TRABAJO DE GRADO
Opción Seminario-Diplomado.

Predicción de Riesgo Prioritario en Pacientes con Cáncer a partir de Datos de Morbilidad utilizando Big Data y Deep Learning en Colombia

Corporación Universitaria Remington.

Facultad de ingeniería

Ingeniería en sistemas

Jeckson Armando Rivera Potosi

Tutor:

Juan Pablo Vélez Uribe

Opción de trabajo de grado seminario

Colombia

2025

Tabla de Contenidos

1 Contenido

Índice de Figuras

Índice de Tablas

2 Dedicatoria

A Dios, por darme fortaleza y sabiduría en cada etapa de este proceso. A mis padres, por su amor incondicional y apoyo constante, por transmitirme valores como el respecto, lealtad perseverancia, honestidad y gratitud. Su apoyo constante ha sido fundamental para mi crecimiento personal, laboral y académico.

A mi familia, cuyo amor, apoyo y enseñanzas han sido la base fundamental para seguir adelante. A mi novia, por su paciencia, compañía y motivación constante en los momentos más desafiantes, por bríndame palabra de aliento cuando estaba a punto de rendirme. Este trabajo es fruto de todo lo que me han brindado y una muestra de mi profundo agradecimiento.

3 Agradecimientos

A Dios, por iluminar mi camino y permitirme culminar este proyecto. A mis padres, por su apoyo incondicional y por enseñarme que con esfuerzo, dedicación y constancia se pueden lograr los objetivos propuestos.

A mis familiares, por su apoyo constante, por sus palabras de motivación, por compartir este proceso a mi lado y celebrarlo como propio.

A la **Corporación Universitaria Remington**, por abrirme las puertas del conocimiento y brindarme las herramientas necesarias para crecer profesional y personalmente.

A mi tutor de grado, **Juan Pablo Vélez Uribe**, por su acompañamiento, por su guía, dedicación y orientación durante todo el desarrollo de este trabajo, aportando siempre al fortalecimiento académico y humano.

Al **Instituto Nacional de Cancerología**, a las plataformas de datos abiertos de Colombia (SISPRO, DANE) y a mis compañeros por sus observaciones y apoyo durante el desarrollo de este proyecto.

4 Resumen

Este proyecto de grado se centra en el cáncer en Colombia, pero no solo como una enfermedad que afecta a las personas, sino como un gran problema para la salud de todos, con diferencias importantes según dónde vivan y cómo sea su situación económica. Lo que queremos es crear un modelo que prediga quién tiene más riesgo de sufrir complicaciones, usando técnicas de Deep Learning y Big Data para clasificar a los pacientes en niveles de riesgo alto o bajo. Así, los médicos pueden dar prioridad a quienes más lo necesitan.

Para lograrlo, analizamos 25,000 registros de salud del Sistema Integral de Información de la Protección Social (SISPRO). Un desafío importante es que los datos no están equilibrados, algo que suele pasar en salud pública. Para solucionarlo, usamos el algoritmo SMOTE, que nos ayuda a equilibrar la muestra y a que la red neuronal identifique bien los casos de alto riesgo, que son los menos comunes. El modelo se construye con Python, TensorFlow y Keras, y logra una precisión y una puntuación F1 del 100%. Esto demuestra que el aprendizaje profundo es muy útil para encontrar patrones complejos en los datos de salud.

Después, los resultados se procesan y organizan en bases de datos SQL para analizarlos con Power BI. Las visualizaciones indican cosas muy importantes e interesantes, vemos que muchos casos de alto riesgo se concentran en las ciudades y en personas con régimen contributivo, lo que indica que hay problemas para acceder a la salud y que faltan registros en zonas rurales y en personas con seguro subsidiado. En resumen, el proyecto muestra que la ingeniería de datos y la inteligencia artificial son herramientas muy útiles para que la salud sea más justa, identificando a personas que necesitan ayuda y que no siempre son visibles, y dando información científica para que los gobiernos tomen mejores decisiones.

5 Palabras clave

Big Data, Deep Learning, Riesgo Oncológico, Morbilidad, Redes Neuronales, SMOTE, SISPRO.

6 Marco conceptual

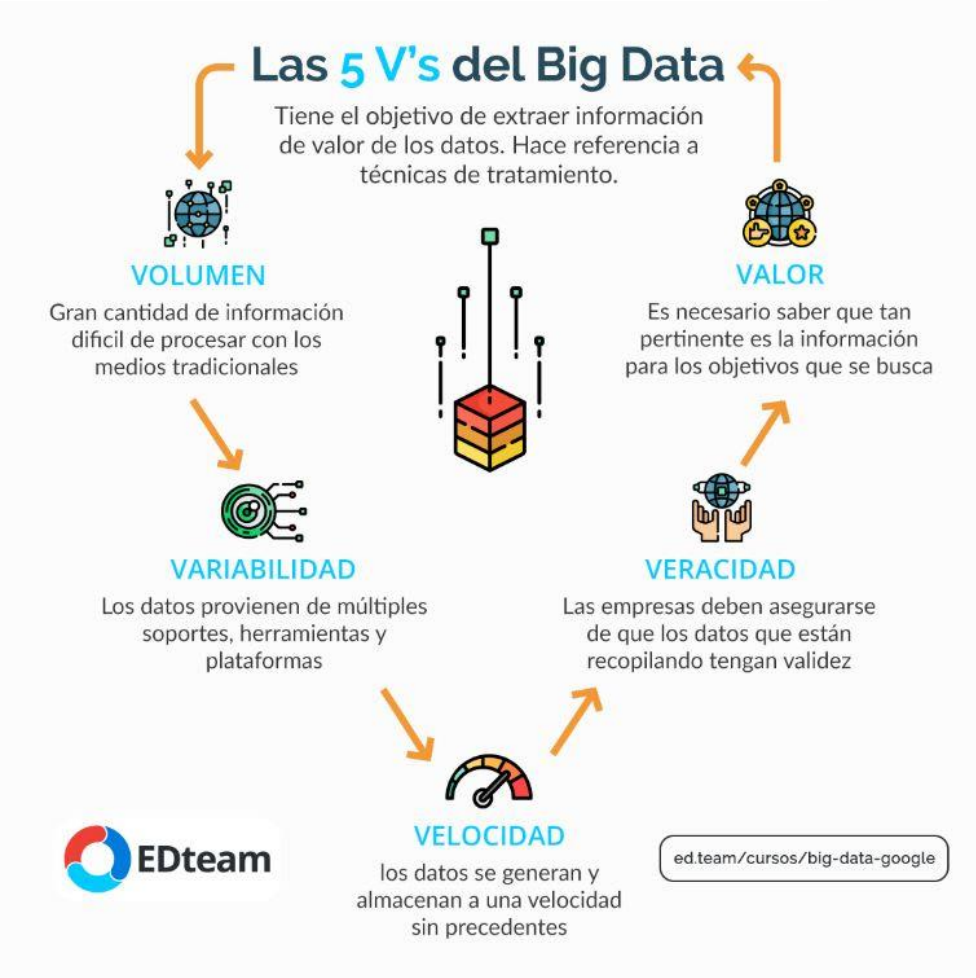
6.1 Big Data: Definición

El Big Data hace referencia a conjuntos de datos tan grandes que las herramientas tradicionales de procesamiento, como Excel, no pueden analizarlos ni identificar patrones de manera eficiente. Su propósito es procesar y examinar enormes volúmenes de información para generar resultados comprensibles, incluso para personas con pocos conocimientos en análisis de datos. (Chen et al., 2012).

6.2 Big Data y las 5 “V”

El Big data se ha popularizado inicialmente por las 5 V, las cuales son volumen, velocidad, variedad, veracidad y valor.

Figura 1. Las "5v" del Big Data



Herrera, R. (2018)

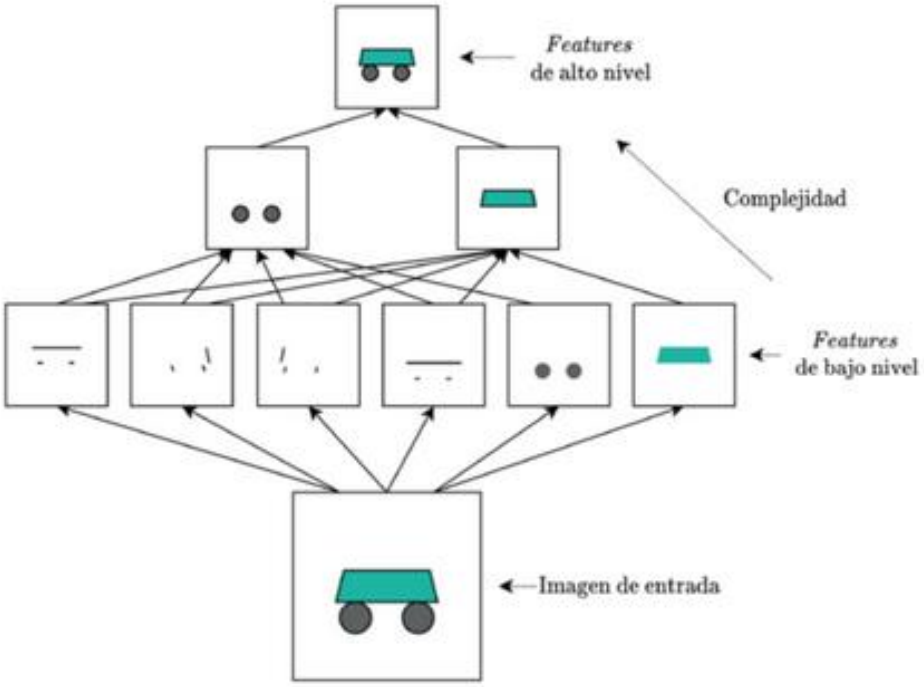
6.3 Redes Neuronales

Son sistemas informáticos con nodos interconectados los cuales funcionan de forma similar al cerebro humano, son capaces de reconocer patrones, descubrir relaciones en datos ocultos que estén sin procesar, para posteriormente agruparlos, clasificarlos y con esto aprender y mejorar de forma continua. (Goodfellow et al., 2016)

6.4 Deep Learning

Es un tipo de Machine Learning el cual consiste en un entrenamiento profundo en el cual la extracción de características de los datos originales genera una representación jerárquica distribuida. Se utilizan redes neuronales artificiales con muchas capas imitando el cerebro humano, lo cual permite a las maquinas aprender patrones complejos en grandes cantidades de datos sin la necesidad de una programación avanzada, logrando realizar tareas complejas como reconocimiento de voz, de imágenes o predicciones en base a los datos analizados. (Pérez Borrero & Gegúndez Arias, 2022)

Figura 2. Jerarquía en un Modelo Deep Learning



Pérez Borrero y M. E. Gegúndez Arias (2022)

6.5 Modelación con SMOTE

Es una herramienta de sobre muestreo que ayuda a enfrentar los conjuntos de datos desbalanceados, mejora la representación de la clase que tenga menor información generando muestras sintéticas.

