Visualización de datos: Caso Cáncer Pulmonar

Jose Perez Mamani, and Henrry Arias Mamani

Abstract—The abstract goes here.

Index Terms—Lung cancer prediction, PCA, visualization analysis.

1 Introduction

El El cáncer pulmonar es una de las principales causas de mortalidad a nivel mundial, con un pronóstico a menudo desfavorable debido a la detección tardía de la enfermedad. A pesar de los avances en la oncología y los tratamientos personalizados, la tasa de supervivencia a cinco años sigue siendo baja. La ciencia de datos ha emergido como una herramienta poderosa para abordar los desafíos en la detección temprana, diagnóstico y personalización del tratamiento en cáncer pulmonar. Mediante el uso de técnicas avanzadas, es posible analizar grandes volúmenes de datos clínicos y genómicos para identificar patrones y desarrollar modelos predictivos que puedan mejorar los resultados para los pacientes .

1.1 Objetivo

El principal objetivo de este artículo es identificar y analizar los factores demográficos y médicos que influencian la supervivencia de los pacientes con cáncer de pulmón. En particular, el artículo intente evaluarse cómo las variables tales como la edad, el género, el estado de tabaquismo, el índice de masa corporal, el nivel de colesterol, la historia familiar acerca de cáncer, y la etapa del cáncer en el momento del diagnóstico afectan el tratamiento y los resultados de los pacientes.

1.2 Problema

A pesar de los avances realizados en el tratamiento de del cáncer de pulmón, el índice de mortalidad continúa siendo elevado. El problema que pretende abordar este articulo radica en la falta de comprensión detallada acerca de cómo varios de factores demográficos, médicos y terapéuticos influyen en la supervivencia de los pacientes con cáncer de pulmón. Aunque existe una serie de estudios previos que exploran sobre este tema, la complejidad de la interrelación entre varios factores, sus afectos al tratamiento y la supervivencia no ha sido suficientemente explorada. Este articulo intenta llenar este vacío ofreciendo un análisis a fondo de varios factores relacionados con la mortalidad de los pacientes que sufren de cáncer de pulmón y proporcionando la base para la mejora de las estrategias de tratamiento. Con la ayuda de este enfoque, podrás centrar tu análisis en la identificación de patrones y correlaciones que puedan ser útiles para la mejora de enfoques terapéuticos y, como resultado, de las tasas de supervivencia.

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received Agosto 18, 2005; revised August 18, 2014.

2 Trabajos relacionados

(author?) [2], se investigó la influencia de factores pronósticos como la edad, el estadiamiento y la extensión del tumor en la supervivencia de mujeres con cáncer de mama, utilizando modelos de riesgos proporcionales de Cox y riesgos competitivos de Fine-Gray. Se analizaron los datos de una cohorte retrospectiva de 524 mujeres en Campinas, Brasil, diagnosticadas entre 1993 y 1995. Se aplicaron estos modelos para evaluar la influencia de los factores mencionados en la mortalidad específica por cáncer de mama y otras causas competidoras. Las curvas de supervivencia estimadas por Kaplan-Meier mostraron diferencias significativas en las muertes por cáncer de mama y por riesgos competidores, aunque la edad no fue un factor significativo en ninguno de los modelos.

El uso de los modelos de Cox y Fine-Gray como muestra **(author?)** [2] ofrece ventajas al permitir una estimación más precisa de los riesgos asociados al cáncer de mama y otras causas de muerte. Mientras que los modelos de Cox son útiles para evaluar el impacto de las covariables en la supervivencia, el modelo de Fine-Gray permite considerar la presencia de riesgos competidores, lo que ofrece una visión más realista del escenario clínico. Sin embargo, una desventaja es que la complejidad de estos modelos puede dificultar su interpretación y aplicación práctica en ciertos contextos. Además, aunque ambos modelos identificaron factores pronósticos similares, la ausencia de significancia de la edad en estos análisis sugiere que otros factores podrían ser más determinantes en la supervivencia de estas pacientes.

