### **Predicción de riesgo de cáncer pulmonar mediante K-Nearest Neighbor y atributos clínicos basados en encuestas**

*Prediction of Lung Cancer Risk Using K-Nearest Neighbor and Clinical Survey-Based Attributes*

Amaro Rodriguez Luz Mia1, [luzamaro@upeu.edu.pe](mailto:luzamaro@upeu.edu.pe)

Aranda Castillo Leidy Aracely 1, [leidy.aranda@upeu.edu.pe](mailto:leidy.aranda@upeu.edu.pe)

Camargo Rojas Maria del Cielo 2, [2223110255@untels.edu.pe](mailto:2223110255@untels.edu.pe) [0009-0009-0033-4434]

Chipana Bandera Benjamin Adriano2, [2223110329@untels.edu.pe](mailto:2223110329@untels.edu.pe) [0009-0002-6869-2691]

Universidad Peruana Unión1

Universidad Nacional Tecnológica de Lima Sur 2

Resumen:

Palabras clave: cancer de pulmón; knn; machine learning; pronóstico.

Abstract:

Keywords: lung cancer; knn; machine learning; prognoses.

Citación:

## INTRODUCCIÓN

El cáncer de pulmón es uno de los mayores desafíos para la salud pública a nivel mundial debido a su alta prevalencia.(1) Es la principal causa de muerte por cáncer, con un 18% de los fallecimientos totales. En mujeres, ocupa el tercer lugar en incidencia (8,4%), después del cáncer de mama (24,5%), mientras que en hombres es el cáncer más común tanto en incidencia (14,3%) como en mortalidad (21,5%) (2,3). Según la Organización Mundial de la Salud (OMS), el cáncer de pulmón es el segundo más frecuente globalmente y causa aproximadamente 1.8 millones de muertes al año.

La distribución del cáncer de pulmón no es uniforme en el mundo y varía considerablemente entre regiones. Estas diferencias se relacionan con factores como el nivel socioeconómico, la exposición a contaminantes ambientales, la predisposición genética y, especialmente, el tabaquismo. De todos estos, el consumo de tabaco sigue siendo el principal factor de riesgo, asociado a más del 80 % de los casos diagnosticados (4). Aunque se han identificado otros determinantes, como la contaminación del aire (tanto en interiores como en exteriores), la exposición a sustancias carcinógenas en el entorno laboral y factores hereditarios, el tabaquismo continúa teniendo el mayor impacto en la incidencia global de esta enfermedad.(4)

En América Latina, el cáncer de pulmón es la principal causa de muerte entre todos los canceres (12% del total), con una tasa de 1 por cada 10 muertes oncológicas Uruguay, Cuba y Argentina presentan las mayores incidencias, correlacionadas con su histórica alta prevalencia de tabaquismo adulto (5). No obstante, se observa una disminución en casos masculinos, atribuida a políticas como el aumento de impuestos al tabaco el cual se ha descrito en algunas regiones, surgiendo la posibilidad en la reducción de casos en los próximos años. (5,6)

El Perú ha experimentado un aumento de nuevos casos registrados de cáncer de pulmón en el último año. Las estadísticas más recientes, al 2023, registraron 471 nuevos diagnósticos. (7–9)

Si bien el cáncer de pulmón no es el de mayor incidencia en el país (se ubica en el séptimo lugar), su impacto en la mortalidad es alarmante, posicionándose en segundo lugar como causa de muerte por cáncer en la población peruana, solo superada por el cáncer de estómago, que sigue siendo el más letal. (8) Esta contradicción de menor incidencia con alta tasa de mortalidad, es consecuencia de los diagnósticos tardíos y las limitaciones en el acceso a tratamiento oportuno. (9)

Con estos datos descritos se resalta la importancia de obtener un diagnóstico precoz del cáncer de pulmón, lo que permitiría una mejora significativa en el pronóstico de los pacientes.(10) Cuando el cáncer es detectado en sus estadios iniciales (I y II) la supervivencia a 5 años puede superar hasta el 60% a diferencia de la detección en estadios avanzados (III y IV), donde esta cifra desciende incluso a menos del 10%. (11)

Este estudio busca proporcionar una herramienta accesible para centros de salud con recursos limitados, donde las encuestas clínicas podrían servir como primer filtro de riesgo antes de derivar a pruebas costosas.

## MATERIALES Y METODOS:

Los datos (data set) fueron obtenidos a partir de repositorio Kaggle, con el nombre de “Lung Cancer Dataset”, este dataset contiene información clínica y de estilo de vida de los pacientes, con el objetivo de predecir la presencia de cancer pulmonar, cuenta con 16 atributos y 4455 instancias, con valores numéricos de los cuales 429 corresponden a casos negativos a cáncer de pulmón y 4026 a casos positivos.

