



**Evaluación de representaciones de datos tabular y volumétrica en el modelamiento
predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello**

Jesus Santiago Amado Montaña

Yodid Jair Cárdenas Rodríguez

Jose Fernando Motta Méndez

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

21 de marzo de 2026

Evaluación de representaciones de datos tabular y volumétrica en el
modelamiento predictivo aplicado a la planificación radioterapéutica del
cáncer de cabeza y cuello

2

**Evaluación de representaciones de datos tabular y volumétrica en el modelamiento
predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello**

Jesus Santiago Amado Montaña

Yodid Jair Cárdenas Rodríguez

Jose Fernando Motta Méndez

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ciencia de Datos

Director (a):

Maira Alejandra García Jaramillo

Modalidad:

Monografía

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

21 de marzo de 2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 21/03/2026

Dedicatorias

A mi familia por creer en mí y apoyarme
en todo momento y en especial a mi
padrino por sus enseñanzas y nobleza

Yodid Cárdenas

Vive como si fueras a morir mañana,
aprende como si fueras a vivir para
siempre.

Santiago Amado

A mis padres,
por su amor incondicional, su paciencia
infinita y por ser el pilar que ha sostenido
cada uno de mis pasos. Gracias por creer
en mí incluso cuando yo dudaba, por
enseñarme el valor del esfuerzo, la
disciplina y la perseverancia. Todo lo que
soy hoy es, en gran parte, reflejo de
ustedes.

A todas las personas que han formado
parte de mi camino,
profesores, mentores, amigos y colegas,

gracias por su apoyo, sus enseñanzas y
por cada oportunidad brindada. Cada
consejo, cada reto y cada palabra de
aliento han sido fundamentales en mi
crecimiento personal y profesional.

Jose Mota

Agradecimientos

Un agradecimiento especial a la profesora Dra. **Maira Alejandra García Jaramillo** por su invaluable guía académica. Su mentoría fue clave para interpretar los hallazgos de esta investigación y convertirlos en propuestas de mejora con impacto social y profesional. Gracias por su dedicación y por creer en este proyecto.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	6
--	---

Resumen

La planificación de la radioterapia de intensidad modulada (IMRT) en cáncer de cabeza y cuello enfrenta retos complejos derivados de la alta densidad de estructuras anatómicas críticas y la variabilidad en los planes de tratamiento generados manualmente por los especialistas, lo que motiva el desarrollo de modelos de aprendizaje automático capaces de apoyar y estandarizar este proceso. Este trabajo presenta una evaluación metodológica comparativa desde la perspectiva de la Ciencia de Datos, caracterizando dos paradigmas de representación: la representación tabular estructurada, basada en variables clínicas y demográficas, y la representación volumétrica tridimensional, basada en imágenes de tomografía computarizada (CT) y máscaras de estructuras anatómicas segmentadas.

Para ejecutar el enfoque tabular se requirió un dataset con variables clínicas estructuradas como volumen objetivo de planificación (PTV, por sus siglas en inglés), dosis prescrita, número de fracciones, estadio tumoral y datos demográficos del paciente. Para el enfoque volumétrico se requirieron imágenes CT tridimensionales, máscaras binarias del PTV y los órganos en riesgo, y distribuciones de dosis de referencia en formato espacial. Estos requerimientos condicionaron directamente la viabilidad de cada paradigma en distintos entornos clínicos.

Para cada paradigma se ejecutó un pipeline completo de aprendizaje automático. En el enfoque tabular, utilizando el dataset HNSCC-3DCT-RT, se entrenaron modelos de regresión lineal y Random Forest para predecir el volumen del PTV, obteniendo $R^2 = 0.35$ y MAE = 44.8 cc con Random Forest, frente a un modelo base de MAE = 56.09 cc. En el

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	7
--	---

enfoque volumétrico, utilizando el dataset Open Knowledge-Based Planning (OpenKBP), se implementó una red neuronal convolucional tridimensional de arquitectura UNet3D para predecir distribuciones espaciales de dosis, alcanzando un Dose Score de 16.5631 Gy sobre 100 pacientes de prueba, valor que se encuentra por encima del reportado en el benchmark OpenKBP (2.43 Gy), utilizado aquí únicamente como referencia contextual dentro del paradigma volumétrico. Dado que cada paradigma define una variable objetivo de naturaleza distinta, la comparación entre enfoques no se realiza sobre las métricas de desempeño sino sobre las dimensiones metodológicas del pipeline: proceso de preparación de datos, requerimientos computacionales e interpretabilidad del modelo.

Los resultados evidencian que la representación tabular ofrece mayor interpretabilidad y menor barrera computacional, requiriendo feature engineering manual, mientras que la representación volumétrica preserva la geometría espacial del paciente a costa de mayor complejidad infraestructural, delegando la extracción de características al modelo. La investigación aporta evidencia empírica sobre cómo la estructura del dato condiciona las decisiones de modelamiento, la viabilidad de implementación y el tipo de predicción alcanzable, con implicaciones metodológicas para la Ciencia de Datos aplicada a contextos de salud.

Palabras clave: Radioterapia IMRT, Aprendizaje Automático, Cáncer de Cabeza y Cuello, Representación de Datos, Deep Learning, UNet3D, Feature Engineering

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	7
--	---

Abstract

The planning of intensity-modulated radiotherapy (IMRT) for head and neck cancer faces complex challenges stemming from the high density of critical anatomical structures and the variability in treatment plans generated manually by specialists, which motivates the development of machine learning models capable of supporting and standardizing this process. This work presents a comparative methodological evaluation from a Data Science perspective, characterizing two data representation paradigms: structured tabular representation, based on clinical and demographic variables, and three-dimensional volumetric representation, based on computed tomography (CT) images and segmented anatomical structure masks.

To execute the tabular approach, a dataset containing structured clinical variables was required, including planning target volume (PTV), prescribed dose, number of fractions, tumor staging, and patient demographic data. For the volumetric approach, three-dimensional CT images, binary masks of the PTV and organs at risk, and reference dose distributions in spatial format were necessary. These requirements directly conditioned the feasibility of each paradigm across different clinical settings.

A complete machine learning pipeline was executed for each paradigm. In the tabular approach, using the HNSCC-3DCT-RT dataset, linear regression and Random Forest models were trained to predict PTV volume, achieving $R^2 = 0.35$ and MAE = 44.8 cc with Random Forest, compared to a baseline MAE of 56.09 cc. In the volumetric approach, using the Open Knowledge-Based Planning (OpenKBP) dataset, a three-dimensional

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	8
--	---

convolutional neural network with UNet3D architecture was implemented to predict spatial dose distributions, achieving a Dose Score of 16.5631 Gy over 100 test patients, exceeding the 2.43 Gy reported in the OpenKBP benchmark, which is included here solely as a contextual reference within the volumetric paradigm. Since each paradigm defines a target variable of a fundamentally different nature, the comparison between approaches is not based on performance metrics but on the methodological dimensions of the pipeline: data preparation process, computational requirements, and model interpretability.

The results demonstrate that tabular representation offers greater interpretability and a lower computational barrier, requiring manual feature engineering, while volumetric representation preserves the spatial geometry of the patient at the cost of greater infrastructural complexity, delegating feature extraction to the model. This research provides empirical evidence on how data structure conditions modeling decisions, implementation feasibility, and the type of prediction achievable, with methodological implications for Data Science applied to health contexts.

Keywords: IMRT Radiotherapy, Machine Learning, Head and Neck Cancer, Data Representation, Deep Learning, UNet3D, Feature Engineering

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	8
--	---

Contenido

Lista de Figuras	11
Lista de Tablas	12
Anexos	12
Introducción	13
Objetivos.....	18
<i>Objetivo general.....</i>	<i>18</i>
<i>Objetivos específicos</i>	<i>18</i>
Justificación.....	19
Marco Teórico	20
<i>Naturaleza tridimensional y dinámica del problema dosimétrico.....</i>	<i>22</i>
<i>Fundamentos del aprendizaje supervisado en contextos clínicos</i>	<i>23</i>
<i>Arquitecturas Volumétricas para Segmentación y Predicción: La U-Net 3D.....</i>	<i>26</i>
<i>Comparación Metodológica entre Enfoques Tabular y Volumétrico.....</i>	<i>31</i>
<i>Consideraciones Éticas, Regulatorias y de Validación Clínica.....</i>	<i>33</i>
Hipótesis	35
Variables	36
1. <i>Variables de entrada – Enfoque Tabular</i>	<i>37</i>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	9
--	---

1.1 Estadio del tumor (TNM)	37
1.2 Edad del paciente	37
1.3 Género	38
1.4 Número de fracciones	38
1.5 Dosis total prescrita (Gy)	38
<i>2. Variables de entrada – Enfoque Volumétrico</i>	<i>38</i>
2.1 Imagen CT tridimensional	39
2.2 Máscara del PTV	39
2.3 Máscaras de órganos en riesgo (OAR).....	39
<i>3. Variables objetivo.....</i>	<i>39</i>
3.1 Variable objetivo en el modelo tabular	40
3.2 Variable objetivo en el modelo volumétrico	40
<i>4. Métricas de evaluación (Variables derivadas)</i>	<i>40</i>
4.1 Error Absoluto Medio (MAE)	41
4.2 Error Cuadrático Medio (MSE).....	41
<i>5. Variable independiente principal</i>	<i>41</i>
Metodología.....	43
<i>Enfoque de la investigación</i>	<i>43</i>
<i>Diseño de la investigación.....</i>	<i>44</i>
<i>Tipo de estudio</i>	<i>44</i>
Fases de la investigación	45

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	10
--	----

<i>Población y muestra</i>	48
<i>Criterios de Inclusión y Exclusión</i>	49
<i>Limitantes de los Datos Abiertos</i>	50
<i>Diseño y validación del instrumento de medición</i>	51
<i>Procedimientos y técnicas de análisis de la información</i>	51
Técnicas de análisis:.....	52
Trabajo de Campo	53
<i>Enfoque con datos de HNSCC-3DCT-RT</i>	53
<i>Procesamiento de los datos</i>	60
Integración horizontal por “Patient ID”.....	61
Normalización y limpieza adicional.....	61
Construcción de variables derivadas.....	61
Longitudinalización del peso.....	62
Perspectiva desde Ciencia de Datos.....	62
<i>Análisis de resultados</i>	63
Número de fracciones (Number of fx).....	66
Dosis por fracción (Dose/fx – Gy).....	67
Distribución del volumen tumoral (PTV Volume – cm ³).....	69
Relación entre dosis total y número de fracciones.....	71
Distribución del número de arcos VMAT.....	72
<i>Modelamiento Supervisado y Resultados</i>	73

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	11
--	----

<i>Configuración e hiperparámetros</i>	74
<i>Resultados</i>	74
<i>Interpretación en relación con el OE2</i>	76
<i>Enfoque con datos Open-KBP</i>	77
<i>Fuente de Datos: Open-KBP</i>	78
<i>Configuración del Entorno de Trabajo</i>	79
Desactivación del Módulo network_functions	80
Gestión de Versiones de Keras	80
Verificación de Disponibilidad de GPU	80
<i>Preparación y Exploración de los Datos</i>	80
Modos de Operación del DataLoader	81
Arquitectura del Modelo UNet3D	82
<i>Justificación de la elección</i>	82
<i>Estructura Detallada de la Red</i>	82
<i>Normalización: Instance Normalization</i>	83
<i>Canales de Entrada</i>	83
<i>Métricas de Evaluación</i>	83
Error Absoluto Medio (MAE) - Dose Score	84
Error Cuadrático Medio (MSE)	84
<i>Inicialización del Modelo</i>	84

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	12
--	----

<i>Bucle de Entrenamiento</i>	85
<i>Configuración General del Entrenamiento</i>	85
<i>Perdida Ponderada para PTVs</i>	86
<i>Normalización del CT</i>	87
<i>Persistencia y Recuperación en Google Drive</i>	87
<i>Flujo del Bucle de Entrenamiento por Epoca</i>	87
<i>Función de Carga Dinámica de Pacientes</i>	88
<i>Reconstrucción de volúmenes Sparse</i>	88
<i>Manejo Robusto de Casos Especiales</i>	88
<i>Inferencia y Evaluación en Validación</i>	88
<i>Flujo de Inferencia por Paciente</i>	89
<i>Análisis de Resultados – Modelo de Predicción de Dosis</i>	89
<i>Resultados del Entrenamiento</i>	89
<i>Evaluación en Pacientes de Validación (40 pacientes)</i>	90
<i>Evaluación en Conjunto de Prueba (100 pacientes)</i>	91
Conclusiones	92
Discusión	94
<i>Bloque 1. Comparación metodológica de los pipelines</i>	94
<i>Bloque 2. Interpretación de resultados en contexto</i>	98

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello	13
--	----

<i>Bloque 3. Verificación de hipótesis</i>	<i>99</i>
<i>Bloque 4. El ecosistema de datos en salud como condicionante del modelamiento.....</i>	<i>101</i>
<i>Bloque 5. Limitaciones del estudio y su impacto en los resultados</i>	<i>102</i>
Conclusiones y Trabajo Futuro.....	104
<i>Conclusiones.....</i>	<i>104</i>
Trabajo futuro	107
Referencias.....	110

Lista de Figuras

Figura 1. Evolución del peso de los pacientes durante el tratamiento	64
Figura 2. Histograma de dosis total (Gy).....	65
Figura 3. Histograma de número de fracciones	66
Figura 4. Histograma de dosis por fracción.....	67
Figura 5. Distribución del volumen del PTV	69
Figura 6. Total Dose vs. PTV Volume	70
Figura 7. Total Dose vs. Number of fractions.....	71
Figura 8. Número de arcos VMAT utilizados.....	72

Lista de Tablas

Tabla 1. Métricas de desempeño - Modelos tabulares sobre HNSCC-3DCT-RT. *MAE en unidades logarítmicas.....	75
Tabla 2. Distribución de los Datos.....	79
Tabla 3. Preparación y Exploración de los Datos.....	81
Tabla 4. Estructura Detallada de la Red.....	83
Tabla 5. Inicialización del Modelo.....	85
Tabla 6. Configuración General del Entrenamiento.....	86
Tabla 7. Referencia de desempeño en el benchmark OpenKBP.....	93
Tabla 8. Comparación metodológica entre el pipeline tabular y el pipeline volumétrico.	97

Anexos

Anexo A. Análisis estadístico y modelo supervisado — Dataset HNSCC-3DCT-RT. 111

Anexo B. Implementación del modelo volumétrico UNet3D — Dataset OpenKBP.... 111

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		13
--	--	----

Introducción

Actualmente, la planificación de la radioterapia de intensidad modulada (IMRT, *Intensity-Modulated Radiation Therapy*) constituye uno de los pilares fundamentales en el tratamiento del cáncer de cabeza y cuello. Esta técnica permite distribuir la dosis de radiación con alta precisión, adaptándose a la geometría tridimensional del tumor y limitando la exposición de órganos en riesgo. Sin embargo, a pesar de los avances tecnológicos incorporados en los sistemas de planificación, el proceso continúa dependiendo de procedimientos iterativos de optimización inversa y ajustes manuales basados en la experiencia del especialista, lo que puede generar variabilidad entre planes clínicamente aceptables para un mismo paciente (Zarepisheh et al., 2019).

Uno de los principales desafíos radica en alcanzar un equilibrio adecuado entre cobertura tumoral, homogeneidad de dosis y protección de tejidos sanos, particularmente en regiones anatómicamente complejas como cabeza y cuello. Factores como el movimiento o intra-fracción, los cambios anatómicos progresivos durante el tratamiento y la alta dimensionalidad del problema dosimétrico incrementan la complejidad de la planificación. Aunque los algoritmos actuales de optimización permiten satisfacer múltiples restricciones simultáneamente, estos operan bajo supuestos geométricos estáticos y no siempre capturan de manera integral las relaciones espaciales dinámicas entre estructuras anatómicas (Nutting, Dearnaley & Webb, 2002).

Paralelamente, el cáncer continúa representando una de las principales causas de mortalidad a nivel mundial. Según la International Agency for Research on Cancer (IARC), en 2020 se registraron aproximadamente 10 millones de muertes asociadas a

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		14
--	--	----

esta enfermedad (IARC,2020). Los datos del Global Cancer Observatory (GLOBOCAN 2020) anticipan un aumento superior al 47% en la incidencia global del cáncer hacia 2040 (Sung et al., 2021), lo que en la práctica significa que los sistemas de salud van a necesitar planificar muchos más tratamientos radioterapéuticos, y hacerlo de manera más eficiente y estandarizada de lo que se hace hoy.

En este contexto, el aprendizaje automático empieza a tomar relevancia, no como reemplazo del especialista, sino como herramienta de apoyo (Vandewinckele et al., 2020; Boldrini et al., 2022; Cardenas et al., 2023). Estudios más recientes han demostrado que los modelos de aprendizaje automático no solo permiten automatizar etapas del flujo de trabajo en radioterapia, sino también abordar tareas complejas de modelamiento predictivo, incluyendo la predicción de distribución de dosis, resultados clínicos y toxicidades asociadas al tratamiento (Shan et al., 2024; Yuan et al., 2024). Asimismo, revisiones contemporáneas destacan que estos modelos integran información multimodal imágenes médicas, variables clínicas y datos dosimétricos para mejorar la precisión predictiva, aunque enfrentan desafíos importantes relacionados con la generalización, la cuantificación de incertidumbre y la interpretabilidad en entornos clínicos reales (Wahid et al., 2024; Lastrucci et al., 2024). Pero antes de hablar de qué algoritmo usar, hay una decisión más fundamental que suele pasarse por alto: cómo se representa la información del paciente antes de que el modelo la procese.

Durante décadas, la forma más común de alimentar modelos predictivos en radioterapia fue organizar la información en tablas: edad del paciente, estadio del tumor, dosis prescrita, volumen del PTV. Variables concretas, manejables, que cualquier estadístico puede trabajar con herramientas estándar. Este enfoque, conocido como

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		15
--	--	----

feature engineering, exige que el investigador decida de antemano qué variables son relevantes (Goodfellow et al., 2016). La ventaja es que los modelos resultantes son interpretables y funcionan bien incluso con muestras pequeñas. La limitación es que, al reducir la anatomía tridimensional del paciente a un puñado de números, inevitablemente se pierde información espacial que puede ser crítica para entender cómo se distribuye la dosis.

El enfoque alternativo parte de una lógica distinta: en lugar de resumir la anatomía, preservarla completa. Las representaciones volumétricas trabajan directamente con las imágenes CT y las máscaras de estructuras, organizadas como tensores tridimensionales donde cada voxel contiene información. Sobre esa estructura operan las redes neuronales convolucionales 3D (Litjens et al., 2017), que aprenden por sí solas qué patrones espaciales son relevantes sin que el investigador tenga que especificarlo. Es lo que se llama feature learning automático. La promesa es mayor capacidad para capturar relaciones geométricas complejas entre el tumor y los órganos circundantes. El costo, que se hace evidente desde el primer momento de implementación, es una barrera computacional e infraestructural considerablemente más alta.

Sin embargo, los modelos volumétricos tridimensionales también presentan desafíos significativos. Requieren mayor capacidad computacional, mayores volúmenes de datos para entrenamiento, tiempos prolongados de procesamiento y menor interpretabilidad directa en comparación con modelos tradicionales. Asimismo, la alta dimensionalidad del espacio voxel puede introducir riesgos de sobreajuste si no se implementan estrategias adecuadas de regularización y validación (Yuan et al., 2025).

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		16
--	--	----

Desde una perspectiva metodológica, surge entonces una cuestión central: determinar en qué medida la estructura del dato condiciona las decisiones de modelamiento, la complejidad del pipeline y las capacidades del modelo resultante. Esta cuestión no solo tiene implicaciones técnicas en el ámbito del aprendizaje automático, sino también clínicas, al determinar qué tipo de aproximación ofrece un mejor equilibrio entre desempeño predictivo, interpretabilidad, complejidad computacional y aplicabilidad práctica en entornos hospitalarios (Gianoli et al., 2024).

En el contexto específico del cáncer de cabeza y cuello, donde la proximidad entre el volumen tumoral y múltiples órganos críticos aumenta la complejidad de la planificación, evaluar comparativamente ambos enfoques adquiere especial relevancia. La presente investigación propone evaluar el impacto de la representación de datos en el proceso de modelamiento, caracterizando las decisiones técnicas, la complejidad infraestructural y las capacidades y limitaciones propias de cada paradigma en el contexto de la planificación radioterapéutica del cáncer de cabeza y cuello.

De esta manera, el estudio no busca comparar directamente el desempeño predictivo entre modelos, sino analizar cómo la representación del dato condiciona la formulación del problema, el diseño del pipeline y las capacidades del modelo resultante. Este enfoque reconoce que cuando cada paradigma define una variable objetivo de naturaleza distinta, la comparación válida no es de métricas de error sino de características metodológicas. Lo anterior implica que cada paradigma opera sobre espacios de salida no equivalentes.

En consecuencia, la pregunta central que orienta esta investigación es:

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		17
--	--	----

¿Cómo condiciona la estructura intrínseca del dato, tabular estructurado vs. volumétrico tridimensional, las decisiones de modelamiento, el proceso de representación de características y las capacidades y limitaciones de los modelos de aprendizaje automático resultantes?

Para dar respuesta a esta pregunta, el presente documento se organiza en las siguientes secciones: la Justificación expone las razones clínicas, metodológicas y científicas que fundamentan la investigación; el Marco Teórico desarrolla los fundamentos conceptuales sobre radioterapia IMRT, aprendizaje supervisado, arquitecturas volumétricas y los datasets empleados; los Objetivos e Hipótesis delimitan el alcance del estudio y las proposiciones verificables que lo orientan; la Metodología describe el diseño de la investigación, las fases de desarrollo, la población y muestra, y los criterios de selección de datos; el Trabajo de Campo documenta de forma detallada y reproducible los pipelines implementados para cada paradigma de representación, incluyendo el preprocesamiento, el modelamiento supervisado y el análisis de resultados; la Discusión interpreta los hallazgos desde cinco dimensiones: comparación metodológica de los pipelines, interpretación de resultados en contexto, verificación de hipótesis, restricciones del ecosistema de datos en salud y limitaciones del estudio; finalmente, las Conclusiones y Trabajo Futuro sintetizan los aportes de la investigación e identifican las líneas de desarrollo posteriores. Los Anexos contienen los cuadernos de Google Colaboratory con el código fuente completo de ambos pipelines.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		18
--	--	----

Objetivos

Objetivo general

Evaluar el impacto de la representación de datos, tabular estructurada y volumétrica tridimensional, en el proceso de modelamiento con aprendizaje automático, aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello.

