Aplicación de Arboles de Decisión para el Diagnostico de Tumores Malignos y Benignos de Cáncer de Mama

German López Rodrigo

I. Resumen

El presente artículo pretende mostrar al lector de cualquier área, las bondades de utilizar el algoritmo de árboles de decisión para la predicción de datos. Para ello se recurre al desarrollo de un caso aplicado a la asistencia en el diagnóstico de Tumores Malignos y Benignos de Cáncer de Mama, el objetivo es que mediante los datos médicos recopilados de una persona podamos diagnosticar si el tumor es Maligno o Benigno.

II. Palabras clave

Árboles de Decisión, Aprendizaje Automático, Cáncer de mama, Algoritmo predictivo.

III. Descripción del problema

Actualmente, el cáncer de mama es considerada cómo la causa más frecuente de muerte por cáncer a nivel mundial. Se estima que una de cada ocho mujeres tiene o va a desarrollar el cáncer de mama en el lapso de su vida, es decir, que el 12 por ciento de la población femenina actual en el mundo va a presentar esta enfermedad [1]. La mayoría de los casos son diagnosticados a partir de un hallazgo anormal en un estudio de control (ecografía mamaria, mamografía y/o resonancia nuclear magnética); sin embargo, algunos casos son detectados por la presencia de determinados hallazgos clínicos. Por este motivo, es importante implementar nuevos algoritmos y técnicas que sean capaces de proporcionar un diagnóstico temprano sobre las pacientes con el fin de que sean tratados de forma eficiente y rápida.

- **Pregunta de investigación**

  ¿Qué factores pueden determinar la detección temprana de Tumores Malignos y Benignos de Cáncer de Mama?

- **Objetivo**

  Determinar un árbol de decisión que ayude en la detección de Tumores Malignos y Benignos de Cáncer de Mama, mediante la implementación del algoritmo C45.

- **Fuente de Datos**

  Cancer [2] es una fuente de datos que consta de 570 registros y 30 características, fue construida a partir de los datos de seguimiento de los pacientes atendidos por el Dr. Wolberg desde 1984, e incluye solo aquellos casos que exhiben invasivo cáncer de mama y sin evidencia de metástasis a distancia en el momento del diagnóstico. Esta fuente de datos es la que será utilizada para cumplir con el objetivo principal del presente artículo de investigación. Cabe aclarar que la fuente de datos fue dividida en dos conjuntos: un conjunto de 400 observaciones para construir el árbol de decisión y otro conjuntó de 170 observaciones para realizar las pruebas del árbol de decisión.

IV. Introducción

El cáncer de mama es un proceso oncológico en el que células sanas de la glándula mamaria degeneran y se transforman en tumorales, proliferando y multiplicándose posteriormente hasta constituir el tumor, está enfermedad se encuentra asociada al envejecimiento y a estilos de vida poco saludables así como a los cambios en los patrones reproductivos y estilos de vida. Actualmente en México el cáncer mamario ocupa el primer lugar como causa de muerte por neoplasia maligna en las mujeres mayores de 25 años y es un grave problema de salud pública en nuestro país [4]. El diagnóstico temprano se realiza a toda mujer con sospecha de patología mamaria mediante una historia clínica completa enfocada a la búsqueda de factores de riesgo; examen clínico completo con énfasis en las glándulas mamarias y zonas linfoportadoras; y estudios de mastografía y/o ultrasonido. De esta forma, con el fin de poder crear un método distinto y más eficaz comparado con los estudios de laboratorio, el presente trabajo propone utilizar un árbol de decisión para la detección temprana de cáncer de mama. Un árbol de decisión es un método de aprendizaje que utiliza un modelo muy similar a los sistemas de predicción basados en reglas, consiste en un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Una de las grandes ventajas de los árboles de decisión es que, el conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol está representado por un conjunto de nodos, hojas y ramas. Los nodos corresponden a los resultados de cada una de las preguntas acerca de los atributos del caso de estudio, las ramas representan los posibles valores del atributo y los nodos hoja corresponden a una decisión, la cual coincide con una de las variables del caso de estudio a resolver.