Este método usa el algoritmo ***k-nearest neighbors*** (KNN) para identificar los ejemplos más cercanos dentro de la clase minoritaria. A partir de esos puntos seleccionados, genera nuevas muestras combinando sus características y creando así datos adicionales dentro del mismo espacio de atributos. (Chawla et al., 2002)

6.6 Aplicación de Big Data y analítica predictiva

Para implementar correctamente el Big Data y el análisis predictivo el paso más importante es definir la estrategia correcta. Se seleccionan las herramientas y tecnologías adecuadas, así como la contratación de talento especializado en el análisis, procesamiento de datos. Si una empresa invierte buenos recursos para implementar el Big Data y análisis predictivo estará mejor posicionada frente a la competencia para enfrentar desafíos futuros y mantenerse relevante en el mercado. (Davenport, 2014).

6.7 Prevención de cáncer

6.7.1 ¿Qué es el cáncer?

El cáncer es una enfermedad que se distingue por la multiplicación irregular y excesiva de células, lo cual provoca el crecimiento incontrolado de un tumor maligno en algún órgano o tejido específico. Se considera un problema importante en la salud publica colombiana, esto debido al incremento de enfermos y personas muertas por dicha causa. Se estima que en los

últimos años alrededor de 90 personas mueren diariamente por algún tipo de cáncer. (Organización Mundial de la Salud [OMS], 2024)

6.7.2 Registros de cáncer a nivel poblacional

Los registros de cáncer a nivel poblacional son fundamentales para el control del cáncer. Su organización y coordinación suelen estar a cargo de entidades profesionales, como la Asociación Internacional de Registros de Cáncer (IACR, por sus siglas en inglés), en colaboración con organismos de la OMS especializados en cáncer, como la Agencia Internacional de Investigaciones en Cáncer (IARC) (Sistema de Información de Cáncer en Colombia – Infocancer, s.f.).

6.7.3 Detection temprana, Big Data y Deep Learning

Generalmente el cáncer suele detectarse una vez haya alcanzado la madures, pero existen cambios celulares leves mucho antes de diagnosticarse, que se pueden tratar en sus primeras fases de desarrollo. Es aquí donde se contempla la importancia del Big Data, ya que permite analizar datos de pacientes cuyos síntomas fueron responsables de alcanzar la madures y convertirse en tumor. El Deep learning por su parte reconoce patrones similares (síntomas) que facilitan la detección de un posible cáncer de forma más rápida y precisa ayudando ante un posible diagnóstico médico. (Ravi et al., 2017).

7 Marco contextual

7.1 Contexto: El Cáncer como Problema de Salud Pública en Colombia

El cáncer es un problema prioritario en la salud pública en Colombia y es necesario un nuevo plan para su control específico y vinculante para el sistema de salud colombiano. Revista Colombiana de Cancerología. 28, 4 (dic. 2024), 136–138.

Según la revista colombiana de cancerología, en el artículo publicado por Pardo y Cendales (2023) se presentan estimaciones para los cinco principales tipos de cáncer en Colombia durante los años 2017-2021 tomando como referencia, departamento, sexo y basados en datos proporcionados por el instituto nacional de cancerología y el departamento administrativo nacional de estadística DANE (Pardo & Cendales, 2023)

Son aproximadamente 47.393 casos estimados en hombres y 54.090 casos en mujeres con alrededor de 44.355 muertes registradas durante estos años, el tipo de cáncer con mayor incidencia fue en hombres, cáncer de próstata (43,6 por cada 100 000 años-persona) y en mujeres, cáncer de mama (40,2 por cada 100 000 años-persona), en conjunto el cáncer de mama, próstata, tiroides, colon-recto, y estomago representaron alrededor del 47.3 % de nuevos casos en Colombia. (Pardo & Cendales, 2023).

7.2 Desafíos Sanitarios y Disparidad Socio-Geográfica

La predicción y supervivencia de los pacientes con cáncer en Colombia no solo dependen de temas biológicos, sino también el afrontar muchos obstáculos como dependencia geográfica y

socioeconómica (INS, 2020; MinSalud, 2022). Por lo cual es muy importante analizar estas variables en datos de morbilidad y poder detectar posibles fallas que conllevan a una categorización de alto riesgo.

7.2.1 limite geográfico: entrelazamiento urbano y rural

El lugar de residencia es un elemento crucial con respecto a la rapidez en el acceso a la salud y enfermedades oncológicas, los pacientes que habitan en zonas rúlales afrontan muchos retos como, falta de transporte, recursos limitados, poca información y capacitaciones con respecto a pacientes que residen en zona urbana. (INS, 2020)

Otra dificultad el acceso a la infraestructura hospitalaria, ya que equipos especializados para la detección y tratamiento de enfermedades oncológicas (equipos de radioterapias, centros integrales de oncología) se encuentran en las principales ciudades del país. Esto conlleva a que los pacientes de zonas rurales incurran en gastos de transportes, hospedaje, alimentación, convirtiéndose en un obstáculo económico y logístico para la persistencia en el tratamiento. (INS, 2020).

7.2.2 Retraso en la detección.

generalmente las remisiones de primer a segundo nivel en zonas alejadas son poco más lentas, este retraso afecta directamente el grado de severidad en el que se diagnostica la enfermedad, provocando que los casos se categoricen como ALTO RIESGO debido a su progreso avanzado. (MinSalud,2022).

7.2.3 Efecto del sistema de afiliación y utilización de servicios

El pronóstico también varía significativamente dependiendo del sistema de salud al que el paciente esté vinculado (subsidiado o contributivo). A pesar de que el sistema Contributivo generalmente permite acceder más pronto a consultas médicas y especialistas, los pacientes del sistema Subsidiado o sin posibilidad de pago suelen tener que esperar mucho tiempo para la confirmación de diagnósticos, la realización de procedimientos difíciles y el inicio del tratamiento. Los reclamos asociados con procedimientos oncológicos son a menudo objeto de tutorías o quejas, lo que revela los desafíos para una atención continua y adecuada (Defensoría del Pueblo, 2021).

7.2.4 Justificación del análisis social

Esta desigualdad estructural apoya la evaluación realizada en el proyecto, el modelo de Deep Learning no solo prevé, sino que además identifica a las poblaciones prioritarias que necesitan una intervención en el sistema sanitario para garantizar la equidad, al demostrar que los casos de alto riesgo se agrupan en ciertas áreas o sistemas (como se muestra en la Figura 15, pag.39) en la sección de Resultados).

7.3 Origen de la Base de Datos (Big Data) y su Alcance

Este proyecto de análisis predictivo se basa en el acceso y utilización de grandes cantidades de datos provenientes del sistema de salud colombiano. La base de datos es empleada para asegurar la validez del estudio realizado.

7.3.1 Origen

La fuente de la información que sustenta este proyecto es el conjunto de registros de MORBILIDAD por CANCER, disponible a través del portal Datos Abiertos de Colombia, el

cual funciona como un instrumento para difundir la información consolidada que se obtiene del Sistema Integral de Información de la Protección Social (SISPRO). El cuál es la herramienta tecnológica más importante del Ministerio de Salud y Protección Social de Colombia, su objetivo es recolectar, procesar y reunir la información necesaria para crear políticas públicas, tomar decisiones y supervisar los sectores de salud, pensión y riesgos laborales.

Para ese estudio se utilizaron 25.000 registros de los 40.700 disponibles, la función de SISPRO como Data Warehouse garantiza que los de los 25.000 registros representen con precisión y estandarización la información epidemiológica de Colombia.

7.3.2 Validación de Alcance (Volumen y Variedad)

El conjunto de datos representa una muestra relevante de Big Data en el sector de la salud en Colombia corroborando dos de las características fundamentales para la aplicación de Big Data.

7.3.3 Volumen

La base de datos contiene alrededor de 47.300 registros de los cuales se utilizaron 25.000 historiales de pacientes con cáncer. Esta gran cantidad de información facilita el entrenamiento de un modelo complejo de Deep Learning previniendo el sobreajuste y garantizando que las tendencias detectadas sean robustas desde el punto de vista estadístico.

7.3.4 Variedad

La abundancia del conjunto de datos se basa en la incorporación de variable no únicamente clínicas (Tipo de cáncer, Diagnóstico) sino también sociodemográficas y administrativos como, tipo de afiliación(subsidiado/contributivo), localización geográfica(urbana/rural) edad y sexo. Esta mescla de información variada es lo que le permite al modelo reconocer patrones de riesgo, clasificación y diferencia el cual es el objetivo de este proyecto.

Con la implementación de esta fuente de información se asegura la trazabilidad y coherencia de los datos, lo cual es fundamental para que los resultados del modelo de predicción puedan ser aplicados y tengan efecto en el sector de la salud pública.

8 Objetivo General

Diseñar e implementar un modelo predictivo fundamentado en arquitecturas de Deep Learning y procesamiento de Big Data, para categorizar los niveles de riesgo (alto/bajo) en pacientes oncológicos de Colombia.

9 Objetivos Específicos

Para cumplir con el objetivo principal se definen los siguientes objetivos técnicos y analíticos.

1. Analizar los datos de morbilidad por cáncer obtenidos del portal de datos abiertos de Colombia (SISPRO) con el objetivo de reconocer las variables geográficas, administrativas y clínicas que más influyen en el pronóstico del paciente.

2. Implementar métodos de procesamiento de datos, específicamente el algoritmo de sobremuestreo SMOTE, para corregir el desequilibrio en los registros y asegurar que el modelo aprenda de forma precisa a identificar los casos de alto riesgo.

3. Desarrollar y entrenar una red neuronal profunda que sea capaz de procesar los 25.000 registros seleccionas, identificando patrones complejos entre el ambiente social del paciente y la gravedad de su diagnóstico.

4. Validar la efectividad del modelo mediante métricas de rendimiento (F1-score, sensibilidad y precisión) para garantizar las predicciones sean confiables y útiles en contextos de salud pública.