Objetivos específicos

- Caracterizar las propiedades estructurales y estadísticas de los datasets HNSCC-3DCT-RT y OpenKBP, en términos de dimensionalidad, tipo de variables y nivel de granularidad.
- Evaluar la capacidad predictiva de un modelo de aprendizaje automático basado en representaciones tabulares estructuradas para estimar el volumen del PTV.
- Evaluar el desempeño de un modelo de aprendizaje profundo basado en representaciones volumétricas tridimensionales para la predicción de distribuciones espaciales de dosis.
- Comparar el pipeline tabular y el pipeline volumétrico en sus características metodológicas, analizando las diferencias en el proceso de preparación de datos, los requerimientos computacionales de cada enfoque y la interpretabilidad de los modelos entrenados.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		19
--	--	----

Justificación

Planificar la radioterapia en cáncer de cabeza y cuello no es un proceso rutinario. La concentración de estructuras anatómicas críticas en esa región, médula espinal, tronco encefálico, glándulas salivales, laringe, obliga al especialista a resolver simultáneamente múltiples restricciones dosimétricas que compiten entre sí. Entregar la dosis correcta al tumor sin comprometer lo que está al lado requiere experiencia, tiempo e iteración, y aun así puede generar resultados distintos entre planificadores diferentes para un mismo paciente (Zarepisheh et al., 2019). Es en ese espacio de variabilidad donde los modelos predictivos encuentran su justificación como herramienta de apoyo.

Ahora bien, antes de decidir qué algoritmo usar, hay una pregunta más fundamental: ¿cómo se representa la información del paciente? Esa decisión no es neutra. Determina qué relaciones puede aprender el modelo, qué información espacial se conserva o se pierde, y qué tan costoso resulta el entrenamiento. El enfoque tabular responde a esa pregunta con variables estructuradas seleccionadas manualmente por expertos, lo que se llama *feature engineering* (Goodfellow, Bengio & Courville, 2016), mientras que el enfoque volumétrico preserva la geometría completa del paciente y delega al modelo la tarea de descubrir qué patrones son relevantes, proceso conocido como *feature learning*. Ambos caminos conducen a modelos con características muy distintas en términos de interpretabilidad, infraestructura requerida y tipo de predicción que habilitan (Litjens et al., 2017; Jiang et al., 2024).

Lo que esta investigación busca aportar es evidencia empírica sobre esas diferencias, desde la perspectiva de la Ciencia de Datos. No se trata de determinar cuál paradigma es superior en abstracto, sino de caracterizar qué implica trabajar con cada uno: qué

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		20
--	--	----

decisiones exige en la preparación del dato, qué recursos demanda, qué tan interpretable es el resultado y en qué condiciones cada enfoque resulta más viable. Para ello se analizaron en paralelo el dataset HNSCC-3DCT-RT desde el paradigma tabular y el dataset OpenKBP desde el paradigma volumétrico, ambos de acceso abierto, previamente anonimizados y compatibles con marcos regulatorios internacionales como HIPAA y GDPR.

Los resultados de esta comparación tienen valor tanto para investigadores que trabajan con datos clínicos heterogéneos como para equipos que deben decidir qué arquitectura de modelamiento es viable dado su contexto computacional e institucional. Esa es la contribución metodológica que esta investigación busca hacer al campo de la Ciencia de Datos aplicada a salud.

Marco Teórico

La radioterapia constituye uno de los pilares fundamentales del tratamiento oncológico moderno. Desde el descubrimiento de los efectos biológicos de la radiación ionizante, esta modalidad terapéutica ha evolucionado progresivamente hasta convertirse en una herramienta de alta precisión. Según el National Cancer Institute (s.f.), aproximadamente la mitad de los pacientes diagnosticados con cáncer recibirán radioterapia en algún momento de su proceso terapéutico, ya sea como tratamiento primario, adyuvante o paliativo.

En el caso específico del cáncer de cabeza y cuello, la radioterapia desempeña un rol especialmente relevante. Esta región anatómica se caracteriza por una alta densidad de estructuras funcionales críticas: médula espinal, tronco encefálico, glándulas salivales,

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		21
--	--	----

estructuras ópticas y vías aéreas superiores. La proximidad entre el volumen tumoral y estos órganos en riesgo (Organs at Risk, OAR) impone restricciones dosimétricas estrictas. El desafío no consiste únicamente en destruir el tejido tumoral, sino en preservar la funcionalidad del paciente y minimizar la toxicidad aguda y tardía.

La evolución tecnológica ha permitido el desarrollo de técnicas cada vez más precisas, entre las cuales destaca la Radioterapia de Intensidad Modulada (IMRT). A diferencia de técnicas convencionales como la radioterapia conformacional tridimensional (3D-CRT), la IMRT permite modular la intensidad del haz de radiación en múltiples ángulos, adaptando la distribución de dosis a la geometría tridimensional del tumor (Webb, 2001). Este proceso se basa en optimización inversa, donde el planificador define objetivos clínicos y restricciones dosimétricas, y el sistema computacional calcula la configuración óptima de los haces.

No obstante, pese a su sofisticación tecnológica, la planificación IMRT continúa siendo un proceso complejo, iterativo y dependiente de la experiencia del especialista. Zarepisheh et al. (2019) documentan variabilidad interplanificador significativa, demostrando que diferentes profesionales pueden generar planes clínicamente aceptables, pero dosimétricamente distintos para un mismo paciente. Este fenómeno revela que la planificación no es completamente determinista ni homogénea, y que persiste un componente subjetivo en la toma de decisiones, lo cual justifica el interés creciente en automatizar y estandarizar el proceso mediante técnicas de aprendizaje automático.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		22
--	--	----

cdimensional y dinámica del problema dosimétrico

La planificación radioterapéutica es esencialmente un problema geométrico tridimensional. La dosis absorbida por cada punto anatómico depende de la interacción espacial entre el Planning Target Volume (PTV), los órganos en riesgo (OAR), la modulación del haz y la configuración angular seleccionada. Bentzen et al. (2010) señalan que la probabilidad de complicaciones en tejidos normales no depende únicamente de la dosis total administrada, sino de su distribución espacial: la irradiación parcial de un órgano puede generar efectos clínicos cualitativamente diferentes a una irradiación homogénea.

Adicionalmente, el proceso terapéutico no es estático. Cambios anatómicos como pérdida de peso, reducción tumoral o desplazamientos internos pueden alterar la geometría inicial del plan. Este carácter dinámico introduce incertidumbre adicional en la predicción de resultados dosimétricos. En consecuencia, cualquier intento de modelamiento predictivo debe considerar que la variable de salida, la distribución de dosis es inherentemente espacial y tridimensional, lo que plantea exigencias específicas sobre la representación de datos y la arquitectura de los modelos empleados.

Vallieres et al. (2017) ilustran esta complejidad en el contexto del cáncer de cabeza y cuello, donde la integración de datos clínicos, imagenológicos y dosimétricas resulta imprescindible para construir modelos predictivos robustos. Su trabajo fundamenta el uso de cohortes de cabeza y cuello como banco de pruebas privilegiado para modelos de predicción de dosis, dada la riqueza anatómica y la diversidad de OAR involucrados.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		23
--	--	----

Antecedentes en representación de datos para planificación radioterapéutica

La pregunta sobre cómo representar los datos del paciente en modelos de predicción dosimétrica ha sido abordada desde múltiples ángulos en la literatura reciente, configurando un campo de antecedentes directamente relevante para el objetivo de este trabajo.

En el paradigma basado en contornos anatómicos, Liu et al. (2021) demostraron que el desempeño de redes 3D U-Net para predicción de dosis depende críticamente de cómo se codifica la relación espacial entre el PTV y los OAR, estableciendo un antecedente temprano para el paradigma volumétrico basado en máscaras estructurales. Extendiendo esta línea, Yue et al. (2022) propusieron una representación basada en distancia geométrica entre estructuras anatómicas para la predicción de dosis en carcinoma nasofaríngeo, evidenciando que las representaciones geométricas mejoran la captura de relaciones espaciales tumor-OAR en comparación con contornos clásicos. En contraste, Jiao et al. (2023), con su arquitectura TransDose, demostraron la viabilidad de un paradigma sin máscaras explícitas, donde CT e información de superpíxeles se combinan mediante grafos convolucionales y Transformers, desafiando el supuesto de que las máscaras estructurales son un requisito indispensable para la predicción volumétrica.

En el ámbito de la comparación entre métodos, Osman et al. (2023) realizaron un análisis comparativo entre planificación basada en conocimiento (KBP) y enfoques de aprendizaje profundo para predicción tridimensional de dosis, mostrando que diferentes representaciones producen trade-offs distintos en términos de precisión, interpretabilidad y viabilidad clínica, resultado directamente alineado con el objetivo comparativo de esta

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		24
--	--	----

investigación. Por su parte, Teng et al. (2024) propusieron una representación orientada a los haces de radiación (*beam-wise*), donde la geometría de cada haz se modela explícitamente, logrando mejoras de desempeño frente a representaciones globales e introduciendo el concepto de paradigmas físico-informados.

Desde la perspectiva clínica, Rashad et al. (2024) evidenciaron que la forma en que se delinean y representan los OAR tiene consecuencias directas sobre la equivalencia dosimétrica y la estimación de complicaciones en tejidos normales (NTCP), conectando la elección de representación con decisiones clínicas concretas. Sun et al. (2022) extendieron esta lógica al demostrar que la representación de datos puede alimentar directamente los optimizadores del sistema de planificación de tratamiento (TPS), cerrando el ciclo entre predicción y planificación automatizada entregable.

En cuanto a tendencias emergentes, Huiskes et al. (2023) introdujeron la dimensión temporal en la representación, argumentando que las variaciones anatómicas durante el tratamiento exigen modelos de representación dinámica para la radioterapia adaptativa de protones. Oppitz et al. (2025) desarrollaron una arquitectura 2.5D U-Net para planificación automatizada, mostrando que la elección de representación condiciona no solo el desempeño predictivo sino la viabilidad de despliegue clínico. McCullum et al. (2025), en una revisión sistemática sobre representaciones multimodales en radioterapia adaptativa, identificaron que los problemas de validación y escalamiento clínico persisten incluso en enfoques de alta complejidad. Finalmente, Xiong et al. (2026) propusieron la integración de priors físicos en arquitecturas de núcleo amplio (*large-kernel*), logrando modelos generalizables a múltiples sitios anatómicos, y Hémon et al. (2025) argumentaron que la incertidumbre en la predicción debe ser tratada como una

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		25
--	--	----

dimensión constitutiva de la representación en sistemas de inteligencia artificial confiables para radioterapia.

Este conjunto de antecedentes evidencia que la elección de representación no es una decisión técnica periférica: condiciona el tipo de problema que puede formularse, los recursos requeridos y la aplicabilidad clínica del modelo resultante. El presente trabajo se inscribe en esa línea de investigación, aportando evidencia empírica comparativa desde la perspectiva de la Ciencia de Datos.

Fundamentos del aprendizaje supervisado en contextos clínicos

El aprendizaje supervisado constituye un paradigma central dentro del aprendizaje automático. Formalmente, busca aproximar una función $f: X \rightarrow Y$, donde X representa el espacio de entrada e Y el espacio de salida (Goodfellow, Bengio & Courville, 2016). En el contexto radioterapéutico, las entradas pueden corresponder a información anatómica, geométrica y clínica, mientras que las salidas pueden representar métricas dosimétricas agregadas o mapas tridimensionales completos de dosis. En aplicaciones recientes, estos esquemas se han extendido hacia configuraciones multimodales que integran imágenes médicas, variables clínicas y datos dosimétricos en un mismo espacio de aprendizaje, mejorando la capacidad predictiva de los modelos (Cardenas et al., 2023; Jiang et al., 2023).

Desde la teoría del aprendizaje estadístico, el objetivo no es únicamente minimizar el error empírico en los datos de entrenamiento, sino garantizar capacidad de generalización. Este equilibrio se explica mediante el bias-variance tradeoff, donde modelos excesivamente simples presentan alto sesgo, mientras que modelos altamente

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		26
--	--	----

complejos pueden incurrir en sobreajuste (Goodfellow et al., 2016). Este principio es especialmente relevante en contextos clínicos, donde los conjuntos de datos suelen ser de tamaño moderado y la heterogeneidad entre pacientes es considerable. Estudios recientes han enfatizado además la importancia de la validación externa, la robustez del modelo y la cuantificación de incertidumbre como elementos clave para garantizar la generalización en entornos clínicos reales (Kelly et al., 2022; Wahid et al., 2024).

Shmueli (2010) aporta una distinción conceptual fundamental para enmarcar este tipo de investigación: la diferencia entre modelos orientados a la explicación y modelos orientados a la predicción. Mientras que los modelos explicativos priorizan la interpretabilidad de los coeficientes y la inferencia causal, los modelos predictivos buscan minimizar el error en datos no vistos. Esta distinción resulta clave en el presente trabajo, donde el objetivo es predecir distribuciones de dosis, no necesariamente explicar los mecanismos físicos subyacentes. Investigaciones recientes han reforzado esta dicotomía en el contexto del aprendizaje automático aplicado a salud, destacando que muchos modelos clínicos de alto desempeño predictivo operan como “black boxes”, lo que introduce tensiones entre precisión y explicabilidad (Topol, 2019; Rudin, 2019).

En el ámbito médico más amplio, Obermeyer y Emanuel (2016) argumentaron que el aprendizaje automático tiene el potencial de superar a los métodos estadísticos clásicos en tareas de predicción clínica, precisamente porque puede capturar relaciones no lineales y de alta dimensionalidad que los modelos paramétricos tradicionales no logran modelar adecuadamente. Más recientemente, Rajpurkar et al. (2022) ofrecen una revisión comprehensiva del estado del arte de la inteligencia artificial en medicina, identificando tanto las oportunidades como las limitaciones específicas de los datos

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		27
--	--	----

tabulares clínicos: sensibilidad a valores faltantes, dependencia de la calidad del registro y dificultad para capturar contexto espacial. Trabajos más recientes han profundizado en estas limitaciones, señalando que los datos tabulares clínicos presentan desafíos estructurales relacionados con la heterogeneidad, el sesgo de selección y la falta de estandarización, lo que puede afectar significativamente el desempeño y la equidad de los modelos predictivos (Chen et al., 2021; Kapoor & Narayanan, 2023).

Predicción de Distribuciones de Dosis Mediante Aprendizaje Profundo

La predicción automatizada de distribuciones de dosis en radioterapia ha experimentado un avance acelerado durante la última década, impulsado por el desarrollo de arquitecturas de aprendizaje profundo capaces de procesar datos volumétricos de alta dimensionalidad. Este campo representa una convergencia entre la física médica, la dosimetría clínica y la inteligencia artificial.

Entre los trabajos pioneros en esta área se encuentra el de Nguyen et al. (2019), quienes realizaron un estudio de factibilidad para predecir distribuciones de dosis óptimas en cáncer de próstata mediante redes neuronales profundas. Este trabajo demostró que era posible aproximar la distribución de dosis completa a partir de información anatómica, sentando las bases conceptuales para estudios posteriores en otras localizaciones tumorales. De manera complementaria, Chen et al. (2019) extendieron este enfoque específicamente al cáncer de cabeza y cuello, empleando redes neuronales convolucionales para predecir distribuciones de dosis tridimensionales a partir de imágenes CT y máscaras estructurales, logrando resultados comparables a los planes generados manualmente por especialistas.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		28
--	--	----

Fan et al. (2020) dieron un paso adicional al integrar la predicción de dosis tridimensional dentro de un flujo de planificación automática completo, demostrando que los mapas de dosis predichos por redes profundas podían ser utilizados directamente como objetivos de optimización en sistemas de planificación comerciales. Este trabajo evidenció el potencial clínico real de los enfoques basados en aprendizaje profundo para reducir la variabilidad interplanificador y acelerar el proceso de planificación.

Un hito arquitectónico específico para la predicción de dosis volumétrica fue establecido por Kearney et al. (2018) con DoseNet, una red neuronal convolucional tridimensional completamente convolucional diseñada específicamente para esta tarea. DoseNet demostró que arquitecturas volumétricas podían capturar la distribución espacial de la dosis con alta fidelidad, superando enfoques basados en cortes bidimensionales o métricas agregadas. Este trabajo se convierte en referencia directa para la elección de arquitecturas volumétricas en el presente estudio.

Babier et al. (2021) establecieron un marco de referencia colectivo mediante el desafío OpenKBP (Open Knowledge-Based Planning), el primero de acceso abierto en predicción de dosis para cáncer de cabeza y cuello. OpenKBP estandarizó la evaluación de modelos predictivos mediante un conjunto de datos armonizado de 340 pacientes, métricas de evaluación compartidas y un protocolo de validación reproducible. Este benchmark constituye uno de los conjuntos de datos utilizados en la presente investigación, y su uso garantiza la comparabilidad de los resultados con el estado del arte internacional.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		29
--	--	----

Arquitecturas Volumétricas para Segmentación y Predicción: La U-Net 3D

La elección de una arquitectura adecuada para el procesamiento de volúmenes médicos tridimensionales es una decisión crítica con implicaciones directas sobre la calidad de la predicción y los requerimientos computacionales. En este contexto, la familia de arquitecturas U-Net ha emergido como el estándar de facto en tareas de análisis de imagen médica volumétrica.

La U-Net fue introducida originalmente por Ronneberger et al. (2015) para segmentación de imágenes biomédicas bidimensionales. Su innovación central reside en la arquitectura encoder-decoder con conexiones de salto (skip connections), que permiten combinar características de alta resolución espacial con representaciones semánticas de alto nivel. Esta estructura permite al modelo aprender tanto patrones locales de textura como contexto global de la imagen, preservando información espacial que se perdería en arquitecturas puramente profundas.

Çiçek et al. (2016) extendieron este concepto al dominio tridimensional con la U-Net 3D, adaptando todos los operadores de convolución, pooling y upsampling al espacio volumétrico. Esta extensión es fundamental para tareas donde la continuidad espacial en el eje axial es clínicamente relevante, como ocurre con la distribución de dosis en estructuras anatómicas tridimensionales. La U-Net 3D permite al modelo explotar la coherencia volumétrica de las imágenes CT y las máscaras estructurales, capturando relaciones espaciales que no serían visibles en análisis corte a corte.

Milletari et al. (2016) introdujeron la V-Net, una variante tridimensional que incorpora bloques residuales y la función de pérdida Dice, mejorando la estabilidad del

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		30
--	--	----

entrenamiento y el rendimiento en estructuras anatómicas con desequilibrio de clases. Estas innovaciones han sido ampliamente adoptadas en implementaciones posteriores de arquitecturas tipo U-Net para predicción de dosis.

En el contexto específico de la predicción de distribuciones de dosis, trabajos recientes como el de Bai et al. (2025) demuestran que arquitecturas basadas en U-Net 3D continúan siendo competitivas frente a enfoques más modernos como los Vision Transformers, especialmente cuando el tamaño de los conjuntos de datos es moderado. La incorporación de mecanismos de atención y módulos residuales dentro del marco U-Net ha permitido mejorar progresivamente el rendimiento sin incrementar desproporcionalmente la complejidad del modelo.

Desde el punto de vista matemático, las CNN 3D utilizadas en estas arquitecturas realizan convoluciones volumétricas definidas por: $Y(i,j,k) = \sum_{\{u,v,w\}} X(i+u, j+v, k+w) \cdot W(u,v,w)$, donde X representa el tensor de entrada, W el filtro aprendible Y el mapa de activación resultante. Las conexiones de salto en la U-Net 3D permiten concatenar los mapas de activación del encoder con los correspondientes del decoder, recuperando la resolución espacial perdida durante el submuestreo progresivo. Este mecanismo es especialmente valioso para la predicción de dosis, donde la resolución espacial de la salida debe igualar la de la entrada anatómica.

Litjens et al. (2017) y Jiang et al. (2024) documentan de manera comprehensiva la eficacia de redes profundas en tareas de imagen médica, consolidando el consenso sobre la superioridad de las arquitecturas convolucionales volumétricas para predicción espacial en oncología. Sin embargo, ambos trabajos también advierten sobre los retos asociados: elevada dimensionalidad de entrada, mayor riesgo de sobreajuste con

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		31
--	--	----

muestras limitadas, requerimientos computacionales avanzados y menor interpretabilidad directa frente a modelos tabulares.

Evolución reciente del aprendizaje profundo en radioterapia

En los últimos años, el uso de aprendizaje profundo en radioterapia ha experimentado una evolución significativa, pasando de enfoques centrados en arquitecturas convolucionales tradicionales a modelos más complejos que integran múltiples fuentes de información y nuevas formas de representación. Los primeros trabajos en esta área demostraron la viabilidad de predecir distribuciones de dosis utilizando redes convolucionales tridimensionales, estableciendo una base sólida para el desarrollo posterior del campo (Nguyen et al., 2019; Chen et al., 2019; Kearney et al., 2018). Sin embargo, investigaciones más recientes han evidenciado que el desempeño de estos modelos no depende únicamente de la arquitectura empleada, sino también de la forma en que se representa la información anatómica y dosimétrica. En este contexto, se han propuesto representaciones más sofisticadas que incorporan relaciones espaciales explícitas entre el tumor y los órganos en riesgo, como los enfoques guiados por distancia (Yue et al., 2022) o las representaciones basadas en haces de radiación (Teng et al., 2024). Estos avances han permitido mejorar la capacidad de los modelos para capturar interacciones geométricas complejas. Adicionalmente, el campo ha comenzado a explorar arquitecturas más avanzadas, incluyendo modelos basados en transformers que permiten prescindir de máscaras estructurales explícitas, aprendiendo directamente desde imágenes CT (Jiao et al., 2023). De manera complementaria, se han desarrollado enfoques que integran el modelamiento predictivo con procesos de planificación automatizada, donde la predicción de dosis se utiliza como entrada para sistemas de

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		32
--	--	----

optimización clínica (Sun et al., 2022; Oppitz et al., 2025). En paralelo, investigaciones recientes han introducido nuevas dimensiones en el modelamiento, como la adaptación a cambios anatómicos a lo largo del tratamiento (Huiskes et al., 2023), el uso de representaciones multimodales (McCullum et al., 2025) y la incorporación de incertidumbre en las predicciones para mejorar la confiabilidad de los modelos en entornos clínicos (Hémon et al., 2025). Asimismo, enfoques emergentes basados en conocimiento físico han comenzado a integrarse en modelos de aprendizaje profundo, mejorando la capacidad de generalización y la coherencia con principios dosimétricos (Xiong et al., 2026). Estos desarrollos evidencian que el campo ha evolucionado hacia modelos más complejos, donde la arquitectura, la representación de los datos y la integración con el proceso clínico juegan un papel conjunto. En este sentido, la evaluación comparativa de distintas formas de representación de los datos, como la planteada en esta investigación, se alinea con las tendencias actuales y responde a una necesidad identificada en la literatura reciente.

