Predictor de cáncer de Pulmón mediante técnicas de aprendizaje supervisado y no supervisado

Cristobal M. Meza[1] and Adamaris L. de Dios R.[1]

[1] *Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias*

Resumen— El cáncer de pulmón es la principal causa de muerte relacionada con el cáncer en todo el mundo, la detección temprana y el diagnóstico del cáncer de pulmón pueden mejorar significativamente las posibilidades de supervivencia para los pacientes, esta situación subraya la necesidad de implementar medidas para asegurar la pronta identificación mediante la aplicación de técnicas de diagnóstico. El aprendizaje automático se ha utilizado cada vez más en el sector médico para la detección del cáncer de pulmón, en este atriculo se hace uso de algoritmos como random forest, ensemble algorithms, Decision Trees (DTs), Support vector machines (SVMs), Naive Bayesian (NB) classifiers, entre otros y debido a la alta cantidad de dimensiones de las bases de datos utilizadas para aplicaciones clínicas, se hace uso del Análisis de Componentes Principales (PCA), el cual reduce la dimensionalidad y mejora el rendimiento de algunos modelos de aprendizaje automático.

Palabras clave— Cancer de pulmon, machine learning, diagnóstico precoz, PCA.

I. Introducción

El cáncer de pulmón es una de las principales causas de mortalidad por cáncer tanto en mujeres como en hombres, en México se registran anualmente al rededor de 10 000 casos nuevos de cáncer pulmonar (CP), de los cuales solo el 5% son detectados en estadios tempranos, por lo que es necesario tomar acciones para asegurar la pronta identificación mediante la implementación de medidas de diagnóstico [1] [2].

Los pacientes identificados en una fase clínica temprana y elegibles para cirugía experimentan una tasa de supervivencia superior al 70% después de someterse a una intervención quirúrgica. Además, investigaciones del *The National Lung Screening Trial* han demostrado una disminución del 20% en la mortalidad mediante el empleo de tomografía de baja dosis.

El propósito de identificar el cáncer de pulmón en sus etapas iniciales es disminuir la mortalidad vinculada a esta enfermedad [3]. Hasta ahora, no se ha establecido ningún programa de detección temprana en México. Aunque el cáncer de pulmón puede manifestarse en edades tempranas, la población con mayor riesgo abarca desde los 50 hasta los 88 años. [4]. Se han desarrollado diversos algoritmos de aprendizaje automático (ML) para aplicaciones clínicas, entre los que se incluyen random forests (RFs), ensemble algorithms, Naive Bayesian (NB) classifiers, Support vector machines (SVMs), neural networks (NNs), Decision Trees (DTs) entre otros [5].

La aplicación de técnicas de aprendizaje automático para relacionar las características clínicas de pacientes con cáncer y sus tasas de supervivencia no solo proporciona insights valiosos, sino que también contribuye a aliviar la carga laboral de los profesionales médicos y mitigar el riesgo de errores humanos, destacando así su potencial impacto positivo en el campo de la atención médica.

II. Materiales y métodos

Esta sección explica nuestro enfoque con respecto a la forma y las técnicas utilizadas para la detección del cáncer de pulmón.

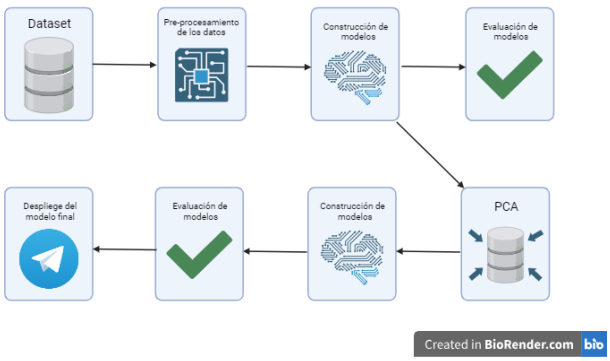

Fig. 1: Flujo de trabajo

Para la implementación de los algoritmos se utilizó Python, en conjunto con sus librerias: pandas, seaborn, scikitlearn [6], imblearn y traitlets. Se trabajó con 309 datos almacenados en un archivo csv de nombre "survey lung cancer" los cuales fueron obtenidos de Kaggle, 270 pacientes diagnosticados con CP y 39 pacientes de control saludables, en la base de datos se encontraron y removieron un total de 33 datos repetidos, los datos extraídos de los pacientes fueron los siguientes: Sexo (Biológico), edad, fumador, dedos amarillos, ansiedad, enfermedades crónicas, fatiga, alergias, jadeos, consumo de alcohol, tos, dificultad para respirar,

