en la cantidad de personas con PCOS y personas que no lo padecen .



En esta parte se puede ver cómo es que el modelo de random forest tiene mejor predicción en comparación a la recta del azar. Esta curva muestra cómo cuando hay pocas personas que no padecen PCOS, hay gran cantidad de diagnósticos relevantes realizados, y cómo esta relación de crecimiento poco a poco va decreciendo según van aumentando las medidas.

A lo largo del desarrollo del modelo, se encontraron con diversas áreas de oportunidad, las cuales si se llevase a cabo un plan de corrección, permitirían un mejor desempeño de los modelos de aprendizaje supervisado. En el caso del diagnóstico del Síndrome de Ovario Poliquístico (PCOS), hay varias áreas de oportunidad para mejorar la asignación que hacen los modelos de aprendizaje automático. Una de las estrategias que podría ser útil para ello es la aplicación de técnicas de remuestreo, como el sobremuestreo de la clase minoritaria o el submuestreo de la clase mayoritaria, con el fin de equilibrar la distribución de clases en el

conjunto de datos. Por ejemplo, existe la estrategia SMOTE por sus siglas en inglés(*Synthetic Minority Over-sampling Technique*) en donde la manera de incrementar los muestreos que están en minoría y disminuir los muestreos que están en mayoría, permitirá que exista un desempeño mucho mayor en pruebas como la curva ROC, (Chawla, 2002). Además, ajustar los umbrales de decisión del modelo puede ser útil para equilibrar la precisión y el recall según las necesidades clínicas específicas. También es importante considerar la optimización de hiperparámetros, como el ajuste de la tasa de aprendizaje o el número de árboles en el caso de modelos de ensemble como random forest, para mejorar el rendimiento del modelo en la detección de casos de PCOS. Al implementar estas estrategias básicas, se puede mejorar la capacidad de los modelos para identificar con precisión el PCOS y proporcionar diagnósticos más confiables en la práctica clínica.

CONCLUSIONES INDIVIDUALES

Daniel Eduardo:

En mi opinión aplicar un método de aprendizaje supervisado fue entretenido y aprendí bastante con el desarrollo de este proyecto. El trabajo en equipo fue sencillo y las actividades realizadas siento que se coordinaron muy bien entre los tres miembros del equipo. En cuanto a los modelos obtenidos, me llamó la atención particularmente obtener el modelo de redes neuronales por medio de la validación cruzada con GridSearch, ya que en mi opinión le agregó un nivel de profesionalidad un poco mayor al esperado porque nos quisimos asegurar de estar empleando el mejor modelo dentro de los hiperparámetros que nosotros consideramos apropiados para la red. Es verdad que si hubiésemos tenido más tiempo para probar diferentes combinación probablemente habríamos obtenido un modelo con mejor desempeño, pero para el tiempo que tuvimos en mi opinión tuvimos un par de modelos muy interesantes, ya que el modelo de redes neuronales tuvo gran éxito para predecir cuando una

mujer padece de PCOS mientras que carecía enormemente para predecir satisfactoriamente cuando no lo padecía, mientras que el modelo random forest resultó en todo lo contrario: predijo con bastante exactitud cuando una mujer no padece de PCOS; sin embargo, tenía fallas a la hora de predecir cuando una mujer sí que lo padecía. Por último me gustaría agregar que considero de suma utilidad el emplear inteligencia artificial y métodos de aprendizaje para predecir diversos resultados en cualquier ámbito, ya sea dentro de la salud, en los negocios, para índices geográficos, entre otros. Sin embargo, considero que, en especial dentro de la rama de la salud, es muy importante entender que la inteligencia artificial no dice la verdad. Es muy probable que cometa errores debidos a un modelo que no está bien entrenado, un método de aprendizaje supervisado erróneo para los tipos de datos o incluso por los mismos sesgos que existan en la base de datos o durante el proceso de ciencia de datos. Por lo tanto hay que ser sumamente críticos con toda la información que obtenemos por medio de inteligencia artificial y usarla como herramienta, más no como verdad absoluta.

Manuel:

A través de esta evidencia se lograron desarrollar competencias pertinentes para el uso, manejo y compresión de modelos de inteligencia artificial no supervisados, así como la diferencia entre los diferentes tipos de inteligencia artificial que se han desarrollado con el tiempo. Lo interesante de el análisis estos modelos generados es cómo han habido avances para poder generar métodos más robustos con estrategias para reducir aquellas anormalidades en los datos, así como el poder brindar la mejor solución a partir de la adecuada comprensión de los datos de entrada, y el uso adecuado del modelo, así como la capacidad poder seleccionar el mejor modelo para el escenario y problema a resolver. Considero que la matriz de confusión, así como las medidas para detectar el desempeño del modelo son parte

fundamental para lograr estos objetivos, ya que permiten que la evaluación del modelo no dependa del contexto de los datos.

Valeria:

En este proyecto, utilizamos métodos de aprendizaje supervisado para la detección del Síndrome de Ovario Poliquístico (PCOS) utilizando datos clínicos recopilados de pacientes. Aprendimos que el aprendizaje supervisado es una buena herramienta en la inteligencia artificial, donde los modelos pueden entrenarse con conjuntos de datos etiquetados para realizar predicciones sobre nuevos datos.

Durante el proceso, aplicamos dos modelos diferentes: redes neuronales y random forest.

Cada modelo tenía sus fortalezas y debilidades en términos de precisión para detectar casos positivos y negativos de PCOS. Mientras que las redes neuronales mostraron un excelente recall en la detección de casos positivos de PCOS, el random forest fue más efectivo en la detección de casos negativos. En general, este estudio nos brinda una comprensión más profunda de cómo aplicar técnicas de aprendizaje supervisado en la medicina y nos hace darnos cuenta de la importancia de evaluar cuidadosamente el rendimiento de los modelos en contextos clínicos.

Anexos:

Notebook:

https://colab.research.google.com/drive/1ZBj469Bg1qYPRsv-S0e8pSs3uxwS3m_1?usp=sharing

Base de datos:

https://drive.google.com/file/d/1iJV5HXXIab-9ja2b2PwgXOH_OaR8inXX/view?usp=sharing

Referencias:

Chawla, N. V., Bowyer, K. W., Hall, L., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. *Journal Of Artificial Intelligence Research/˜The œJournal Of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Hanasi, M. (2023, 29 octubre). El poder de la calidad de los datos en la salud: Cómo la calidad transforma la atención médica 2024 |. *Datos Maestros™*. https://datosmaestros.com/como-la-calidad-de-los-datos-transforma-la-salud/

Martí, A. (2021, 19 mayo). *Detectar problemas en la piel (incluso cáncer) usando la cámara del móvil: eso busca Google con uno de. . .* Xataka. https://www.xataka.com/otros/detectar-problemas-piel-incluso-cancer-usando-camara-movil-eso-busca-google-uno-sus-ultimos-proyectos

Polycystic ovary syndrome (PCOS). (2020, 11 julio). Kaggle. https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

Síndrome del ovario poliquístico: MedlinePlus enciclopedia médica. (s. f.). https://medlineplus.gov/spanish/ency/article/000369.htm