En el proceso de preprocesamiento de los datos, se llevó a cabo una transformación en los valores de los atributos 'GENDER' y 'LUNG\_CANCER'. Para el atributo 'GENDER', se asignó el valor numérico '1' a los registros correspondientes a pacientes masculinos ('M') y el valor '2' a aquellos correspondientes a pacientes femeninas ('F'). En cuanto al atributo 'LUNG\_CANCER', se transformaron los valores de 'YES' a '2' y los de 'NO' a '1', facilitando así la codificación de los datos para su posterior análisis.

El conjunto de datos no requirió un proceso de limpieza, ya que no presentaba valores faltantes ni variables irrelevantes o indescifrables, lo que facilitó su análisis sin la necesidad de intervenciones adicionales. Los atributos incluyen variables demográficas, síntomas y factores de riesgo asociados con el cáncer de pulmón. La variable objetivo es la presencia o ausencia de cáncer de pulmón.

En la Tabla 1 se presenta información de los Atributos contenidos en el conjunto de datos, el tipo de datos, así como el Rango de Valores. Al finalizar este paso, se creó el archivo: “CANCERLUNG.csv”

**Tabla 1:** Información de los atributos del conjunto de datos

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Tipo** | **Rango de Valores** |
| Gender | Nominal | 1 = Masculino, 2 = Femenino |
| Age | Número entero | [1, Más > |
| Smoking | Binomial | Sí = 2, No = 1 |
| Yellow fingers | Binomial | Sí = 2, No = 1 |
| Anxiety | Binomial | Sí = 2, No = 1 |
| Peer\_pressure | Binomial | Sí = 2, No = 1 |
| Chronic Disease | Binomial | Sí = 2, No = 1 |
| Fatigue | Binomial | Sí = 2, No = 1 |
| Allergy | Binomial | Sí = 2, No = 1 |
| Wheezing | Binomial | Sí = 2, No = 1 |
| Alcohol | Binomial | Sí = 2, No = 1 |
| Coughing | Binomial | Sí = 2, No = 1 |
| Shortness of Breath | Binomial | Sí = 2, No = 1 |
| Swallowing Difficulty | Binomial | Sí = 2, No = 1 |
| Chest pain | Binomial | Sí = 2, No = 1 |
| Lung Cancer | Binomial | Sí = 2, No = 1 |

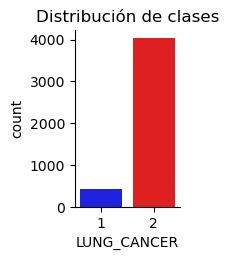
**Entrenamiento del algoritmo KNN**

Primero, para la construcción del modelo de clasificación, emplearemos el algoritmo de aprendizaje automático k-NN (k-Nearest Neighbors), que se basa en la proximidad de los puntos de datos para realizar predicciones sobre la pertenencia a una determinada categoría.

Para ello, primero se va a entrenar al algoritmo y reducir su tasa de error en función al valor de la variable n\_neighbors, osea, el número de datos vecinos más cercanos mediante los cuales se realizará la predicción. Asimismo, para integrar el modelo con la interfaz del formulario, utilizaremos el conjunto de herramientas proporcionado por el framework Flask, facilitando la conexión y la interacción entre el usuario y el modelo

Este código realiza un análisis de datos sobre cáncer de pulmón. Utiliza un archivo CSV con información médica para **entrenar un modelo de clasificación** que predice si un caso es benigno o maligno.

**Figura 1.** Distribución de datos por el resultado



Segundo, se **importan las bibliotecas necesarias** para el manejo de datos, gráficos y aprendizaje automático. Luego, se **lee el archivo CANCERLUNG.csv**, y se separan los datos entre las características (variables predictoras) y el resultado esperado (diagnóstico de cáncer).

Después, se divide el conjunto de datos en dos partes: una para entrenamiento y otra para prueba. Con estos datos, se **entrena un modelo de clasificación KNN** (K-Nearest Neighbors) usando inicialmente 1 vecino para predecir. Se evalúa el rendimiento del modelo con una **matriz de confusión** y un **reporte de clasificación**, que muestra métricas como precisión y sensibilidad.

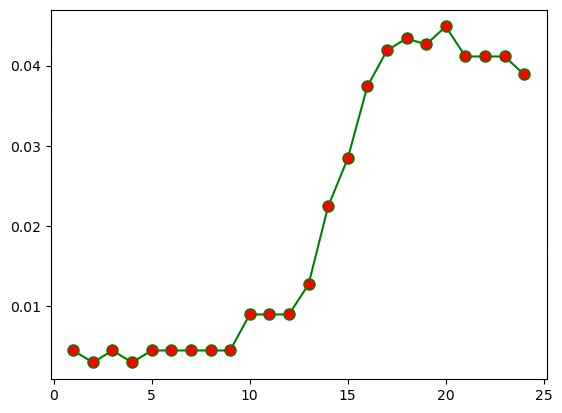
A continuación, se realiza un análisis para **ver cómo varía la tasa de error** al cambiar el número de vecinos (de 1 a 25). Esto se grafica para ayudar a elegir el mejor valor.