presión social, dificultad para tragar y dolor de pecho. Los datos fueron almacenados en un archivo csv con valores de 0 a 1 a excepción de la edad, siendo 0 falso y 1 verdadero. Nuestra muestra cuenta con una edad media de 62.7 años con una desviación estándar de 8 años y un 87% de ellos cuenta con CP.
Para el entrenamiento de los modelos se usaron 357 datos seleccionados aleatoriamente, mientras que para su evaluación se usaron 119. De acuerdo con el mapa de correlación de estos ninguna combinación tiene valores cercanos a 1, por lo que es poco probable que sean linealmente dependientes.

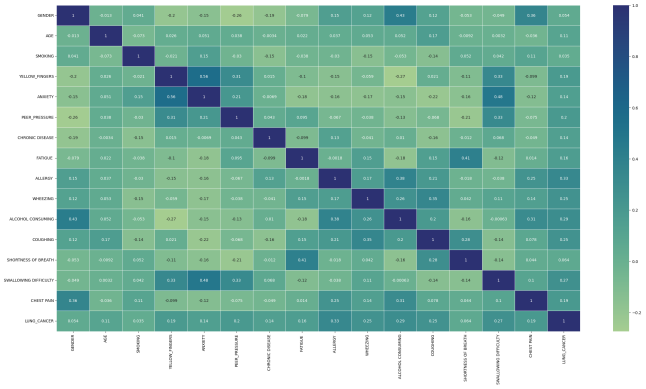

Fig. 2: Mapa de correlación

Para la predicción se usaron distintos algoritmos: KNN, SMV, Random forest, DecisionTreeClassifier, MLP Classifier Y PCA.

a. Evaluación de modelos

1. Primera iteración de los modelos

El código se encuentra en anexos.
KNN
Se utilizaron 20 ticks, y predijo correctamente un 95% de los datos con un recall de 94% y una media armónica de 94%.

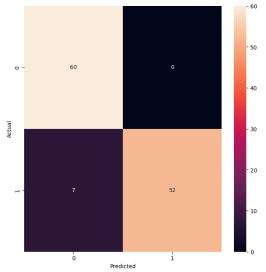

Fig. 3: KNN confusion matrix

SVM

Random forest

DecisionTreeClassifier
Se utilizaron los parámetros por defecto y obtuvo una precisión del 91.5%
MLP Classifier
Se utilizaron los siguientes parametros:

```
MLPClassifier(hidden_layer_sizes=(50, ),
max_iter=500, alpha=1e-4, solver='sgd',
verbose=10,random_state=54,
learning_rate='adaptive')
```

y se obtuvo una precisión del 86%

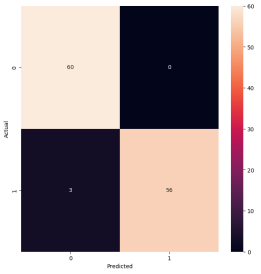

Fig. 4: SVM confusion matrix

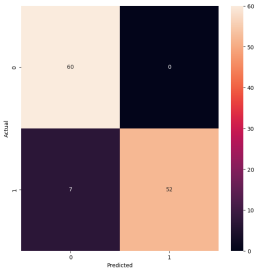

Fig. 5: Random forest confusion matrix

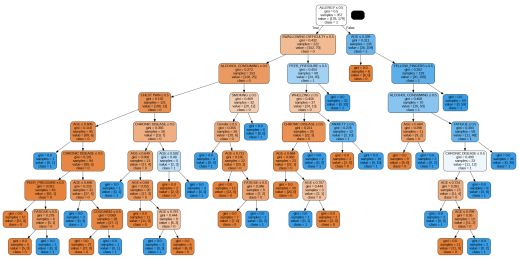

Fig. 6: DecisionTreeClassifier

2. PCA

Debido a la alta cantidad de dimensiones con las que cuentan nuestros datos, y, al propósito final del modelo, se hace uso del Análisis de Componentes Principales (PCA), el cual se define como una técnica de reducción de dimensionalidad que se utiliza con frecuencia para reducir la dimensionalidad de conjuntos de datos extensos [7]. Los componentes resultantes fueron: Sexo (Biológico), dedos amarillos, ansiedad, consumo de alcohol y presión social.

