Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos

Comparative analysis of algorithms of decision trees in the processing of biological data

Luis Charris, Cesar Henriquez, Stiven Hernandez, Luis Jimeno, Oscar Guillen, Silvia Moreno.
Universidad Simón Bolívar, Barranquilla-Colombia

Resumen En este trabajo se evalúa el desempeño de varios algoritmos de árboles de decisión, para así encontrar por medio de comparaciones, cuales son más efectivos en el análisis de datos biológicos. Los árboles de decisión son un modelo de clasificación utilizado en la inteligencia artificial, cuya principal característica es su aporte visual a la toma de decisiones. Para poner a prueba el rendimiento en el proceso de clasificación de los árboles de decisión, se utilizarán datos biológicos de pacientes reales, estos datos serán analizados en el software WEKA. Con esta comparación lo que se busca también es determinar la pertinencia de los árboles de decisión, es decir si estos pueden ser una buena herramienta para diagnósticos médicos. Estas comparaciones nos llevaran a aclarar que algoritmos son los más eficaces y apropiados para el análisis de dichos datos, y así llegar a una buena conclusión.

Palabras clave: Weka, árboles de decisión, inteligencia artificial.

Abstract in this work the performance of several algorithms of decision trees is evaluated, to find through comparisons, which are more effective in the analysis of biological data. Decision trees are a classification model used in artificial intelligence, whose main feature is its visual contribution to decision making. To test the performance in the classification process of the decision trees, biological data of real patients will be used, this data will be analyzed in the software WEKA. With this comparison, what is also sought is to determine the relevance of decision trees, is whether they can be a good tool for medical diagnostics. These comparisons will lead us to clarify which algorithm is the most efficient and appropriate for the analysis of said data, and thus arrive at a good conclusion.

Key words: Weka, decision trees, artificial intelligence.

Introducción

A través del tiempo se han desarrollado una gran cantidad de métodos para el análisis de datos, los cuales principalmente están basados en técnicas estadísticas. Sin embargo, a medida que la información almacenada crece considerablemente, los métodos estadísticos tradicionales han empezado a enfrentar problemas de eficiencia y escalabilidad. Debido a que la mayor parte de esta información es histórica y procede de fuentes diversas, parece clara la inminente necesidad de buscar métodos alternativos para el análisis de este tipo de datos y a partir de ellos, poder obtener información relevante y no explícita.

El análisis e interpretación de los datos en la mayoría de los casos se hacen de forma manual, es decir los especialistas analizan y elaboran un informe o una hipótesis acerca de dichos datos, para luego llegar a una conclusión y a partir de esta tomar decisiones importantes y significativas. Estos procesos a menudo son muy lentos y caros, además cuando el volumen de datos es exageradamente grande sobrepasa la capacidad humana, entonces se hace muy difícil su análisis sin ayuda de las herramientas adecuadas. Así también con ayudas de estas herramientas podemos llegar a un diagnóstico preciso[1].

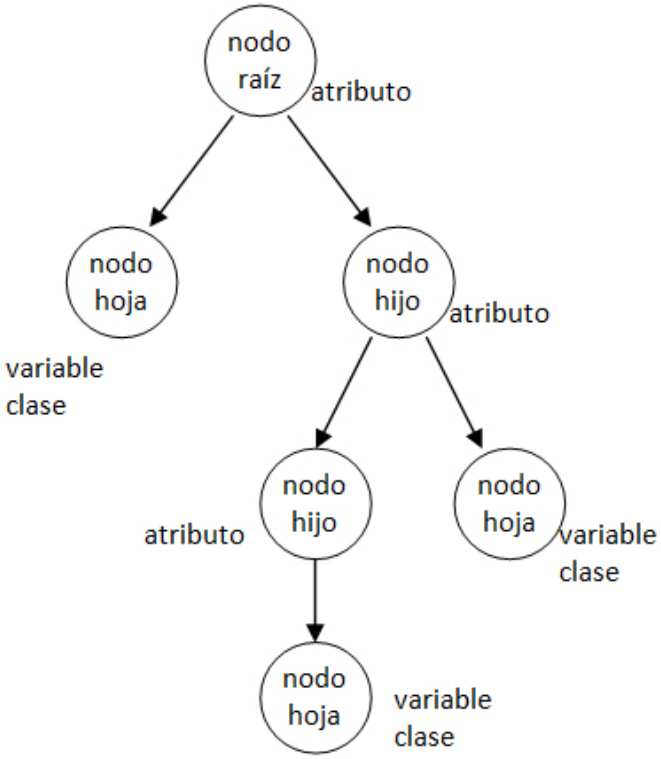
Para el caso de la medicina, es posible aplicar métodos alternativos, debido a la gran cantidad de padecimientos involucrados, las sintomatologías y los pacientes. Lo ideal sería que los médicos pudieran contar con el apoyo de una herramienta que les permita analizar los datos sintomatológicos de cada uno de sus pacientes para poder determinar con base en casos anteriores, el diagnóstico más acertado, así como el tratamiento óptimo a seguir, lo cual representaría un soporte y ayuda para el médico. Una herramienta alternativa para la predicción y clasificación de grandes cantidades de datos que es utilizada ampliamente en el área de la inteligencia artificial son los árboles de decisión.