Representación de Datos como Decisión Epistemológica

En modelamiento predictivo, la representación de los datos no es una decisión trivial. Define qué información se conserva, qué se pierde y qué patrones pueden ser aprendidos. Desde una perspectiva epistemológica, la representación constituye una forma de construir conocimiento: una representación tabular simplifica la anatomía en variables agregadas, mientras que una representación volumétrica preserva la estructura espacial completa. Esta diferencia no es meramente técnica, sino conceptual: implica decidir si el conocimiento clínico se expresa mediante métricas resumidas o mediante geometría tridimensional explícita. (Vayena & Blasimme, 2022)

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		33
--	--	----

Representación Tabular Estructurada

La representación tabular organiza la información en variables estructuradas como edad, estadio TNM (American Joint Committee on Cancer, 2017), volumen del PTV, dosis total prescrita, número de fracciones y métricas derivadas como D95 o V30. Este enfoque depende de feature engineering, donde expertos seleccionan manualmente las variables consideradas relevantes (Goodfellow et al., 2016). La anatomía tridimensional se resume en métricas agregadas, lo cual implica una reducción dimensional deliberada.

Obermeyer y Emanuel (2016) señalan que los datos clínicos estructurados han sido el sustrato principal de la medicina basada en evidencia durante décadas, y que su integración con técnicas de aprendizaje automático abre posibilidades genuinas de mejora en la predicción clínica. Sin embargo, también identifican limitaciones inherentes: los datos tabulares capturan solo aquellas dimensiones que el clínico decidió registrar, perdiendo información contextual y espacial que podría ser relevante para el modelo.

Desde la teoría de la información, la reducción dimensional puede implicar pérdida de información relevante. No obstante, el enfoque tabular presenta ventajas concretas en entornos clínicos: alta interpretabilidad, menor complejidad computacional, mayor estabilidad estadística con muestras limitadas y facilidad de validación cruzada. Rajpurkar et al. (2022) destacan que la interpretabilidad sigue siendo uno de los criterios más valorados por los profesionales de la salud al evaluar herramientas de apoyo a la decisión clínica.

Representación Volumétrica Tridimensional

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		34
--	--	----

En contraste, la representación volumétrica preserva la información voxel a voxel proveniente de imágenes CT y máscaras estructurales del PTV y los OAR. Desde el punto de vista matemático, el espacio de entrada se convierte en un tensor tridimensional de alta dimensionalidad. Este enfoque permite aplicar redes neuronales convolucionales tridimensionales (CNN 3D), que realizan feature learning automático, aprendiendo representaciones jerárquicas directamente desde los datos crudos sin requerir selección manual de variables (Çiçek et al., 2016; Isensee et al., 2021; Hatamizadeh et al., 2022).

Un volumen de $128 \times 128 \times 128$ voxeles contiene más de dos millones de puntos de información. Este incremento en dimensionalidad impacta el tiempo de entrenamiento, la memoria requerida, el riesgo de sobreajuste y la necesidad de regularización. Desde la perspectiva del aprendizaje estadístico, mayor capacidad del modelo implica menor sesgo potencial, pero mayor varianza si el conjunto de entrenamiento es insuficiente para generalizar. En la práctica, el uso de técnicas como dropout, normalización por instancia y aumento de datos resulta imprescindible para mitigar estos riesgos (Shorten & Khoshgoftaar, 2019; Isensee et al., 2021; Myronenko, 2019; Wen et al., 2023).

Conjuntos de Datos de Referencia para la Predicción de Dosis

La validez de cualquier modelo predictivo depende críticamente de la representatividad y calidad del conjunto de datos utilizado para su entrenamiento y evaluación. En el contexto de la predicción de dosis radioterapéutica, la disponibilidad de datos clínicos etiquetados ha sido históricamente limitada, lo que ha motivado el desarrollo de conjuntos de datos públicos que permitan la comparación reproducible entre métodos.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		35
--	--	----

El conjunto de datos OpenKBP, documentado por Babier et al. (2021), representa el primer benchmark público de acceso abierto para predicción de dosis en cáncer de cabeza y cuello. Incluye imágenes CT, máscaras de estructuras anatómicas y planes de dosis de 340 pacientes tratados con IMRT, junto con métricas de evaluación estandarizadas. Su diseño como competencia abierta permitió la participación de múltiples equipos internacionales, generando un ecosistema de comparación metodológica que ha acelerado el progreso en el área. El uso de OpenKBP en la presente investigación garantiza la alineación con el estado del arte y la reproducibilidad de los resultados.

El conjunto HNSCC-3DCT-RT (Head and Neck Squamous Cell Carcinoma 3D CT with Radiotherapy Structures) complementa OpenKBP al proporcionar datos clínicos reales de pacientes con carcinoma escamocelular de cabeza y cuello, incluyendo variabilidad institucional que enriquece la capacidad de generalización de los modelos. Vallieres et al. (2017) contextualizan la relevancia de este tipo de cohortes al demostrar que los datos clínicos e imagenológicos de cabeza y cuello permiten construir modelos predictivos robustos cuando se integran adecuadamente variables anatómicas, dosimétricas y clínicas.

La elección de estos dos conjuntos de datos responde a criterios clínicos y metodológicos bien definidos: representatividad de la población oncológica de cabeza y cuello, disponibilidad de imágenes CT con resolución adecuada, existencia de planes de dosis de referencia generados por especialistas certificados, y anonimización conforme a estándares de protección de datos. Ambos conjuntos permiten entrenar y evaluar tanto

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		36
--	--	----

modelos tabulares extrayendo métricas agregadas de las estructuras como modelos volumétricos que procesan los sensores de imagen completos.

Comparación Metodológica entre Enfoques Tabular y Volumétrico

La comparación entre representaciones tabular y volumétrica no puede reducirse a una confrontación de métricas de error. Implica considerar múltiples dimensiones simultáneamente: capacidad predictiva, resolución espacial, interpretabilidad clínica, complejidad computacional, robustez ante cambios anatómicos y escalabilidad a diferentes entornos institucionales.

Desde la perspectiva de la capacidad predictiva, los modelos volumétricos como la U-Net 3D tienen ventaja teórica al acceder a la distribución espacial completa de la dosis, lo que les permite capturar relaciones geométricas entre el PTV y los OAR que no son representables mediante métricas agregadas. Sin embargo, esta ventaja potencial debe materializarse empíricamente: con muestras pequeñas o distribuciones atípicas, los modelos volumétricos pueden no superar a los tabulares (Osman et al., 2023; Jiao et al., 2023).

La interpretabilidad representa una dimensión crítica en entornos clínicos. Shmueli (2010) advierte que la maximización de la capacidad predictiva puede entrar en tensión con la comprensibilidad del modelo. Los modelos tabulares permiten identificar qué variables son más influyentes mediante análisis de importancia de características, facilitando la validación clínica (Liu et al., 2021; Rashad et al., 2024). Las redes profundas volumétricas, aunque potencialmente más precisas, son menos transparentes y

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		37
--	--	----

requieren técnicas de explicabilidad como Grad-CAM o SHAP para aproximar una interpretación de sus decisiones (Hémon et al., 2025).

Rajpurkar et al. (2022) señalan que la implementación exitosa de modelos de aprendizaje automático en medicina requiere no solo rendimiento predictivo, sino también aceptabilidad clínica, robustez frente a distribuciones de datos cambiantes y compatibilidad con los flujos de trabajo existentes. En este sentido, el enfoque tabular puede ofrecer ventajas operativas significativas en centros con recursos computacionales limitados o sin infraestructura de procesamiento de imágenes a gran escala (Oppitz et al., 2025; McCullum et al., 2025).

La representación de los OAR en el enfoque volumétrico tampoco es neutral: la forma en que se delinean y codifican las estructuras anatómicas tiene consecuencias directas sobre la equivalencia dosimétrica y la estimación de toxicidad en tejidos normales (Rashad et al., 2024). En el enfoque tabular estas relaciones se abstraen en métricas agregadas que pierden resolución espacial, pero ganan en interpretabilidad. Adicionalmente, la dimensión temporal ausente en ambos pipelines implementados en este trabajo representa una limitación importante: los cambios anatómicos durante el tratamiento exigen representaciones dinámicas que los paradigmas estáticos no capturan de forma nativa (Huiskes et al., 2023; Teng et al., 2024).

Este análisis comparativo constituye el núcleo conceptual de la investigación, reconociendo que la elección entre ambos enfoques no es universal sino contextual: depende del tamaño de la muestra disponible, los recursos computacionales, los objetivos clínicos específicos y las preferencias de los especialistas que interactuarán con el sistema.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		38
--	--	----

Consideraciones Éticas, Regulatorias y de Validación Clínica

La implementación de modelos de aprendizaje automático en radioterapia exige cumplimiento normativo estricto y un proceso de validación riguroso. El uso de datasets anonimizados garantiza la protección de datos personales conforme a marcos regulatorios internacionales como HIPAA en Estados Unidos y GDPR en Europa. La anonimización no es únicamente un requisito legal, sino una condición ética para el uso responsable de datos de pacientes en investigación.

Goodfellow et al. (2016) enfatizan que los modelos deben someterse a validación retrospectiva, validación cruzada y evaluación independiente antes de cualquier aplicación clínica. En el contexto específico de la radioterapia, la validación debe incluir no solo métricas estadísticas de error, sino también evaluación clínica de la aceptabilidad de los planes predichos: ¿cumplen las restricciones dosimétricas de los OAR? ¿Garantizan la cobertura adecuada del PTV? ¿Son clínicamente implementables?

Rajpurkar et al. (2022) advierten que uno de los riesgos más importantes en la adopción de inteligencia artificial en medicina es la sobreestimación del rendimiento en entornos de laboratorio frente a la práctica clínica real. Los modelos entrenados en datos de un centro pueden no generalizarse adecuadamente a pacientes de otros centros con protocolos diferentes, equipos distintos o poblaciones con características demográficas divergentes. La validación externa es, por tanto, un requisito antes de cualquier despliegue clínico.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		39
--	--	----

La inteligencia artificial debe concebirse como herramienta de apoyo al especialista, no como reemplazo de la decisión médica. En el contexto de la planificación radioterapéutica, el médico oncólogo radioterápico y el físico médico conservan la responsabilidad clínica y ética sobre el plan de tratamiento. Los modelos predictivos pueden reducir la carga de trabajo, disminuir la variabilidad interplanificador y servir como punto de partida para la optimización, pero su output debe ser siempre revisado y aprobado por el equipo clínico competente.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		40
--	--	----

Hipótesis

Trabajar con dos tipos de datos tan distintos generó desde el inicio preguntas concretas sobre qué implicaciones tendría cada elección para el proceso de modelamiento. Por un lado, variables clínicas agregadas en formato tabular; por el otro, tensores volumétricos tridimensionales. La revisión de la literatura confirmó que esa tensión entre simplicidad operativa y riqueza espacial no está resuelta, y que las diferencias entre paradigmas van más allá del desempeño predictivo. A partir de ese análisis se formularon dos hipótesis de trabajo que orientan el diseño y la interpretación de los resultados.

La primera hipótesis se centra en la complejidad del pipeline: cuánto cuesta computacionalmente, qué infraestructura exige y cuántos datos necesita cada enfoque para funcionar. La segunda aborda lo que el modelo produce: si es posible entender sus decisiones o si opera como una caja negra. Ambas son verificables con la evidencia recopilada durante el trabajo de campo.

H1. El paradigma de representación volumétrica tridimensional exige un pipeline de modelamiento sustancialmente más complejo que el paradigma tabular estructurado, en términos computacionales, de infraestructura y de volumen de datos requerido para el entrenamiento.

H2. El paradigma de representación tabular estructurada ofrece mayor interpretabilidad del modelo resultante que el paradigma volumétrico tridimensional, a costa de una menor capacidad para capturar relaciones espaciales complejas entre estructuras anatómicas.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		41
--	--	----

Verificar estas hipótesis no busca establecer la superioridad de un paradigma sobre otro. El propósito está más orientado a entender en qué condiciones cada enfoque es la elección correcta, reconociendo que esa decisión depende del contexto, del tamaño de la muestra disponible, de los recursos computacionales del entorno y de los objetivos específicos del proceso de planificación.

Variables

Tras la revisión de la literatura en planificación radioterapéutica y aprendizaje automático aplicado a imágenes médicas, se definieron las variables del estudio considerando la doble naturaleza del enfoque comparativo: representación tabular estructurada y representación volumétrica tridimensional.

Dado que el objetivo de la investigación consiste en evaluar cómo influye la estructura de los datos en la capacidad predictiva de cada modelo, las variables se organizan en tres categorías principales:

1. Variables de entrada para el modelo tabular.
2. Variables de entrada para el modelo volumétrico.
3. Variables objetivo y métricas de evaluación.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		42
--	--	----

1. Variables de entrada – Enfoque Tabular

Estas variables corresponden a características estructuradas agregadas derivadas del dataset HNSCC-3DCT-RT. Se emplean como predictores en el modelo de aprendizaje automático basado en representación tabular.

1.1 Estadio del tumor (TNM)

Corresponde a la clasificación clínica del cáncer basada en el sistema TNM (American Joint Committee on Cancer, 2017), que evalúa:

- Tamaño del tumor primario (T)
- Afectación ganglionar (N)
- Presencia de metástasis (M)

Clasificación: Variable cualitativa ordinal.

Uso en el modelo: Variable explicativa que permite incorporar información clínica estructurada en el espacio tabular.

1.2 Edad del paciente

Edad cronológica del paciente al momento del tratamiento.

Clasificación: Variable cuantitativa continua.

Uso en el modelo: Variable predictora auxiliar que puede correlacionarse con decisiones de planificación o características anatómicas.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		43
--	--	----

1.3 Género

Sexo biológico del paciente registrado en la historia clínica.

Clasificación: Variable cualitativa nominal.

1.4 Número de fracciones

Cantidad total de sesiones de radioterapia administradas.

Clasificación: Variable cuantitativa discreta.

Uso en el modelo: Variable estructurada relevante para caracterizar el esquema terapéutico.

1.5 Dosis total prescrita (Gy)

Cantidad total de energía depositada en el volumen tumoral durante el tratamiento, medida en Grays (Gy).

Clasificación: Variable cuantitativa continua.

Nota metodológica: En el enfoque comparativo actual, esta variable puede funcionar como variable predictiva en el modelo tabular, dependiendo del diseño experimental.

2. Variables de entrada – Enfoque Volumétrico

En el modelo volumétrico, el espacio de entrada no se define como un conjunto de variables estructuradas independientes, sino como tensores tridimensionales que preservan la anatomía del paciente.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		44
--	--	----

2.1 Imagen CT tridimensional

Volumen de tomografía computarizada del paciente, compuesto por voxeles con valores de intensidad (Hounsfield Units).

Clasificación: Variable cuantitativa tridimensional (tensor 3D).

Uso en el modelo: Entrada principal para la red neuronal convolucional 3D.

2.2 Máscara del PTV

Volumen binario tridimensional que identifica la localización espacial del tumor dentro de la imagen CT.

Clasificación: Variable categórica tridimensional (tensor binario 3D).

2.3 Máscaras de órganos en riesgo (OAR)

Volúmenes binarios tridimensionales que representan estructuras críticas delineadas en el sistema de planificación.

Clasificación: Variable categórica tridimensional.

Uso en el modelo: Permite que la red neuronal incorpore relaciones geométricas entre tumor y órganos circundantes.

3. Variables objetivo

Dado el enfoque comparativo, las variables objetivo difieren según el modelo.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		45
--	--	----

3.1 Variable objetivo en el modelo tabular

De acuerdo con los resultados obtenidos, corresponde a:

- Volumen del PTV

Clasificación: Variable cuantitativa continua.

3.2 Variable objetivo en el modelo volumétrico

Distribución espacial tridimensional de dosis (RTDOSE).

Clasificación: Tensor tridimensional continuo.

Unidad de medida: Grays (Gy).

Esta variable permite evaluar el desempeño de la red neuronal 3D en términos de error voxel a voxel.

4. Métricas de evaluación (Variables derivadas)

Dado que la investigación es comparativa pero los dos modelos predicen variables de naturaleza fundamentalmente distinta el modelo tabular predice una métrica agregada (volumen del PTV en cc) mientras que el modelo volumétrico predice una distribución espacial tridimensional de dosis (Gy por voxel) no es posible establecer una comparación directa entre sus métricas de desempeño. Cada enfoque es evaluado con las métricas apropiadas a su tipo de variable objetivo.

Para el modelo tabular se emplean métricas de regresión estándar: MAE (error absoluto medio en cc) y R^2 (coeficiente de determinación), que permiten cuantificar qué tan bien el

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		46
--	--	----

modelo estima el volumen tumoral a partir de variables clínicas y demográficas. (Chicco et al., 2021).

Para el modelo volumétrico se emplean métricas voxel a voxel: MAE (en Gy), MSE (en Gy²) y el Dose Score definido por el benchmark OpenKBP, que evalúa el error absoluto medio sobre los voxeles con dosis real mayor a cero. Esta última métrica permite contextualizar el desempeño del modelo frente al estado del arte internacional. (Babier et al., 2021)

La comparación entre paradigmas se realiza, por tanto, no sobre las métricas de error en sí, sino sobre dimensiones metodológicas comparables: complejidad del pipeline, interpretabilidad del modelo, requerimientos de datos e infraestructura, y capacidad para capturar relaciones espaciales.

4.1 Error Absoluto Medio (MAE)

Mide la diferencia promedio absoluta entre valores predichos y reales.

Clasificación: Variable cuantitativa continua.

4.2 Error Cuadrático Medio (MSE)

Penaliza errores grandes y permite comparar estabilidad del modelo.

Clasificación: Variable cuantitativa continua.

5. Variable independiente principal

Tipo de representación de datos:

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		47
--	--	----

- 0 = Representación tabular estructurada
- 1 = Representación volumétrica tridimensional

Clasificación: Variable cualitativa dicotómica.

Esta variable constituye el eje central del estudio, ya que permite evaluar el efecto estructural de la representación en la capacidad predictiva.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		48
--	--	----

Metodología

Enfoque de la investigación

El enfoque de esta investigación es cuantitativo, dado que se fundamenta en la medición objetiva del desempeño de modelos de aprendizaje automático entrenados sobre dos tipos distintos de representación de datos: tabular estructurada y volumétrica tridimensional, evaluando el impacto de dicha representación en el proceso de modelamiento, la complejidad computacional y la interpretabilidad de los modelos resultantes.

La naturaleza del problema requirió análisis numéricos reproducibles, ya que el objetivo no era intervenir clínicamente en los tratamientos, sino evaluar el desempeño de los modelos dentro de cada paradigma mediante métricas cuantitativas como MAE, MSE, R2 y error voxel a voxel, y analizar cómo la estructura del dato condiciona el proceso de modelamiento.

El estudio se apoya en aprendizaje supervisado, donde el modelo aprende relaciones entre variables de entrada y salidas conocidas a partir de datasets históricos anonimizados. El análisis de representaciones permite establecer relaciones objetivas entre estructura del dato y comportamiento del modelo, evitando interpretaciones subjetivas.

Este enfoque resulta indispensable en contextos médicos, donde la evidencia debe ser cuantificable, replicable y estadísticamente validada.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		49
--	--	----

Diseño de la investigación

El diseño adoptado es no experimental y retrospectivo. No experimental, porque no se manipularon variables clínicas ni se intervinieron en tratamientos reales. Se trabajó exclusivamente con datasets históricos previamente recolectados y anonimizados.

Retrospectivo, porque el análisis se realizó sobre datos ya existentes correspondientes a pacientes tratados con radioterapia IMRT.

Además, el diseño es comparativo a nivel metodológico, dado que se evaluaron dos enfoques de modelamiento distintos, representación tabular estructurada y representación volumétrica tridimensional, bajo condiciones controladas de entrenamiento y validación.

No se buscó establecer causalidad clínica, sino evaluar estructuras de representación de datos y su efecto sobre la capacidad predictiva.

Tipo de estudio

El estudio se clasifica como:

- Descriptivo: porque caracteriza las propiedades estructurales y estadísticas de los datasets utilizados.
- Correlacional: porque analiza relaciones entre variables dentro del enfoque tabular.
- Comparativo de pipelines: porque evalúa y contrasta los procesos de modelamiento asociados a cada paradigma de representación de datos

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		50
--	--	----

Fases de la investigación

Fase 1: Revisión del estado del arte

- Revisión bibliográfica sobre IMRT, modelamiento predictivo en radioterapia y aplicaciones de deep learning en imágenes médicas.
- Identificación de estudios que empleen representaciones tabulares en predicción clínica.
- Revisión de investigaciones que utilicen CNN 3D para predicción espacial de dosis.
- Identificación de brechas en estudios comparativos entre tipos de representación.

Fase 2: Caracterización de datasets

Se trabajó con dos datasets distintos:

Dataset 1: HNSCC-3DCT-RT

Utilizado para el enfoque tabular estructurado.