V. DESARROLLO DE CONTENIDO

Para poder diagnosticar si el tumor de cáncer de mama de una persona se trata de un tumor maligno o benigno, se utilizó un árbol de decisión de 30 variables independientes y una variable dependiente. El diagnostico se estableció como la variable dependiente la cual estaría en función de las características del tumor: radio, textura, perímetro, área, suavidad, compacidad, concavidad, puntos cóncavos, simetría y dimensión fractal.

El árbol de decisión es construido a partir de un algoritmo programado en lenguaje Python, a continuación se explica el proceso que sigue el algoritmo para la construcción de dicho árbol. El algoritmo consta de tres etapas principales:

- **Etapa de Preprocesamiento de Datos**

  En esta etapa el algoritmo se encarga de obtener que variables independientes son categóricas y cuales son numéricas con el fin de poder darles el tratamiento adecuado.

- **Etapa de Procesamiento de Datos**

  En esta etapa el algoritmo se encarga de realizar la construcción del árbol de decisión para esto el primer paso que realiza el algoritmo es calcular la ganancia de información para cada atributo de la fuente de datos con el fin de identificar el atributo más significativo. La ecuación número 1 representa la formula para calcular la ganancia de información.

$$Gan.Inf(S,A) = Entropia(S) - \sum_{v \varepsilon V(A)} \frac{|Sv|}{|S|} Entropia(Sv) \quad (1)$$

  Donde:
  S : es una colección de objetos.
  A : son los atributos de los objetos.
  V(A) : Conjunto de valores que A puede tomar.

  Cabe aclarar que para obtener la ganancia de información, primero se tuvo que calcular la entropía de un sub-conjunto de datos el cual depende del atributo al que se quiere obtener la ganancia de información. La ecuación número 2 representa la formula para calcular la entropía.

$$Entropia = \sum_{i=1}^{n} -p_i log_2 p_i \quad (2)$$

  Donde:
  S: es una colección de objetos.
  Pi : es la probabilidad de los posibles valores.
  i: las posibles respuestas de los objetos.

  Una vez obtenido el atributo con mayor ganancia de información el algoritmo procede por ir construyendo el árbol, para esto al conocer el mejor atributo se asigna como nodo raíz del árbol, sus ramas corresponden a los valores que puede llegar a tener. Si el atributo seleccionado es categórico los valores corresponden a las diferentes clases del atributo, en caso de que sea numérico se realiza un calculo numérico para identificar los limites que nos proporcionen la mayor ganancia de información. La ecuación número 3 representa la formula para calcular los limites.

$$\frac{1}{2}(v_i + v_{i+1}) \leq x \leq \frac{1}{2}(v_i + v_{i+1}) \quad (3)$$

  Los nodos corresponden a los demás atributos de la fuente de datos, entonces el algoritmo para elegir que atributo sigue en la estructura del árbol, realiza de forma recursiva los anteriores pasos hasta llegar alguno de los siguientes casos base:

  – Si todas (o casi todas) las instancias del subconjunto son de la misma clase, el nodo pasa a ser una hoja con el nombre de la clase.

  – Si se han agotado todos los atributos pero aún quedan instancias del subconjunto que no pertenecen a la misma clase, el nodo pasa a ser una hoja con el nombre de la clase más común.

  – No quedan instancias en el subconjunto entonces el nodo pasa a ser una hoja con el nombre de la clase más común.

  Una vez que el algoritmo cae en alguno de esos casos base, el árbol de decisión se encuentra construido totalmente. Entonces el algoritmo procede por almacenar el árbol en un archivo JSON con el objetivo que se pueda visualizar de una forma simple.

- **Etapa de Resultados**

  En esta última etapa el algoritmo se encarga de evaluar el árbol de decisión recién construido, para esto requiere de otra fuente de datos que solo contenga las variables independientes. Lo que realiza el algoritmo es pasar la nueva fuente de datos por la etapa de preprocesamiento, una vez finalizada la etapa de preprocesamiento se procede por ir evaluando de forma recursiva las variables independientes en el árbol de decisión con la finalidad de poder predecir el valor de la variable dependiente según las variables independientes de entrada.