Marco teórico

Árboles de Decisiones:

Un árbol de decisión es un modelo de predicción utilizado en diferentes ámbitos. Su objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema[2].

El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El árbol se representa por nodos, donde el nodo principal raíz es el atributo a partir del cual se inicia el proceso de clasificación. Los nodos internos o nodos hijos son preguntas acerca del atributo o problema. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver.



Un algoritmo de generación de árboles de decisión consta de 2 etapas: la primera corresponde a la inducción del árbol y la segunda a la clasificación. En la primera etapa se construye el árbol de decisión a partir del conjunto de entrenamiento; comúnmente cada nodo interno del árbol se compone de un atributo de prueba y la porción del conjunto de entrenamiento presente en el nodo es dividida de acuerdo con los valores que pueda tomar ese atributo.

La construcción del árbol inicia generando su nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de entrenamiento en dos o más subconjuntos; para cada partición se genera un nuevo nodo y así sucesivamente. Cuando en un nodo se tienen objetos de más de una clase se genera un nodo interno; cuando contiene objetos de una

clase solamente, se forma una hoja a la que se le asigna la etiqueta de la clase.

En la segunda etapa del algoritmo cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja, a partir de la que se determina la membresía del objeto a alguna clase. El camino a seguir en el árbol lo determinan las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente en él.

Existen muchos algoritmos para generar árboles de decisión, algunos que se pueden encontrar en el software WEKA son los siguientes:

El árbol CART es un método de regresión usado para predecir valores de variables continuas, pero cuando los supuestos para aplicar este modelo no se cumplen sus conclusiones pueden ser erróneas. Los árboles de regresión CART es un método muy fácil de interpretación de resultados. Los CART utilizan datos históricos, los cuales se usan para construir árboles de regresión que permiten la clasificación y la predicción de nuevos datos, estos tienen como ventaja que pueden manipular con facilidad variables numéricas, sus principales características es su robustez a outliers o valores atípicos, la invariancia en la estructura de sus árboles de clasificación a transformación monótonas de las variables independientes, y la interoperabilidad [3].

El árbol de decisión REETTree este método de aprendizaje por medio de árbol de decisiones es muy fácil y muy rápido de utilizar. Este árbol se construye mediante la información de varianzas, y se poda utilizando los criterios de reducción de errores. Este árbol de decisión clasifica solo atributos numéricos una única vez, los valores restantes son obtenidos de futuras instancias, dividiendo dichas instancias en segmentos de información[4].

El RandomTree es un árbol dibujado al azar de una serie de árboles posibles. En este contexto y en otras fuentes de información tomaremos el "al azar" como que cada árbol de estudio de árboles tiene una posibilidad igual de ser probado. Otro modo de decir esto sería que la distribución de árboles es uniforme. El proceso del RandomTree es un proceso que produce random tree de permutaciones arbitrarias[5].

C4.5 es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan [2]. C4.5 es una extensión del algoritmo ID3 desarrollado anteriormente por Quinlan. Los árboles de decisión generador por C4.5 pueden ser usados para clasificación, y por esta razón, C4.5 está casi siempre referido como un clasificador estadístico, loara nuestro proyecto usaremos una implementación open source en lenguaje de programación Java del algoritmo C4.5 que es el J48 en la herramienta weka.