En el articulo **(author?)** [3], se implementado una combinación de Análisis de Componentes Principales (PCA) y la técnica de remuestreo Synthetic Minority Oversampling Technique (SMOTE) para mejorar la precisión en la predicción en un conjunto de datos sobre cáncer de pulmón. PCA se utilizó inicialmente para reducir la dimensionalidad del conjunto de datos, comprimiendo el espacio de características y eliminando atributos redundantes e irrelevantes. Posteriormente, aplicamos SMOTE para equilibrar la distribución de clases, generando nuevas muestras sintéticas en la clase minoritaria. Esta combinación

metodológica no solo facilita un modelo de clasificación más eficiente, sino que también incrementa la diversidad del dominio de las muestras, mejorando así el rendimiento del clasificador Naïve Bayes aplicado en la etapa final.

Los resultados obtenidos en el articulo **(author?)** [3] confirman la eficacia de la metodología propuesta, mostrando mejoras significativas en varias métricas de evaluación, incluyendo la precisión global, la tasa de falsos positivos, la precisión y el recall. En particular, se observó que la aplicación secuencial de PCA seguida por SMOTE permitió una mejora considerable en la exactitud del modelo, aumentando de un 60% a más del 80% después de las iteraciones de SMOTE. Estos hallazgos sugieren que la combinación de técnicas de reducción de dimensionalidad y remuestreo puede ser una estrategia efectiva para abordar problemas de desequilibrio de clases y características redundantes en conjuntos de datos médicos.

3 Marco Teórico

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística de reducción de dimensionalidad utilizada para transformar un conjunto de variables posiblemente correlacionadas en un conjunto de variables no correlacionadas, conocidas como componentes principales. Cada componente principal es una combinación lineal de las variables originales, capturando la mayor variabilidad posible en los datos. PCA se basa en el cálculo de vectores propios y valores propios de la matriz de covarianza del conjunto de datos, donde los vectores propios determinan la dirección de las nuevas dimensiones y los valores propios indican la cantidad de varianza capturada por cada componente. Esta técnica es especialmente útil en aplicaciones con datos de alta dimensionalidad, como el análisis de imágenes y genética, al reducir el número de dimensiones y mejorar el rendimiento de los modelos predictivos [4].

t-Distributed Stochastic Neighbor Embedding (t-SNE) es una técnica de reducción de dimensionalidad no lineal que se utiliza principalmente para la visualización de datos de alta dimensionalidad. A diferencia de técnicas como PCA, que conservan la variabilidad global de los datos, t-SNE se enfoca en preservar las relaciones locales, es decir, la proximidad entre puntos en el espacio original. Lo hace al modelar las distancias entre pares de puntos usando distribuciones de probabilidad tanto en el espacio original como en el espacio de menor dimensión, y luego minimizando la divergencia de Kullback-Leibler entre estas distribuciones. Esto permite que t-SNE revele estructuras complejas y patrones ocultos en los datos que pueden ser difíciles de discernir con otras técnicas de reducción de dimensionalidad. Es ampliamente utilizado en la visualización de conjuntos de datos en áreas como la genética, la visión por computadora y el procesamiento del lenguaje natural [5].

El análisis de correlación tetracórica es una técnica estadística utilizada para estimar la correlación entre dos variables latentes continuas a partir de datos observados en forma de variables dicotómicas (binarias). Este tipo de correlación es útil cuando las variables subyacentes se consideran continuas, pero solo se pueden observar en dos categorías. La correlación tetracórica asume que las variables subyacentes tienen una distribución normal bivariante, y a partir de esta suposición, se estima la correlación que podría existir si se midieran de manera continua. Este método se aplica en áreas como psicometría, donde los cuestionarios a menudo generan respuestas binarias que pueden ser indicativas de rasgos subyacentes continuos [6].

3.1 Categorización

En términos médicos, el Índice de Masa Corporal (BMI, por sus siglas en inglés) y el nivel de colesterol están relacionados en el sentido de que ambos son indicadores importantes de la salud cardiovascular y metabólica.