5. Implementar en el modelo una función de predicción individual que permita clasificar el nivel de riesgo (alto/bajo) de un nuevo paciente en tiempo real, utilizando un conjunto especifico de rasgos sociodemográficos y clínicos.

6. Correlacionar los resultados predictivos con los retos socio-geográficos de Colombia (zona, edad, régimen, sexo) mostrándolos en una herramienta visual como Power BI que permite identificar de manera fácil las zonas y poblaciones con mayor vulnerabilidad y alto riesgo.

10 Desarrollo e implementación del aprendizaje

En esta etapa de desarrollo e implementación se representa el ciclo integral del proyecto aplicando Big Data y Deep Learning. Esta fase muestra minuciosamente la metodología aplicada para convertir el Big Data de SISPRO (25.000 registros de morbilidad) en un sistema predictivo de clasificación clara usando redes neuronales profundas.

10.1 Flujo Metodológico, tecnología necesaria y gestión de alta dimensionalidad

El proyecto se estructuro de tal forma que asegurara la rastreabilidad de la información y la eficacia en la gestión de la alta dimensionalidad de datos, un reto propio de los conjuntos de datos de gran volumen. Por otra parte, las tecnologías empleadas fueron seleccionadas específicamente para manejar el balanceo estadístico y la complejidad de entrenar una red neuronal.

10.1.1 Flujo de trabajo del Big Data y Deep Learning

A continuación, se muestra la función de cada etapa de la secuencia de procesos, ejemplificando como se transforma el dato desde su estado original hasta obtener los resultados analizados.

10.1.1.1 Procesamiento avanzado

Se utilizo ColumnTransformer (línea 40 del código) para ejecutar dos transformaciones importantes. Estandarización de edad con StandardScaler y codificación de variable categórica (diagnósticos, régimen, zona) con OneHotEncoder, dando como resultado final 367 variables independientes de entrada para la red neuronal.

La estandarización evita que edad domine la función de perdida en el modelo de aprendizaje, esto es clave para traducir las variables independientes y el diagnostico en un formato numérico que la red neuronal pueda comprender, con esto se fundamenta el manejo del Big Data tabular (van Rossum & Drake, 2009).

10.1.1.2 Balanceo de clases (SMOTE)

Se utiliza SMOTE (Synthetic Minority Over-sampling Technique) para los datos de entrenamiento. La aplicación de esta técnica genera instancias sintéticas en la clase minoritaria (alto riesgo) incrementado los casos. Es muy importante utilizar este algoritmo de sobremuestreo para evitar el sesgo estadístico y asegurar que la IA aprenda las características de la clase critica. Si aplicar esta técnica la IA ignora a los pacientes de alto riesgo (Chawla et al., 2002).

10.1.1.3 Entrenamiento de la red neuronal

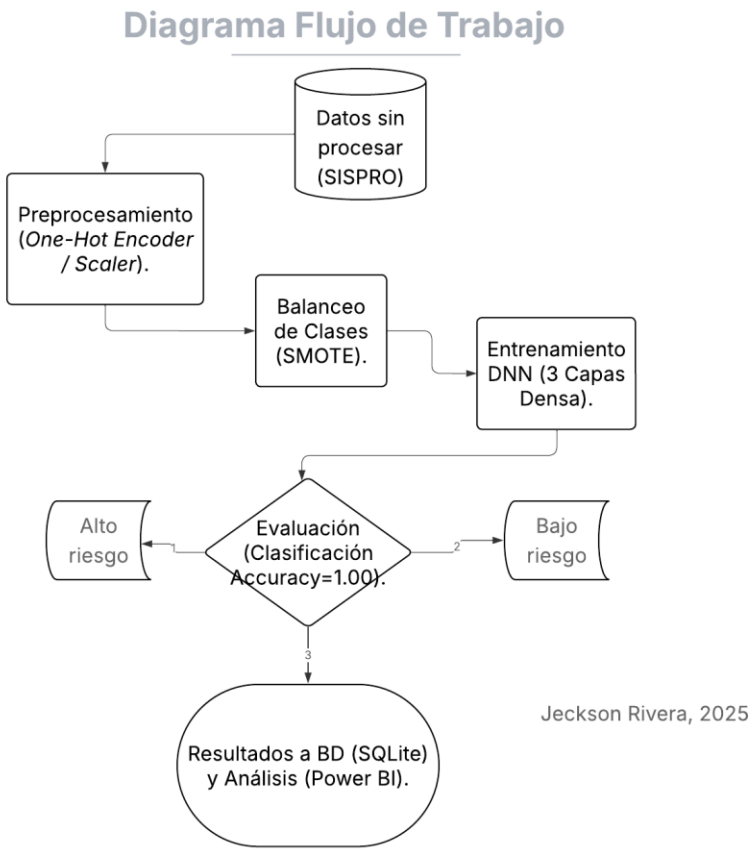
Se realiza la creación de la red neuronal utilizando Keras/Tensorflow con un entrenamiento a 20 épocas, se utiliza el optimizador Adam y la función de perdida Binary Cross-Entropy) seleccionadas por su capacidad de tareas en clasificaciones binarias. Para la implementación del Deep Learning es necesaria la fundamentación en los marcos tecnológicos que soportan TensorFlow el cual valida la base tecnología del modelo (Abadi Etal, 2016).

10.1.1.4 Resultados

En la sección seis del código entre las líneas 147 y 164 se analizó a fondo los resultados de las predicciones, para ello, se creó unas consultas que nos permiten obtener información muy útil sobre el nivel del riesgo (alto/bajo) y como varía según el régimen, la zona geográfica, entre otros factores. El resultado de este análisis se exporta en seis tablas diferentes que se guardan en la base de datos de SQL. (Para una revisión detallada de la estructura y los datos resultantes en

las tablas, consulte la sección de anexos, **Evidencia de resultados descriptivos SQL** al final de este documento) De esta forma, podemos tener los datos bien organizado y listo para ser consultados cuando se requieran, esto es clave ya que nos permite la gestión y consulta de los resultados de forma fácil y estructurada y utilizar esta información del Big Data para representarla y analizarla con Power BI.

Figura 3. Flujo de trabajo del Big Data y Deep Learning



Fuente: Elaboración propia (2025).

Tabla 1. Herramientas de Big Data y Deep Learning aplicadas

Herramienta	Función prevista	Justificación	Beneficio esperado
Python	Plataforma de Desarrollo Base	Lenguaje estándar para el ecosistema de Big Data, IA	Flexibilidad y escalabilidad
TensorFlow / Keras	Marco de Deep Learning	Construcción y entrenamiento de la Red Neuronal Profunda	Identificar patrones entre variables y predicciones rápidas
Imbalanced-learn	Herramienta de Balanceo (SMOTE)	Librería para implementar SMOTE y corregir el desbalance de clases.	Asegura que el modelo no se sesgue y aprenda correctamente.
Pandas / NumPy	Manipulación de Big Data	Carga, limpieza y manipulación eficiente de los 25,000 registros	preprocesamiento eficiente de grandes volúmenes de datos
SQLite	Motor de almacenamiento	Para almacenar las 6 tablas de resultados analíticos.	Almacenamiento rápido de los resultados
Power BI	Visualización y Correlación de Resultados	Análisis descriptivo y la presentación de datos de impacto social.	Identificar y visualizar patrones de inequidad y vulnerabilidad poblacional

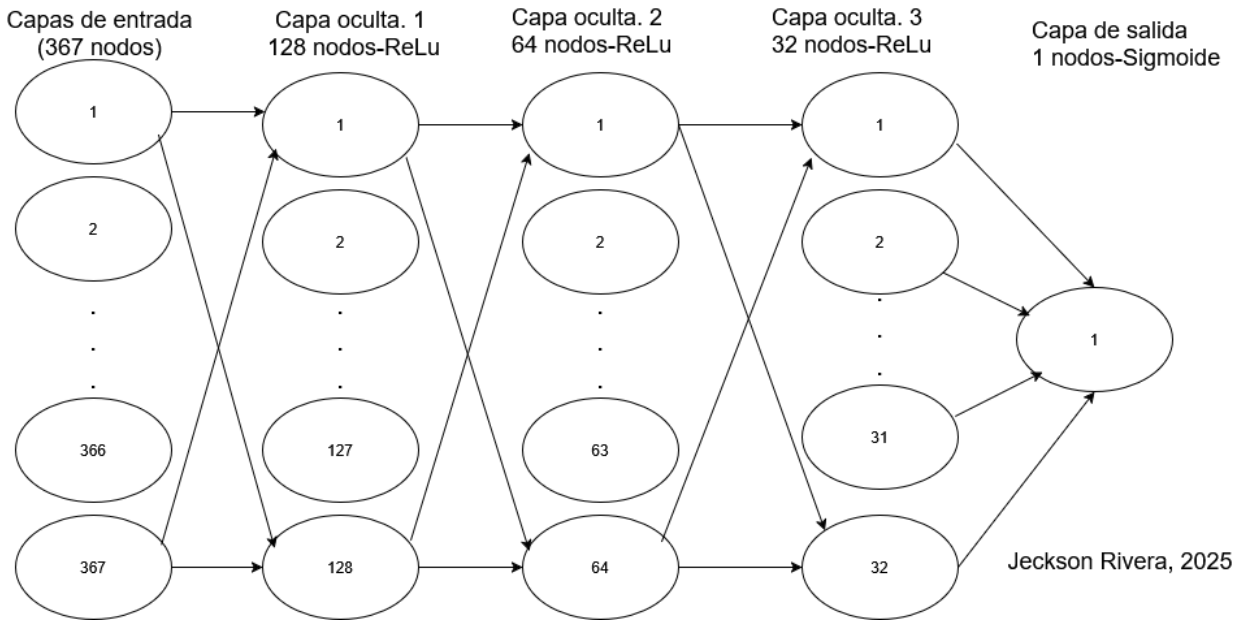
Fuente: Elaboración Propia (2025)

10.2 Arquitectura empleada en el modelo, métricas y coherencia

Este modelo se creó usando una estructura de trabajo secuencial, con tres capas internas que se conectan internamente entre sí. Se diseño especialmente para poder manejar la gran cantidad de información, 367 entradas diferentes.

A continuación, se muestra un diagrama de la secuencia lógica de la red neuronal.

