Clasificación de pacientes con síntomas de COVID-19 mediante árboles de decisión como una aplicación del aprendizaje automático

Trujillo González, Julio Enrique; Vejerano García, Carolina

Julio Enrique Trujillo González
julio.trujillo@up.ac.pa
Universidad de Panamá., Panamá

Carolina Vejerano García
carolina.yvg@gmail.com
Universidad de Panamá. , Panamá

Guacamaya
Universidad de Panamá, Panamá
ISSN-e: 2616-9711
Periodicidad: Semestral
vol. 8, núm. 1, 2023
solismu@yahoo.com

Recepción: 14 Junio 2023
Aprobación: 02 Julio 2023

URL: http://portal.amelica.org/ameli/journal/212/2124647006/

DOI: https://doi.org/10.48204/j.guacamaya.v8n1.a4319

Resumen: El objetivo de este trabajo fue clasificar los pacientes con síntomas de COVID-19 utilizando los árboles de decisión. Donde el paquete sklearn fue de gran utilidad para obtener el modelo, su representación gráfica y las variables con mayor importancia. El modelo nos permite clasificar el 74.78% de los pacientes con síntomas.

Palabras clave: Árboles de decisión, Medicina, COVID-19, Aprendizaje automático.

Abstract: The objective of this work was to classify patients with symptoms of COVID-19 using decision trees. Where the sklearn package was very useful to obtain the model, its graphic representation and the most important variables. The model allows us to classify 74.78% of patients with symptoms.

Keywords: Decision Trees, Medicine, COVID-19, Machine learning.

Introducción

Los árboles de decisión son un modelo de aprendizaje automático que representa un conjunto de reglas de decisión y las posibles consecuencias de esas decisiones en una estructura de árbol. Cada nodo interno del árbol representa una pregunta o prueba sobre una característica del dato de entrada, y las ramas que salen de cada nodo representan posibles respuestas o valores para esa característica. Los nodos hoja del árbol contienen las etiquetas de clasificación o los valores de predicción resultantes.

La construcción de un árbol de decisión implica dividir el conjunto de datos de entrenamiento en subconjuntos más pequeños de manera recursiva, basándose en las características que mejor discriminan las clases o predicen los valores objetivo. Esto se logra mediante la evaluación de métricas de impureza o ganancia de información, como el índice de Gini, la entropía o el error de clasificación. El objetivo es maximizar la pureza o la ganancia de información en cada división.

Una vez construido el árbol, se puede utilizar para realizar clasificación o predicción sobre nuevos datos de entrada. Esto se logra siguiendo el camino desde la raíz hasta un nodo hoja, siguiendo las pruebas en cada nodo interno de acuerdo con los valores de las características del dato de entrada. La etiqueta de clasificación o el valor de predicción asociado con el nodo hoja alcanzado se asigna al dato de entrada.

Los árboles de decisión son ampliamente utilizados en diversas aplicaciones debido a su capacidad para representar y comprender decisiones complejas de manera intuitiva. Además, son interpretables y permiten

identificar las características más relevantes en el proceso de toma de decisiones. Sin embargo, pueden ser sensibles a variaciones en los datos de entrenamiento y pueden sufrir de sobreajuste si no se controla su crecimiento o se aplican técnicas de poda adecuadas (Bishop, 2006., Breiman, 2017., Mitchell, 2007., Quinlan,1993). La construcción de árboles de decisión se ha utilizado en diversas aplicaciones, incluyendo en la toma de decisiones empresariales, el análisis de mercados financieros, la clasificación de imágenes y la medicina.

En medicina, los árboles de decisión se han utilizado para en la clasificación de enfermedades y en la toma de decisiones clínicas. Por ejemplo, Citodiagnóstico del cáncer de mama (Cruz-Ramírez *et al.,* 2007): en este estudio los autores utilizaron árboles de decisión para clasificar pacientes. El estudio encontró que los árboles de decisión eran altamente precisos en la clasificación de pacientes y que podrían ser útiles en la toma de decisiones clínicas.