- BMI: Es una medida del peso en relación con la altura. Un BMI alto, que indica sobrepeso u obesidad, se asocia con un mayor riesgo de desarrollar enfermedades cardiovasculares, diabetes tipo 2 y otros problemas de salud.
- Colesterol: El colesterol elevado, especialmente el LDL (colesterol "malo"), puede aumentar el riesgo de enfermedades cardíacas y accidentes cerebrovasculares.
- Categorías de BMI
  - Bajo peso: BMI $<$ 18.5
  - Normal: BMI 18.5 - 24.9
  - Sobrepeso: BMI 25 - 29.9
  - Obesidad: BMI $\geq$ 30
- Categorías de colesterol
  - Colesterol total
    - Deseable: ¡ 200 mg/dL
    - Límite superior: 200-239 mg/dL
    - Alto: $\geq$ 240 mg/dL
  - LDL colesterol
    - Óptimo: ¡ 100 mg/dL
    - Cercano al óptimo: 100-129 mg/dL
    - Alto: 130-159 mg/dL
    - Muy alto: $\geq$ 160 mg/dL
  - HDL colesterol (colesterol "bueno")
    - Bajo (riesgo alto): ¡ 40 mg/dL
    - Normal: 40-59 mg/dL
    - Alto (beneficioso): $\geq$ 60 mg/dL

Estas categorías pueden ayudar a identificar patrones en los datos y establecer posibles correlaciones entre BMI y niveles de colesterol [7][8][9].

4 Análisis de tareas

Después de identificar las principal retos a los que se enfrentan los expertos y comprender como se estructuran los datos, realizamos una serie de cuestiones que se deben de investigar. Ha quedado claro que los expertos están interesados en comprender la dinámica del cancer pulmonar mediante análisis de patrones. A partir de las revisiones en clase, compilamos la siguiente lista de tareas analíticas:

- Factores que influyen en el cáncer pulmonar (T1): ¿Qué factores influyen en el diagnostico del cáncer

pulmonar? ¿Cómo afectan estos factores al diagnostico de pacientes? ¿Por qué algunos factores tienen un impacto más significativo que otros el diagnostico de la enfermedad?
- Análisis de pacientes con diagnósticos y comportamiento (T2): ¿Cómo varían los comportamientos y patrones entre pacientes con diferentes diagnósticos de cáncer pulmonar? ¿Existen diferencias notables en el pronóstico y evolución entre distintos grupos de pacientes?
- Distribución de variables clave (T3): ¿Cuál es la distribución de las variables principales en los pacientes con cáncer pulmonar? ¿Cómo influye esta distribución en la dinámica y el tratamiento del cáncer pulmonar?

5 Propuesta

5.1 Análisis de datos

El conjunto de datos utilizado en este artículo se descargó de Kaggle. Registra un total de 56000 pacientes y un total de 16 síntomas de los pacientes. La Table 1 muestra la atribución de los datos.

5.2 El sistema Cancer pulmonar

Basándonos en las tareas analíticas descritas en la section 4, hemos desarrollado una propuesta para explorar datos espacio-temporales sobre el diagnóstico de cáncer pulmonar. Esta propuesta permite consultar, filtrar y visualizar dichos datos de manera efectiva. Los módulos y la arquitectura de la solución se ilustran en la Figure 1.

En primer lugar, creamos un data frame para utilizarlo como repositorio central de los datos. Eliminamos la columna 'id', ya que es un identificador correlativo que no se utilizará en el procesamiento posterior. Durante el preprocesamiento, verificamos la presencia de datos duplicados utilizando el método duplicated seguido de sum para obtener la cantidad total. No se encontraron datos duplicados.

En la Table 5realizamos un análisis del tipo de atributos disponibles, lo que nos permitirá aplicar técnicas de categorización más adelante.En la Table 3 se detalla el tipo de variables presentes en el conjunto de datos, y en la Table 2 confirmamos que no existen valores nulos.