Otro tipo de árbol que se desprende de este mismo es el J48 Consolidated que se utiliza para construir un árbol consolidado del C4.5 [5]. En este se construye un solo árbol basado en un conjunto de sub muestras, se añaden nuevas opciones a la clase J48 para establecer el método RE-sampling: que en un ámbito estadístico posee una serie de métodos que permiten en pocas palabras estimar muestras estadísticas, realizar test de significancia y la validación de modelos, todo esto para la generación de muestras que se utilizarán en el proceso de consolidación.

El LMT (Logistic Model Tree) proporciona una descripción muy buena de los datos. Consiste en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. Como en los árboles de decisión ordinarios, una prueba sobre uno de los atributos es asociado con cada nodo interno [6].

Siguiendo con los arboles tenemos el M5P (Árbol de regresión) en este árbol de decisiones se utiliza un criterio estándar llamado M5 que es un árbol basado en árbol de decisión numérico tipo model tree. Se caracteriza por; Construir árboles mediante un algoritmo inductivo de árbol de decisión, toma de decisiones de enrutado en nodos los cuales son tomados a partir de los valores de los atributos y cada hoja tiene asociada una clase que permite calcular el valor estimado de la instancia mediante una regresión lineal [7].

Estado del arte

Para afianzar la formulación de nuestro tema investigativo, estableceremos una serie de comparaciones con otros trabajos paralelos al de nosotros para aumentar nuestra compresión hacia nuestro proyecto, por ejemplo, se revisó la investigación de Rocío Erandi Barrientos Martínez [8] que trabajó en la temática de “Árboles de decisión como herramienta en el diagnóstico médico” donde evalúo el desempeño de tres de los algoritmos más representativos para la construcción de árboles de decisión. Donde argumenta ante mano que los árboles de decisión son un modelo de clasificación utilizado en la inteligencia artificial, cuya principal característica es su aporte visual a la toma de decisiones. Para poner a prueba el rendimiento en el proceso de clasificación de los árboles de decisión, se utilizaron dos bases de datos que contenían datos médicos de pacientes reales. Estos datos

corresponden a la sintomatología que un médico especialista considera para el diagnóstico de cáncer de seno. Una de las bases de datos contenía 692 casos recopilados de las observaciones de un solo médico y la otra, contenía 322 casos recopilados de la observación de 19 especialistas. En suma, se busca determinar la pertinencia de los árboles de decisión, es decir, si pueden ser una herramienta de apoyo para el diagnóstico médico.

Otro artículo que leímos fue el de ”Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares” por Guillermo Roberto Solarte Martínez [9]. En este artículo se presenta una descripción de los árboles de decisión y del algoritmo ID3 (Inducción Decision tree) para determinar si se debe o no aplicar fármacos a paciente con enfermedades cardiovasculares. En esta investigación se demuestra empíricamente que es posible diagnosticar la necesidad de administrar fármacos en pacientes con síntomas de enfermedad cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias, mediante la utilización de árboles de decisión con el algoritmo ID3 (Induction Decision tree) implementado en el lenguaje Java.

Otro artículo muy similar al anterior pero usando otro tipo de árbol de decisiones es el de “Técnicas de ml en medicina cardiovascular” por Alexis Kevin de la Hoz Manotas [10] en el cual menciona que la ciencia médica maneja grandes cantidades de información. Técnicas avanzadas de Machine Learning (ml) como árboles de decisión, máquinas de soporte vectorial y regresión logística pueden ser utilizadas para descubrir patrones ocultos en los datos. Modelos desarrollados a partir de estas técnicas serán de gran utilidad para la ciencia médica, permitiendo tomar decisiones eficaces. Este artículo permite observar los resultados obtenidos en relación con la capacidad de precisión de las técnicas de Machine Learning, luego de someterlas a pruebas mediante un conjunto de datos relacionados con enfermedades cardiovasculares suministrados por el repositorio uci. Después de validar las técnicas mencionadas con el repositorio uci, se obtiene como resultado que la regresión logística ofrece los mayores niveles de precisión. Cabe resaltar que las técnicas de máquinas de soporte vectorial (svm) y árboles de decisión (ad) ofrecen resultados aceptables; sin embargo, no se encuentran al nivel de los resultados obtenidos por la regresión logística.