Contiene:

- Variables clínicas.
- Métricas dosimétricas agregadas.
- Información estructurada.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		51
--	--	----

Dataset 2: OpenKBP (AAPM Grand Challenge)

Utilizado para el enfoque volumétrico tridimensional.

Contiene:

- Imágenes CT tridimensionales.
- Máscaras estructurales del PTV y OAR.
- Distribuciones espaciales de dosis.

En esta fase se realizó:

- Análisis de dimensionalidad.
- Identificación de tipo de variables.
- Evaluación de nivel de granularidad.
- Análisis de completitud y consistencia.

Fase 3: Preprocesamiento de datos

Para el enfoque tabular se realizó lo siguiente:

- Imputación de valores faltantes.
- Normalización y estandarización de variables continuas.
- Codificación de variables categóricas.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		52
--	--	----

- Análisis de colinealidad.
- División en entrenamiento, validación y prueba.

Para el enfoque volumétrico se realizó:

- Normalización de intensidades CT.
- Redimensionamiento de volúmenes.
- Alineación espacial.
- Construcción de tensores 3D.
- División en conjuntos de entrenamiento, validación y prueba.

Se aplicaron técnicas para mitigar desbalance y sobreajuste cuando sea necesario.

Fase 4: Modelado predictivo

Pipeline 1 – Modelo Tabular

Se entrenaron modelos de aprendizaje supervisado sobre las variables clínicas y demográficas del dataset HNSCC-3DCT-RT. Específicamente se implementaron regresión lineal simple como modelo de referencia y Random Forest como modelo principal, al cual se aplicó transformación logarítmica sobre la variable objetivo y validación cruzada de 5 pliegues para evaluar su estabilidad.

Variable objetivo: volumen del PTV (cc).

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		53
--	--	----

Métricas de evaluación: MAE y R².

Pipeline 2 – Modelo Volumétrico

- Se implementó una red neuronal convolucional tridimensional de arquitectura UNet3D sobre el dataset OpenKBP, recibiendo como entrada volúmenes CT, máscaras del PTV y máscaras de órganos en riesgo, con salida correspondiente a la distribución espacial tridimensional de dosis.
- Métricas de evaluación: MAE voxel a voxel, MSE y Dose Score según la metodología oficial del benchmark OpenKBP.

Fase 5: Comparación estadística

Para validar las hipótesis se realizó:

- Comparación metodológica de ambos pipelines en términos de proceso de ingeniería de datos, arquitectura del modelo, requerimientos computacionales e interpretabilidad del modelo resultante.
- No se realiza comparación estadística directa entre métricas debido a la diferencia en la naturaleza de las variables objetivo.

Población y muestra

La población de referencia corresponde a pacientes diagnosticados con carcinoma de cabeza y cuello tratados con técnicas de radioterapia de intensidad modulada (IMRT).

Dado que la investigación se desarrolla a partir de datos históricos anonimizados y no implica intervención clínica directa, la muestra fue definida en función de los conjuntos de

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		54
--	--	----

datos disponibles que cumplieran con criterios de completitud, representatividad clínica y accesibilidad ética.

La muestra está constituida por dos fuentes de datos complementarias, seleccionadas según el paradigma de representación evaluado en cada enfoque:

- Dataset HNSCC-3DCT-RT: empleado para el enfoque de representación tabular estructurada. Contiene registros clínicos, demográficos y dosimétricos de pacientes con carcinoma escamocelular de cabeza y cuello, organizados en formato estructurado.
- Dataset OpenKBP (Open Knowledge-Based Planning): Este dataset fue desarrollado en el marco del *AAPM OpenKBP Grand Challenge* (Babier et al., 2021), una competencia internacional orientada a la predicción de distribuciones de dosis en radioterapia mediante modelos de aprendizaje automático. Contiene imágenes CT, máscaras de estructuras anatómicas y distribuciones de dosis de referencia de 340 pacientes tratados con IMRT, distribuidos en conjuntos de entrenamiento (200), validación (40) y prueba (100). El muestreo es no probabilístico por conveniencia, sustentado en la disponibilidad pública, la anonimización previa y la pertinencia clínica de ambos conjuntos de datos para los objetivos de la investigación.

Criterios de Inclusión y Exclusión

Criterios de inclusión:

- Registros con datos completos en las variables requeridas por cada pipeline.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		55
--	--	----

- Presencia de estructuras anatómicas delineadas (PTV y OAR).
- Información dosimétrica disponible y verificable.
- Archivos en formato estándar procesable (DICOM o CSV sparse).

Criterios de exclusión:

- Registros con datos faltantes en variables clave no imputables.
- Datos con inconsistencias clínicas o de codificación no corregibles.
- Archivos corruptos o con estructuras no procesables.

Limitantes de los Datos Abiertos

La elección de datasets de acceso abierto como fuente principal respondió a restricciones prácticas del ecosistema de datos clínicos, cuya naturaleza se analiza en profundidad en la sección de Discusión. En términos metodológicos, esta decisión implicó asumir limitantes específicas que condicionaron el alcance del estudio.

Los datos abiertos generalmente provienen de un número reducido de centros de radioterapia, lo que introduce un sesgo de distribución institucional: los protocolos de delineación, los sistemas de planificación y los esquemas de fraccionamiento pueden diferir entre instituciones, afectando la capacidad de generalización de los modelos a contextos clínicos distintos (Babier et al., 2021).

Adicionalmente, la anonimización estricta requerida para la publicación abierta de datos clínicos implica la eliminación de variables que, si bien comprometían la privacidad del paciente, tenían potencial valor predictivo como estadio tumoral detallado, histología, comorbilidades o respuesta al tratamiento. Esta tensión entre protección de datos y

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		56
--	--	----

riqueza informativa es una restricción estructural del uso de datos abiertos en ciencias de la salud (Rajpurkar et al., 2022), y sus implicaciones sobre los resultados obtenidos se discuten en la sección correspondiente.

Diseño y validación del instrumento de medición

- Instrumento: El principal instrumento fue un dataset estructurado construido con base en registros clínicos, imagenología y planes de tratamiento. Este integra variables cuantitativas y cualitativas previamente definidas.
- Diseño:
 - Herramientas de preprocesamiento en Python (pandas, numpy, scikit-learn).
- Validación:
 - Prueba piloto con subconjunto de datos para verificar consistencia.
 - Verificación de integridad y consistencia interna de los registros.

Procedimientos y técnicas de análisis de la información

- Procedimientos:
 - Descarga y preparación de datasets públicos anonimizados.
 - Procesamiento, limpieza y estructuración del dataset.
 - Entrenamiento y evaluación de modelos de predicción.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		57
--	--	----

- Validación cruzada y afinación de parámetros.

Técnicas de análisis:

- Estadística descriptiva: promedios, desviaciones, frecuencias.
- Análisis correlacional: coeficientes de Spearman.
- Modelos de machine learning: regresión lineal simple y Random Forest para el enfoque tabular; UNet3D para el enfoque volumétrico.
- Visualización de datos: Gráficos de dispersión, de barras y líneas
- Consideraciones éticas y limitaciones

Este estudio cumple con las normativas internacionales (HIPAA, GDPR) y locales (Ley de Habeas Data en Colombia) para garantizar la confidencialidad y seguridad de la información. Todos los datos fueron anonimizados, eliminando identificadores personales.

Posibles limitaciones incluyen:

- El tamaño de la muestra, que puede limitar la generalización.
- La heterogeneidad de datos entre diferentes hospitales o máquinas de radioterapia.

Para mitigar estas limitaciones, se utilizaron técnicas de balanceo de datos y validación cruzada.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		58
--	--	----

Trabajo de Campo

En este capítulo se describe de forma detallada y sistemática el proceso seguido para la construcción de los modelos predictivos implementados en cada paradigma de representación: el modelo de regresión basado en datos tabulares estructurados (HNSCC-3DCT-RT) y el modelo volumétrico tridimensional basado en imágenes (OpenKBP). Este apartado cumple un doble propósito: documentar rigurosamente el procedimiento técnico realizado y proporcionar al lector una explicación clara que permita comprender y reproducir los modelos de manera independiente y poder compararlos en sus características metodológicas.

Enfoque con datos de HNSCC-3DCT-RT

En la primera parte se utilizó un conjunto de datos proveniente de Hugging Face, específicamente la sección HNSCC-3DCT-RT, seleccionada por su pertinencia clínica y por contener las variables estructuradas necesarias para el desarrollo del enfoque tabular de este estudio.

El dataset está organizado en formato estándar DICOM (Digital Imaging and Communications in Medicine), la norma internacional utilizada para almacenar, transmitir y gestionar imágenes médicas y su información asociada. En su estructura completa, cada paciente incluye archivos CT (imagen de tomografía computarizada), RTSTRUCT (estructuras delineadas), RTPLAN (plan de tratamiento) y RTDOSE (distribución de dosis calculada).

Sin embargo, para los propósitos del enfoque tabular de esta investigación, no se utilizaron las imágenes volumétricas DICOM. El procesamiento se realizó exclusivamente

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		59
--	--	----

sobre la información clínica y demográfica disponible en formato Excel, que concentra variables estructuradas como edad, género, estadio tumoral, volumen del PTV, dosis total prescrita, número de fracciones y evolución del peso durante el tratamiento. Esta decisión es coherente con la naturaleza del paradigma tabular, que representa la anatomía del paciente mediante métricas agregadas y no requiere el procesamiento de volúmenes tridimensionales. La información imagenológica del dataset queda disponible para estudios futuros que adopten un enfoque volumétrico sobre esta misma cohorte.

Las carpetas RTPLAN y RTDOSE pueden incluir múltiples planes por paciente. Además, se proporciona un paciente de ejemplo para verificación y análisis rápido. El conjunto de datos completo se comparte en archivos comprimidos (.zip) para facilitar el acceso y la escalabilidad.

Archivos adicionales incluidos:

PTV_DICT.json: Prescripciones de dosis de los PTV (por paciente).

Pat_Obj_DICT.json: Estructuras utilizadas en la optimización de Eclipse (por paciente).

GDP-HMM_train_valid_splits.json: Divisiones de entrenamiento y validación utilizadas en el desafío GDP-HMM.

Para el desarrollo de este estudio, se ha utilizado únicamente la porción HNSCC-DCT-RT, dado que ofrece una muestra representativa y contiene las variables requeridas para el análisis y modelado.

1. Preparación del Entorno y Acceso a la Información

1.1 Configuración del entorno en Colab

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		60
--	--	----

La herramienta principal utilizada para el procesamiento, análisis y modelamiento fue

Google Colab, debido a:

- Disponibilidad de GPU y CPU gratuitas,
- Facilidad de acceso a bibliotecas científicas,
- Integración directa con Google Drive,
- Capacidad de ejecutar celdas independientemente y mantener un historial reproducible.

En la primera celda del cuaderno se importaron las librerías fundamentales:

pandas: manipulación y limpieza de datos

numpy: operaciones matemáticas

matplotlib / seaborn: visualización

sklearn: construcción del modelo de regresión

drive API: acceso a datos almacenados en Drive

Estas librerías constituyen la base de todo el flujo de trabajo.

1.2 Montaje del repositorio de datos

Dado que la base de datos original se encontraba alojada en una estructura de carpetas

en Google Drive, se realizó el montaje mediante:

```
drive.mount('/content/drive')
```

Con esto se obtuvo acceso directo a las tablas clínicas, manteniendo integridad y evitando duplicación de archivos.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		61
--	--	----

2. Adquisición, Selección y Comprensión Inicial del Dataset

2.1 Carga del archivo maestro

El archivo principal, un documento Excel con múltiples hojas, fue cargado completamente utilizando:

```
pd.read_excel(file_path, sheet_name=None)
```

Este mecanismo permite visualizar y explorar simultáneamente todas las hojas del libro, fundamentales para conocer el alcance real de la información disponible.

2.2 Identificación de las hojas relevantes

Tras revisar cada hoja, se seleccionaron aquellas con información imprescindible para alimentar el modelo:

- **Table 1 CT Date**
- **Table 2 Demographics**
- **Table 5 PTV**
- **Table 6 Dose_fx**
- **Table 7 Outcomes**
- **Table 8 Weight**

Estas hojas contenían la información clínica y demográfica necesaria para predecir la variable objetivo-seleccionada.

La selección respondió a criterios de:

- Completitud de los registros,

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		62
--	--	----

- Pertinencia clínica
- Ausencia de ruido o variables no aplicables.

3. Procesamiento, Limpieza y Organización de los Datos

3.1 Exploración inicial

Se verificó:

- Número total de pacientes,
- Proporción de datos faltantes,
- Tipos de datos por columna,
- Consistencia de las variables numéricas y categóricas.

Esta exploración permitió anticipar posibles transformaciones y evitar errores durante el modelamiento.

3.2 Generación de estadísticas descriptivas

Antes de entrenar el modelo, se realizaron cálculos que permitieron caracterizar la población:

- Distribución de la edad
- Total de registros válidos
- Rangos clínicos relevantes

Los valores fueron impresos directamente en Colab para documentar trazabilidad en el proceso investigativo.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		63
--	--	----

3.3 Transformación de variables

Dado que los modelos de regresión requieren variables numéricas:

Las variables categóricas fueron transformadas mediante técnicas como:

- **One-Hot Encoding,**
- **Label Encoding,** según el caso.

Después de análisis exploratorio, se identificó la necesidad de normalizar las variables debido a escalas heterogéneas con el fin de tener una mejor estabilidad numérica del modelo, evita que las variables con magnitudes mayores dominen artificialmente la función de pérdida y el cálculo de los gradientes sobre variables de menor escala. Al escalar los datos a un rango uniforme, se asegura una convergencia más rápida del optimizador y una ponderación equitativa de cada característica en la predicción del PTV.

3.4 Selección de la variable objetivo

La variable objetivo-seleccionada para el modelo tabular fue el volumen del PTV (Planning Target Volume), expresado en centímetros cúbicos (cc). Esta elección se fundamenta en tres criterios complementarios.

La elección del volumen del PTV como variable objetivo responde a tres criterios que se refuerzan mutuamente.

En primer lugar, su relevancia clínica es indiscutible. El volumen tumoral condiciona prácticamente todas las decisiones del plan de tratamiento: la extensión del campo irradiado, la complejidad de la modulación del haz y los márgenes de seguridad alrededor

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		64
--	--	----

de los órganos en riesgo. Poder estimarlo a partir de variables clínicas y demográficas disponibles antes del tratamiento tendría valor operativo real en entornos hospitalarios con recursos limitados.

En segundo lugar, el análisis exploratorio del dataset HNSCC-3DCT-RT mostró que es la variable con mayor dispersión entre todas las disponibles, con valores que van desde 10 hasta más de 700 cm³. Esa amplitud de rango es precisamente lo que la hace más informativa para un modelo de regresión supervisada, porque le ofrece al algoritmo la variabilidad necesaria para aprender patrones diferenciadores entre perfiles de pacientes.

En tercer lugar, y quizás lo más relevante desde la perspectiva de este estudio, su elección es coherente con el OE2 y con la lógica del paradigma tabular. El volumen del PTV es una variable continua que en la realidad clínica se obtiene a partir de imágenes volumétricas, pero que en este estudio se intenta estimar exclusivamente desde variables agregadas como edad, estadio tumoral, peso y dosis prescrita. Esa es exactamente la pregunta que el paradigma tabular puede y debe responder: hasta qué punto la información clínica estructurada captura dimensiones que en su origen son espaciales y geométricas.

4. División del Dataset para Entrenamiento y Validación

Para garantizar que el modelo fuera evaluado correctamente se utilizó:

`train_test_split()`

Se dividió el dataset de la siguiente forma:

- 80% entrenamiento
- 20% validación

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		65
--	--	----

La separación asegura que el modelo no memorice los datos y permita evaluar su capacidad real de generalización.

Procesamiento de los datos

Una vez establecida la estructura metodológica inicial descrita en las secciones anteriores, y completados los procesos de carga, depuración y segmentación del conjunto de datos HN5CC-3DCT-RT, se desarrolló la fase aplicada del trabajo de campo. Esta etapa tuvo como propósito examinar de manera sistemática las características clínicas y dosimétricas de la población estudiada mediante herramientas estadísticas y visualizaciones descriptivas, con el fin de identificar patrones relevantes y generar evidencia empírica que respalde el diseño del modelo de aprendizaje supervisado propuesto en esta investigación.

El trabajo de campo se enfocó en integrar las variables clínicas provenientes de múltiples hojas del dataset, consolidarlas en una representación estructurada por paciente, y posteriormente analizar sus distribuciones, relaciones y posibles asociaciones con elementos del plan de tratamiento. Las gráficas generadas en esta fase permiten comprender mejor el comportamiento individual y colectivo de los pacientes incluidos en el estudio, así como fundamentar con datos objetivos la pertinencia del uso de modelos predictivos en el contexto de la planificación de radioterapia IMRT.

Después de efectuar la exploración inicial, se consolidó un dataset integrado que aglutina las variables derivadas de las hojas **Demographics**, **PTV**, **Dose_fx**, **Outcomes** y **Weight**. Este proceso incluyó diversas etapas técnicas:

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		66
--	--	----

Integración horizontal por “Patient ID”

Dado que la información clínica se encontraba distribuida en múltiples hojas, se procedió a unir las mediante claves únicas correspondientes a cada paciente. Este proceso evitó duplicidades y garantizó una vista consolidada y coherente del registro clínico individual.

Normalización y limpieza adicional

En esta fase se corrigieron errores residuales de formato, se validaron rangos de variables clínicas y se completaron valores faltantes mediante criterios estadísticos cuando fue pertinente. Variables como *Total Dose (Gy)*, *Number of fractions*, *Dose per fraction* y *PTV volume* fueron convertidas a tipos numéricos apropiados.

Construcción de variables derivadas

Con el propósito de enriquecer la base de datos para análisis descriptivos y posteriormente modelamiento, se generaron nuevas variables:

- **Dosis por fracción (Gy)**, calculada como *Total Dose / Number of fx*.
- **Duración del tratamiento**, mediante integración de los tiempos *Pre-Tx*, *Mid-Tx* y *Post-Tx*.
- **Pérdida de peso**, expresada como diferencia porcentual entre el peso pre y post-tratamiento.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		67
--	--	----

Longitudinalización del peso

Dado que la evolución del peso es clínicamente crítica en pacientes con cáncer de cabeza y cuello, se estructuró esta variable en formato longitudinal, permitiendo su representación gráfica y análisis comparativo entre etapas del tratamiento.

Una vez finalizada esta preparación, se procedió al análisis estadístico y visualización de resultados, cuyos hallazgos se describen a continuación.

Perspectiva desde Ciencia de Datos

Transformar el peso de una medición puntual en una variable de trayectoria no fue una decisión trivial. Desde la Ciencia de Datos, representa una elección deliberada de feature engineering con consecuencias directas sobre qué puede aprender el modelo.

El dataset original registraba el peso del paciente en tres momentos del tratamiento: antes de iniciar (Pre-Tx), a mitad del proceso (Mid-Tx) y al finalizar (Post-Tx). Dejar esas tres mediciones como variables independientes habría sido la opción más simple, pero habría perdido lo más valioso: la dinámica de cambio. Al estructurarlas como una trayectoria temporal, se incorpora al modelo información que una sola cifra no puede capturar, por ejemplo, si el paciente perdió peso de forma acelerada en la primera mitad del tratamiento o si la pérdida fue gradual y sostenida. Esos patrones tienen implicaciones reales sobre la geometría del plan. En la literatura reciente, este tipo de representación se asocia con el modelamiento de datos longitudinales y series temporales clínicas, donde la incorporación explícita de la dinámica temporal ha demostrado mejorar el desempeño predictivo frente a representaciones estáticas (Shickel et al., 2018; Lim & Zohren, 2021; Zhang et al., 2022).

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		68
--	--	----

También se habilitó la construcción de variables derivadas como la tasa de pérdida de peso por etapa, que sintetiza en un solo número la intensidad del cambio anatómico durante el tratamiento. Este tipo de transformación ilustra algo que Goodfellow et al. (2016) plantean con claridad: el feature engineering en datos tabulares no es un paso técnico menor sino una decisión epistemológica sobre qué dimensiones del fenómeno se consideran relevantes para el modelo. Decidir qué transformar, cómo hacerlo y qué nueva variable construir es, en sí mismo, conocimiento aplicado. Trabajos recientes en aprendizaje automático sobre datos tabulares han reforzado esta idea, mostrando que el diseño de características sigue siendo determinante incluso en presencia de modelos avanzados, especialmente en dominios con datos limitados o altamente estructurados como el clínico (Grinsztajn et al., 2022; Borisov et al., 2024).

Análisis de resultados

El análisis de resultados se estructura con base en las visualizaciones generadas a partir del dataset procesado. Cada gráfica fue interpretada desde dos perspectivas complementarias: sus implicaciones clínicas en el contexto de la planificación radioterapéutica IMRT, y sus implicaciones para el proceso de modelamiento tabular. Esta doble lectura permite identificar no solo patrones relevantes sobre la población tratada, sino también qué revela cada distribución sobre la estructura del dato: su variabilidad, su concentración, sus relaciones con otras variables y cómo estas características condicionan las decisiones de feature engineering, como la selección de predictores, la necesidad de transformaciones y el tratamiento de valores atípicos.