No.	Atributo	Descripción	Ejemplo
1	id	Identificador único de paciente	1, 2, 3, ..., 1048575
2	age	Edad del paciente	4, ..., 104
3	gender	Sexo del paciente	masculino y femenino
4	country	País o región donde reside el paciente	Protugal, Alemania, etc
5	diagnosis_date	Fecha en la que al paciente se le diagnosticó cáncer de pulmón	2014-06-03, ...,2024-05-31
6	cancer_stage	Estadío del cáncer de pulmón en el momento del diagnóstico	estadio; I, II, III, IV
7	family_history	Indica si hay antecedentes familiares de cáncer	true, false
8	smoking_status	Condición de fumador del paciente	fumador actual, exfumador, nunca fumó, fumador pasivo
9	bmi	Índice de Masa Corporal del paciente en el momento del diagnóstico	16,..., 45
10	cholesterol_level	Nivel de colesterol del paciente	150, ...,300
11	hypertension	Indica si el paciente tiene hipertensión	0, 1
12	asthma	Indica si el paciente tiene asma	0, 1
13	cirrhosis	Indica si el paciente tiene cirrosis hepática	0, 1
14	other_cancer	Indica si el paciente ha tenido algún otro tipo de cáncer además del diagnóstico primario	0, 1
15	treatment_type	Tipo de tratamiento que recibió el paciente	cirugía, quimioterapia, radiación, combinado
16	end_treatment_date	Fecha en la que el paciente completó su tratamiento contra el cáncer o falleció	2014-06-03, ...,2024-05-31
17	Fsurvived	Indica si el paciente sobrevivió	0, 1

TABLE 1: Descripción de características de la base de datos

6 Diseño Visual

7 Resultados

8 Discusión

Según la imagen, podemos concluir que una persona con obesidad debería someterse a un chequeo preventivo de cáncer, ya que presenta una mayor probabilidad de desarrollar esta enfermedad.

9 Conclusion

The conclusion goes here.

Appendix A
Colab TrabajoCancerPulmon v2.ipynb

Detalle de la aplicación

- Lenguaje de Programación: Python
- Servidor: Google
- Librerias Principales; sklearn, matplotlib, pandas,seaborn, numpy

Codigo Colab

Característica	Cantidad
age	0
bmi	0
gender	0
country	0
treatment_type	0
smoking_status	0
family_history	0
cancer_stage	0
diagnosis_date	0
end_treatment_date	0
survived	0
other_cancer	0
cirrhosis	0
asthma	0
hypertension	0
cholesterol_level	0

TABLE 2: Tipos de valores nulos

Tipo de variable	Atributo
Continuos	age, bmi, cholesterol_level
Discretos	survived, other_cancer, cirrhosis, asthma, hypertension , smoking_status, cancer_stage, _date, end_treatment_date
Categóricos	gender, country, treatment_type diagnosis, family_history

TABLE 3: Tipos de variable por atributos

	unique	min	max
id	56000	1	56000
age	79	15	101
bmi	291	16	45
gender	2	NAN	NAN
country	27	NAN	NAN
treatment_type	4	NAN	NAN
smoking_status	4	NAN	NAN
family_history	2	NAN	NAN
cancer_stage	4	NAN	NAN
diagnosis_date	3651	2014-06-02	2024-05-30
end_treatment_date	4097	2014-12-02	2026-05-15
survived	2	0	1
other_cancer	2	0	1
cirrhosis	2	0	1
asthma	2	0	1
hypertension	2	0	1
cholesterol_level	151	150	300

TABLE 4: Valores únicos, mínimos y máximos por atributo

Tipo de dato	Atributo
float64	age, bmi
Category	gender, country, treatment_type, smoking_status, family_history, cancer_stage
datetime64[ns]	diagnosis_date, end_treatment_date
int64	survived, other_cancer, cirrhosis, asthma, hypertension, cholesterol_level