Entre otras investigaciones tenemos una realizada por Frank Dávila Hernández [11], en su investigación llamada “Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas” que consiste en disminuir el error médico y mejorar los procesos de salud, lo que es prioridad de todo el personal sanitario. En este contexto surgen los "Sistemas Clínicos de Soporte para la toma de Decisiones" (CDSS), los cuales son un componente fundamental en la informatización de la capa clínica. Con la evolución de las tecnologías gran cantidad de datos han podido ser estudiados y clasificados a partir de la minería de datos. Una de las principales ventajas de la utilización de esta, en los CDSS, ha sido su capacidad de generar nuevos conocimientos. Con este fin se propone, mediante la combinación de dos modelos matemáticos, cómo se puede contribuir al diagnóstico de enfermedades usando técnicas de minería de datos. Para mostrar los modelos utilizados se tomó como caso de estudio la hipertensión arterial. El desarrollo de la investigación se rige por la metodología más utilizada actualmente en los procesos de Descubrimiento de Conocimiento en Bases de Datos: CRISP-DM 1.0, y se apoya en la herramienta de libre distribución WEKA 3.6.2, de gran prestigio entre las utilizadas para el modelado de minería de datos. Como resultados se obtuvieron diversos patrones de comportamiento con relación a los factores de riesgo a sufrir hipertensión mediante técnicas de minería de datos.

Citando otro trabajo tenemos el de Grettel Pérez Díaz [12] que se dedicó a la temática de “Sistema inteligente para pronóstico de supervivencia de paciente con trasplante renal” que se basó en obtener un sistema basado en el conocimiento híbrido para la predicción del tiempo de supervivencia del injerto renal de pacientes del Hospital Universitario “Dr. Arnaldo Milián Castro”. Este se desarrolla a partir de la edición de una base de casos obtenida como resultado de la ingeniería de conocimiento, utilizando WEKA se determinan los métodos de aprendizaje que mejores resultados generan en el pronóstico del rasgo objetivo que es continuo y representa tiempo de supervivencia del injerto.

Otra persona que se basó en los árboles de decisiones para sus investigaciones fue Ana Aguilera y Alberto Subero [13] en su proyecto llamado “Modelos de clasificación en marcha patológica usando árboles de regresión logística” tomando como objeto de estudio la Hemiplejía Espástica (HE) que es un tipo de parálisis cerebral, en donde los miembros superior e inferior del mismo lado están comprometidos. El Dr. Gage ha sugerido que esta patología pudiera clasificarse en al menos cuatro grupos considerando el patrón cinemático en 3 planos (sagital, transversal y coronal). Tradicionalmente, esta clasificación de pacientes patoló- gicos es realizada por

médicos especialistas basándose en el estudio físico del paciente y en el análisis clínico de la marcha, esto último, mediante exámenes complementarios reportados sobre registros cinéticos, cinemáticas y electromiografía. La aplicación de técnicas automáticas de clasificación mediante métodos computarizados constituye un soporte en esta tarea, no como reemplazo al especialista sino como una herramienta de apoyo al diagnóstico.

Leímos también sobre la investigación de Javier Trujillanoa [3], que consiste en "Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio, en resumen, se basó en Realizar una aproximación a la metodología de árboles de decisión tipo CART (Classification and Regression Trees) desarrollando un modelo para calcular la probabilidad de muerte hospitalaria en infarto agudo de miocardio (IAM). Método: Se utiliza el conjunto mínimo básico de datos al alta hospitalaria (CMBD) de Andalucía, Cataluña, Madrid y País Vasco de los años 2001 y 2002, que incluye los casos con IAM como diagnóstico principal.

Otro proyecto que tuvimos en cuenta fue uno llamado "Determinación de la eficacia de la braquit de salida. Mediante este entrenamiento el modelo ajusta los pesos de las neuronas ocultas para optimizar la salida. La ventaja de minería frente a los nomogramas es que posee terapia en tratamiento de cáncer basada en minería de datos" [14] dicho proyecto consistía en usar la minería de datos en vez de nomogramas ya que existe gran variedad de algoritmos, que tienen la capacidad de aprender de la experiencia. Están formados por nodos de ingreso, nodos ocultos y nodos a capacidad de resolver relaciones no lineales complejas entre las variables, sin necesidad de hacer ninguna suposición previa respecto a dichas relaciones.