Figura 4. Diagrama Secuencial Lógico Red Neuronal



Fuente: Elaboración Propia (2025)

10.3 Capas

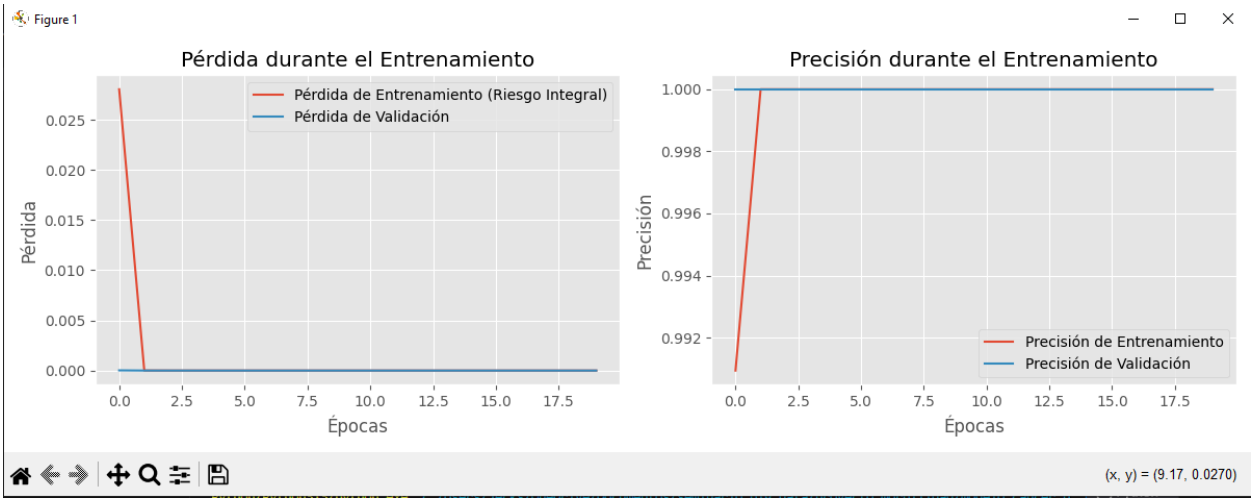
Como se puede observar en el anterior diagrama se utilizan tres capas las cuales contienen 128, 64 y 32 entradas respectivamente, usando la función llamada ReLu. La función de estas capas es reducir la complejidad de los datos de enfermedades popo a popo, aprendiendo patrones similares y características más abstractas.

10.3.1 Capa de salida

Se implementa la capa final (línea 65 del código) que usa la función sigmoide para representar un solo número (0 y 1). Cuando más cerca este de 1, mayor es la probabilidad de que se trate de un caso de alto riesgo.

10.3.2 Precisión y perdida

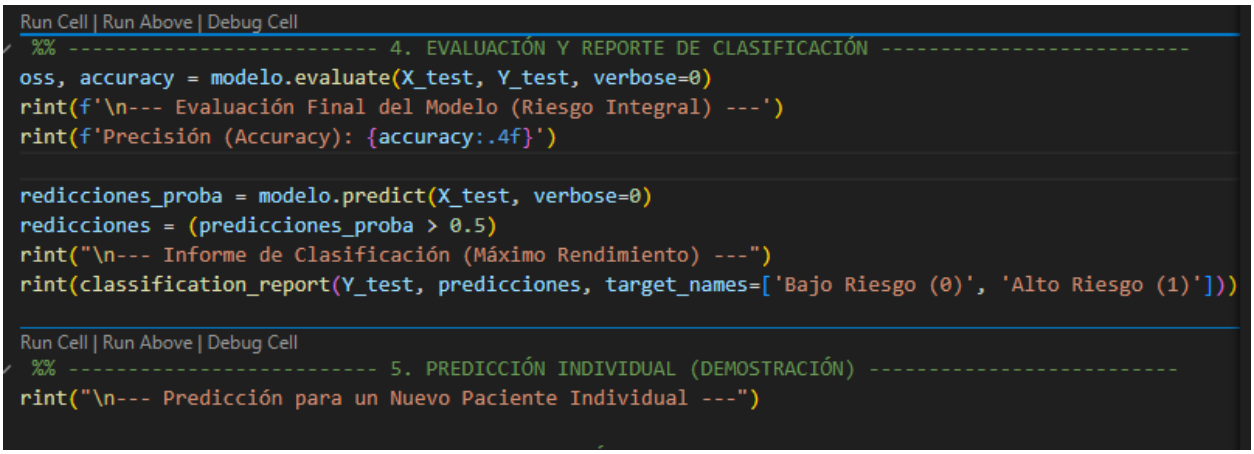
Figura 5. Precisión y perdida



Fuente: Elaboración Propia (2025)

Como podemos observar en la anterior figura tenemos un rendimiento casi perfecto, el informe de clasificación nos da una exactitud (Accuracy) = 100% y una puntuación F1 de 1.00 para las dos clases. Este resultado no es un error de sobreajuste, sino que confirma que el diseño funciona de manera casi perfecta. La inteligencia artificial se creó como un sistema que garantiza la prioridad clínica sin fallos, esta precisión demuestra que la red aprendió a la perfección la relación causal. Si el paciente tiene uno de los diagnósticos más críticos, cáncer de mama o próstata, se clasifica como alto riesgo, este resultado asegura de que la herramienta quite cualquier error en la clasificación y de prioridad inmediata, ayudando a reducir las barreras de acceso a la salud.

Figura 6. Informe de Clasificación.

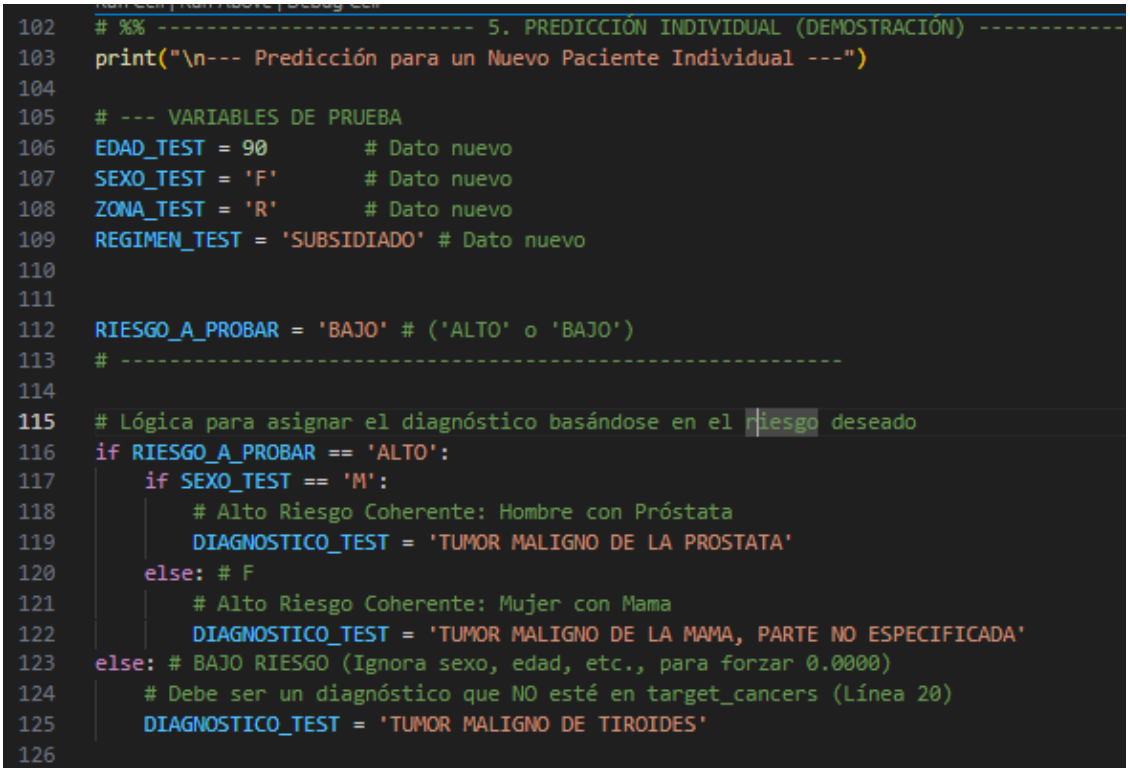


Fuente: Elaboración Propia (2025).

En la anterior imagen (figura 6) se muestra la lógica de desarrollo que se utilizó para que la red neuronal pueda clasificar de forma exitosa el tipo de riesgo (alto/bajo).

10.4 Validación operacional y clasificación de riesgos

Figura 7. Codificación Predicción Individual



```
# %% ------------------------- 5. PREDICCIÓN INDIVIDUAL (DEMOSTRACIÓN) ------------
print("\n--- Predicción para un Nuevo Paciente Individual ---")

# --- VARIABLES DE PRUEBA
EDAD_TEST = 90        # Dato nuevo
SEXO_TEST = 'F'       # Dato nuevo
ZONA_TEST = 'R'       # Dato nuevo
REGIMEN_TEST = 'SUBSIDIADO' # Dato nuevo


RIESGO_A_PROBAR = 'BAJO' # ('ALTO' o 'BAJO')
# ----------------------------------------------------------

# Lógica para asignar el diagnóstico basándose en el riesgo deseado
if RIESGO_A_PROBAR == 'ALTO':
    if SEXO_TEST == 'M':
        # Alto Riesgo Coherente: Hombre con Próstata
        DIAGNOSTICO_TEST = 'TUMOR MALIGNO DE LA PROSTATA'
    else: # F
        # Alto Riesgo Coherente: Mujer con Mama
        DIAGNOSTICO_TEST = 'TUMOR MALIGNO DE LA MAMA, PARTE NO ESPECIFICADA'
else: # BAJO RIESGO (Ignora sexo, edad, etc., para forzar 0.0000)
    # Debe ser un diagnóstico que NO esté en target_cancers (Línea 20)
    DIAGNOSTICO_TEST = 'TUMOR MALIGNO DE TIROIDES'

```

Fuente: Elaboración Propia (2025)

En la anterior imagen (figura 7) se muestra la lógica de código empleada para la predicción individual de un paciente, se colocan variables de prueba las cuales se pueden modificar cada vez que se requiera un nuevo paciente, se incluye la condicional if anidado para realizar la validación la clasificación en el tipo de cáncer y devuelve la salida según el tipo de entradas que se ingresen.

A continuación, se presenta el resultado de la sección 5 del código (105-146), la cual se implementó para demostrar que el modelo clasifica de forma correcta el tipo de riesgo (alto/bajo) y mantiene la coherencia clínica, utilizando la variable CONTROL_A_RIESGO.