TABLE 5: Tipos de datos por atributos

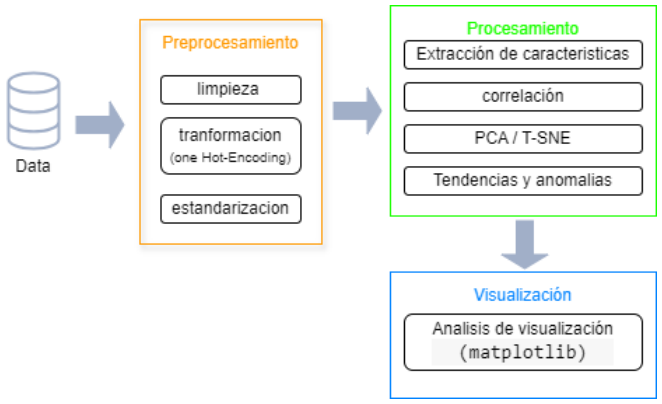

Fig. 1: Descripción general de pipeline de la propuesta para diagnosticar cancer pulmonar

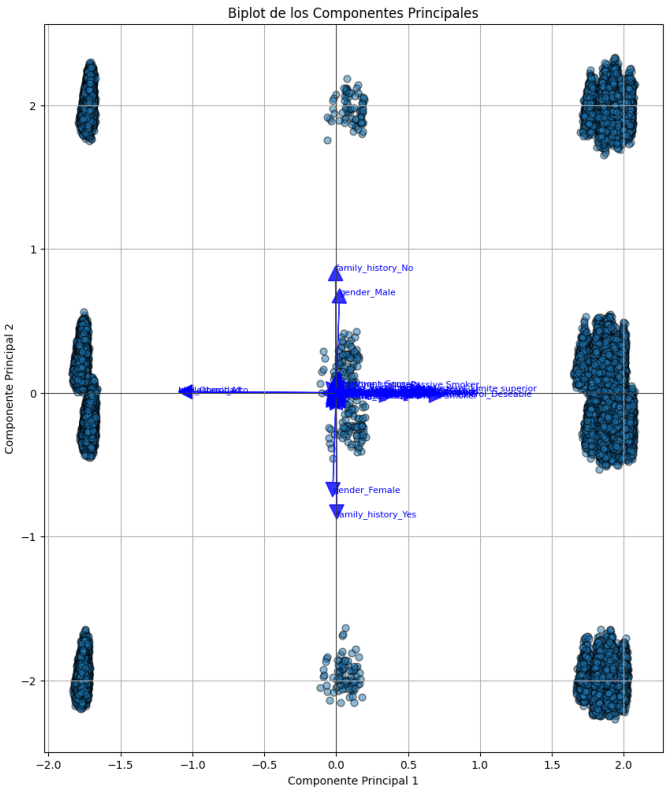

Fig. 2: Resultado de la aplicación de PCA para reducir las variables a dos. Los vectores muestran el peso de cada variable por componente.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] R. O. Ferraz and D. C. Moreira-Filho, "Análise de sobrevivência de mulheres com câncer de mama: modelos de riscos competitivos," *Ciência & Saúde Coletiva*, vol. 22, no. 11, pp. 3743–3754, Nov. 2017. [Online]. Available: https://doi.org/10.1590/1413-812320172211.05092016.

[3] Author(s) Name, "An Analysis of Machine Learning Models for Efficient Data Processing," *International Journal of Computer Applications*, vol. 77, no. 3, pp. 33–38, 2013. [Online]. Available: DOI: 10.48550/arXiv.1403.1949.