Finalmente, se **elige un valor óptimo de vecinos** (en este caso 2), se vuelve a entrenar el modelo con ese valor, y se genera nuevamente la matriz de confusión y el reporte, esta vez con etiquetas claras: "Benigno" y "Maligno".

**Tabla 2.** Reporte de clasificación inicial con n\_neighbors=1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Benigno | 0.97 | 1.00 | 0.98 | 115 |
| Maligno | 1.00 | 1.00 | 1.00 | 1222 |
| Acurracy |  |  | 1.00 | 1337 |
| Macro avg | 0.98 | 1.00 | 0.99 | 1337 |
| Wighted avg | 1.00 | 1.00 | 1.00 | 1337 |

**Figura 2**. Reporte de la tasa de error



Luego, integra ese modelo en una **aplicación web** para que los usuarios puedan hacer predicciones desde un formulario en HTML.Para eso primero, se importan bibliotecas necesarias para el manejo de datos, visualización y creación del modelo. Se carga el archivo DSCancerDePulmon.csv y se separan las variables predictoras del resultado (LUNG\_CANCER).

Con esos datos, se entrena un modelo de tipo K-Nearest Neighbors (KNN) utilizando como valor óptimo de vecinos el número 2 (que fue hallado previamente en otro análisis).

Después se utiliza **Flask**, un microframework web en Python, para construir la parte que interactúa con el usuario. Se define una ruta principal (/) que acepta solicitudes tipo GET (para mostrar el formulario) y POST (para procesar los datos ingresados).

Cuando alguien llena y envía el formulario, los datos ingresados (como género, edad, síntomas, hábitos, etc.) son recolectados y validados para asegurarse de que estén en el rango correcto.

Si los datos son válidos, el modelo predice si el diagnóstico es benigno o maligno y con qué **probabilidad**. Esa información se convierte en un mensaje que se envía de vuelta al usuario.

Si los datos son incorrectos o hay un error, se genera un mensaje de advertencia explicando que hubo un problema.

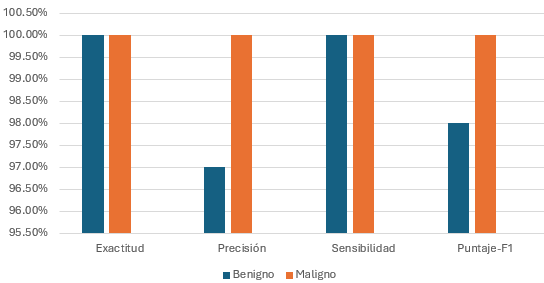
Finalmente, se ejecuta la aplicación web, permitiendo que el formulario funcione localmente y que el modelo pueda responder con predicciones en tiempo real.

## III. RESULTADOS

3.1 Desempeño del modelo

El modelo de machine larning llamado KNN\_LungCancer, donde en la Figura 3 se observa una exactitud = 98%, precisión = 97%, sensibilidad = 100% y puntaje-f1 = 98% para Benigno ;y exactitud = 100%, precisión =100%, sensibilidad = 100% y puntaje-f1 = 100% para maligno.

**Figura 3.** Desempeño del modelo KNN\_LungCancer



3.2 Predicciones del modelo

Con relación a este modelo, con se observa en la Tabla 3 se desarrolló predicciones utilizando para ellos un comando knn\_mm.predict([[]]):

**Tabla 3.** Predicciones con el modelo KNN\_LungCancer

|  |  |
| --- | --- |
| Comando | Resultado |
| print(knn\_mm.predict([[1,69,1,2,2,1,1,2,1,2,2,2,2,2,2]])) | Maligno |
| print(knn\_mm.predict([[2,59,1,1,1,2,1,2,1,2,1,2,2,1,2]])) | Benigno |
| print(knn\_mm.predict([[2,63,1,2,1,1,1,1,1,2,1,2,2,1,1]])) | Benigno |
| print(knn\_mm.predict([[1,74,2,1,1,1,2,2,2,1,1,1,2,2,2]])) | Maligno |

En relación con el modelo en la tabla 4 comparamos la probabilidad de clasificación de benigno o maligno utilizando knn\_mm.predict([[]]):