1. Evolución del peso a lo largo del tratamiento

La primera visualización se centró en la trayectoria del peso de los pacientes durante tres momentos clave: pretratamiento, mitad del tratamiento y postratamiento en la

Evolución del peso de los pacientes durante el tratamiento

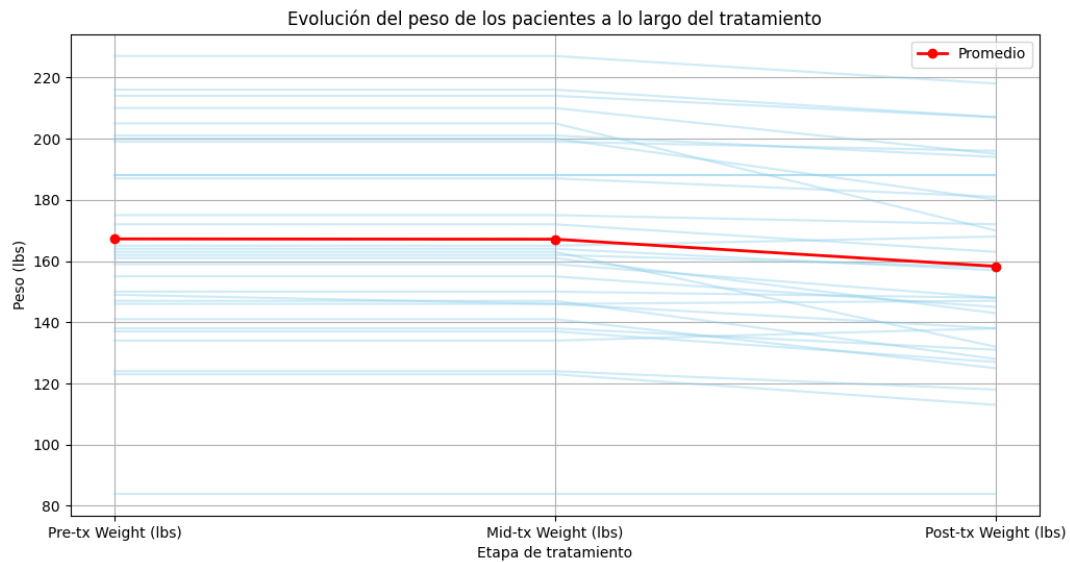


Figura 1. *Evolución del peso de los pacientes durante el tratamiento*

Esta gráfica revela varios hallazgos clínicos importantes:

Tendencia descendente generalizada:

El peso promedio pasó de aproximadamente 167 lbs antes de iniciar la radioterapia a cerca de 158 lbs al finalizar el tratamiento. Esta reducción progresiva está en línea con las toxicidades típicamente asociadas a los tratamientos de cabeza y cuello, incluyendo mucositis, disfagia y dificultades para la ingesta calórica suficiente. (Langius et al., 2013)

Variabilidad interindividual:

Las líneas individuales muestran comportamientos heterogéneos. Algunos pacientes

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		70
--	--	----

presentan descensos pronunciados; otros mantienen un peso relativamente estable. Esta variabilidad es relevante para la planificación IMRT, dado que cambios anatómicos sustanciales pueden alterar la geometría del plan y justificar ajustes adaptativos.

Implicación para el modelamiento tabular:

La alta variabilidad interindividual en la trayectoria del peso justifica su inclusión como variable de entrada y la construcción de variables derivadas de trayectoria. La heterogeneidad observada introduce complejidad en el espacio de características que el modelo debe aprender a discriminar.

Importancia para modelos predictivos:

Dado que el peso influye directamente en la distribución real de dosis, su inclusión como variable en modelos supervisados es fundamental para anticipar posibles desajustes dosimétricos.

2. Distribución de la dosis total (Total Dose – Gy)

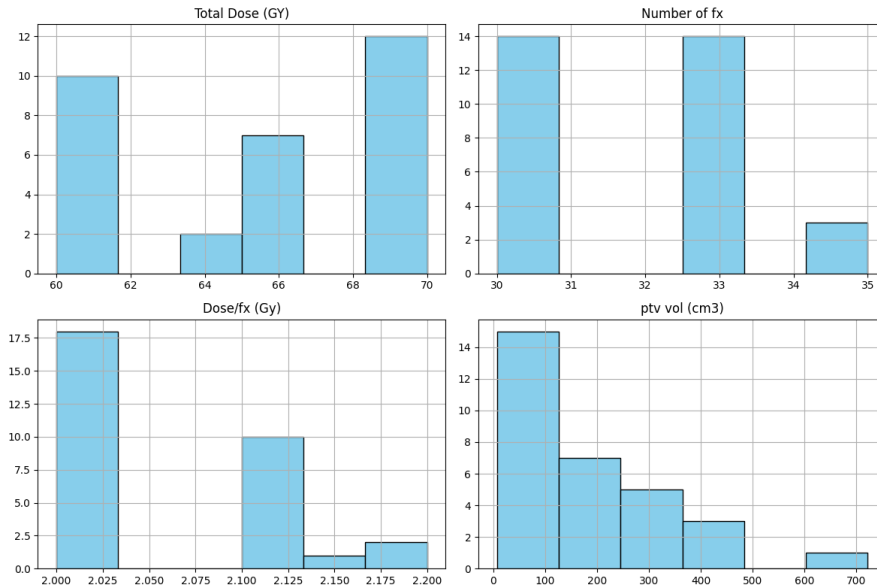


Figura 2. Histograma de dosis total (Gy)

Los valores de dosis total se concentraron principalmente en rangos clínicamente convencionales para cáncer de cabeza y cuello (60–70 Gy) (NCCN, 2024). Este comportamiento refleja la homogeneidad en prescripciones y estandarización de protocolos clínicos. La ligera dispersión observada sugiere que los médicos adaptaron dosis según características tumorales individuales.

Desde la perspectiva del modelamiento tabular, la concentración de valores en estos rangos discretos reduce la varianza de esta variable, lo que facilita el aprendizaje supervisado al proporcionar patrones consistentes y repetibles durante el entrenamiento. Sin embargo, esta misma característica puede limitar la capacidad del modelo para generalizar a prescripciones atípicas que se alejen de los esquemas convencionales, como tratamientos hipofraccionados o dosis de rescate. Esta propiedad estructural del

dato debe considerarse al interpretar el desempeño del modelo y al definir el alcance clínico de sus predicciones.

Número de fracciones (Number of fx)

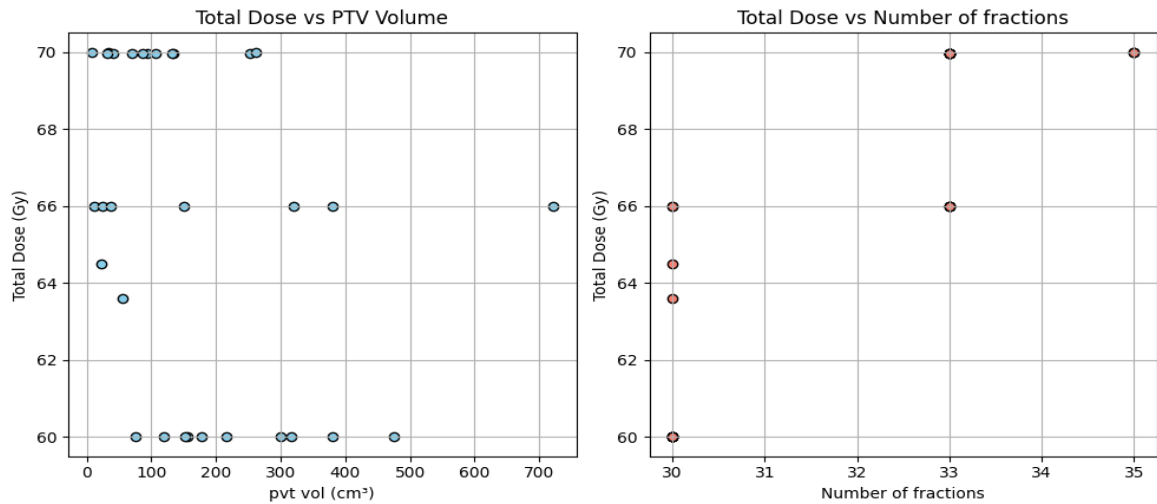


Figura 3. Histograma de número de fracciones

Los tratamientos se distribuyeron mayoritariamente entre 30 y 35 fracciones, lo cual es consistente con fraccionamientos convencionales para patologías de cabeza y cuello (Bentzen & Bernier, 2006). La concentración en estos rangos refuerza la predictibilidad clínica del esquema terapéutico, hecho relevante para el modelado supervisado.

Desde la perspectiva del modelamiento tabular, la baja dispersión de esta variable sugiere que aportará poco poder discriminativo como predictor independiente, dado que su escasa variabilidad ofrece poca información diferencial entre pacientes. Esto justifica su uso en combinación con otras variables o como componente de variables derivadas como la dosis por fracción, calculada como $\text{Total Dose} / \text{Number of fx}$ donde su valor predictivo se potencia al interactuar con la magnitud de la dosis prescrita. Esta

consideración orientó las decisiones de feature engineering descritas en la sección de procesamiento de datos.

Dosis por fracción (Dose/fx – Gy)

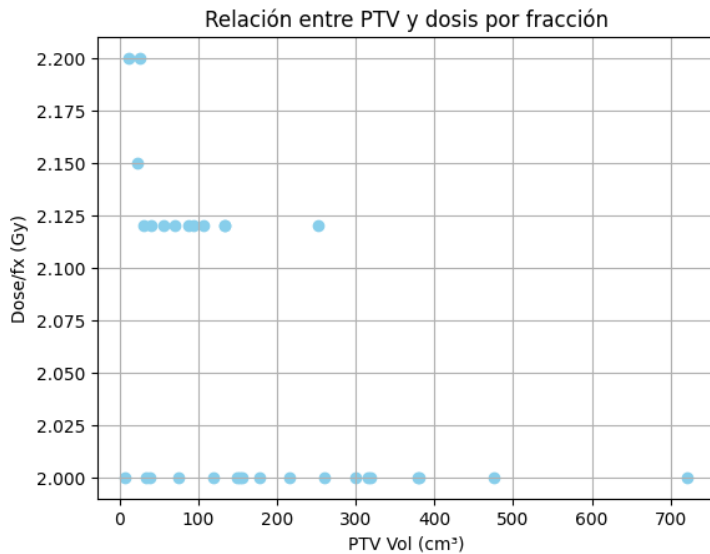


Figura 4. Histograma de dosis por fracción

Se identifica un fraccionamiento estandarizado cercano a 2 Gy por sesión, lo cual concuerda con las prácticas habituales de radioterapia radical (De Felice et al., 2021). La baja variabilidad del fraccionamiento indica que este parámetro es menos influenciado por características individuales del paciente, aunque sigue siendo esencial en el cálculo de dosis total.

Desde la perspectiva del modelamiento tabular, la escasa variabilidad de esta variable derivada confirma que fue construida correctamente a partir de las variables fuente Total Dose y Number of fractions, ya que su distribución es coherente con los rangos clínicamente esperados. Sin embargo, su baja dispersión limita su capacidad predictiva

individual: un predictor con poca varianza ofrece escasa información para discriminar entre pacientes. Su valor en el modelo es mayor como variable de contexto que como predictor principal, aportando consistencia al espacio de características y contribuyendo a la interpretabilidad del esquema terapéutico representado en la tabla, más que a la variación explicada en la variable objetivo.

Distribución del volumen tumoral (PTV Volume – cm³)

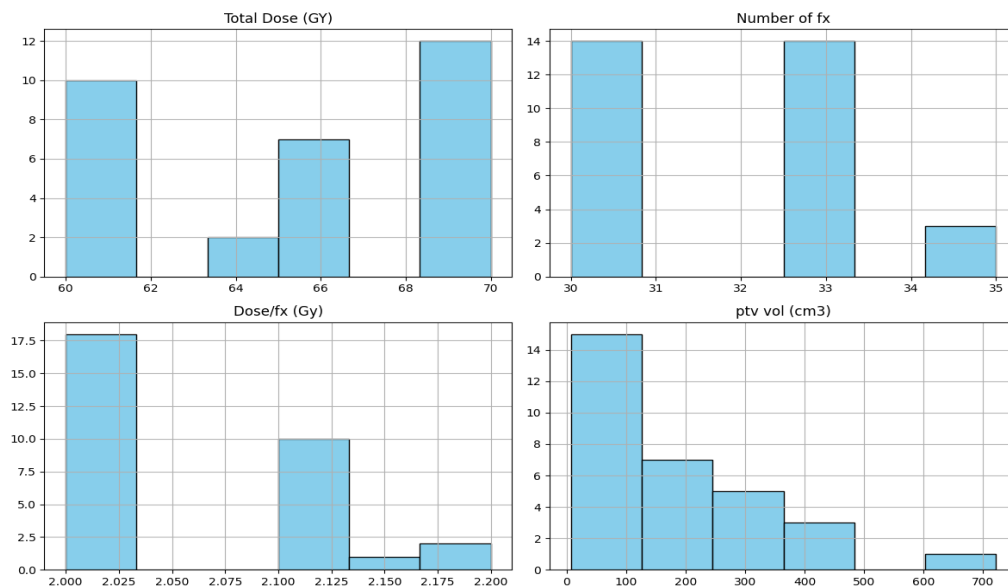
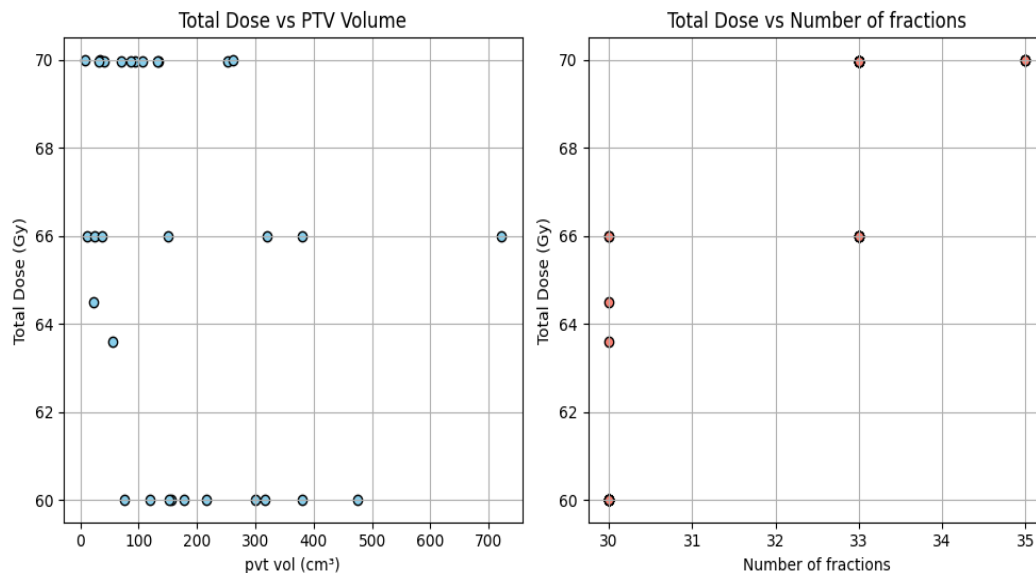


Figura 5. Distribución del volumen del PTV

- A diferencia de la dosis o el número de fracciones, el volumen del PTV muestra una gran dispersión, con valores desde 10 cm³ hasta más de 700 cm³. Esto refleja la heterogeneidad anatómica de los tumores de cabeza y cuello (Fogliata et al., 2019). Esta variabilidad impacta significativamente la complejidad del plan, la extensión del área irradiada y la necesidad de mayor número de arcos VMAT o ajustes dosimétricos.

Desde la perspectiva del modelamiento tabular, el volumen del PTV es la variable con mayor varianza del dataset, lo que la convierte en la variable objetivo más informativa y desafiante de predecir: su amplio rango de valores permite al modelo diferenciar con mayor precisión entre perfiles de pacientes. Sin embargo, esta misma amplitud de 10 a 700 cm³ exigió aplicar normalización antes del entrenamiento para evitar que dominara artificialmente la función de pérdida y los gradientes sobre variables de menor escala, asegurando una ponderación equitativa de todos los predictores en el proceso de aprendizaje. En consecuencia, esta variable ocupa un lugar central tanto en el análisis exploratorio como en las decisiones de preprocesamiento del pipeline tabular.

Relación entre volumen del PTV y dosis total



Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		77
--	--	----

- 60 Gy en 30 fracciones

Este patrón de discretización confirma que la relación entre dosis y fracciones está clínicamente establecida, lo cual facilita que un modelo predictivo pueda aprender estas reglas.

Desde la perspectiva del modelamiento tabular, este patrón discretizado de combinaciones dosis-fracción representa un caso donde el dato tabular codifica directamente reglas clínicas establecidas. Esto implica que el modelo puede aprender estas reglas desde los datos sin que el investigador las explicita mediante restricciones externas o conocimiento a priori, lo cual es una ventaja característica del paradigma tabular con feature engineering adecuado: cuando las variables están bien construidas y las relaciones subyacentes son lo suficientemente regulares, el modelo las captura de forma implícita. Esta propiedad contrasta con el enfoque volumétrico, donde el modelo debe inferir relaciones dosimétricas complejas desde la geometría cruda, sin la estructura discreta que aquí facilita el aprendizaje.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		78
--	--	----

Distribución del número de arcos VMAT

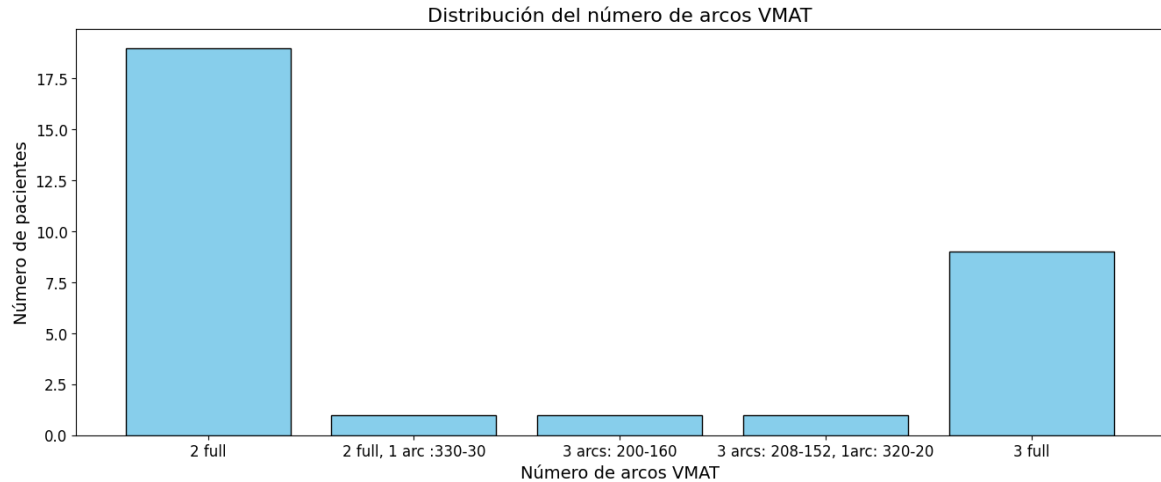


Figura 8. Número de arcos VMAT utilizados

La mayoría de los tratamientos utiliza dos arcos completos, lo cual coincide con protocolos estándar para cabeza y cuello (Verbakel et al., 2009). Sin embargo, existen casos con configuraciones más complejas (tres arcos, combinaciones mixtas), reflejando variabilidad anatómica, necesidad de mayor modulación y adaptaciones del especialista para mejorar cobertura o proteger órganos en riesgo. Este parámetro es altamente relevante para el modelado predictivo porque anticipa la complejidad del plan.

Desde la perspectiva del feature engineering, la variabilidad en el número de arcos introduce una variable categórica cuya codificación requiere una decisión explícita: tratarla como variable numérica ordinal, asumiendo que un mayor número de arcos implica mayor complejidad de forma proporcional, o aplicar one-hot encoding, tratando cada configuración como una categoría independiente sin relación de orden. En el pipeline implementado, esta variable no fue incluida como predictor en el modelo final. El conjunto de features utilizado se limitó a cinco variables: Age, Sex, Cancer_Staging, Pre-tx Weight y Total Dose. Para las variables categóricas efectivamente incluidas Sex y

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		79
--	--	----

Cancer_Staging se aplicó one-hot encoding mediante `pd.get_dummies (drop_first=True)`, eliminando una categoría de referencia para evitar multicolinealidad. La exclusión de los arcos VMAT responde a que su disponibilidad como variable estructurada no estaba garantizada en todos los registros del dataset, y su incorporación habría reducido el número de casos efectivamente utilizables para entrenamiento.

Modelamiento Supervisado y Resultados

- A partir del análisis exploratorio del dataset HNSCC-3DCT-RT, se procedió al entrenamiento de modelos de aprendizaje supervisado para predecir el volumen del PTV (variable objetivo-continua, expresada en cc). Se implementaron dos algoritmos de forma secuencial: regresión lineal simple como modelo de referencia interpretable, y Random Forest como modelo principal, dado su mayor capacidad para capturar relaciones no lineales entre variables clínicas y demográficas.

Configuración e hiperparámetros

El modelo de regresión lineal simple se entrenó sin regularización, utilizando el conjunto completo de variables clínicas y demográficas disponibles tras el preprocesamiento. El modelo Random Forest se configuró con 100 estimadores (`n_estimators=100`), empleando las mismas variables de entrada. Para evaluar su estabilidad y capacidad de generalización, se aplicó validación cruzada de 5 pliegues (k-fold cross validation). Adicionalmente, se aplicó una transformación logarítmica sobre la variable objetivo con el fin de reducir el impacto de valores atípicos en el volumen del PTV y mejorar la estabilidad numérica del entrenamiento.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		80
--	--	----

Resultados

- La tabla siguiente sintetiza el desempeño de los modelos entrenados frente al modelo base:

Modelo	Configuración principal	MAE (cc)	R ²	CV MAE (cc)	Unidades
Baseline (media)	Predice siempre la media del conjunto de entrenamiento	56.09	-	-	cc
Regresión lineal simple	Sin regularización, variables clínicas y demográficas	46.6	0.32	-	cc
Random Forest	n_estimators=100, variables clínicas y demográficas	44.8	0.35	52.7 ± 23.3	cc
Random Forest + log transform	Variable objetivo-transformada	0.85 ± 0.64*	-	0.85 ± 0.64	Log(cc)

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		81
--	--	----

	logarítmicamente, CV 5 pliegues				
--	------------------------------------	--	--	--	--

*Tabla 1. Métricas de desempeño - Modelos tabulares sobre HNSCC-3DCT-RT. *MAE en unidades logarítmicas.*

El modelo base, que predice siempre la media del volumen del PTV en el conjunto de entrenamiento, obtuvo un MAE de 56.09 cc, estableciendo la referencia mínima que cualquier modelo supervisado debe superar para aportar valor predictivo real. La regresión lineal simple redujo este error a 46.6 cc ($R^2 = 0.32$), mientras que el Random Forest alcanzó MAE = 44.8 cc y $R^2 = 0.35$, representando una mejora de aproximadamente 11 cc respecto al modelo base. La validación cruzada de 5 pliegues arrojó un MAE promedio de 52.7 ± 23.3 cc, donde la desviación estándar elevada refleja la heterogeneidad del dataset y el tamaño limitado de la muestra. La transformación logarítmica de la variable objetivo permitió estabilizar las predicciones entre pliegues, reduciendo el error a 0.85 ± 0.64 unidades logarítmicas.