Los árboles de decisión son particularmente útiles en medicina porque permiten integrar múltiples variables y síntomas para determinar la mejor opción de tratamiento para el paciente. Este método se ha extendido porque son muy fáciles de interpretar, ya que los resultados se presentan en un formato visual y es posible entender el proceso de toma de decisiones. Además, los árboles de decisión son muy eficientes para el procesamiento de grandes cantidades de dados y suelen proporcionar una buena precisión en la clasificación de pacientes.

Sin embargo, también es importante destacar que los árboles de decisión pueden tener ciertas limitaciones. Por ejemplo, si los datos de entrenamiento no son representativos o si existen variables importantes que no se han considerado, el árbol de decisión puede ser menos preciso. Además, es posible que este método sea vulnerable a sobreajuste, lo que significa que pueden generar modelos muy específicos para el conjunto de datos de entrenamiento y que no se generalizan bien para otros conjuntos de datos.

Materiales y métodos

La base de datos

La base de datos COVID-19 Symptoms Checker tiene como objetivo identificar si alguna persona tiene una enfermedad por coronavirus en función de algunos síntomas estándar registrados. Los síntomas se basan en las pautas de la Organización Mundial de la Salud, las cuales podemos mencionar: la fatiga, expectoración, nariz tapada, fiebre, tos seca, entre otras (Michelen *et al.,* 2020). Esta base de datos se puede encontrar en la página web Kaggle (https://www.kaggle.com/).

El conjunto de datos contiene siete variables principales que tendrán mayor impacto en si alguien tiene la enfermedad del coronavirus o no, las variables su definición es la siguiente:

TABLA 1
Variables de la base de datos COVID-19 Symptoms Checker

Variable Observada	Definición
Country	Lista de países que visitó la persona
Age	Clasificación del grupo de edad para cada persona, según el estándar de grupo de edad de OMS
Symptoms	Según la OMS, 5 son los síntomas principales de OCVID-19, fiebre, cansancio, dificultad para respirar, tos seca y dolor de garganta
Experience any other symptoms	Dolores, congestión nasal, secreción nasal, diarrea y otros
Severity	El nivel de gravedad, leve, moderado, severo
Contact	Ha tenido la persona contacto con otro paciente de COVID-19

Fuente. Propia

El Algoritmo de clasificación

Los árboles de clasificación son esencialmente una serie de preguntas diseñadas para asignar una clasificación. Una definición más detallada se puede encontrar en el libro "Data Mining: Practical Machine Learning Tools and Techniques" de Witten *et al.* (2016), donde se define el árbol de decisión como "una estructura de árbol en la que cada nodo interno representa una prueba en una variable de entrada, cada rama representa el resultado de la prueba y cada nodo hoja representa una clase".

El criterio de selección

En el aprendizaje automático, el criterio de selección es una medida utilizada para evaluar la calidad de una división en un árbol de decisión. dos de los criterios de selección más comúnmente utilizados son Gini y Entropía.

• El criterio Gini se utiliza para medir la impureza de una partición. En términos simples, mide qué tan homogéneo es el conjunto de datos dentro de una partición.

$$I_G = 1 - \sum_{i=1}^{c} p_i^2$$

##:proporción de los ejemplares que están a lo largo de clase # para un nodo en particular.

• Por otro lado, el criterio de selección de entropía mide la incertidumbre o desorden de una partición. La entropía se refiere a la cantidad de información necesaria para describir la distribución de probabilidades de las clases.

$$I_H = -\sum_{i}^{c} p_i \log_2 p_i$$

##:proporción de los ejemplares que están a lo largo de clase # para un nodo en particular.

La elección del criterio de selección depende de la tarea y los datos específicos. en general, el criterio de Gini funciona mejor en conjuntos de datos más pequeños y menos complejos, mientras que la Entropía puede ser más adecuada para conjuntos de datos más grandes y complejos.

Resultados y Discusión

La base de datos sobre los síntomas de COVID-19 (31680 filas y 12 columna) tomando en consideración a España, para comparar los resultados obtenido por Trujillo González, J. E., & Martínez Valderrama, I. V. (2022) que utilizaron una estructura de red neuronal artificial para clasificar los pacientes que presentan un cuadro grave de coronavirus.