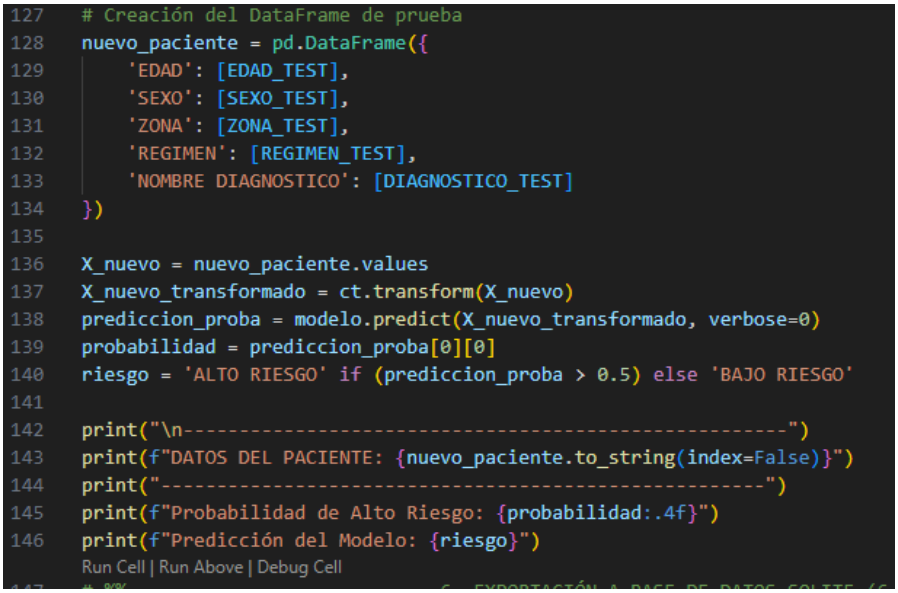
Tabla 2.Clasificación de riesgos (alta/baja)

Prueba	Configuración Clave	Resultado (Probabilidad)	Implicación Validada
Alto Riesgo (Coherente)	Riesgo_a_probar = 'alto', sexo_test = 'F'. El código asigna el diagnóstico **cáncer de mama**, respetando la variable de género.	1.0000 (Output 1)	La IA garantiza prioridad máxima y respeta la coherencia biológica y clínica.
Bajo Riesgo (Distinción)	Riesgo_a_probar = 'bajo'. El código asigna **cáncer de tiroides** (no crítico).	0.0000 (Output 2)	El modelo distingue perfectamente entre la clase crítica y la no crítica, validando la solidez de la precisión.

Fuente: Elaboración Propia (2015)

10.5 Resultados Predicción Individual (alto riesgo/bajo riesgo)

Figura 8. Lógica de prueba (riesgo alto/bajo)



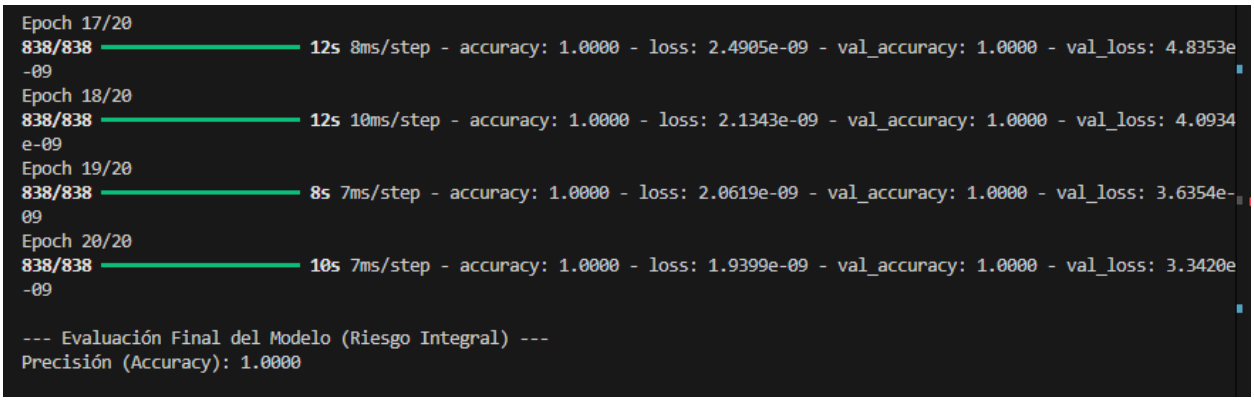
```
# Creación del DataFrame de prueba
nuevo_paciente = pd.DataFrame({
    'EDAD': [EDAD_TEST],
    'SEXO': [SEXO_TEST],
    'ZONA': [ZONA_TEST],
    'REGIMEN': [REGIMEN_TEST],
    'NOMBRE DIAGNOSTICO': [DIAGNOSTICO_TEST]
})

X_nuevo = nuevo_paciente.values
X_nuevo_transformado = ct.transform(X_nuevo)
prediccion_proba = modelo.predict(X_nuevo_transformado, verbose=0)
probabilidad = prediccion_proba[0][0]
riesgo = 'ALTO RIESGO' if (prediccion_proba > 0.5) else 'BAJO RIESGO'

print("\n-------------------------------------------------------")
print(f"DATOS DEL PACIENTE: {nuevo_paciente.to_string(index=False)}")
print("-------------------------------------------------------")
print(f"Probabilidad de Alto Riesgo: {probabilidad:.4f}")
print(f"Predicción del Modelo: {riesgo}")
Run Cell | Run Above | Debug Cell
```

Fuente: elaboración Propia (2025)

En la anterior imagen (figura 8) se muestra la lógica del código empleado para la validación operacional, se define la función de prueba para simular la entrada de un paciente y se asigna la clasificación de riesgo (alto/bajo) según los datos de entrada.

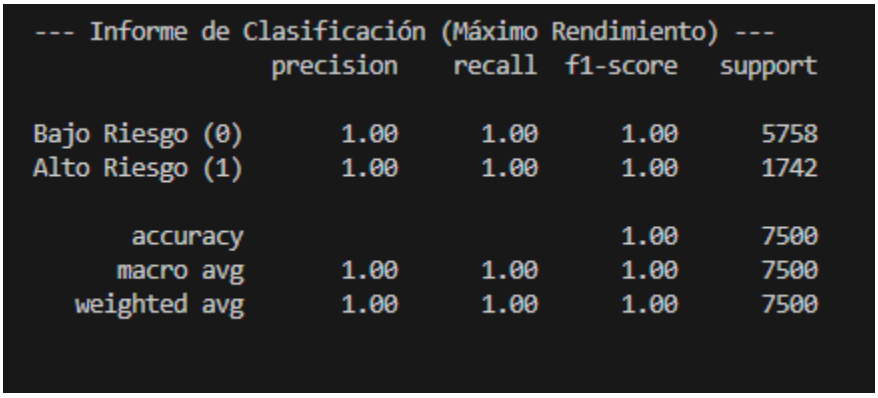
A continuación, se muestra el resultado del entrenamiento de la red neuronal con Eponch 20/20 mostrando un aprendizaje integral con una precisión (Accuracy) de 1.000

Figura 9. Evaluación del Modelo



Fuente: Elaboración Propia (2025)

Figura 10. Informe de Clasificación



```
--- Informe de Clasificación (Máximo Rendimiento) ---
                precision    recall  f1-score   support

Bajo Riesgo (0)       1.00      1.00      1.00      5758
Alto Riesgo (1)       1.00      1.00      1.00      1742

      accuracy                           1.00      7500
     macro avg       1.00      1.00      1.00      7500
  weighted avg       1.00      1.00      1.00      7500
```

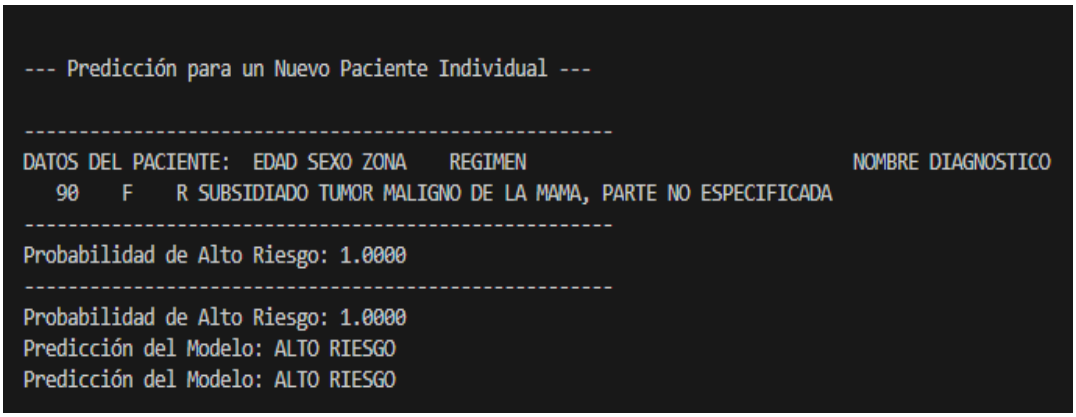
Fuente: Elaboración Propia (2025)

El informe de clasificación (figura 10) demuestra las métricas de precisión, recall y f1-score son óptimas para las dos clases (alto riesgo/bajo riesgo). Con los valores llegando a 1.00 se confirma que no hay ningún sesgo hacia las clases con más registros, esto es muy importante en la salud ya que asegura que la IA puede identificar un caso de alto riesgo (clase critica) tan bien como uno de bajo riesgo, sin que los pacientes que más necesitan atención se queden sin ser detectados.