Este siguiente proyecto fue más llamativo ya que se basa en una temática un poco más conocida que es la del dengue tenemos una investigación realizada por la MSc Beatriz Vega Riverón [15] cuya investigación es sobre "Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad". Este trabajo se enfoca en la aplicación de la técnica clasificatoria de árboles de regresión y clasificación (ARC), para hallar reglas de decisión que permitan clasificar un paciente con dengue en las diversas formas de la enfermedad a partir de características clínicas y de laboratorio. El desempeño se evaluó sobre la base de la capacidad del método de reducir la tasa de error global y su habilidad de clasificar correctamente a los pacientes.

En la siguiente investigación realizada por Sonia Lilia Mestizo Gutiérrez [16] a través de su artículo "Árboles de decisión y redes bayesianas para el análisis de genes involucrados en la enfermedad de Alzheimer", mencionando que los árboles de decisiones representan decisiones anidadas que sirven para clasificar los datos. Cuando se utiliza un árbol de decisión sobre los datos, se obtienen reglas que permiten clasificarlos. Un árbol se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver.

Materiales y métodos

Los datos y la información que se usaron para realizar nuestra investigación consistían en una base de datos creada a partir de unas pruebas de audiología, a esta base de datos se le aplico los diferentes métodos que poseen los árboles de clasificación para comprobar la eficacia de cada uno de ellos. En estos datos se indica un diagnóstico para cada paciente y las características de la persona como su edad, tipo de tímpano, si ha presentado mareos y otros. Aquí el objetivo es determinar cuáles de esos atributos sirven más para predecir el diagnóstico obtenido [17].
https://archive.ics.uci.edu/ml/datasets/Audiology+%28Standardized%29

Otros datos que también se usaron fueron acerca de una base de datos la cual contenía estudios sobre cáncer de próstata, pero a esa base de datos se le aplicaron algoritmos que podemos encontrar en los árboles de regresión, los cuales presentan alguna diferencia comparados con los arboles de clasificación. En este estudio el objetivo es predecir el valor del PSA (Antígeno Prostático Específico) en base a los valores de las demás características del paciente[18].
https://web.stanford.edu/~hastie/ElemStatLearn/

Una de las principales diferencia que podemos encontrar entre estos dos tipos de árboles es que cuando la "variable respuesta" o para ser más claros cuando nuestra variable de interés es numérica hablamos de árboles de regresión, mientras que las variables categóricas se analizan usando arboles de clasificación, pero en cualquier caso, el

funcionamiento de estos dos tipos de árboles es relativamente similar.

Por tal motivo si queremos explicar y predecir características u observaciones pertenecientes a los objetos de una clase cuyas bases pueden ser variables explicativas o cualitativas utilizaremos árboles de clasificación y por otro lado para un modelo explicativo y predictivo para una variable cuantitativa dependiente cuyas bases sean variables cuantitativas de igual forma, utilizaremos árboles de regresión.

Resultados

Arboles de clasificación (examen de audiología):

J-48.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	139	69.5 %
Incorrectly Classified Instances	61	30.5 %
Kappa statistic	0.6339	
Mean absolute error	0.0348	
Root mean squared error	0.1403	
Relative absolute error	47.8765 %	
Root relative squared error	73.9898 %	
Total Number of Instances	200	

Observaciones:

El algoritmo de árbol de decisión J-48 logra ser un análisis prácticamente óptimo de los datos que se le fueron ingresados, cuyas características fueron modificadas para que nos presentara un árbol menos extenso y comprensible, alcanzando un 69.5% de clasificación de las variables.

RandomTree.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	90	45 %
Incorrectly Classified Instances	110	55 %
Kappa statistic	0.3303	
Mean absolute error	0.056	
Root mean squared error	0.1703	
Relative absolute error	76.9768 %	
Root relative squared error	89.8101 %	
Total Number of Instances	200	

Observaciones:

El algoritmo de árbol de decisión RandomTree por mucho que se le trabajo a sus propiedades buscando un óptimo resultado y un árbol más simplificado, solo logro analizar un 45% de los datos introducidos.