[4] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

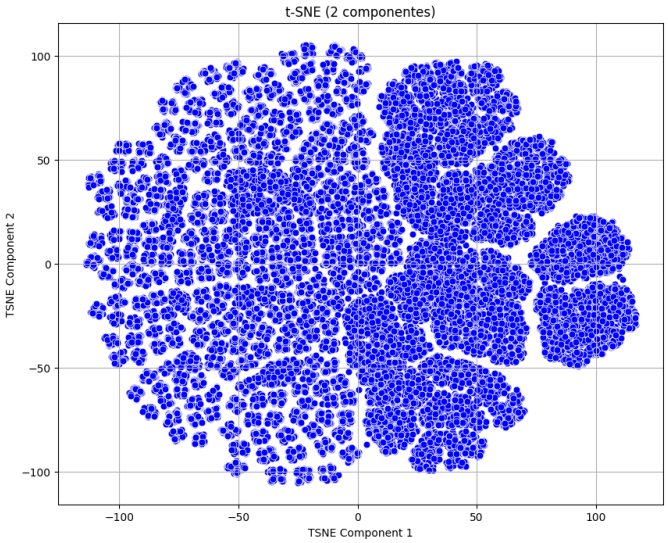

Fig. 3: Resultado de la aplicación de t-SNE. Se pueden apreciar agrupaciones de características.

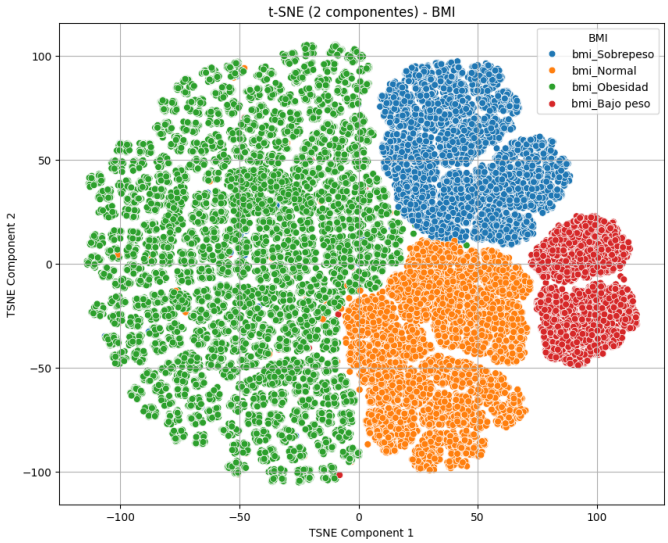

Fig. 5: Clusters de acuerdo al índice de masa corporal (BMI)