3. Segunda iteración de los modelos

Para la segunda iteración únicamente se hizo uso del modelo DecisionTreeClassifier, por su facilidad para implementarse en el día a día de la población, este obtuvo una precisión del 84% con un f1-score de 0.84.

4. Comparación de los modelos

Para todos los modelos se utilizó una muestra aleatoria de 357 datos de entrenamiento y 119 de prueba, solo

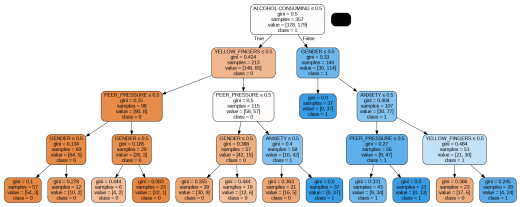

Fig. 7: DecisionTreeClassifier after PCA

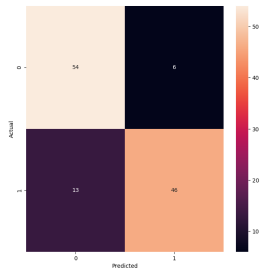

Fig. 8: Confusion matrix

se utilizaron los datos con dimensionalidad reducida en DecisionTreeClassifier-PCA.

$$Presicion = \frac{P}{P+FP} \times 100$$

$$Recall = \frac{P}{P+FN} \times 100$$

$$F1-Score = \frac{2 \times Presicion \times Recall}{Presicion + Recall}$$

Donde P = Positivos correctamente predichos FP = Falsos positivos y FN = Falsos Negativos

Model	precision	recall	f1-score
KNN	0.92	0.92	0.92
SVM	0.95	0.95	0.95
Random forest	0.93	0.92	0.92
DecisionTreeClassifier	0.93	0.92	0.92
DecisionTreeClassifier-PCA	0.85	0.84	0.84

b. Implementación

Para el uso práctico de estos algoritmos, se diseñó un chatbot dentro de la plataforma Telegram. Este chatbot emplea tanto la primera como la segunda iteración del DecisionTreeClassifier debido a su facilidad de implementación y su destacada precisión.

El chatbot fue desarrollado utilizando la plataforma "manychat", aprovechando preguntas de tipo verdadero y falso que son evaluadas en el modelo para generar respuestas. Además de su funcionalidad de respuesta inmediata, además el chatbot realiza el registro y almacenamiento de los datos proporcionados por el usuario.