REPTree.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	48	24 %
Incorrectly Classified Instances	152	76 %
Kappa statistic	0	
Mean absolute error	0.0727	
Root mean squared error	0.1896	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	200	

Observaciones:

El algoritmo de árbol de decisión REPTree no es un buen algoritmo para analizar los datos en los cuales estamos trabajando ya que por más que se intentó aumentar el porcentaje de variables evaluadas el algoritmo no presentaba mejores resultados.

DecisionStump.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	94	47 %
Incorrectly Classified Instances	106	53 %
Kappa statistic	0.3061	
Mean absolute error	0.0598	
Root mean squared error	0.1732	
Relative absolute error	82.2089 %	
Root relative squared error	91.3553 %	
Total Number of Instances	200	

Observaciones:

El algoritmo de árbol de decisión DecisionStump con los datos introducidos y con la configuración de las propiedades de análisis solo nos evaluó el 47% de las variables y otra desventaja que encontramos fue que no presentaba ni un esquema del árbol ni tampoco el árbol mismo.

SimpleCART.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	48	24 %
Incorrectly Classified Instances	152	76 %
Kappa statistic	0	
Mean absolute error	0.0717	
Root mean squared error	0.1895	
Relative absolute error	98.6842 %	
Root relative squared error	99.9371 %	
Total Number of Instances	200	

Observaciones:

El algoritmo de árbol de decisión SimpleCart presento casi las mismas desventajas que el algoritmo REPTree con un bajo porcentaje a l momento de analizar las variables y con un árbol de un solo nivel, todo esto con configuraciones para un mejor muestreo de datos.

LMT.

Estadísticas del análisis para el mejor de los casos:

Correctly Classified Instances	160	80	%
Incorrectly Classified Instances	40	20	%
Kappa statistic	0.768		
Mean absolute error	0.02		
Root mean squared error	0.1193		
Relative absolute error	27.5167 %		
Root relative squared error	62.9206 %		
Total Number of Instances	200		

Observaciones:

El algoritmo de árbol de decisión LMT resulto ser el algoritmo más eficiente al momento de interpretar el tipo de datos que presentaban los datos de estudio presentando un árbol bien resumido y con un 80% de las variables analizadas.

Arboles de regresión (examen de próstata):

M5P.

Estadísticas del análisis para el mejor de los casos:

Correlation coefficient	0.6123
Mean absolute error	0.7521
Root mean squared error	0.942
Relative absolute error	83.8664 %
Root relative squared error	81.4852 %
Total Number of Instances	97

Observación:

El algoritmo de árbol de regresión M5P resulto ser muy eficiente al momento de interpretar los atributos y variables que presentaba la base de datos de estudio presentando un árbol bien muy bien detallado y con un aproximado del 0.61 de frecuencia de las variables analizadas.

REPTree.

Estadísticas del análisis para el mejor de los casos:

Correlation coefficient	0.6297
Mean absolute error	0.7491
Root mean squared error	0.9159
Relative absolute error	83.5307 %
Root relative squared error	79.2276 %
Total Number of Instances	97

Observaciones:

El algoritmo de árbol de regresión REPTree es un buen algoritmo para analizar los datos de esta clase de bases de datos en los cuales estamos trabajando ya que presento un óptimo desempeño al momento de analizar las variables evaluadas en cual presento un aproximado de 0.62 de frecuencia de eficacia.

Conclusiones

El impacto que se desea obtener con el proyecto en aplicación de árboles de decisiones y de regresión como herramienta para el pronóstico de condiciones médicas es tomar un óptimo manejo del software Weka, tomando este como referencia en nuestra investigación y nos permita usarlo como apoyo para poder realizar un análisis exhaustivo y altamente comparativo de los datos analizados y de cómo los algoritmos internamente llevan a cabo sus funciones y distintos métodos de análisis[19].

Después de someter las dos distintas bases de datos que teníamos como material de investigación, a cada uno de los dos tipos de árboles de estudio, llegamos a la conclusión que al momento de analizar datos más que todo de variables cualitativas que fue el caso de los datos del estudio de audiología, los arboles de clasificación son los más competentes para estos datos, más precisamente el árbol de clasificación Logistic Model tree o LMT[20 - 23] que estadísticamente resulto ser el tipo de árbol que presento los resultados más eficientes en sus estadísticas con un promedio del 80% de clasificaciones correctas al momento de ejecutarse sobre los datos, cuyas variables respuesta o de interés fueron la variable Tymp() y la variable speech(), que correponden al tipo de tímpano y si la persona presenta problemas del habla [24-25].