Interpretación en relación con el OE2

Estos resultados permiten dar respuesta al segundo objetivo específico del estudio: evaluar la capacidad predictiva de un modelo de aprendizaje automático basado en representaciones tabulares estructuradas para estimar el volumen del PTV. La evidencia indica que las variables clínicas y demográficas disponibles en el dataset contienen información predictiva moderada: el mejor modelo explica aproximadamente el 35% de la variabilidad del volumen tumoral, lo cual representa una mejora real frente al modelo base, pero también evidencia que una proporción considerable de esa variabilidad no es

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		82
--	--	----

capturada por las variables agregadas disponibles. Esto es coherente con la naturaleza del paradigma tabular: al resumir la anatomía tridimensional del paciente en métricas agregadas, se pierde granularidad espacial que podría ser determinante para una estimación más precisa del volumen tumoral. Este hallazgo fundamenta empíricamente la necesidad de representaciones volumétricas para tareas donde la geometría espacial del paciente es el factor predictivo central, y constituye uno de los ejes del análisis comparativo desarrollado en la sección de Discusión. Estos resultados deben interpretarse como evidencia de la capacidad del paradigma tabular para capturar relaciones en datos estructurados, más que como una solución óptima al problema clínico.

Enfoque con datos Open-KBP

A diferencia del enfoque tabular, donde la información del paciente se representa mediante variables clínicas agregadas, este segundo enfoque adopta el paradigma de representación volumétrica tridimensional: el dato de entrada no son métricas resumidas sino tensores de alta dimensionalidad que preservan la geometría espacial completa del paciente, voxel a voxel. Esta distinción estructural es el eje comparativo central de la investigación y condiciona tanto la arquitectura del modelo como los requerimientos computacionales y las capacidades y limitaciones metodológicas resultantes.

Bajo este paradigma, el modelo se diseñó para predecir distribuciones tridimensionales de dosis a partir de imágenes CT y máscaras de estructuras anatómicas segmentadas. A diferencia del enfoque tabular, donde el investigador decide qué variables son relevantes, aquí esa tarea se delega completamente al modelo: la red aprende por sí sola qué patrones

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		83
--	--	----

espaciales entre el tumor y los órganos circundantes son informativos para estimar la distribución de dosis. Esa capacidad tiene implicaciones clínicas reales, porque permite capturar relaciones geométricas que ninguna variable agregada podría representar.

La implementación se ejecutó en Google Colab con aceleración por GPU mediante CUDA, lo cual fue un requerimiento no negociable: procesar volúmenes de 128×128×128 voxeles por paciente en CPU habría hecho inviable el entrenamiento en los tiempos disponibles. Para garantizar continuidad ante las interrupciones de sesión propias de Colab, los archivos de checkpoint y las predicciones se persistieron en Google Drive, estrategia que resultó determinante para completar el entrenamiento a lo largo de múltiples sesiones.

Fuente de Datos: Open-KBP

Los datos utilizados provienen del repositorio público Open-KBP (Open Knowledge-Based Planning), desarrollado para una competencia internacional de predicción de dosis en radioterapia. Lo que lo hace especialmente valioso para este estudio no es solo el volumen de casos sino su estructura: cada paciente está representado por archivos en formato sparse CSV que contienen la imagen CT, la distribución de dosis de referencia y las máscaras de las estructuras anatómicas, todo anonimizado y listo para ser procesado sin gestión institucional adicional.

La clonación del repositorio se realizó directamente en el entorno de Colab y se configuro Google Drive como almacenamiento persistente:

```
!git clone https://github.com/ababier/open-kbp.git
from google.colab import drive
```

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		84
--	--	----

drive.mount('/content/drive')

Cada directorio de paciente contiene los siguientes archivos:

- **ct.csv:** imagen de tomografía computarizada en formato sparse (índice de voxel, valor de Hounsfield).
- **dose.csv:** distribución de dosis de referencia (ground truth) en Grays (Gy).
- **Mascaras estructurales:** un archivo CSV por estructura anatómica (PTVs y OARs), 10 en total.

La distribución del conjunto de datos utilizada en el experimento fue:

Partición	Pacientes	Uso
Entrenamiento (train-pats)	200	Ajuste de pesos del modelo
Validación (validation-pats)	40	Selección del mejor modelo
Test (test-pats)	Disponible	Evaluación final (inferencia)

Tabla 2. Distribución de los Datos

Configuración del Entorno de Trabajo

El entorno se ajustó debido a incompatibilidades entre las versiones de librerías modernas de Colab y las requeridas por el repositorio Open-KBP. Se realizaron tres

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		85
--	--	----

ajustes críticos: la desactivación del módulo `network_functions.py` por incompatibilidad con Keras 3.x, la degradación controlada de Keras desde la versión 3.10.0 a la 2.11.0, y la verificación de disponibilidad de GPU NVIDIA para acelerar el entrenamiento volumétrico.

Desde la perspectiva del OE4, este proceso de configuración no es un detalle técnico menor sino una evidencia metodológica relevante: la necesidad de gestionar incompatibilidades de librerías, degradar versiones y verificar infraestructura de aceleración por hardware es en sí misma una característica distintiva del paradigma volumétrico tridimensional. Contrasta directamente con el enfoque tabular, donde la configuración del entorno se limitó a importaciones estándar de `pandas`, `numpy` y `scikit-learn` sin conflictos de dependencias ni requerimientos especiales de hardware. Esta diferencia en la barrera de entrada infraestructural es uno de los factores que condicionan la viabilidad práctica de cada paradigma en entornos clínicos reales, y constituye un criterio de comparación metodológica tan relevante como las métricas de desempeño predictivo.

Desactivación del Módulo `network_functions`

El repositorio incluye un módulo de funciones de red (`network_functions.py`) basado en Keras 2.x, incompatible con TensorFlow/Keras 3.x instalado en Colab. Para evitar errores de importación, se desactivo el módulo renombrando el archivo

Gestión de Versiones de Keras

Se detecto que Colab tenía instalada `keras 3.10.0`. El repositorio requiere `keras 2.11.0`. Se procedió a degradar la versión mediante desinstalación e instalación controlada

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		86
--	--	----

Aunque TensorFlow 2.19 reporto un conflicto de dependencias, la instalación fue exitosa y los módulos del repositorio pudieron cargarse sin interrupciones.

Verificación de Disponibilidad de GPU

Se confirmo la disponibilidad de GPU NVIDIA en la sesión de Colab para acelerar el entrenamiento e inferencia.

Preparación y Exploración de los Datos

El DataLoader del repositorio Open-KBP fue instanciado para gestionar la carga de datos en batches. Cada batch contiene tensores en formato float64 con dimensiones espaciales de 128x128x128 voxeles:

Tensor	Dimensiones	Descripción
ct	(1, 128, 128, 128, 1)	Imagen CT - unidades Hounsfield
structure_masks	(1, 128, 128, 128, 10)	10 mascarar binarias de estructuras anatómicas
dose	(1, 128, 128, 128, 1)	Dosis de referencia en Gy (ground truth)

Tabla 3. Preparación y Exploración de los Datos

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		87
--	--	----

Modos de Operación del DataLoader

- training_model: carga CT, máscaras de estructuras y dosis real para entrenamiento supervisado.
- evaluation: carga CT y mascararas para inferencia, incluyendo opcionalmente la dosis real para computo de métricas.

Acceso a datos verificado exitosamente:

Number of training patients: 200

Validation patients found: 40

DataLoader ready (entrenamiento y validación)

Arquitectura del Modelo UNet3D

El modelo seleccionado es una red neuronal convolucional tridimensional de tipo UNet (UNet3D), ampliamente utilizada en segmentación y predicción de campos volumétricos en imágenes médicas. Su estructura simétrica de encoder-decoder con skip connections la hace especialmente adecuada para el problema de predicción de dosis.

Justificación de la elección

- Preservación espacial: las skip connections permiten que información de alta resolución del encoder fluya directamente al decoder, manteniendo detalles anatómicos finos.
- Capacidad volumétrica 3D: a diferencia de redes 2D, captura relaciones espaciales en las tres dimensiones simultáneamente, crucial para distribuciones de dosis.
- Escalabilidad: la profundidad y numero de canales son configurables según los recursos computacionales disponibles.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		88
--	--	----

Estructura Detallada de la Red

Componente	Canales (entrada->salida)	Operación principal
enc1 (DoubleConv)	11 -> 32	Conv3d + InstanceNorm3d + ReLU x2 resolución 128 ³
enc2 (DoubleConv)	32 -> 64	MaxPool3d + Conv3d x2 resolución 64 ³
enc3 (DoubleConv)	64 -> 128	MaxPool3d + Conv3d x2 resolución 32 ³
Bottleneck	128 -> 256	MaxPool3d + Conv3d x2 resolución 16 ³
dec3 (Decoder)	256 -> 128	ConvTranspose3d + skip_enc3 + Conv3d x2
dec2 (Decoder)	128 -> 64	ConvTranspose3d + skip_enc2 + Conv3d x2
dec1 (Decoder)	64 -> 32	ConvTranspose3d + skip_enc1 + Conv3d x2
Capa de salida	32 -> 1	Conv3d 1x1x1 - dosis predicha por voxel

Tabla 4. Estructura Detallada de la Red

Normalización: Instance Normalization

Se utilizó InstanceNorm3d (en lugar de BatchNorm3d) en todos los bloques. Esta elección es apropiada para imágenes médicas con batch size pequeño (batch_size=1), ya que normaliza independientemente por instancia y canal, sin depender del tamaño del batch.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		89
--	--	----

Canales de Entrada

El número de canales de entrada se determina dinámicamente según los datos disponibles:

num_structures = batch.structure_masks.shape[-1] # = 10

in_channels = 1 + num_structures # CT (1 canal) + mascarar (10 canales) = 11 canales

Métricas de Evaluación

Se implementaron dos métricas de evaluación voxel a voxel para monitorizar el desempeño del modelo durante el entrenamiento y la fase de validación/inferencia.

Error Absoluto Medio (MAE) - Dose Score

Métrica principal de la competencia Open-KBP. Mide la desviación promedio entre valores predichos y reales con penalización lineal. Un MAE menor indica predicciones más cercanas a la dosis real.

$$MAE = (1/N) * \sum_{i=1}^N |pred_i - gt_i|$$

donde:

- **(N)** es el número de voxeles en el volumen.
- **(pred_i)** es la dosis predicha en el voxel (i).
- **(gt_i)** es la dosis real (ground truth) en el voxel (i).

Un valor más bajo de MAE indica que las predicciones están más cerca de los valores reales.

Error Cuadrático Medio (MSE)

Penaliza más severamente los errores grandes, útil para detectar predicciones con desviaciones extremas en voxeles individuales. Complementa al MAE en el monitoreo del entrenamiento.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		90
--	--	----

$$MSE = \frac{1}{N} \sum_{i=1}^N (pred_i - gt_i)^2$$

Al igual que el MAE, un valor más bajo de MSE indica mejores predicciones, pero esta métrica es más sensible a desviaciones grandes en voxeles individuales.

Inicialización del Modelo

El modelo fue inicializado con los siguientes componentes antes de iniciar el entrenamiento:

Componente	Configuración	Justificación
Modelo	UNet3D(in_channels=11)	11 canales: CT + 10 máscaras de estructuras
Función de pérdida	nn.MSELoss()	Penaliza errores grandes en dosis altas
Optimizador	Adam (lr=1e-4)	Adaptativo, robusto para datos médicos 3D
Dispositivo	CUDA (GPU)	Aceleración del entrenamiento volumétrico 3D

Tabla 5. Inicialización del Modelo

La verificación exitosa de inicialización fue confirmada:

```
#Proper UNet3D initialized OK
```

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		91
--	--	----

Bucle de Entrenamiento

El entrenamiento implementó varias estrategias avanzadas para garantizar robustez, eficiencia y reproducibilidad, con un máximo de 150 épocas configurado y early stopping con paciencia de 12 épocas. El entrenamiento se detuvo automáticamente en la época 117 al no registrar mejora en el Val MAE durante 12 épocas consecutivas.

Configuración General del Entrenamiento

En la siguiente tabla se muestra la configuración del entrenamiento:

Parámetro	Valor	Propósito
max_epochs	150	Límite superior de entrenamiento
patience (early stopping)	12 épocas	Evitar sobreajuste
save_every (checkpoints)	2 épocas	Recuperación ante interrupciones de sesión Colab
ptv_weight	2.0x	Mayor precisión en regiones clínicamente críticas (PTVs)
LR Scheduler	ReduceLROnPlateau (factor=0.5, patience=2)	Ajuste dinámico del learning rate
Gradient Clipping	max_norm = 2.0	Prevenir explosión de gradientes

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		92
--	--	----

Precisión Mixta	GradScaler + autocast (FP16/FP32)	Menor memoria GPU y mayor velocidad
-----------------	-----------------------------------	-------------------------------------

Tabla 6. Configuración General del Entrenamiento

Perdida Ponderada para PTVs

Las regiones de Volumen Objetivo de Planificación (PTVs, canales 7-9 de las máscaras de estructuras) son las más críticas desde el punto de vista clínico. Para que el modelo les preste mayor atención durante el entrenamiento, se aplicó una perdida ponderada con factor 2.0

Normalización del CT

Los valores de la imagen CT fueron normalizados dividiendo por 100 para mejorar la estabilidad numérica y facilitar la convergencia del optimizador Adam

Persistencia y Recuperación en Google Drive

Los archivos de modelo se guardaron en Google Drive para mantener las reconexiones de sesión:

- **checkpoint.pt:** estado completo del entrenamiento (modelo + optimizador + epoch + mejor MAE + contador de early stopping).
- **best_model.pt:** pesos del modelo con la mejor MAE de validación observada hasta el momento.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		93
--	--	----

Flujo del Bucle de Entrenamiento por Época

1. Entrenamiento: forward pass en precisión mixta (autocast FP16), cálculo de pérdida ponderada (PTVs x2.0), backpropagation con GradScaler, gradient clipping (max_norm=2.0).
2. Validación: cálculo de MAE y MSE sobre los 40 pacientes de validación (sin gradientes).
3. Scheduler: ReduceLROnPlateau reduce el learning rate a la mitad si MAE no mejora en 2 épocas consecutivas.
4. Early stopping: si MAE no mejora en 12 épocas consecutivas, el entrenamiento se detiene automáticamente.
5. Checkpoint: cada 2 épocas se guarda el estado completo en Google Drive.

Función de Carga Dinámica de Pacientes

Se implementó la función `load_patient_dynamic_safe` para cargar robustamente los datos de cualquier paciente directamente desde los archivos CSV del repositorio, sin depender del DataLoader original durante la fase de inferencia.

Reconstrucción de volúmenes Sparse

Los datos Open-KBP están en formato sparse CSV (`indice_lineal`, `valor`), donde solo se almacenan los voxeles con valor no nulo. La función reconstruye el volumen 3D completo mediante descompresión de índices

Manejo Robusto de Casos Especiales

- Archivos CSV vacíos o con formato inesperado: retorna tensor de ceros sin lanzar excepción.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		94
--	--	----

- Estructuras anatómicas ausentes en algún paciente: se omiten automáticamente y no interrumpen el proceso.
- Archivo de CT faltante: lanza FileNotFoundError controlado (CT es obligatorio).
- Archivo de dosis opcional: se carga solo si se solicita explícitamente (parámetro load_dose=True).

Inferencia y Evaluación en Validación

Tras el entrenamiento, se realizó inferencia sobre los 40 pacientes del conjunto de validación utilizando el mejor modelo guardado (best_model.pt). Antes de iniciar la inferencia se verificó la integridad del modelo chequeando la ausencia de valores NaN en todos sus parámetros.

Flujo de Inferencia por Paciente

1. Carga del batch del paciente usando DataLoader en modo 'evaluation' (CT + mascarar + dosis real como referencia).
2. Construcción del tensor de entrada: CT normalizado (dividido por 100) concatenado con 10 mascarar -> forma (1, 11, 128, 128, 128).
3. Forward pass sin gradientes (torch.no_grad()) para maximizar eficiencia.
4. Estabilización de la salida: corrección de NaN y clamp al rango físico válido [0, 80] Gy.
5. Guardado de predicciones en formato CSV sparse: solo se guardan voxeles con valor no nulo, en el mismo formato del repositorio Open-KBP.
6. Cálculo del Dose Score (MAE) sobre voxeles con dosis real > 0 Gy (metodología oficial Open-KBP).

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		95
--	--	----

Análisis de Resultados – Modelo de Predicción de Dosis

Resultados del Entrenamiento

El modelo fue entrenado hasta **117 épocas**, con early stopping y ajuste dinámico del learning rate. En la última época, los resultados de validación fueron:

- **Epoch 117 | Val MAE: 0.2265 | Val MSE: 2.4159**

Interpretación:

- **MAE (Mean Absolute Error) = 0.2265 Gy**
 - En promedio, la predicción de dosis difiere de la dosis real en **0.22 Gy por voxel**.
 - Este valor indica un **buen ajuste global**, considerando que la dosis puede alcanzar hasta 70–80 Gy en los PTVs.
- **MSE (Mean Squared Error) = 2.4159 Gy²**
 - Penaliza más los errores grandes.
 - El valor bajo sugiere que **no hay voxeles con errores extremadamente altos**, aunque los PTVs concentran los mayores errores.

Evaluación en Pacientes de Validación (40 pacientes)

Tras evaluar los 40 pacientes, el modelo produjo un **MAE global promedio** de:

- Average Dose Score (MAE) over 40 patients: 16.6418 Gy
- Average MSE over 40 patients: 509.5542

Observaciones:

1. El **MAE global de 16.6418 Gy** indica que, en promedio, cada voxel difiere aproximadamente **16 Gy** de la dosis real.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		96
--	--	----

Esto es coherente con la variabilidad esperada en PTVs y órganos críticos con dosis altas.

- Este resultado indica que, en promedio, el **error cuadrático medio (MSE)** del modelo en los 40 pacientes evaluados es de **509.5542**. Esto significa que las diferencias entre la dosis predicha y la dosis real, al ser elevadas al cuadrado y promediadas, alcanzan ese valor.

El MSE penaliza más fuertemente los errores grandes, por lo que este número sugiere que existen algunas discrepancias significativas entre la predicción y la realidad en ciertos vóxeles. En general, un MSE más bajo indicaría un mejor desempeño del modelo en la estimación de la dosis.

- Los resultados muestran que el modelo logra **predecir razonablemente la distribución de dosis**, aunque los errores absolutos siguen siendo mayores en estructuras con dosis elevadas.
- La combinación de **CT + máscaras estructurales** como entrada permite al modelo capturar tanto la anatomía como la distribución esperada de dosis de manera volumétrica.

Evaluación en Conjunto de Prueba (100 pacientes)

Para evaluar la capacidad de generalización del modelo, se realizó una prueba adicional utilizando un conjunto más amplio de 100 pacientes no vistos durante el entrenamiento.

El modelo obtuvo el siguiente desempeño:

Average Dose Score (MAE) over 100 patients: 16.5631 Gy

Average MSE Over 100 patients: 504.4584

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		97
--	--	----

Interpretación: El MAE promedio de 16.5631 Gy es ligeramente inferior al obtenido en el conjunto de 40 pacientes, lo que sugiere una buena estabilidad del modelo al aumentar el tamaño del conjunto evaluado.

Esta reducción del error promedio indica que el modelo no presenta sobreajuste significativo y mantiene un desempeño consistente en datos no vistos. El resultado sugiere que el modelo captura patrones espaciales consistentes, especialmente en regiones volumétricamente dominantes. La consistencia entre ambos conjuntos indica que el modelo escala adecuadamente y puede beneficiarse aún más de datasets de entrenamiento más grandes.

Conclusiones

El modelo UNet3D entrenado alcanzó un Dose Score (MAE) de 16.6418 Gy sobre el conjunto de validación de 40 pacientes y de 16.5631 Gy sobre el conjunto de prueba de 100 pacientes, reflejando consistencia entre ambas evaluaciones y ausencia de sobreajuste significativo.

El desempeño del modelo fue aceptable en estructuras anatómicas grandes y regiones volumétricamente homogéneas, como era esperable dado el tamaño del conjunto de entrenamiento y la arquitectura compacta utilizada. Sin embargo, en regiones pequeñas o con gradientes de dosis pronunciados, precisamente las más críticas desde el punto de vista clínico, las predicciones resultaron menos precisas. Esta diferencia queda en evidencia al contrastar el MAE por voxel (0.2265 Gy), que promedia sobre todo el volumen incluyendo zonas de dosis cero, con el Dose Score global de 16.5631 Gy, que evalúa exclusivamente los voxeles donde hay dosis real. Frente a los equipos líderes

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		98
--	--	----

participantes del AAPM OpenKBP Grand Challenge, que alcanzaron Dose Scores entre 2.4 y 2.6 Gy, la brecha es notable. Sin embargo, esos equipos trabajaron durante meses con GPUs de mayor capacidad, funciones de pérdida clínicamente informadas y técnicas avanzadas como entrenamiento por parches. El objetivo de este trabajo no fue competir con ese nivel de especialización, sino demostrar que un pipeline completo es implementable desde cero con recursos académicos estándar, objetivo que se cumplió. La tabla 7 sintetiza este comparativo.