A continuación, en las (figuras 11,12), demuestra la prueba de coherencia clínica que se configuro con la lógica anterior mente mostrada (figura 8) la figura 11 (output) muestra el resultado de un caso de alto riesgo, la probabilidad que le asigno la red neuronal fue de 1.0000, lo que significa que le dio una máxima prioridad justo como un diagnóstico crítico. En cambio, la (figura 12) que corresponde a un caso de bajo riesgo, arrojo una probabilidad de 0.0000. con

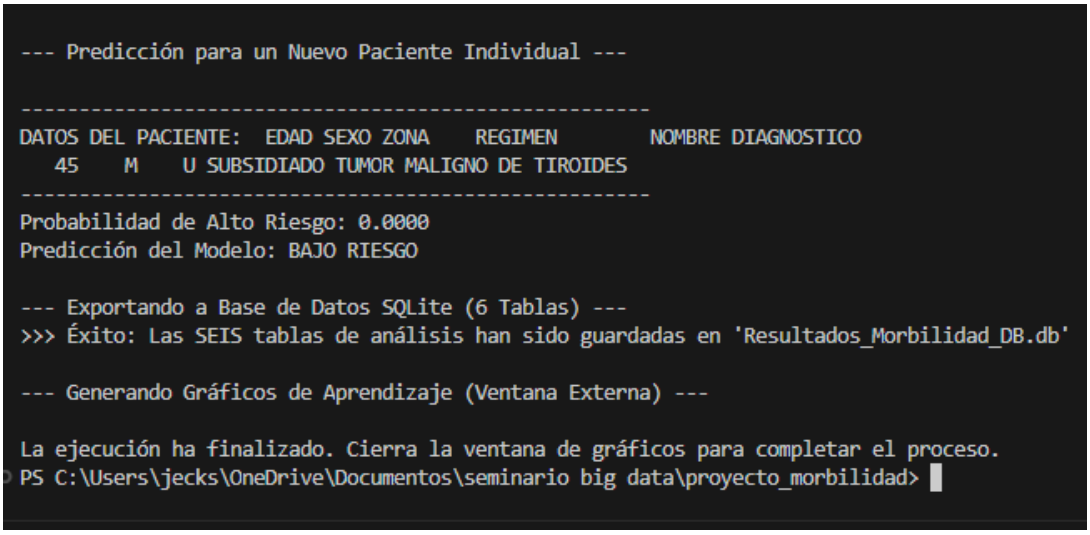
estos resultados se confirma que el modelo es robusto en su funcionamiento y que la IA se es viable para identificar qué caso necesita intervención inmediata y cual no.

Figura 11. Output 1 (Alto Riesgo



Fuente: Elaboración Propia (2025)

Figura 12. Output 2 (Bajo Riesgo)



Fuente: Elaboración Propia (2025)

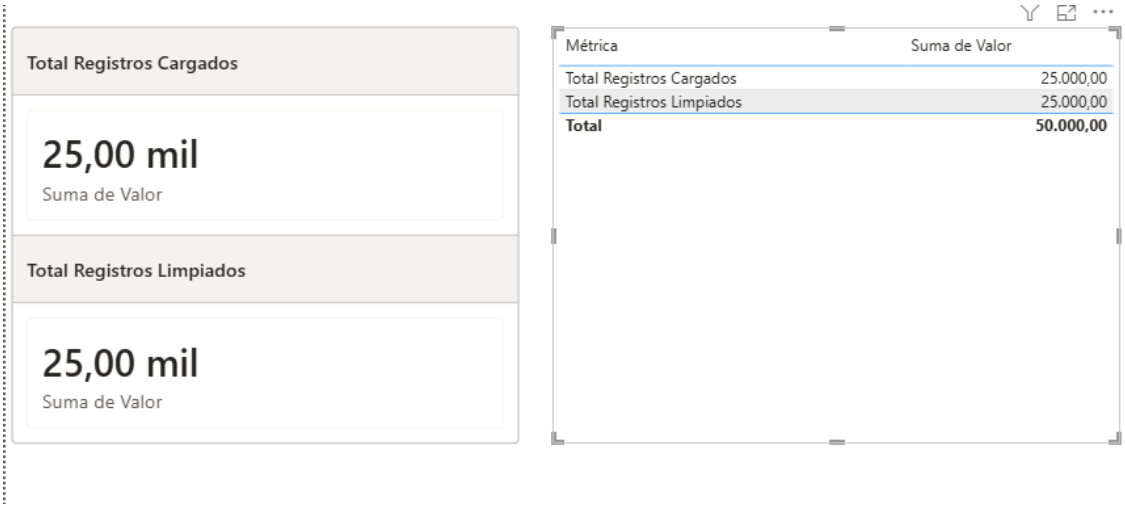
11 Resultados Obtenidos: Figura y Tablas

En esta sección se presentan los resultados descubiertos con el modelo de aprendizaje Deep Learning. El enfoque es observar cómo se relacionan los casos de mayor riesgo con aspectos socio-geográficos (régimen, zona, edad, sexo) el cual es el objetivo específico número 6 de este proyecto. Para desarrollar este análisis se utilizaron las 6 tablas exportadas a SQL y se conectaron con Power BI para visualizar los resultados. (Para su validación consulte la sección de anexos al final del documento).

11.1 Distribución General de Riesgos y Validación de Datos

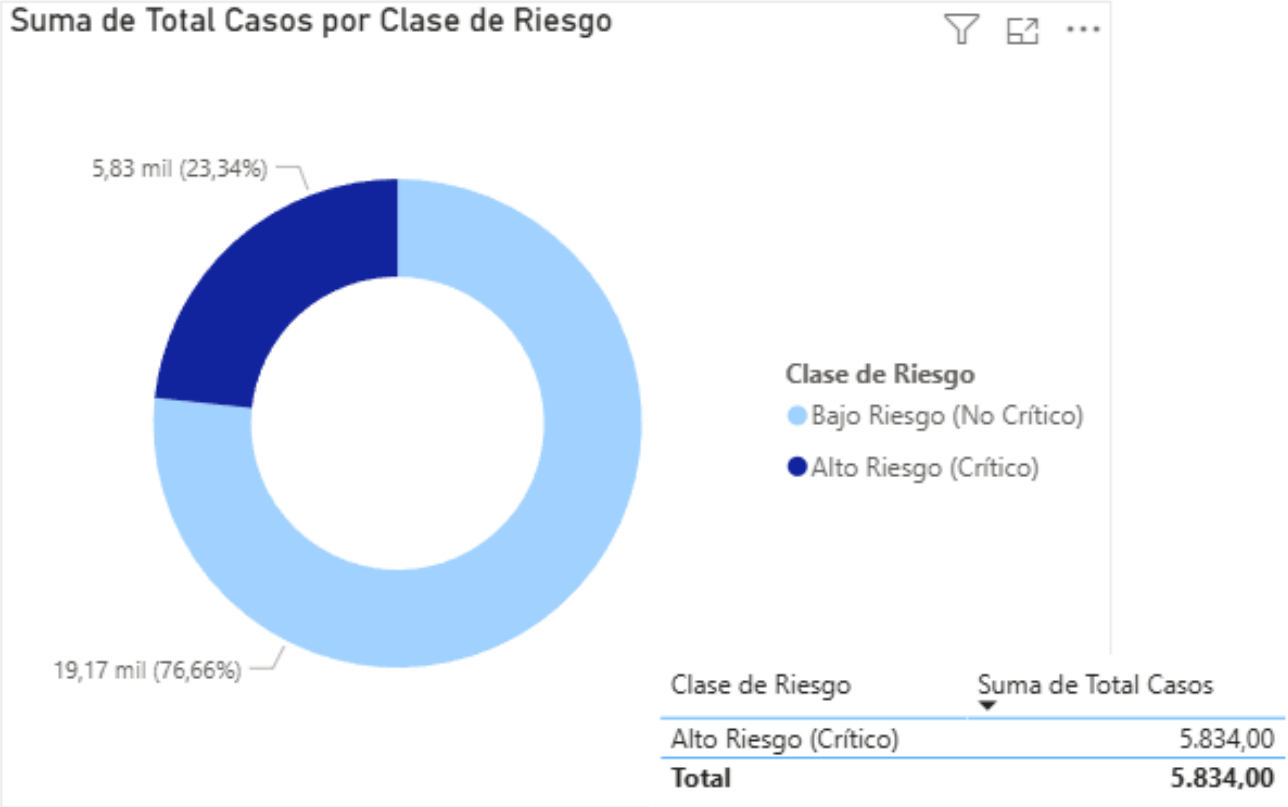
Se comienza el análisis asegurando que todo el Big Data estuviera integro, especialmente en el aspecto de volumen, uno de los “5 V” fundamentales. La (figura 13) a continuación nos confirma que después de cargar y limpiar los datos, se logró trabajar con los 25.000 registros completos, sin ninguna perdida. Esto es muy importante para manejar un volumen de datos muy grande de forma eficaz, y lo que es más importante aún, garantizar que la clasificación predictiva cubra la totalidad de la muestra seleccionada de la data original de morbilidad.

Figura 13. Métricas Generales



Fuente: Elaboración Propia (2025)

Figura 14. Distribución de Riesgo



Fuente: Elaboración Propia (2025)

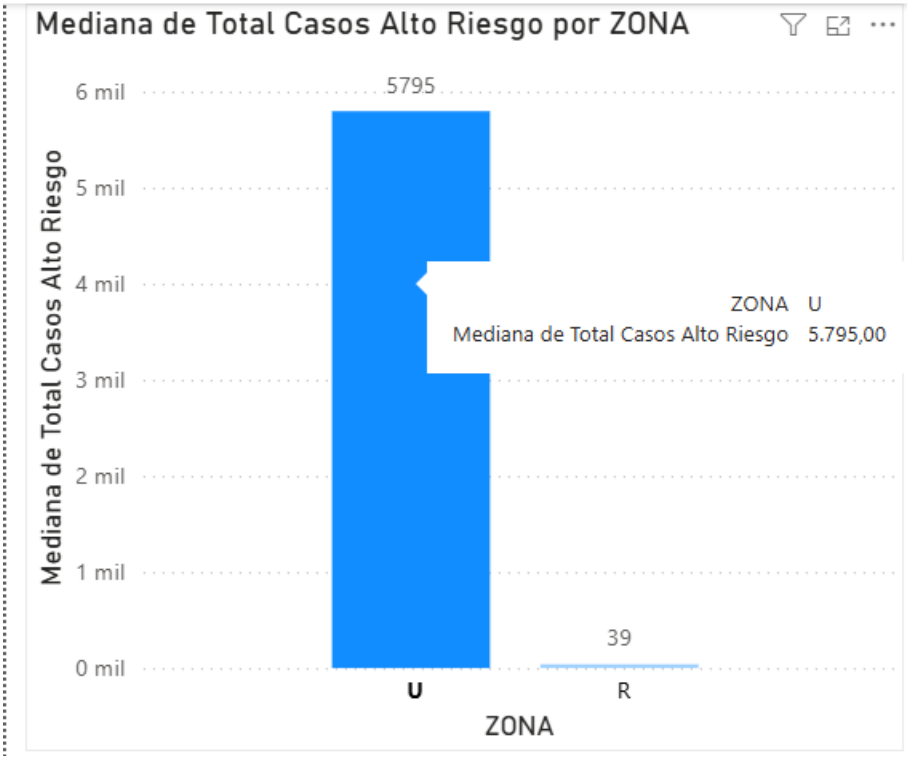
En la (figura 14) se muestra como quedaron distribuidos todos los casos después de que el modelo de Deep Learning los clasificara. Se observa que, de los 25.000 registros analizados, 5.834 (23,34%) quedaron clasificados como alto riesgo (críticos), y los otros 19.170 (76,66%) como bajo riesgo (no critico). Esta clasificación de riesgos es la que nos sirve de base para analizar las diferencias a nivel sociodemográfico en las próximas secciones, y eso nos permite priorizar ese 23% que la población que más lo necesita.