Tabla 4. Probabilidad de clasificar benigno o maligno

|  |  |  |
| --- | --- | --- |
| Comando | Probabilidad de benigno | Probabilidad de maligno |
| print(knn\_mm.predict\_proba([[1,69,1,2,2,1,1,2,1,2,2,2,2,2,2]])) | 0% | 100% |
| print(knn\_mm.predict\_proba([[2,59,1,1,1,2,1,2,1,2,1,2,2,1,2]])) | 100% | 0% |
| print(knn\_mm.predict\_proba([[2,63,1,2,1,1,1,1,1,2,1,2,2,1,1]])) | 100% | 0% |
| print(knn\_mm.predict\_proba([[1,74,2,1,1,1,2,2,2,1,1,1,2,2,2]])) | 0% | 100% |

## IV. DISCUSIÓN

## V. CONCLUSIÓN

## VI. AGRADECIMIENTO

Agradecimiento a nuestros docentes Daniel Levano y Flor Cerdán.

## VII. REFERENCIAS

1. Cordova Sanchez GA, Ugarte Palacios CV, Ugarte Palacios NA, Morales Labre KO. Cáncer de pulmón y su importancia en el diagnóstico primario. RECIAMUC. 2022 May 4;6(2):208–17.

2. Giraldo-Osorio A, Ruano-Ravina A, Rey-Brandariz J, Arias-Ortiz N, Candal-Pedreira C, Pérez-Ríos M. Lung cancer mortality trends in Colombia, 1985-2018. Revista Panamericana de Salud Publica/Pan American Journal of Public Health. 2022;46.

3. World Health Organization. Lung Cáncer . 2023.

4. Nogueira JF, Mota AL, Araújo APF, Figueiredo BQ de, Santos GM, Silva LCS e, et al. Epidemiological profile of lung cancer in Brazil between 2013 and 2020. Research, Society and Development [Internet]. 2021 Dec 11 [cited 2025 May 18];10(16):e203101623566–e203101623566. Available from: https://rsdjournal.org/index.php/rsd/article/view/23566

5. Piñeros M, Laversanne M, Barrios E, Cancela M de C, de Vries E, Pardo C, et al. An updated profile of the cancer burden, patterns and trends in Latin America and the Caribbean. The Lancet Regional Health - Americas [Internet]. 2022 Sep 1 [cited 2025 Apr 5];13. Available from: https://www.thelancet.com/action/showFullText?pii=S2667193X22001119

6. Carolina Pimentel Duarte Novo1, Luciane Zanin2, Viviane Perin3, Marcia Regina Campos Costa da Fonseca4. Mortalidad por cáncer de pulmón: perfil y tendencias tras la entrada en vigor de la Ley Antitabaco. [cited 2025 May 18]; Available from: http://seer.uftm.edu.br/revistaeletronica/index.php/refacs

7. Rodriguez-Morales AJ, Lopardo G, Verbanaz S, Orduna T, Lloveras S, Azeñas-Burgoa JM, et al. Latin America: Situation and preparedness facing the multi-country human monkeypox outbreak. The Lancet Regional Health - Americas [Internet]. 2022 Sep 1 [cited 2025 May 18];13:100318. Available from: https://www.thelancet.com/action/showFullText?pii=S2667193X22001351

8. International Agency for Research on Cancer. Statistics at a glance, 2022 Top 5 most frequent cancers in Perú Number of new cases 72 827.

9. Ruiz R, Galvez-Nino M, Poquioma E, Limache-García A, Amorin E, Olivera M, et al. EDITORIAL: LUNG CANCER WORLDWIDE Lung Cancer in Peru A Perspective From Peru. Journal of Thoracic Oncology [Internet]. 2020 [cited 2025 May 18];15(6):891–8. Available from: https://doi.org/10.1016/j.jtho.2020.01.018

10. col Arroyo-Hernández M AHM, Arroyo-Hernández M, Neumól M, Felipe Alva-López L, Radiól M, Rendón A, et al. Clinical Practice Guideline for the early detection and timely referral of lung cancer patients. Salud Publica Mex. 2022;64:530-538. Hospital Central Militar Ciudad de México, México [Internet]. 2022 [cited 2025 May 18];64(10):530–8. Available from: https://doi.org/10.21149/13919

11. National Cancer Institute (Surveillance E and ERP. SEER 5-year Relative survival Rates, 2014-2020 [Internet]. 2024 [cited 2025 Apr 6]. Available from: https://seer.cancer.gov/statistics-network/explorer/application.html?site=47&data\_type=4&graph\_type=5&compareBy=stage&chk\_stage\_104=104&chk\_stage\_105=105&chk\_stage\_106=106&chk\_stage\_107=107&series=age\_range&chk\_age\_range\_1=1&chk\_age\_range\_9=9&chk\_age\_range\_141=141&chk\_age\_range\_157=157&sex=1&race=1&advopt\_precision=1&advopt\_show\_ci=on&hdn\_view=0&advopt\_show\_apc=on&advopt\_display=2#resultsRegion0