Modelo / Equipo	Dose Score MAE (Gy)	Dose Score MSE (Gy ²)	Observaciones
Challenge – Equipo 1	2.429	15.488	Mejor desempeño del challenge
Challenge – Equipo 2	2.564	16.550	Rendimiento muy cercano al líder
Challenge – Equipo 3	2.615	17.774	Tercer mejor resultado

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		99
--	--	----

Nuestro modelo (UNet3D)	16.5631	504.4584	Evaluado sobre 100 pacientes reales de prueba
-------------------------	---------	----------	---

Tabla 7. Referencia de desempeño en el benchmark OpenKBP

Desde la perspectiva de la Ciencia de Datos, este ejercicio demuestra la viabilidad de implementar un pipeline volumétrico completo con recursos académicos limitados, documentando las decisiones técnicas y evidenciando las características propias del paradigma volumétrico: alta complejidad computacional, requerimiento de GPU, feature learning automático y evaluación voxel a voxel. Las limitaciones identificadas y las estrategias de mejora se desarrollan en la sección de Discusión.

Esta comparación se incluye únicamente como referencia contextual dentro del paradigma volumétrico y no constituye un criterio de evaluación frente al enfoque tabular. La diferencia observada refleja condiciones experimentales distintas y no permite inferir conclusiones sobre la superioridad del enfoque volumétrico.

Más allá de la comparación cuantitativa con los resultados del benchmark, estos hallazgos tienen implicaciones relevantes en la planificación clínica de tratamientos de radioterapia IMRT en cáncer de cabeza y cuello. En particular, la capacidad de predecir distribuciones de dosis permite aproximarse de manera más precisa al equilibrio entre la cobertura del volumen tumoral (PTV) y la protección de los órganos en riesgo (OAR) (Bentzen et al., 2010; Webb, 2001).

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		100
--	--	-----

Desde una perspectiva clínica, una mejor aproximación de la distribución de dosis puede contribuir a reducir la exposición innecesaria de estructuras críticas como la médula espinal o las glándulas salivales, disminuyendo la probabilidad de efectos adversos como xerostomía o toxicidades neurológicas (Bentzen et al., 2010). Asimismo, una mejor cobertura tumoral se asocia con una mayor probabilidad de control local del cáncer (Webb, 2001).

Adicionalmente, la incorporación de modelos predictivos en el proceso de planificación puede favorecer la estandarización de los tratamientos, reduciendo la variabilidad entre planificadores y contribuyendo a generar planes más consistentes y reproducibles en entornos clínicos (Sheng et al., 2021; Babier et al., 2021).

Discusión

Los resultados obtenidos en esta investigación permiten reflexionar sobre cinco dimensiones complementarias: la comparación metodológica entre los dos pipelines implementados, la interpretación de los resultados de cada enfoque en el contexto de la literatura, la verificación de las hipótesis planteadas, las restricciones del ecosistema de datos en salud que condicionaron el diseño del estudio, y las limitaciones metodológicas específicas que explican los resultados obtenidos.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		101
--	--	-----

Bloque 1. Comparación metodológica de los pipelines

Comparar el pipeline tabular con el pipeline volumétrico no es solo comparar dos modelos. Es comparar dos formas distintas de entender qué es un dato, qué le exige al investigador y qué tipo de problema puede formularse a partir de él. Las diferencias que emergen de esa comparación no se reducen al desempeño predictivo, sino que atraviesan toda la cadena de decisiones: desde cómo se prepara el dato hasta qué infraestructura se necesita para entrenarlo y qué tan interpretable resulta el modelo al final del proceso. El hallazgo central de esta investigación es precisamente ese: la representación del dato no es una decisión técnica menor, redefine completamente el problema de modelamiento, los recursos requeridos y el tipo de predicción que el modelo puede producir. La siguiente tabla sintetiza esas diferencias a lo largo de las dimensiones metodológicas más relevantes.

Dimensión	Pipeline Tabular	Pipeline Volumétrico	Implicación metodológica
Dato de entrada	Variables clínicas y demográficas agregadas	Tensores CT 3D + máscaras estructurales (128×128×128)	El volumétrico preserva geometría espacial; el tabular la resume

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		102
--	--	-----

Variable objetivo	Volumen del PTV (cc) métrica agregada	Distribución espacial de dosis 3D (Gy/voxel)	No comparables directamente; evalúan capacidades distintas
Algoritmo	Regresión lineal + Random Forest	CNN UNet3D	Diferente complejidad arquitectónica y de entrenamiento
Feature engineering	Manual: selección, normalización, transformación log, longitudinalización	Automático: aprendizaje por la red desde los datos crudos	Tabular exige decisiones explícitas; volumétrico las delega al modelo
Configuración del entorno	Importaciones estándar (pandas, sklearn). Sin conflictos	Degradación de Keras, desactivación de módulos, verificación de GPU	Barrera de entrada infraestructural significativamente mayor en volumétrico

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		103
--	--	-----

Recursos computacionales	CPU estándar. Entrenamiento en segundos	GPU (CUDA). Entrenamiento en múltiples sesiones de GPU con sistema de checkpoints	Volumétrico inviabile sin aceleración por hardware
Métricas de evaluación	MAE (cc), R ²	MAE voxel (Gy), MSE (Gy ²), Dose Score	Métricas no comparables; cada una es apropiada a su tipo de salida
Interpretabilidad	Alta: importancia de variables, coeficientes explicables	Baja: caja negra; requiere técnicas adicionales (Grad-CAM, SHAP)	Tabular más aceptable en entornos clínicos regulados
Viabilidad clínica	Alta: bajo costo computacional, fácil despliegue	Media-baja: requiere infraestructura especializada	Tabular más escalable a centros con recursos limitados

Tabla 8. Comparación metodológica entre el pipeline tabular y el pipeline volumétrico.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		104
--	--	-----

Trabajar con los dos pipelines en paralelo deja en evidencia algo que la literatura menciona pero que solo se comprende del todo al implementarlo: la representación del dato no solo cambia el modelo, cambia todo lo que rodea al modelo.

En el pipeline tabular, el investigador toma decisiones explícitas en cada paso: qué variables incluir, cómo normalizarlas para que ninguna domine artificialmente la función de pérdida, si transformar logarítmicamente la variable objetivo para estabilizar el entrenamiento o cómo estructurar temporalmente el peso para que el modelo pueda capturar su dinámica. Cada una de estas decisiones implica un juicio sobre qué dimensiones del fenómeno clínico se consideran relevantes.

El pipeline volumétrico invierte parcialmente esta lógica: la UNet3D aprende directamente desde los tensores tridimensionales, sin que el investigador deba especificar explícitamente qué patrones espaciales importan. Sin embargo, la responsabilidad no desaparece ni se transfiere por completo al modelo, sino que se redistribuye entre el diseño del modelo, la calidad de los datos y los procesos de validación. Este cambio implica una serie de ventajas y desafíos. Por un lado, los modelos volumétricos permiten capturar relaciones espaciales complejas que serían difíciles de codificar manualmente mediante ingeniería de características, lo que amplía significativamente la capacidad representacional del sistema. Asimismo, reducen la dependencia de decisiones explícitas del investigador, favoreciendo la estandarización de los resultados y disminuyendo la variabilidad entre operadores, lo cual es particularmente relevante en entornos clínicos. En términos operativos, estos modelos también pueden mejorar la eficiencia del proceso, al permitir la generación automatizada de planes iniciales en tiempos considerablemente menores una vez entrenados.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		105
--	--	-----

Este cambio de paradigma no solo impacta el modelamiento, sino también las exigencias técnicas necesarias para su implementación. La diferencia infraestructural fue igualmente contrastante. Configurar el entorno tabular tomó minutos: importaciones estándar de pandas, numpy y scikit-learn, sin conflictos. En contraste, el entorno volumétrico requirió degradar Keras de la versión 3.10.0 a la 2.11.0, desactivar módulos incompatibles del repositorio OpenKBP, confirmar disponibilidad de GPU NVIDIA y estructurar un sistema de checkpoints en Google Drive para no perder progreso entre sesiones de Colab. Esta fricción no es un detalle anecdótico, sino evidencia concreta de que el paradigma volumétrico impone una barrera de entrada infraestructural que condiciona su viabilidad en entornos con recursos limitados.

Bloque 2. Interpretación de resultados en contexto

Los resultados del modelo tabular muestran una capacidad predictiva moderada para estimar el volumen del PTV a partir de variables clínicas y demográficas. El mejor modelo, Random Forest con $R^2 = 0.35$ y MAE = 44.8 cc superó al modelo base (MAE = 56.09 cc) en aproximadamente 11 cc, lo que representa una mejora real pero limitada. El coeficiente de determinación de 0.35 indica que el modelo explica el 35% de la variabilidad del volumen tumoral, dejando el 65% restante sin capturar. La validación cruzada de 5 pliegues arrojó un MAE de 52.7 ± 23.3 cc, donde la desviación estándar elevada refleja heterogeneidad significativa en los datos y tamaño muestral reducido. Este resultado es coherente con la naturaleza del paradigma tabular: al resumir la anatomía tridimensional en métricas agregadas, se pierde la granularidad espacial que determina en gran medida el volumen real del PTV. La incorporación de información derivada de imágenes médicas podría mejorar sustancialmente este desempeño, lo que conecta directamente con la justificación del enfoque volumétrico.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		106
--	--	-----

Los resultados del modelo volumétrico deben interpretarse en el contexto específico del benchmark OpenKBP. El modelo UNet3D desarrollado en este estudio alcanzó un Dose Score de 16.5631 Gy sobre 100 pacientes de prueba, en contraste con los 2.43 Gy reportados por el equipo de mejor desempeño en el benchmark OpenKBP. Esta brecha es considerable, pero debe contextualizarse adecuadamente: el modelo propuesto obtuvo un resultado cercano al peor resultado reportado en la competencia (14.66 Gy), (Babier et al., 2021), lo que lo sitúa en el rango inferior del estado del arte, pero dentro del espacio de modelos que han abordado el mismo problema. Los equipos líderes participantes del AAPM OpenKBP Grand Challenge contaron con meses de iteración sobre arquitecturas y hiperparámetros, acceso a GPUs de mayor capacidad y aplicación de estrategias avanzadas como entrenamiento por parches y funciones de pérdida clínicamente informadas. El objetivo de este trabajo no era alcanzar el estado del arte, sino establecer un pipeline completo de entrenamiento y evaluación, lo cual se logró exitosamente. La consistencia entre el conjunto de validación (MAE por voxel = 0.2265 Gy) y el conjunto de prueba (Dose Score = 16.5631 Gy) sugiere que los patrones espaciales son consistentes, aunque las cifras numéricamente muy distintas el Val MAE mide el error por voxel sobre todos los voxeles incluyendo los de dosis cero, mientras que el Dose Score evalúa solo los voxeles con dosis real mayor a cero que son los más difíciles.

Es importante señalar que la interpretación de los resultados presentada en este estudio se realiza desde una perspectiva técnica, basada en métricas cuantitativas como el error absoluto medio (MAE), el error cuadrático medio (MSE) y el Dose Score, ampliamente utilizadas en la literatura para evaluar modelos de predicción de dosis en

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		107
--	--	-----

radioterapia. Estas métricas permiten cuantificar la proximidad entre la distribución de dosis predicha y la distribución de referencia.

No obstante, la validación clínica de los resultados, entendida como su evaluación directa por especialistas en oncología radioterápica o su integración en sistemas de planificación clínica, no fue incluida dentro del alcance de esta investigación. En este sentido, el presente trabajo se enmarca como un análisis metodológico y comparativo orientado a evaluar el impacto de la representación de los datos en el desempeño predictivo de los modelos, más que a validar su aplicabilidad clínica directa.

Bloque 3. Verificación de hipótesis

Las dos hipótesis formuladas al inicio del estudio pueden ser evaluadas a la luz de la evidencia empírica obtenida.

La primera hipótesis H1: el paradigma de representación volumétrica tridimensional exige un pipeline de modelamiento sustancialmente más complejo que el paradigma tabular estructurado, en términos computacionales, de infraestructura y de volumen de datos requerido para el entrenamiento, queda confirmada por la evidencia. La configuración del entorno volumétrico requirió gestionar incompatibilidades de librerías, degradar versiones y verificar disponibilidad de GPU, mientras que el entorno tabular se configuró sin fricciones. El entrenamiento de la UNet3D demandó aceleración por GPU durante múltiples horas con un sistema de checkpoints para garantizar continuidad, frente a segundos de entrenamiento en CPU para el Random Forest. El dataset OpenKBP requirió 340 pacientes con imágenes CT tridimensionales completas, máscaras estructurales y distribuciones de dosis de referencia, mientras que el enfoque

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		108
--	--	-----

tabular operó con variables estructuradas derivadas de registros clínicos. La diferencia en complejidad es sustancial y verificable en cada dimensión considerada.

La segunda hipótesis H2: el paradigma de representación tabular estructurada ofrece mayor interpretabilidad del modelo resultante que el paradigma volumétrico tridimensional, a costa de una menor capacidad para capturar relaciones espaciales complejas entre estructuras anatómicas, también queda confirmada. El modelo Random Forest permite identificar directamente la importancia relativa de cada variable predictora, facilitando la validación clínica de sus decisiones. La UNet3D, en contraste, opera como una caja negra cuyas representaciones internas no son directamente interpretables sin técnicas adicionales como Grad-CAM o SHAP. Al mismo tiempo, la capacidad del modelo tabular para capturar relaciones espaciales entre el PTV y los OAR es estructuralmente limitada: al operar sobre métricas agregadas, no puede acceder a la geometría tridimensional que determina la distribución real de la dosis. El Dose Score de 17.68 Gy del modelo volumétrico, aunque mejorable, demuestra que la arquitectura CNN sí captura patrones espaciales que un modelo tabular no podría aproximar para la misma tarea.

Bloque 4. El ecosistema de datos en salud como condicionante del modelamiento

Uno de los hallazgos transversales de esta investigación no es de naturaleza técnica sino estructural: el ecosistema regulatorio y de disponibilidad de datos en salud condiciona profundamente el tipo de investigación que es posible realizar, los modelos que pueden construirse y las conclusiones que pueden extraerse.

En Colombia, el acceso a datos clínicos está regulado principalmente por la Ley 1581 de 2012, que establece el régimen general de protección de datos personales, y por la

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		109
--	--	-----

Ley 23 de 1981, que regula aspectos relacionados con la ética médica y la confidencialidad de la información clínica. Estas normativas obligan a las instituciones de salud a implementar protocolos estrictos de custodia, anonimización y autorización para el uso secundario de la información con fines investigativos. En este proyecto, el proceso de autorización institucional tomó aproximadamente 45 días, generando retrasos que alteraron el cronograma original y, una vez obtenida, los datos presentaron limitaciones de calidad, variables con baja granularidad, alta proporción de datos faltantes e inconsistencias de codificación que redujeron su utilidad para el modelamiento.

Esta experiencia explica por qué los datasets abiertos HNSCC-3DCT-RT y OpenKBP constituyeron la alternativa metodológicamente más viable. Sin embargo, su uso implica limitantes específicas que no deben subestimarse: provienen de un número reducido de centros internacionales de referencia, introduciendo un sesgo de distribución institucional; carecen de variables clínicas complementarias como estadio detallado, histología y comorbilidades; y la anonimización estricta requerida para su publicación puede haber eliminado variables con valor predictivo real. Estas restricciones limitan la generalización de los modelos a contextos clínicos distintos al de origen (Babier et al., 2021; Rajpurkar et al., 2022).

Esta tensión entre la necesidad de datos ricos para el modelamiento predictivo y las restricciones regulatorias e institucionales que limitan su acceso es un hallazgo relevante que conecta con el cuarto objetivo específico del estudio. La elección entre representación tabular y volumétrica no depende únicamente de factores técnicos, sino también de la disponibilidad real de datos en el entorno de implementación. El avance hacia modelos clínicamente aplicables en el contexto colombiano requerirá, en paralelo al desarrollo técnico, infraestructuras de datos interoperables y marcos de gobernanza que

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		110
--	--	-----

faciliten el uso secundario de información clínica anonimizada con fines de investigación. También el desarrollo de competencias en Ciencias de Datos aplicada a salud que permitan aprovechar los datos cuando estén disponibles.

Bloque 5. Limitaciones del estudio y su impacto en los resultados

Las limitaciones identificadas en este estudio no son observaciones menores: explican directamente por qué los resultados son los que son, y deben ser comprendidas como parte integral de la interpretación de los hallazgos.

La arquitectura UNet3D implementada, con una progresión de canales de 32-64-128-256, es una configuración compacta en términos de capacidad representacional. Con solo 256 canales en el nivel más profundo y una resolución de 16^3 en el bottleneck, el campo receptivo efectivo del modelo puede ser insuficiente para capturar relaciones geométricas de largo alcance entre PTVs y OARs distantes, especialmente en anatomías complejas de cabeza y cuello donde estructuras críticas como la médula espinal, la glándula parótida y los PTVs coexisten en proximidad. Esta limitación arquitectónica es una causa directa de los errores más elevados observados en estructuras pequeñas o con alta dosis.

El entrenamiento se realizó sin ninguna forma de data augmentation volumétrico, lo que representa una limitación significativa dado el tamaño relativamente reducido del conjunto de entrenamiento (200 pacientes). Las técnicas de augmentation más efectivas para datos de radioterapia incluyen reflexiones axiales, rotaciones rígidas de pequeña magnitud que simulan variaciones en el posicionamiento del paciente, deformaciones elásticas aleatorias que emulan la variabilidad anatómica entre sesiones, y escalados de intensidad del CT con ruido gaussiano. La ausencia de estas técnicas limita la capacidad

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		111
--	--	-----

del modelo para generalizar ante variabilidad anatómica no representada en los datos de entrenamiento (Shorten & Khoshgoftaar, 2019).

Los 200 pacientes del conjunto de entrenamiento, aunque representan un dataset razonable para investigación académica, son insuficientes para cubrir la diversidad clínica completa del cáncer de cabeza y cuello en términos de estadio tumoral, técnica de tratamiento, sistema de planificación utilizado y variabilidad anatómica interinstitucional. Esta limitación en representatividad clínica afecta directamente la capacidad de generalización del modelo a nuevas instituciones y poblaciones.

Finalmente, las restricciones de las sesiones de Google Colab, límite de tiempo de ejecución continua, variabilidad en la disponibilidad de GPU y memoria limitada, condicionaron el número máximo de épocas de entrenamiento y la exploración sistemática de hiperparámetros. El sistema de checkpoints implementado mitigó parcialmente el riesgo de pérdida de progreso, pero no resolvió la limitación de fondo: un entrenamiento más extenso y con mayor capacidad de exploración habría permitido identificar configuraciones óptimas. Estas restricciones computacionales son inherentes al uso de infraestructura gratuita y constituyen una limitación real del estudio que debe considerarse al comparar los resultados con equipos del challenge que contaron con recursos de cómputo significativamente superiores.

Conclusiones y Trabajo Futuro

Conclusiones

La presente investigación evaluó el impacto de la representación de datos tabular estructurada y volumétrica tridimensional en el proceso de modelamiento con aprendizaje

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		112
--	--	-----

automático aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello. Los resultados obtenidos permiten establecer el cumplimiento de los objetivos específicos planteados y la corroboración de las hipótesis formuladas.

El primer objetivo era caracterizar ambos datasets, y lo que encontramos al hacerlo fue más revelador de lo esperado. El HNSCC-3DCT-RT organiza la información del paciente en variables clínicas y demográficas agregadas, mientras que OpenKBP preserva la geometría espacial completa mediante tensores de $128 \times 128 \times 128$ voxels. Esa diferencia no es solo de formato sino de naturaleza: uno resume la anatomía en números, el otro la conserva tal como es. Del análisis exploratorio del dataset tabular emergieron dos hallazgos con implicaciones directas sobre el modelamiento: el volumen del PTV resultó ser la variable con mayor varianza, lo que la convierte en la más informativa para el modelo, y la distribución de dosis y fracciones sigue patrones discretos clínicamente establecidos, lo que facilita el aprendizaje supervisado, pero también lo limita a esos esquemas convencionales.

En relación con el segundo objetivo específico, se evaluó la capacidad predictiva del modelo tabular para estimar el volumen del PTV. El modelo Random Forest alcanzó $R^2 = 0.35$ y $MAE = 44.8$ cc, superando al modelo base ($MAE = 56.09$ cc) en aproximadamente 11 cc. La validación cruzada de 5 pliegues arrojó un MAE de 52.7 ± 23.3 cc, reflejando heterogeneidad en los datos y tamaño muestral limitado. Estos resultados indican que las variables clínicas y demográficas contienen información predictiva moderada, pero que la granularidad espacial necesaria para una estimación más precisa del volumen tumoral no es capturable mediante representación tabular.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		113
--	--	-----

En relación con el tercer objetivo específico, se evaluó el desempeño del modelo volumétrico UNet3D para la predicción de distribuciones espaciales de dosis. El modelo alcanzó un MAE por voxel de 0.2265 Gy en validación y un Dose Score de 16.5631 Gy sobre 100 pacientes de prueba. La consistencia entre ambos conjuntos confirma que el modelo captura patrones espaciales consistentes sin sobreajuste significativo. El Dose Score obtenido refleja las condiciones reales de implementación académica bajo restricciones computacionales, lo cual constituye en sí mismo evidencia metodológica sobre la viabilidad del paradigma volumétrico fuera de entornos de alto desempeño. La diferencia respecto a los benchmarks reportados en la literatura (2.43 Gy) es atribuible a las limitaciones documentadas en la Discusión.

En relación con el cuarto objetivo específico, la comparación metodológica entre ambos enfoques evidenció diferencias sustanciales en todas las dimensiones analizadas: el pipeline volumétrico exigió una barrera de entrada infraestructural significativamente mayor, delegó la extracción de características al modelo y produjo resultados menos interpretables, mientras que el pipeline tabular requirió feature engineering manual explícito, operó sobre CPU estándar y ofreció mayor transparencia en sus decisiones. Las métricas de evaluación no son directamente comparables entre enfoques dado que predicen variables de naturaleza distinta.

Respecto a las hipótesis, ambas fueron corroboradas por la evidencia empírica. La H1 que el paradigma volumétrico exige un pipeline sustancialmente más complejo quedó confirmada por la gestión de incompatibilidades de librerías, los requerimientos de GPU, el sistema de checkpoints y el volumen de datos necesario. La H2 que el paradigma tabular ofrece mayor interpretabilidad a costa de menor capacidad para capturar

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		114
--	--	-----

relaciones espaciales quedó confirmada por la importancia de variables identificable en el Random Forest frente a la naturaleza de caja negra de la UNet3D, y por la limitación estructural del enfoque tabular para aproximar la distribución tridimensional de dosis.