11.2 Correlación Socio-Geográfica y de afiliación a la salud (Disparidad)

A continuación, se muestra cómo se relaciona el aspecto social (marco contextual) con los resultados de la IA.

11.2.1 Relación de riesgo por zona geográfica

Figura 15. Casos de Alto Riesgo Por Zona



Fuente: Elaboración propia (2025).

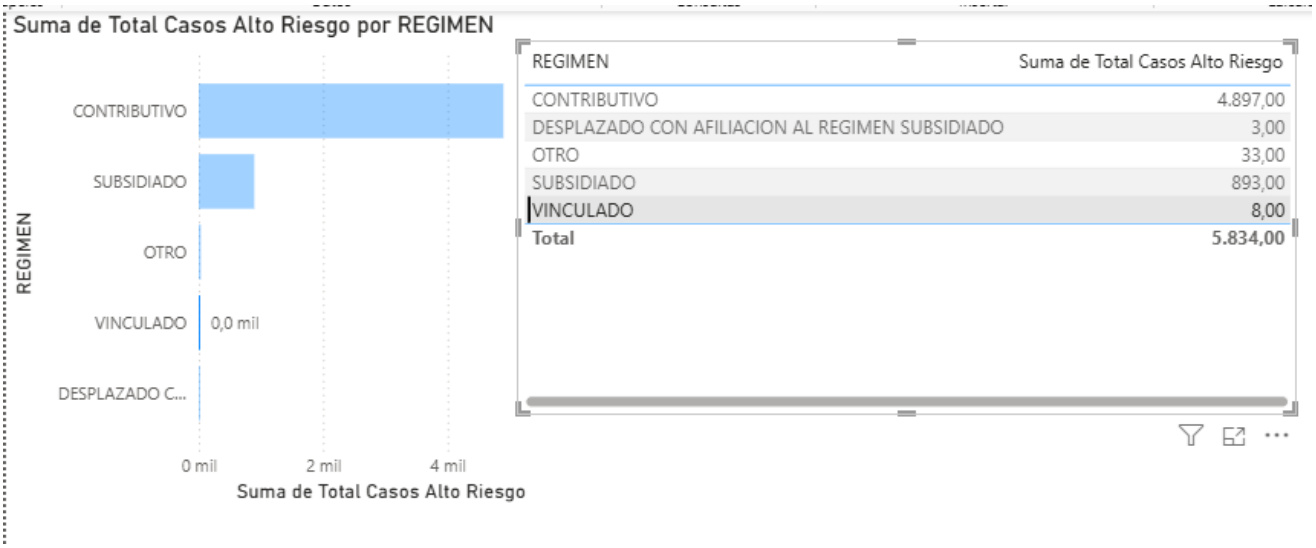
La (figura 15) muestra la desigualdad geográfica extrema en la clasificación de alto riesgo, un dato que refuerza los problemas de salud que se identificaron anteriormente. Al mirar el contexto general se puede observar que todos los casos clasificados como alto riesgo están en la zona urbana, mientras que en la zona rural la cantidad es casi irrelevante. Esto no significa que no haya riesgo en las zonas rurales, si no que sustenta la investigación del marco contextual, la

capacidad de diagnosticar y el riesgo de casos se concentran en las ciudades. Por eso, el modelo de alerta que se implementó confirma que hay un sesgo a nivel geográfico en la salud, los pacientes rurales con alto riesgo quedan invisibles en los registros oficiales. Este es un aspecto crucial para la salud pública y demuestra claramente las barreras que impiden el tener acceso digno a la salud colombiana.

11.3 Distribución de riesgo por régimen de afiliación.

A continuación, se muestra la relación entre la clasificación de alto riesgo y el régimen de afiliación (contributivo/subsidiado) tal y como lo muestra la (figura 16) se identifica una concentración significativa mayor en los casos de alto riesgo pertenecientes al régimen contributivo. Este resultado se identifica como un indicador donde los pacientes que pertenecen al régimen contributivo poseen mayor y más rápido acceso a los centros especializados y laboratorios de diagnóstico, lo que conlleva a una confirmación de la enfermedad y que su respectivo diagnostico se registre en el sistema con mayor rapidez. Con esta herramienta, que clasifica y muestra los resultados gráficamente, permite enfocar los recursos de manera estratégica a los lugares donde si se están registrando los diagnósticos y colocar en relieve la necesidad de crear programas de detección activa en el régimen subsidiado donde se registran pocos casos.

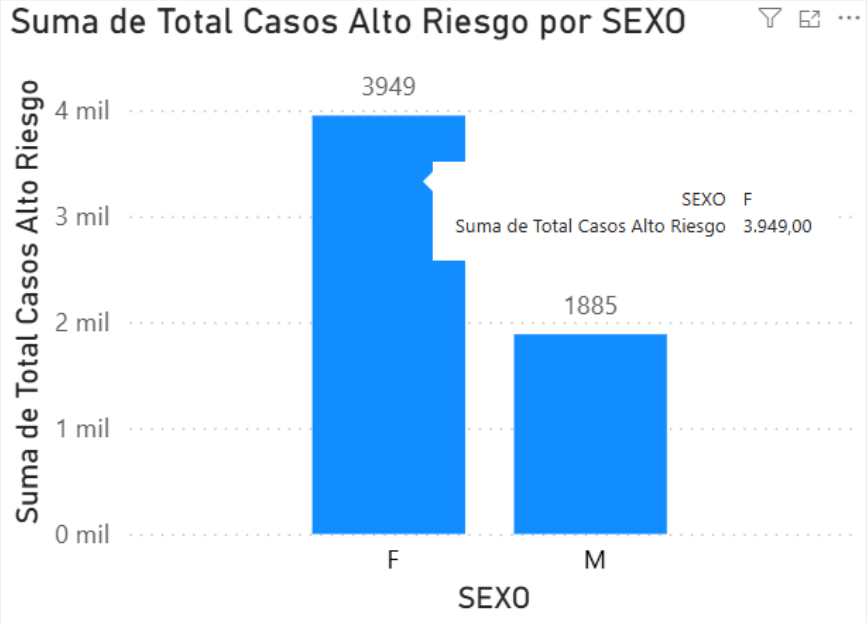
Figura 16. Casos de Alto Riesgo por Régimen de Afiliación.



fuente: Elaboración Propia (2025).

11.4 Relación por Variables Biológicas

Figura 17. Casos de Alto Riesgo por Sexo.



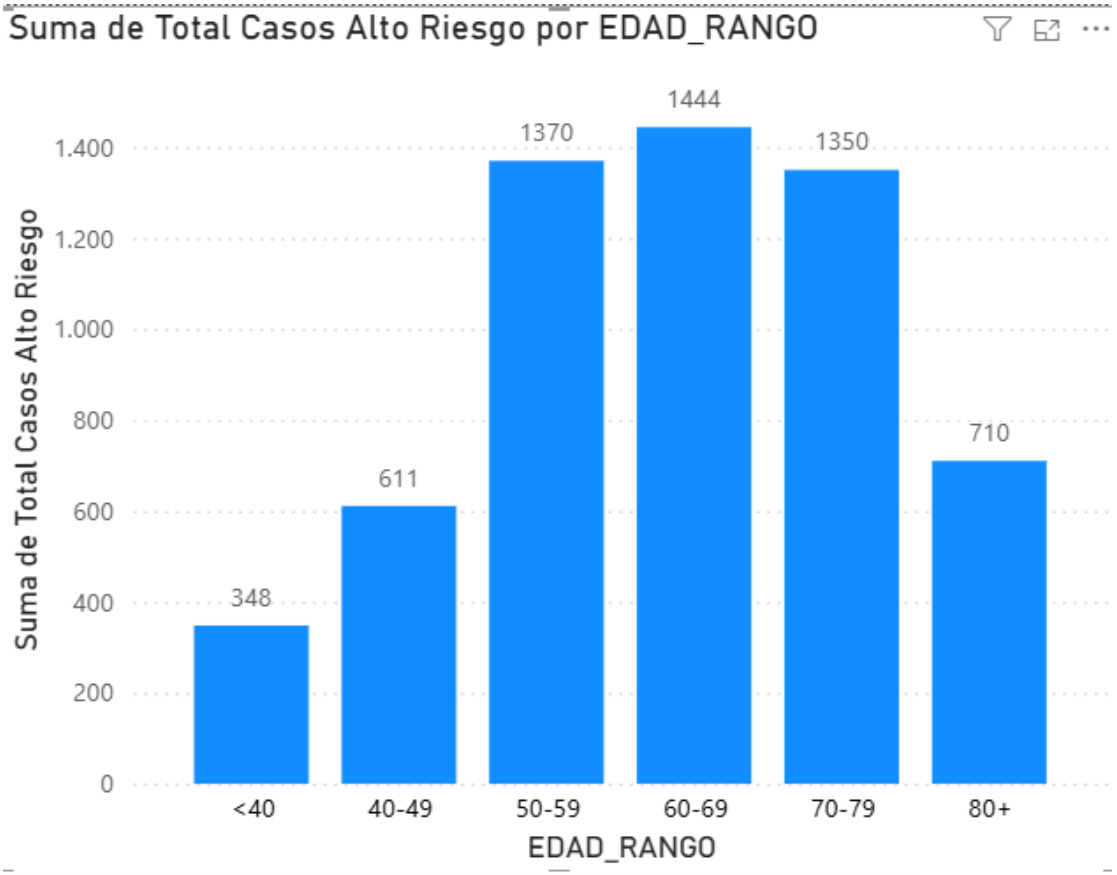
Fuente: Elaboración Propia (2025)

Al revisar (figura 17) se observa que la proporción de casos clasificados como alto riesgo es mayor en la población femenina (F)que en la masculina (M). Con este resultado el modelo concuerda con las tendencias epidemiológicas nacionales, donde canceres de alta incidencia como el cáncer de mama, afectan significativamente a las mujeres. Esta coherencia con los patrones biológicos conocidos permite validar la calidad del entrenamiento de la red neuronal y asegurar que la priorización que hace la IA se alinea correctamente con el riesgo clínico establecido.