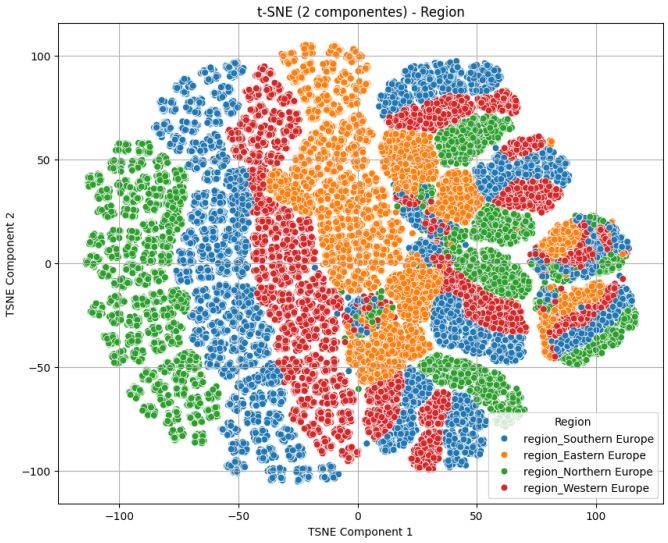

Fig. 4: Clusters según la región de Europa

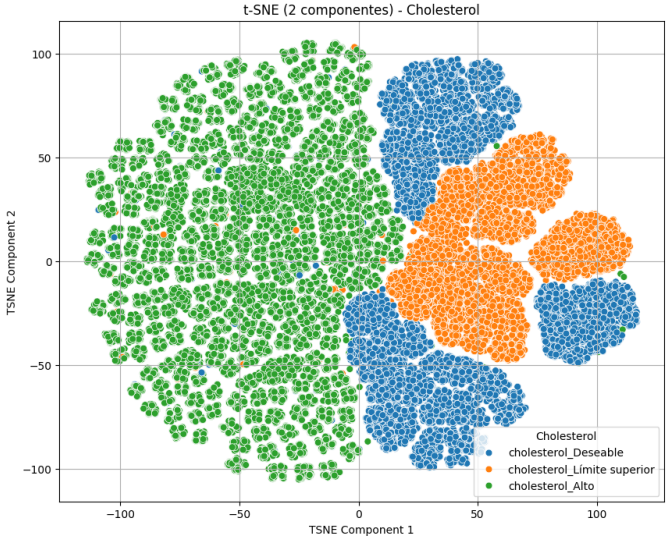

Fig. 6: Clusters de acuerdo al nivel de colesterol

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[6] U. Olsson, "Maximum likelihood estimation of the polychoric correlation coefficient," *Psychometrika*, vol. 44, no. 4, pp. 443–460, 1979.

[7] American Heart Association (AHA), "Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults," *American College of Cardiology*, 2018. [Online]. Available: https://www.heart.org/en/professional/education/guidelines-for-treatment-of-blood-cholesterol

[8] Centers for Disease Control and Prevention (CDC), "Defining Adult Overweight and Obesity," *CDC*, 2020. [Online]. Available: https://www.cdc.gov/obesity/adult/defining.html

[9] National Institutes of Health (NIH), "Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults," *NIH*, 1998. [Online]. Available: https://www.nhlbi.nih.gov/files/docs/guidelines/obesity.pdf

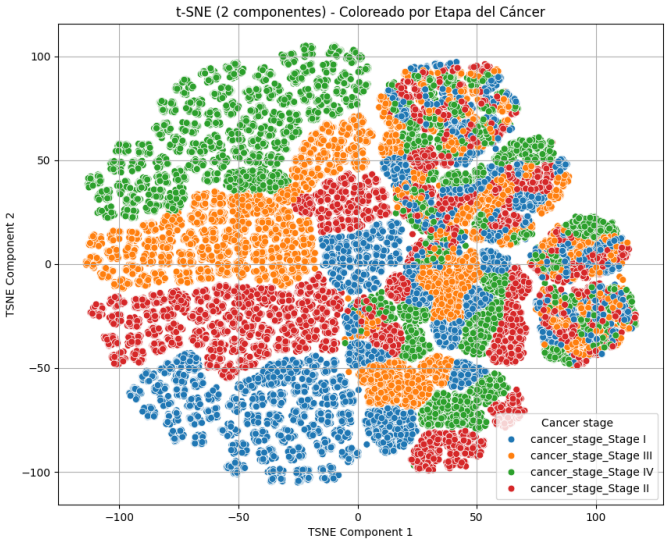

Fig. 7: Clusters de acuerdo a la etapa diagnosticada de cáncer

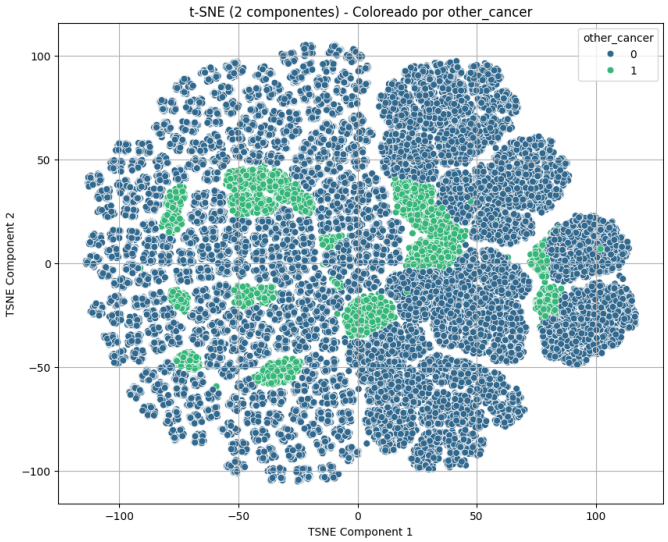

Fig. 8: Clusters de acuerdo a la presencia o no de otros tipos de cáncer