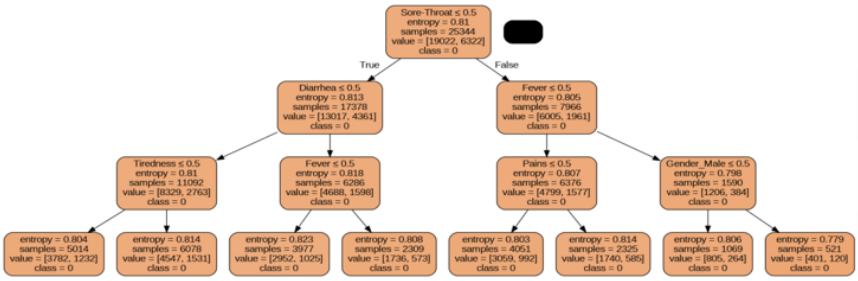

FIGURA 1

Árbol de decisión para clasificar pacientes COVID-19

Fuente. Propia

En la figura 1 hemos tomado en consideración la profundidad del árbol igual a 3, ya que un árbol con muchos nodos pierde la idea de tener una representación fácil de interpretar.

Tomaremos en consideración el accuracy, que es una medida comúnmente utilizada para evaluar la precisión de un modelo de clasificación (Müller, A. C., & Guido, S. 2016; Bishop, C. M., 2006). En nuestro modelo obtuvimos un accuracy 0.7477904040404041, es decir, que el modelo clasificó correctamente el 74.78% de las muestras en el conjunto de datos de prueba.

Lo interesante de aplicar este tipo de análisis es que hemos obtenido la contribución de cada variable a reducir la impureza en el conjunto de entrenamiento.

TABLA 2

Importancia de los síntomas

Síntomas	Importancia
Fever	0.260
Sore thoat	0.184
Tiredness	0.155
Diarrhea	0.154
Gender	0.147
Pains	0.099

Fuente. Propia

Conclusión

El propósito de este artículo es igual a presentado por Trujillo González, J. E., & Martínez Valderrama, I. V. (2022), era identificar de forma temprana las personas que pueden tener un cuadro grave de COVID-19 y teniendo en cuenta esta información elaborar estrategias y solicitar los recursos para la atención oportuna de los pacientes.

El modelo de árbol de decisión, se utilizó el criterio de selección de entropy por la complejidad y una profundad máxima de 3, obteniendo una clasificación de 74.78% de los pacientes del conjunto de prueba.

Como algoritmo de aprendizaje automático no existe una regla para seleccionar el número de la profundidad máxima, de forma empírica se podría decir que es cuando no una mejor significativa del accuracy.

El modelo presentado y el anterior basado en redes neuronales son complementarios para otros modelos basados en ecuaciones diferenciales o análisis estadísticos.

Referencias

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Breiman, L. (2017). Classification and regression trees. Routledge.

Cruz-Ramírez, N., Acosta-Mesa, H. G., Carrillo-Calvet, H., & Barrientos-Martínez, R. E. (2007). Comparison of the Performance of Seven Classifiers as Effective Decision Support Tools for the Cytodiagnosis of Breast Cancer: A Case Study. Analysis and Design of Intelligent Systems using Soft Computing Techniques, 79-87.

Michelen, M., Jones, N. & Stavropoulou, C. (2020). In patients of COVID-19, what are the symptoms and clinical features of mild and moderate cases? The Centre for Evidence- Based Medicine develops, promotes and disseminates better evidence for healthcare. https://www.cebm.net/covid-19/in-patients-ofcovid-19-what-are-the-symptoms-and-clinical-features-of-mild-andmoderate-case/

Mitchell, T. M. (2007). Machine learning (Vol. 1). New York: McGraw-hill.

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

Quinlan, J. R. (1993). Program for machine learning. C4. 5.

Trujillo González, J. E., & Martínez Valderrama, I. V. (2022). UN MODELO MATEMÁTICO DE CLASIFICACIÓN DE PACIENTES CON SINTOMAS COVID-19. Tecnociencia, 24(2), 66–77. Recuperado a partir de https://revistas.up.ac.pa/index.php/tecnociencia/article/view/3071

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.