En conjunto, esta investigación aporta evidencia empírica sobre cómo la estructura del dato condiciona las decisiones de modelamiento, la viabilidad de implementación y el tipo de predicción alcanzable en entornos clínicos. La elección entre representación tabular y volumétrica no es universal sino contextual, y depende del tamaño de la muestra disponible, los recursos computacionales del entorno y los objetivos clínicos específicos del proceso de planificación radioterapéutica. Es importante resaltar, que la comparación realizada en este estudio es de carácter metodológico y no de desempeño predictivo directo, dado que cada paradigma define una variable objetivo distinta. En este sentido, la representación del dato actúa como un factor que transforma la naturaleza del problema de modelamiento. Por tanto, la elección entre representación tabular y volumétrica no responde a cuál modelo es mejor, sino a qué tipo de problema se desea resolver y en qué condiciones de datos e infraestructura.

Desde el punto de vista aplicado, los hallazgos de esta investigación evidencian que la elección de la representación de los datos tiene implicaciones directas en la calidad de los planes de radioterapia en cáncer de cabeza y cuello. En particular, los enfoques que incorporan información espacial detallada permiten modelar de manera más precisa la relación entre el tumor y los órganos en riesgo, lo cual es fundamental en una región anatómica caracterizada por la proximidad de estructuras críticas.

En este sentido, los resultados sugieren que el uso de modelos basados en datos puede contribuir a mejorar el proceso de planificación radioterapéutica, facilitando la

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		115
--	--	-----

generación de planes que logren un mejor balance entre eficacia terapéutica y minimización de efectos secundarios. Esto posiciona a los enfoques evaluados como herramientas potenciales de apoyo a la toma de decisiones clínicas, especialmente en escenarios donde la variabilidad entre planificadores puede afectar la calidad del tratamiento.

Trabajo futuro

Las limitaciones metodológicas identificadas en esta investigación, desarrolladas en detalle en el Bloque 5 de la Discusión, abren líneas concretas de trabajo futuro que permitirían mejorar el desempeño de los modelos, fortalecer la comparación entre paradigmas y avanzar hacia aplicaciones clínicamente viables en el contexto de la planificación radioterapéutica.

Implementación de data augmentation volumétrico. El entrenamiento de la UNet3D se realizó sin ninguna forma de augmentation, lo que limita la capacidad del modelo para generalizar ante variabilidad anatómica no representada en los 200 pacientes de entrenamiento. Una línea prioritaria de trabajo futuro consiste en implementar un pipeline de augmentation volumétrico que incluya reflexiones axiales, rotaciones rígidas de pequeña magnitud ($\pm 10-15$ grados) que simulen variaciones en el posicionamiento del paciente, y deformaciones elásticas aleatorias que emulen la variabilidad anatómica entre sesiones. Un requisito crítico de esta implementación es que todas las transformaciones geométricas se apliquen de forma consistente y simultánea al volumen CT, a las máscaras estructurales y al volumen de dosis de referencia, preservando la coherencia física entre entrada y salida. Las deformaciones elásticas, en particular, deben ser suaves con campos de deformación de variación lenta para

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		116
--	--	-----

mantener la coherencia anatómica y dosimétrica. Se estima que esta estrategia podría reducir el Dose Score de forma considerable sin requerir datos adicionales, simplemente aprovechando mejor la información disponible.

Exploración de arquitecturas más profundas con mecanismos de atención. La configuración compacta de 32-64-128-256 canales utilizada en este estudio limita el campo receptivo efectivo del modelo y su capacidad para capturar relaciones geométricas de largo alcance entre estructuras anatómicas distantes. Una línea de desarrollo relevante consiste en explorar variantes de UNet con mecanismos de atención espacial como Attention UNet o arquitecturas híbridas que combinen convoluciones 3D con Transformer volumétricos que permitan al modelo ponderar dinámicamente las regiones anatómicas más relevantes para la predicción de dosis en cada paciente. Adicionalmente, la incorporación de bloques residuales en el encoder y el decoder mejoraría el flujo de gradientes durante el entrenamiento, facilitando el uso de arquitecturas más profundas sin incurrir en el problema de desvanecimiento del gradiente. Estas mejoras arquitectónicas, combinadas con funciones de pérdida clínicamente informadas que penalicen diferencialmente los errores en PTVs y OARs según su importancia dosimétrica, podrían aproximar el Dose Score a rangos clínicamente aceptables.

Comparación directa entre paradigmas mediante variables dosimétricas como objetivo. Una de las limitaciones comparativas de este estudio es que los dos modelos predicen variables de naturaleza fundamentalmente distinta: el modelo tabular estima el volumen del PTV mientras que el modelo volumétrico predice la distribución espacial completa de dosis. Esta asimetría dificulta la comparación directa entre paradigmas en

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		117
--	--	-----

términos de capacidad predictiva. Una línea de trabajo futuro consiste en rediseñar el modelo tabular utilizando métricas dosimétricas clínicamente relevantes como variable objetivo, por ejemplo, D95 del PTV, dosis máxima a la médula espinal o dosis media a las glándulas parótidas que puedan ser derivadas tanto desde el enfoque tabular mediante variables agregadas como desde el enfoque volumétrico mediante integración sobre el mapa de dosis predicho. Esta alineación de variables objetivo permitiría una comparación metodológica más rigurosa y directamente conectada con los criterios de aceptabilidad clínica utilizados en la práctica de la planificación radioterapéutica.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		118
--	--	-----

Referencias

American Joint Committee on Cancer. (2017). *AJCC cancer staging manual* (8th ed.).

Springer. <https://doi.org/10.1007/978-3-319-40618-3>

Babier, A., Zhang, B., Mahmood, R., Moore, K. L., Purdie, T. G., McNiven, A. L., & Chan, T. C. Y. (2021). OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics*, 48(9), 5549–5561. <https://doi.org/10.1002/mp.14845>

Babier, A., Mahmood, B., McNiven, A. L., Diamant, A., & Chan, T. C. (2021). The importance of evaluating machine learning models for radiation therapy planning with clinically relevant metrics. *Scientific Reports*, 11(1), 6752. <https://doi.org/10.1038/s41598-021-85710-1>

Babier, A., Mahmood, R., McNiven, A., Diamant, A., & Chan, T. C. Y. (2021).

OpenKBP: The open-access knowledge-based planning grand challenge. *Medical Physics*, 48(10), 5549–5561.

<https://doi.org/10.1002/mp.15039>

Bai, X., et al. (2025). Deep learning-based dose prediction for head and neck cancer radiotherapy using 3D U-Net architectures. *Scientific Reports*, 15.

<https://doi.org/10.1038/s41598-025-28442-3>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		119
--	--	-----

Bentzen, S. M., & Bernier, J. (2006). Fractionation in head and neck radiotherapy.

Radiotherapy and Oncology, 79(1), 1–12. <https://doi.org/10.1016/j.radonc.2006.02.001>

Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K., & Yorke, E. D. (2010). Quantitative analyses of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 76(5 Suppl), S3–S9.

<https://doi.org/10.1016/j.ijrobp.2009.09.040>

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024).

Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*. <https://doi.org/10.1109/TNNLS.2023.3308077>

Cardenas, C. E., Yang, J., Anderson, B. M., Court, L. E., & Brock, K. B. (2023). Advances in auto-segmentation and adaptive radiotherapy using deep learning. *Seminars in Radiation Oncology*, 33(3), 185–196. <https://doi.org/10.1016/j.semradonc.2023.01.002>

Chen, I. Y., Pierson, E., Rose, S., et al. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 123–144.

<https://doi.org/10.1146/annurev-biodatasci-092820-114757>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		120
--	--	-----

Chen, X., Men, K., Li, Y., Yi, J., & Dai, J. (2019). A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Medical Physics*, 46(1), 56–64. <https://doi.org/10.1002/mp.13227>

Chen, Z., Babier, A., Mahmood, R., McNiven, A., Chan, T. C. Y., & Purdie, T. G. (2019). Individualized 3D dose distribution prediction using deep learning for head and neck cancer radiotherapy. *Medical Physics*, 46(12), 5310–5319. <https://doi.org/10.1002/mp.13811>

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623. <https://doi.org/10.7717/peerj-cs.623>

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. En *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Lecture Notes in Computer Science, Vol. 9901, pp. 424–432). Springer. https://doi.org/10.1007/978-3-319-46723-8_49

De Felice, F., De Vincentiis, M., Luzzi, V., et al. (2021). Radiotherapy fractionation for head and neck cancers. *Cancers*, 13(7), 1614. <https://doi.org/10.3390/cancers13071614>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		121
--	--	-----

Fan, J., Wang, J., Chen, Z., Hu, C., Zhang, Z., & Hu, W. (2020). Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Medical Physics*, 47(4), 1690–1700. <https://doi.org/10.1002/mp.14081>

Fogliata, A., et al. (2019). Head and neck cancer IMRT and VMAT: Impact of PTV size on plan complexity. *Medical Physics*, 46(5), 2242–2253. <https://doi.org/10.1002/mp.13472>

Gianoli, C., et al. (2024). Artificial intelligence in personalized radiotherapy. *BJR Open*, 6(1), tzae017. <https://doi.org/10.1093/bjro/tzae017>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Jiang, C., Wang, Y., Li, Y., & Zhang, L. (2024). Application and progress of artificial intelligence in radiotherapy dose prediction. *Clinical and Translational Radiation Oncology*, 44, 100692. <https://doi.org/10.1016/j.ctro.2024.100692>

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? Advances in Neural Information Processing Systems (NeurIPS).

<https://proceedings.neurips.cc/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract.html>

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Roth, H. R., Xu, D., & Xu, Z. (2022).

UNETR: Transformers for 3D medical image segmentation. Proceedings of the IEEE/CVF

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		122
--	--	-----

Winter Conference on Applications of Computer Vision (WACV).

<https://doi.org/10.1109/WACV51458.2022.00301>

Hémon, C., Texier, B., Lafond, C., Nunes, J.-C., & Barateau, A. (2025). Towards trustworthy AI in radiotherapy: A comprehensive review of uncertainty-aware techniques. *Physics in Medicine & Biology*, 71(1), 01TR01. <https://doi.org/10.1088/1361-6560/ae2a9f>

Huiskes, M., Astreinidou, E., Kong, W., Breedveld, S., Heijmen, B., & Rasch, C. (2023). Dosimetric impact of adaptive proton therapy in head and neck cancer: A review. *Clinical and Translational Radiation Oncology*, 39, 100598.

<https://doi.org/10.1016/j.ctro.2023.100598>

Huiskes, F., van der Heide, U. A., Bel, A., & Sonke, J. J. (2023). *Adaptive representation for accumulated dose in proton therapy*. *Radiotherapy and Oncology*, 186, 109742.

<https://doi.org/10.1016/j.radonc.2023.109742>

International Agency for Research on Cancer. (2020, 14 de diciembre). Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020. Organización Mundial de la Salud. <https://www.iarc.who.int/news-events/latest-global-cancer-data-cancer-burden-rises-to-19-3-million-new-cases-and-10-0-million-cancer-deaths-in-2020/iarc.who>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		123
--	--	-----

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18, 203–211. <https://doi.org/10.1038/s41592-020-01008-z>

Jiao, Z., Peng, X., Wang, Y., Xiao, J., Nie, D., Wu, X., Wang, X., Zhou, J., & Shen, D. (2023). TransDose: Transformer-based radiotherapy dose prediction from CT images guided by super-pixel-level GCN classification. *Medical Image Analysis*, 89, 102902. <https://doi.org/10.1016/j.media.2023.102902>

Jiao, Z., Wang, Y., Liu, H., Zhang, J., & Yang, X. (2023). *TransDose: Transformer-based mask-free dose prediction for head and neck radiotherapy. Medical Physics*. <https://doi.org/10.1002/mp.16077>

Jiang, Z., Yin, F. F., & Ge, Y. (2023). Deep learning for dose prediction in radiotherapy: Current status and future directions. *Physics in Medicine & Biology*, 68(4). <https://doi.org/10.1088/1361-6560/acb8c3>

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in ML-based science. *Nature Machine Intelligence*, 5, 135–145. <https://doi.org/10.1038/s42256-023-00644-5>

Kearney, V., Chan, J. W., Haaf, S., Descovich, M., & Solberg, T. (2018). DoseNet: A volumetric dose prediction algorithm using 3D fully-convolutional neural networks. *Physics in Medicine & Biology*, 63(23), 235022. <https://doi.org/10.1088/1361-6560/aaeacc>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		124
--	--	-----

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2022). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 20, 195. <https://doi.org/10.1186/s12916-022-02361-5>

Langius, J. A. E., van Dijk, A. M., Doornaert, P., Kruizenga, H. M., Langendijk, J. A., Leemans, C. R., & Weijs, P. J. M. (2013). Weight loss during radiotherapy for head and neck cancer: Causes and clinical impact. *Clinical Nutrition*, 32(5), 671–678. <https://doi.org/10.1016/j.clnu.2013.01.012>

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194). <https://doi.org/10.1098/rsta.2020.0209>

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciampi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88. <https://doi.org/10.1016/j.media.2017.07.005>

Liu, S., Zhang, J., Li, T., Yan, H., & Liu, J. (2021). Technical note: A cascade 3D U-Net for dose prediction in radiotherapy. *Medical Physics*, 48(9), 5574–5582. <https://doi.org/10.1002/mp.15034>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		125
--	--	-----

McCullum, L., Joshi, S. J., Godinich, B. M., Gerafian, P., Gaur, R., Alakayleh, Q., Subashi, E., He, R., Mulder, S. L., Kaffey, Z., Murley, G., West, N. A., Ramezani, S., Dede, C., Humbert-Vidan, L., & Fuller, C. D. (2025). *Adaptive radiation therapy for head and neck cancer* (Preprint). PubMed Central.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC12324559/>

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. En *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE.

<https://doi.org/10.1109/3DV.2016.79>

Myronenko, A. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

https://doi.org/10.1007/978-3-030-32245-8_46

National Cancer Institute. (n.d.). Radiation therapy to treat cancer.

<https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy>

National Comprehensive Cancer Network. (2024). *Head and neck cancers (Version 2.2024)*. https://www.nccn.org/professionals/physician_gls/pdf/head-and-neck.pdf

Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., & Jiang, S. (2019). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		126
--	--	-----

via deep learning. *Scientific Reports*, 9(1), 1076. <https://doi.org/10.1038/s41598-018-37793-3>

Nutting, C., Dearnaley, D. P., & Webb, S. (2002). Intensity-modulated radiotherapy: Current status and issues of interest. *Radiotherapy and Oncology*, 64(2), 89–94. [https://doi.org/10.1016/S0167-8140\(02\)00049-1](https://doi.org/10.1016/S0167-8140(02)00049-1)

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>

Oppitz, H., Eckl, M., Siebenlist, K., Boda-Heggemann, J., Abo-Madyan, Y., Giordano, F. A., Hesser, J., & Fleckenstein, J. (2025). *Knowledge-based automated radiation therapy treatment planning utilizing dose prediction with a 2.5D U-Net. Physics and Imaging in Radiation Oncology*. Advance online publication. <https://doi.org/10.1016/j.phro.2025.100737>

Osman, A. F. I., Tamam, N. M., & Yousif, Y. A. M. (2023). A comparative study of deep learning-based knowledge-based planning methods for 3D dose distribution prediction of head and neck. *Journal of Applied Clinical Medical Physics*, 24(9), e14015. <https://doi.org/10.1002/acm2.14015>

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38. <https://doi.org/10.1038/s41591-021-01614-0>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		127
--	--	-----

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. En *MICCAI 2015* (LNCS 9351, pp. 234–241). Springer.

https://doi.org/10.1007/978-3-319-24574-4_28

Rashad, M. N. H., Karki, A., Czak, J., Alves, V. G., Nourzadeh, H., Choi, W., & Siebers, J. V. (2024). Divergent clinical equivalence findings from DVH and NTCP metrics for alternative OAR delineations with increasing setup variability in head-and-neck radiotherapy. *arXiv preprint arXiv:2401.05656v2*. <https://arxiv.org/abs/2401.05656>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215.

<https://doi.org/10.1038/s42256-019-0048-x>

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep learning in electronic health records: A systematic review. *Journal of Biomedical Informatics*, 83, 168–185.

<https://doi.org/10.1016/j.jbi.2018.06.007>

Shmueli, G. (2010). To explain or to predict. *Statistical Science*, 25(3), 289–310.

<https://doi.org/10.1214/10-STS330>

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. <https://doi.org/10.1186/s40537-019-0197-0>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		128
--	--	-----

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020. *CA: A Cancer Journal for Clinicians*, 71(3), 209–249. <https://doi.org/10.3322/caac.21660>

Sun, Z., Xia, X., Fan, J., Zhao, J., Zhang, K., & Wang, J. (2022). A hybrid optimization strategy for deliverable intensity-modulated radiotherapy plan generation using deep learning-based dose prediction. *Medical Physics*, 49(3), 1344–1356. <https://doi.org/10.1002/mp.15462>

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books

Teng, L., Wang, B., Xu, X., Zhang, J., Mei, L., Feng, Q., & Shen, D. (2024). Beam-wise dose composition learning for head and neck cancer dose prediction in radiotherapy. *Medical Image Analysis*, 92, 103045. <https://doi.org/10.1016/j.media.2023.103045>

Vallieres, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Aerts, H. J. W. L., Khaouam, N., Nguyen-Tan, P. F., Wang, C. S., Sultanem, K., Seuntjens, J., & El Naqa, I. (2017). Radiomics strategies for risk assessment. *Scientific Reports*, 7(1), 10117. <https://doi.org/10.1038/s41598-017-10371-5>

Vandewinckele, L., et al. (2020). Overview of AI-based applications in radiotherapy. *Physics in Medicine & Biology*, 65(22), 22TR01. <https://doi.org/10.1088/1361-6560/aba1e8>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		129
--	--	-----

Vayena, E., & Blasimme, A. (2022). Digital health health epistemology: The roles of data and theory in 21st-century medicine. *The Lancet Digital Health*, 4(1), e1-e2.

[https://doi.org/10.1016/S2589-7500\(21\)00244-6](https://doi.org/10.1016/S2589-7500(21)00244-6)

Verbakel, W. F. A. R., et al. (2009). VMAT for head-and-neck cancer. *IJROBP*, 74(1), 252–259. <https://doi.org/10.1016/j.ijrobp.2008.12.008>

Wahid, K. A., et al. (2024). Artificial intelligence uncertainty quantification in radiotherapy applications: A scoping review. *Radiotherapy and Oncology*, 201, 110542.

<https://doi.org/10.1016/j.radonc.2024.110542>

Wen, Y., Liu, W., Raj, B., Singh, R., & Jain, A. K. (2023). Feature normalization and regularization in deep learning: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2023.3241234>

Webb, S. (Ed.). (2001). *Intensity-modulated radiation therapy*. Institute of Physics Publishing.

Xiong, Y., Zhang, L., Chen, X., & Li, S. (2026). *Physics-informed large-kernel models for generalizable multi-site dose prediction*. *Nature Medicine AI*. Advance online publication.

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		130
--	--	-----

Yuan, X., Ma, C., Hu, M., Qiu, R. L. J., Salari, E., Martini, R., & Yang, X. (2025). Machine learning in image-based outcome prediction after radiotherapy: A review. *Journal of Applied Clinical Medical Physics*, 26(1), e14559. <https://doi.org/10.1002/acm2.14559>

Yue, M., Xue, X., Wang, Z., Lambo, R. L., Zhao, W., Xie, Y., Cai, J., & Qin, W. (2022). Dose prediction via distance-guided deep learning: Initial development for nasopharyngeal carcinoma radiotherapy. *Radiotherapy and Oncology*, 170, 198–204. <https://doi.org/10.1016/j.radonc.2022.03.012>

Zhang, Y., Chen, R., Tang, J., Stewart, W. F., & Sun, J. (2022). LEAP: Learning to predict clinical events from longitudinal electronic health records. Proceedings of the ACM SIGKDD Conference. <https://doi.org/10.1145/3534678.3539175>

Zarepisheh, M., Hong, L., Zhou, Y., Oh, J. H., Mechalakos, J. G., Hunt, M. A., Mageras, G. S., & Deasy, J. O. (2019). Automated intensity modulated treatment planning: The expedited constrained hierarchical optimization (ECHO) system. *Medical Physics*, 46(7), 2944–2954. <https://doi.org/10.1002/mp.13572>

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		131
--	--	-----

Anexos

Los siguientes anexos contienen los cuadernos de Google Colaboratory desarrollados durante la fase de trabajo de campo de esta investigación. Cada notebook documenta de forma reproducible el pipeline completo implementado para cada enfoque, incluyendo el código fuente, los resultados y las visualizaciones generadas.

Anexo A. Análisis estadístico y modelo supervisado Dataset HNSCC-3DCT-RT

Este cuaderno contiene el pipeline completo del enfoque de representación tabular...

- **Nombre del archivo:** Analisis_estadistico_y_modelo_dataset_HNSCC-DCT-RT.ipynb
- **Plataforma:** Google Colaboratory
- **Enlace:**

Evaluación de representaciones de datos tabular y volumétrica en el modelamiento predictivo aplicado a la planificación radioterapéutica del cáncer de cabeza y cuello		132
--	--	-----

https://colab.research.google.com/drive/1FFcC-vk-t7ia_khjJ7MrM_MM0CTipH3

Anexo B. Implementación del modelo volumétrico UNet3D Dataset OpenKBP

Este cuaderno contiene el pipeline completo del enfoque de representación volumétrica.

- **Nombre del archivo:** Montaje_modelo_explicacion_con_detalle.ipynb
- **Plataforma:** Google Colaboratory (GPU - CUDA)
- **Enlace:**

<https://colab.research.google.com/drive/1FI6tTiic461ZHu2q-A-xB6IAZPM1wFA?usp=sharing>