VI. Resultados

A continuación se presentan los resultados obtenidos al ejecutar el algoritmo anteriormente descrito sobre la fuente de datos Cancer [2]. Debido a que en este tipo de algoritmo el único resultado que arrojan es el arbol de decision construido, a continuacion se muestra el arbol de decision conrtruido para la fuente de datos Cancer [2].

```
{
"Root": "perimeter_worst",
"Decision":{
  "<= 101.65":{
    "Root": "concave points_worst",
    "Decision":{
      "<= 0.18075":{
        "Root": "area_se",
        "Decision":{
          "<= 48.975":{
            "Root": "texture_worst",
            "Decision":{
              "<= 33.34999": "B",
              "> 33.34999":{
                "Root": "
                    smoothness_worst",
                "Decision":{
                  "<= 0.14125": "B",
                  "> 0.14125":{
                    "Root": "
                        fracDimensioWorst
                        ",
                    "Decision":{
                      "<= 0.07906": "B",
                      "> 0.07906": "M"
                    }
                  }
                }
              }
            }
          },
          "> 48.975":{
            "Root": "symmetry_worst",
            "Decision":{
              "<= 0.20785": "M",
              "> 0.20785": "B"
            }
          }
        }
      },
      "> 0.18075": "M"
        }
  },
  "> 101.65":{
    "Root": "concave points_worst",
    "Decision":{
      "<= 0.15075":{
        "Root": "area_worst",
        "Decision":{
          "<= 957.45":{
            "Root": "texture_mean",
            "Decision":{
              "<= 20.245":{
                "Root": "
                    smoothness_worst",
                "Decision":{
                  "<= 0.13985": "B",
                  "> 0.13985":{
                    "Root": "
                        compactness_worst
                        ",
                    "Decision":{
                      "<= 0.31930": "B",
                      "> 0.31930": "M"
                    }
                  }
                }
              },
              "> 20.245":{
                "Root": "
                    smoothness_worst",
                "Decision":{
                  "<= 0.11230":{
                    "Root": "
                        fractalDimensionS
                        ",
                    "Decision":{
                      "<= 0.001956": "M"
                          ,
                      "> 0.001956": "B"
                    }
                  },
                  "> 0.11230": "M"
                }
              }
            }
          },
          "> 957.45":{
            "Root": "concavitWorst",
            "Decision":{
              "<= 0.1907":{
                "Root": "
                    fracDimensioWorst",
                "Decision":{
                  "<= 0.074605": "B",
                  "> 0.074605": "M"
                }
              },
              "> 0.1907": "M"
            }
          }
        }
      },
      "> 0.15075":{
        "Root": "concavity_mean",
```

```
        "Decision":{
          "<= 0.085885":{
            "Root": "
                fractal_dimension_worst",
            "Decision":{
              "<= 0.079495": "B",
              "> 0.079495": "M"
            }
          },
          "> 0.085885": "M"
        }
      }
    }
  }
}
```

VII. CONCLUSIÓN

De acuerdo a los resultados obtenidos se puede concluir que es posible construir con precisión árboles de decisión a partir de datos médicos, ya que los porcentajes de clasificación, es decir el número de casos que clasificó correctamente, tienen un margen de error mínimo y es posible que pueda mejorar su eficiencia con la ayuda del experto, ajustando los datos mismos, esto es, agregando variables o cambiando sus parámetros.

REFERENCES

[1] "Cáncer, el asesino silencioso de la mujer mexicana". Conversus. Instituto Politécnico Nacional. México. 2006. p. 47

[2] Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu, Version 1. Retrieved 19 Marzo, 2020. From https://www.kaggle.com/sarahvch/breast-cancer-wisconsin-prognostic-data-set.

[3] Córdoba Villalobos, José Ángel. "Introducción". Programa de acción: cáncer de mama. 20072012. Secretaría de Salud. México. 2007.

[4] Compendio de patología mamaria. Secretaría de Salud. México..2002. p. 1314.

[5] Córdoba Villalobos, José Ángel. "Introducción". Programa de acción: cáncer de mama. 20072012. Secretaría de Salud. México. 2007.

[6] NOM041SSA22002 para la prevención, diagnóstico, tratamiento, control y vigilancia epidemiológica del cáncer de mama. Secretaría de Salud. México. Septiembre 2003

[7] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. Classification and Regression Trees. Chapman and Hall, 1984.

[8] Quinlan, J. R. C4.5 Programs for Machine Learning. Morgan Kaufmann, 1993.