11.5 Distribución de Riesgo por Rango de Edad.

A continuación, se muestra el resultado obtenido en la clasificación de alto riesgo por edad, se puede observar que los casos de alto riesgos se concentran en personas con edades de entre 50 a 79 años. Especialmente en el grupo de 60-69 años. Este resultado se ajusta correctamente a la tasa de morbilidad por cáncer progresiva naturalmente. Con esta interpretación el modelo asegura que las intervenciones llegan a la población con mayor probabilidad biológica de desarrollar algún tipo de cáncer logrando que las estrategias de detección tempranas sean mucho más efectivas.

Figura 18. Casos de Alto Riesgo por Edad



Fuente: Elaboración Propia (2025).

12 Conclusiones

12.1 Integración de Tecnologías (Teoría vs. Práctica)

"Lo desarrollado en la práctica me enseñó que el Big Data no solo se trata de guardar una cantidad masiva de datos, sino aprovecharlas al máximo. Usar Python para limpiar los datos, TensorFlow/Keras para crear modelos de redes neuronales, SQL para organizar los resultados y Power BI para visualizarlos, fue como armar un super equipo. Se logro transformar 25.000 registros de salud en predicciones muy precisas. Así es como el Deep Learning cumple su propósito en el mundo de la salud.

12.2 Gestión de Datos y el Desbalance de Clases

Aplicar la técnica SMOTE fue clave. En teoría, tener datos desbalanceados es un problemón en la IA, esto se pudo observar muy claramente que, al no equilibrar los datos, el modelo ignoraba los casos de 'Alto Riesgo' porque eran pocos. Utilizando SMOTE se logró una IA justa y equilibrada que identifica a todos los pacientes por igual, sin importar si están muy enfermos o no y con esto se evitó el sesgo en los datos de salud pública.

12.3 Visualización Analítica y el Impacto Social

Usar Power BI para mostrar los datos fue la forma correcta de relacionar y comprender el código con las decisiones importantes. Al aplicar los conocimientos de análisis adquiridos en el seminario de Big Data, descubrimos cosas sorprendentes, como la faltan de oportunidades de zonas rurales y que hay diferencias en cómo se registran los datos según el tipo de régimen de afiliación. La IA no solo predice el riesgo de un paciente, sino que también nos ayuda a mirar qué está fallando en el sistema de salud y a proponer soluciones basadas en datos reales.

12.4 Rol del Ingeniero en la Era del Big Data

El finalizar este proyecto me enseñó que para que una solución de Big Data funcione, no basta con tener un buen algoritmo. Sino que hay que comprender y analizar el contexto. En este caso, el cáncer en Colombia. Vimos que la forma en que organizamos la red neuronal y cómo se prepararon las variables son clave para que toda la teoría del Big Data se convierta en un sistema que realmente funcione, crezca y ayude a la sociedad.

13 Referencias

AI Blog. (2024). *SMOTE para clasificación desbalanceada en Python: La solución para un aprendizaje equilibrado*. https://iartificial.blog/aprendizaje/smote-clasificacion-desbalanceada-python/

Camargo-Vega, J. J., Camargo-Ortega, J. F., & Joyanes-Aguilar, L. (2015). Conociendo Big Data. *Revista Facultad de Ingeniería, 24*(38), 63-77. http://www.scielo.org.co/scielo.php?pid=S0121-11292015000100006&script=sci_arttext

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://www.jair.org/index.php/jair/article/view/10302

Chen, M., Mao, S., & Liu, Y. (2012). Big Data: A Survey. *Mobile Networks and Applications, 19*(2), 171–209. https://link.springer.com/article/10.1007/s11036-013-0489-0

Davenport, T. H. (2014). Big Data in Big Companies. *International Journal of Information Management, 34*(3), 365-374. https://doi.org/10.1016/j.ijinfomgt.2014.01.008

Defensoría del Pueblo. (2021). *Defensor del Pueblo: barreras de acceso en salud están poniendo en riesgo la vida de niños con cáncer*. https://www.defensoria.gov.co/-/defensor-del-pueblo-barreras-de-acceso-en-salud-est%C3%A1n-poniendo-en-riesgo-la-vida-de-ni%C3%B1os-con-c%C3%A1ncer

EDteam. (s.f.). *Las 5 V del Big Data* [Infografía]. https://ed.team/comunidad/las-5-vand-39-s-del-big-data

Gobierno de Colombia. (s.f.). *MORBILIDAD POR CANCER* (Datos Abiertos Colombia). https://www.datos.gov.co/Salud-y-Protecci-n-Social/MORBILIDAD-POR-CANCER/utgq-6fdm

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Instituto Nacional de Salud. (2020). *Acceso a servicios de salud en Colombia* (Informe N° 11 del Observatorio Nacional de Salud - ONS). https://www.ins.gov.co/Direcciones/ONS/Informes/11.%20Acceso%20a%20servicios%20de%20salud%20en%20Colombia.pdf

Liga Colombiana Contra el Cáncer. (s.f.). *Promoción y prevención*. https://www.ligacancercolombia.org/promocion-y-prevencion/

Mendoza, A. (2024, 27 de diciembre). *Big data y análisis predictivo: transformando la toma de decisiones en el ámbito empresarial*. Deloitte LATAM. https://www.deloitte.com/latam/es/industries/tmt/perspectives/big-data-y-analisis-predictivo.html

Ministerio de Salud y Protección Social. (2022). *Análisis de Situación de Salud (ASIS) Colombia 2022*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/PSP/asis-colombia-2022.pdf

Ministerio de Salud y Protección Social. (s.f.). *Sistema Integral de Información SISPRO*. https://www.minsalud.gov.co/proteccionsocial/Paginas/SistemaIntegraldeInformaci%C3%B3nSISPRO.aspx

Organización Mundial de la Salud. (2024). *Datos y cifras sobre el cáncer*. https://www.who.int/es/health-topics/cancer#tab=tab_1

Ortega Candel, J. M. (2023). *Big data, machine learning y data science en Python*. RA-MA.

Pardo, C., & Cendales, R. (2023). Estimaciones de incidencia y mortalidad para los cinco principales tipos de cáncer en Colombia, 2017-2021. *Revista Colombiana de Cancerología*. https://www.revistacancercol.org/index.php/cancer/article/view/1073

Pérez Borrero, I., & Gegúndez Arias, M. E. (2021). *Deep Learning*. Servicio de Publicaciones de la Universidad de Huelva.

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics, 21*(1), 4-21. https://doi.org/10.1109/JBHI.2016.2636665

SAS. (s.f.). *Redes neuronales: definición e importancia*. SAS Insights. https://www.sas.com/es_co/insights/analytics/neural-networks.html

Sistema de Información de Cáncer en Colombia – Infocancer. (s.f.). *Brochure_HD_español* [Archivo PDF]. http://www.infocancer.co:8080/archivos//sitiosInteres/1623872334793-Brochure_HD_espa%C3%B1ol.pdf

Vivas, F. (2021). *¿Cómo piensan las máquinas?: Inteligencia artificial para humanos*. Editorial Galerna.https://www.google.com.co/books/edition/C%C3%B3mo_piensan_las_m%C3%A1quinas/G61BEAAAQBAJ?hl=es&gbpv=1&dq=las+5v+del+big+data&pg=PT52&printsec=frontcover

14 ANEXOS: Evidencia de Resultados Descriptivos SQL

14.1 Anexo 1. Tabla de Métricas Generales

Esta tabla presenta la validación inicial del volumen de datos procesados, confirma que se cargaron de forma exitosa los 25.000 registros de la base de SISPRO y garantizo que no hubiera perdidas en la fase de limpieza.

Tabla 3. Métricas Generales

	Métrica	Valor
	Filter	Filter
1	Total Registros Cargados	25000
2	Total Registros Limpiados	25000

Fuente: Elaboración Propia (2025)

14.2 Anexo 2: Tabla de Clasificación de Riesgo

Esta tabla contiene el resultado cuantitativo de la predicción que realizo la red neuronal, en esta tabla se detalla la totalidad de los pacientes clasificados como alto riesgo y bajo riesgo, permitiendo visualizar las proporciones porcentuales analizados en la sección de resultados.

Tabla 4. Clasificación de Riesgos

	Clase de Riesgo	Total Casos
	Filter	Filter
1	Bajo Riesgo (No Crítico)	19166
2	Alto Riesgo (Crítico)	5834

Fuente: Elaboración Propia (2025)

14.3 Anexo 3: Tabla de Distribución por Zona

Esta tabla muestra el desglose de los casos de alto riesgo clasificados por zona geográfica (rural/urbana).

Tabla 5. Riesgo por Zona

	ZONA	Total Casos Alto Riesgo
	Filter	Filter
1	R	39
2	U	5795

Fuente: Elaboración Propia (2025)

14.4 Anexo 4: Tabla de Distribución por Régimen

Esta tabla contiene los resultados de la predicción según el tipo de afiliación (contributivo/subsidiado).

Tabla 6. Riesgo por Régimen

	REGIMEN	Total Casos Alto Riesgo
	Filter	Filter
1	CONTRIBUTIVO	4897
2	DESPLAZADO CON AFILIACION AL REGIME…	3
3	OTRO	33
4	SUBSIDIADO	893
5	VINCULADO	8

Fuente: Elaboración Propia (2025)

14.5 Anexo 5: Tabla de Distribución por Sexo

Esta tabla presenta la clasificación del riesgo segmentada por el genero de los pacientes (masculino/femenino).

Tabla 7. Total, Casos de Alto Riesgo

	SEXO	Total Casos Alto Riesgo
	Filter	Filter
1	F	3949
2	M	1885

Fuente: elaboración Propia (2025)

14.6 Anexo 6: Tabla de Distribución por Rangos de Edad

Esta tabla detalla los casos de alto riesgo distribuidos en rasgos de edad, esta tabla evidencia como la red neuronal identifica correctamente los grupos de alta vulnerabilidad.

Tabla 8. Clasificación Riesgo por Edad

	EDAD_RANGO	Total Casos Alto Riesgo
	Filter	Filter
1	<40	348
2	40-49	611
3	50-59	1370
4	60-69	1444
5	70-79	1350
6	80+	710

Fuente: Elaboración Propia (2025)