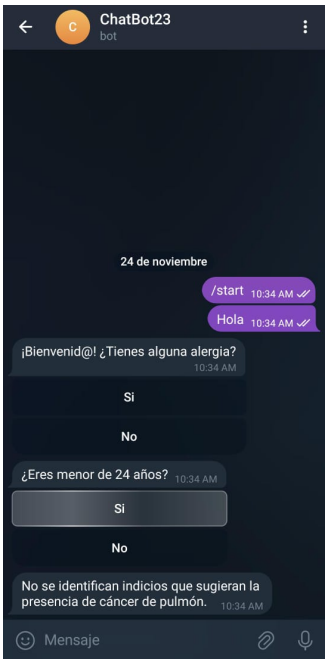

Fig. 9: Conversación con el chatbot

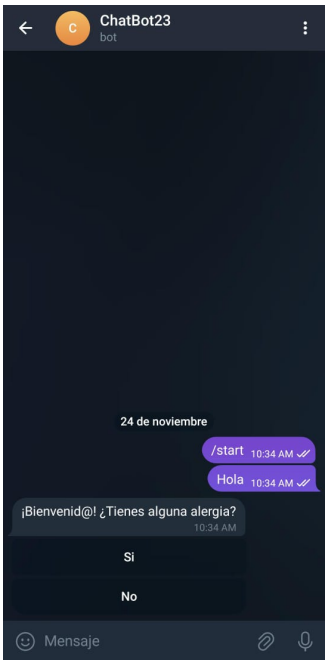

Fig. 10: Conversación con el chatbot

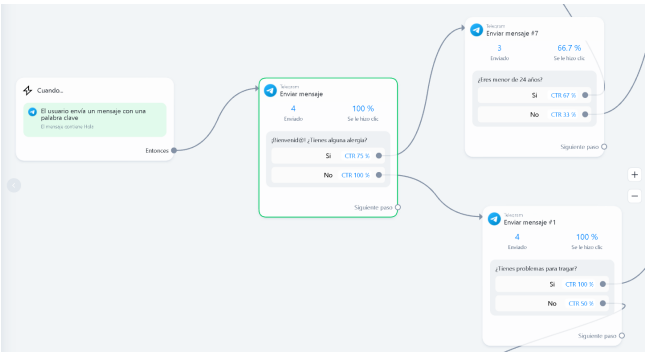

Fig. 11: Diagrama del chatbot

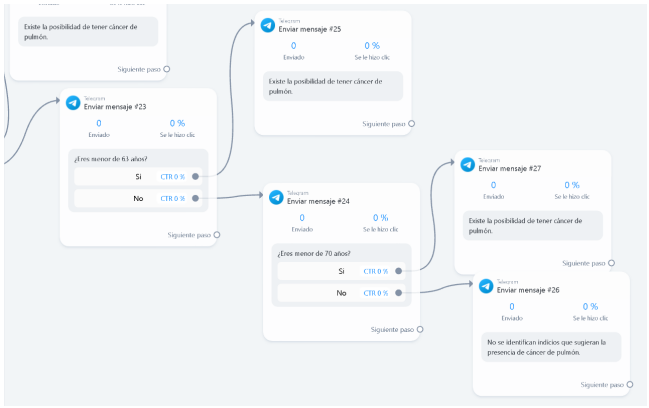

Fig. 12: Diagrama del chatbot

III. Conclusiones

La constatación de que el cáncer de pulmón es una de las formas más letales a nivel mundial, con su impacto significativo en la población mexicana, resalta la urgencia de medidas preventivas y de detección temprana. En este trabajo se destaca la aplicación de algoritmos de aprendizaje automático para el análisis de datos clínicos, que proporcionan una herramienta valiosa para comprender y predecir el cáncer de pulmón. La diversidad de algoritmos utilizados, junto con la incorporación de técnicas como PCA para reducir la dimensionalidad, resaltan no solo la eficacia de los métodos de detección y predicción del cáncer de pulmón, sino también la necesidad de integrar soluciones prácticas y accesibles en la atención médica diaria.

A. Anexos

Tanto los datos como el código utilizado en este trabajo, así como algunas figuras extra y la presentación, se encuentran en el repositorio: /lung-cancer-predictor de GitHub [8].

References

[1] ARROYO-HERNANDEZ Marisol; ZINSER-SIERRA JWyVG. 2019 Detección temprana de cáncer de pulmón en México. *Salud pública Méx [online]* **61**, 3. (doi:https://doi.org/10.21149/10326).

[2] Adams SJ, Mikhael P, Wohlwend J, Barzilay R, Sequist LV, Fintelmann FJ. 2023 Artificial intelligence and machine learning in lung cancer screening. *Thoracic Surgery Clinics* **33**, 4, 401–409. (doi: https://doi.org/10.1016/j.thorsurg.2023.03.001). Lung Screening: Updates and Access.

[3] Linh VTN, Kim H, Lee MY, Mun J, Kim Y, Jeong BH, Park SG, Kim DH, Rho J, Jung HS. 2024 3d plasmonic hexaplex paper sensor for label-free human saliva sensing and machine learning-assisted early-stage lung cancer screening. *Biosensors and Bioelectronics* **244**, 115779. (doi:https://doi.org/10.1016/j.bios.2023.115779).

[4] Tanoue LT Tanner NT GMSG. 2015 Lung cancer screening. *Am J Respir Crit Care Med.* pp. 19–33. (doi:https://doi.org/10.1164/rccm.201410-1777ci).

[5] Altuhaifa FA, Win KT, Su G. 2023 Predicting lung cancer survival based on clinical data using machine learning: A review. *Computers in Biology and Medicine* **165**, 107338. (doi:https://doi.org/10.1016/j.compbiomed.2023.107338).

[6] Liu L, et al. scikit-learn. https://scikit-learn.org/stable/index.html.

[7] A T. 2016 Principal component analysis - a tutorial. *Int J Appl Pattern Recognit* p. 197. (doi:https://doi.org/10.1504/ijapr.2016.079733.).

[8] Meza CM. lung-cancer-predictor. https://github.com/CristobalMe/lung-cancer-predictor/tree/main.