Después al trabajar sobre otro estudio el cual era sobre el antígeno Prostático Específico el cual manejaba variables cuantitativas, nos damos cuenta que los arboles de regresión son los indicados para analizar este tipo de datos, cuyo árbol más eficaz fue el árbol de modelo M5 o el M5P[21 - 27], el cual en sus estadísticas alcanzo un aproximado del 0.62 de frecuencia al analizar los datos, cuya variable de interés fueron las variables de volume y el peso de la próstata (lcavol() y lweight()).

Referencias

[1] "Inteligencia Artificial Proyecto final," p. 65.

[2] D. Saur, "on Arboles de Decisi ´ on," pp. 121–133, 2003.

[3] J. Trujillano, A. Sarria-santamera, A. Esquerda, M. Badia, M. Palma, and J. March, "Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio," *Gac. Sanit.*, vol. 22, no. 1, pp. 65–72, 2008.

[4] L. Rokach and O. Maimon, "Decision Trees," *Data Min. Knowl. Discov. Handb.*, pp. 165–192, 2010.

[5] J. Villena, "Minería de datos," pp. 1–34, 2004.

[6] J. A. Alvarado Valencia, A. Carrillo, J. Forero, L. Caicedo, and J. C. Ureña, "XXVI Simposio Internacional de Estadística 2016 Sincelejo, Sucre, Colombia, 8 al 12 de Agosto de 2016," *XXVI Simp. Int. Estadística 2016*, pp. 1–4, 2016.

[7] "PARA LAS UPZ 79 CALANDAIMA , 65 PARA LAS UPZ 79 CALANDAIMA , 65," 2017.

[8] R. Barrientos, N. Cruz, H. Acosta, I. Rabatte, M. del C. Gogeascoechea, P. Pavón, and S. Blázquez, "Árboles De Decisión Como Herramienta En El Diagnóstico Médico," 2009.

[9] G. R. S. Martínez and J. A. S. Mejía, "Árboles De Decisiones En El Diagnóstico De Enfermedades Cardiovasculares," *Sci. Tech.*, vol. XVI, no. 49, pp. 104–109, 2011.

[10] A. Kevin, D. Hoz, U. J. Martínez-palacio, and F. E. Mendoza-palechor, "Técnicas de ml en medicina cardiovascular."

[11] F. Centro de Cibernética Aplicada a la Medicina. and Y. Sánchez Corales, "Revista cubana de informática médica," *Rev. Cuba. Informática Médica*, vol. 4, no. 2, pp. 174–183, 2001.

[12] M. Garc, L. Consultante, S. Ignacio, and C. Romero, "Trabajo de diploma Sistema inteligente para pronóstico de supervivencia," 2013.

[13] A. Aguilera and A. Subero, "Modelos de clasificación en marcha patológica usando árboles de regresión logística," *Multiciencias*, 2012.

[14] "Determinacion De La Eficacia De La Braquiterapia En Tratamiento De Cáncer Basada En Mineria De Datos," pp. 456–460, 2008.

[15] B. V Riverón, C. Lizet Sánchez Valdés, C. José Cortiñas Abrahantes, O. C. Peraza, C. Daniel González Rubio, and M. C. Peraza, "Classification of dengue hemorrhagic fever using decision trees in the early phase of the disease [Clasificación de dengue hemorrágico utilizando árboles de decisión en la fase temprana de la enfermedad]," *Rev. Cubana Med. Trop.*, vol. 64, no. 1, pp. 35–42, 2012.

[16] M. I. A. S. Lilia and M. Gutiérrez, "TESIS Árboles de decisión y redes bayesianas para el análisis de genes involucrados en la enfermedad de Alzheimer Doctor en Investigaciones Cerebrales Dr . Nicandro Cruz Ramírez," 2015.

[17] J. C. Goddard, J. M. Cornejo, F. M. Martínez, A. E. Martínez, H. L. Rufiner, and R. C. Acevedo, "Redes Neuronales y Árboles de Decisión: Un enfoque híbrido," pp. 1–7.

[18] O. Juarez and E. Castells, "Modelos de árbol de regresión bayesiano: un estudio de caso.," *Rev. Investig. Operacional*, vol. 31, no. 2, pp. 109–125, 2010.

[19] H. J. García and J. M. M. López, "Técnicas De Análisis De Datos Aplicaciones Prácticas Utilizando Microsoft Excel Y Weka," pp. 5–43, 2012.

[20] A. I. Aguilera, L. D. Cala, and A. R. Subero, "Modelo basado en metaclasificadores para diagn ´ ostico en marcha patol ´ ogica mediante an ´ alisis cin ´ etico Metaclassifiers – based model for pathological gait diagnosis using kinetic analysis," pp. 7–16, 2010.

[21] J. Molina and J. García, "Técnicas de Minería de Datos basadas en Aprendizaje Automático," *Técnicas de Análisis de Datos*, pp. 96–266, 2008.

[22] García Barrera, M., Córdova Rivera, L., & Rodríguez, A. (2018). La confidencialidad, el prestigio y la reputación como valores intangibles de la conciliación médica para el profesionista de la salud. Justicia, 23(34), 358-371. https://doi.org/10.17081/just.23.34.2896

[23] Pinillos PY, Herazo BY, Vidarte CJ, Crissien QE, Suarez PD.Niveles de actividad física y sus determinantes en mujeres adultas de Barranquilla, Revista Ciencia e Innovación en Salud, vol. 2, n°. 1, 2014. DOI:10.17081/innosa.2.1.7210.17081/innosa.2.1.68

[24] Valdés DO, Chávez PE, Torres BF. "Comportamiento de las crisis hipertensivas en un grupo de pacientes hipertensos", Revista Ciencia e Innovación en Salud, vol. 2, n°. 1, 2014. DOI:10.17081/innosa.2.1.72

[25] B. Londoño González and P. Sánchez, "Algoritmo Novedoso Para la Detección de Tareas Repetitivas en el Teclado", Investigacion e Innovación en Ingenierias, vol. 3, no. 2, 2015. DOI: 10.17081/invinno.3.2.2031

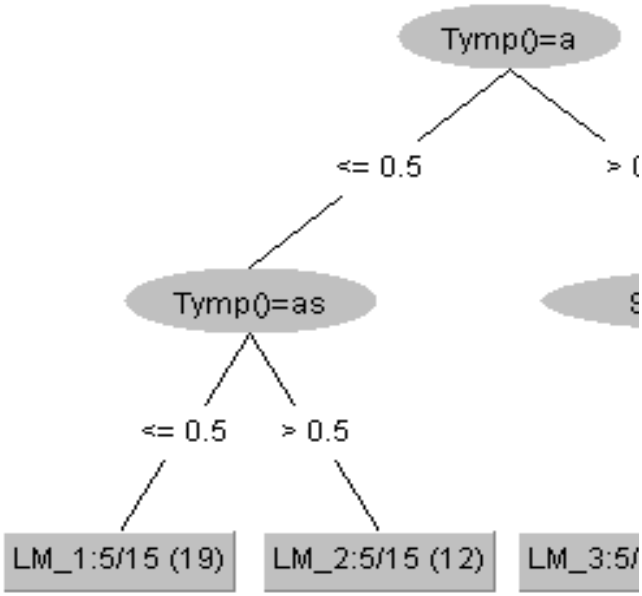
[26] Velez C, Vidarte JA. "Discapacidad y Determinantes Sociales de la Salud Estructurales e Intermedios: Diferencias por Género", Revista Ciencia e Innovación en Salud, vol. 2, n°. 2, 2014. DOI:10.17081/innosa.2.2.42

[27] E. Martelo, M. Manotas and B. Vallejo, "Prototipo De Una Aplicación Móvil Con Realidad Aumentada Para Mostrar Puntos De Información

De Ubicación De La Universidad Simón Bolívar En Barranquilla Colombia Mediante El Uso Del Navegador Móvil Junaio", Investigación e Innovación en Ingenierías, vol. 2, no. 2, 2014. DOI: https://doi.org/10.17081/invinno.2.2.2048

ANEXO 1

ARBOL DE CLASIFICACION:
LMT:



ARBOL DE REGRESION:
M5P: