Los árboles que predicen el futuro

Brahyan Esteban Rios Soto Universidad Eafit Colombia Berioss@eafit.edu.co	Jonathan Smith Julio Diaz Universidad Eafit Colombia jsjuliod@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

Los datos son algo de suprema importancia ya que estos nos dan información para tomar decisiones de la manera más segura y precisa posible. Así tener una predicción del éxito académico es de suma importancia; teniendo ese propósito, usaremos arboles de decisión y tomando como éxito el puntaje de la prueba Saber Pro predecir el resultado de los exámenes y con esto tomar medidas pertinentes para el mejoramiento de dicho resultado

Palabras clave

Árboles de decisión, aprendizaje automático, éxito académico, predicción de los resultados de los exámenes

1. INTRODUCCIÓN

La tecnología es un factor esencial hoy en día, en Colombia la transformación digital en la educación se conoce como educación 4.0 que busca entender a través de medios tecnológicos aspectos de suma importancia en la educación Colombiana, grandes avances se han hecho como lo es la deserción y diferentes factores de motivación en los estudiantes, no obstante al momento de la predicciones el éxito académico se han quedado cortos y con esta problemática en mente buscamos poder llegar a predecir el éxito académico utilizando arboles de decisión ya que es una herramienta excepcional en problemas lineales y no lineales lo nos da más herramientas para lograr nuestro objetivo

1.1. Problema

Predecir el éxito académico, con el fin de tomar medidas que puedan influir de forma significativa y positiva a los resultados académicos de la presente y futuras generaciones

2. TRABAJOS RELACIONADOS

2.1 Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales

Gracias a lo que hicieron pudieron predecir los resultados académicos de los estudiantes que cursaban Estructuras de datos y algoritmos I y II del Instituto Superior Politécnico José Antonio Echeverría usando como base calificaciones anteriores entre otros registros previos, como es descrito en el tirulo se utilizaron redes neuronales diseñadas en MATLAB, la precisión fue de un 75% a un 78%. (Álvarez et al. 2016)

2.2 Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia

Poder predecir el puntaje de unos estudiantes de Barranquilla en la Saber Pro usando minería de datos, se usa una minería de extracción de datos conocida como KDD, no mencionan una precisión exacta, pero se puede concluir que ronda el 90% (Garcia el al. 2019).

2.3 ANÁLISIS DEL DESEMPEÑO DE ESTUDIANTES DE INGENIERÍA EN LAS PRUEBAS SABER-PRO

Aplicando la metodología de minería de datos CRISP-DM, se realiza un estudio de los resultados obtenidos en las pruebas Saber-Pro de estudiantes de ingeniería en Antioquia (Colombia).
A partir de 108 variables académicas, económicas y socio demográficas se realizan 3 modelos analíticos:
1) agrupación de los tipos de estudiantes
2) selección de los factores que más influyen en el desempeño de las pruebas
3) predicción del desempeño en las pruebas a partir de las variables seleccionadas.

2.4 Evaluación del resultado académico de los estudiantes a partir del análisis del uso de los Sistemas de Control de Versiones

El objetivo de este trabajo es evaluar si el resultado académico de los estudiantes se puede predecir monitorizando su actividad en uno de estos sistemas. Para tal efecto, hemos construido un modelo que predice el resultado de los estudiantes en una práctica de la asignatura Ampliación de Sistemas Operativos, perteneciente al segundo curso del grado en Ingeniería Informática de la Universidad de León. Para obtener la predicción, el modelo analiza la interacción del estudiante con un repositorio Git. Para diseñar el modelo, se evalúan varios modelos de clasificación y predicción utilizando la herramienta MoEv. Esta herramienta permite entrenar y validar diferentes modelos de clasificación y obtener el más adecuado para un problema concreto.

3.2 Alternativas de algoritmos de árbol de decisión

3.2.1 ID3

Este algoritmo funciona de forma *recursiva*, los árboles que crea son binarios lo que quiere decir que hay dos opciones por cada atributo y el funcionamiento se basa en la entropía.

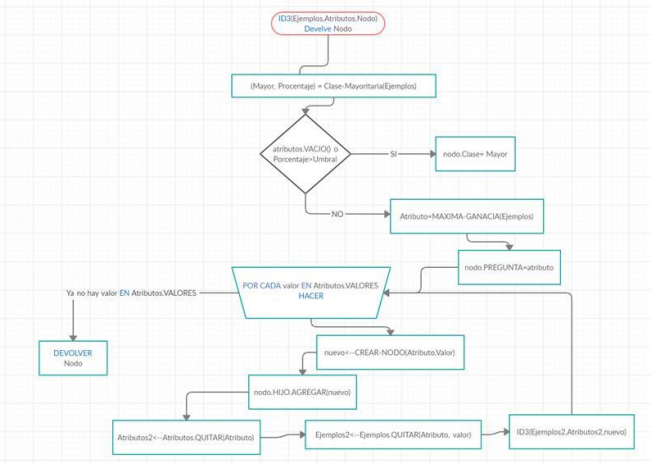
Ganancia de atributos = Se refiere a la cantidad de información que se puede extraer de esta, a su vez es inversamente proporcional a la Entropía; al maximizar la ganancia se minimiza la Entropía

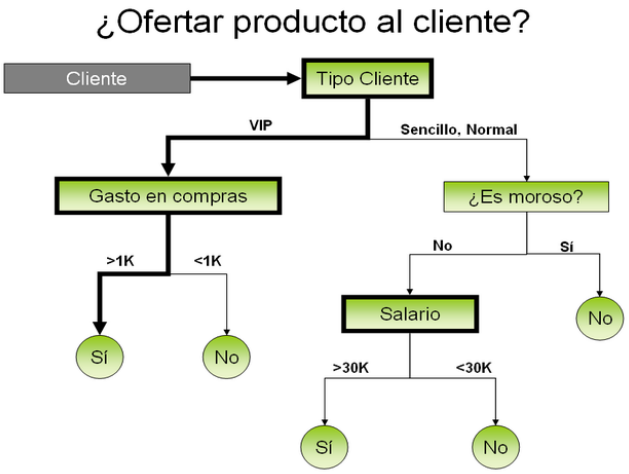
El algoritmo funciona insertando un conjunto de ejemplos, un conjunto de *atributos* y un *nodo* el cual puede ser vacío de ese modo se creara uno nuevo, lo que el algoritmo hace es buscar cuál de los *atributos* da mayor información después de eso se creara un nodo con ese atributo para posteriormente crear los *nodos hijos* de ese *atributo*, esto utilizando los ejemplos que contengan ese *atributo*; para después volver a llamar la función *recursiva* pero esta vez sin el anterior *atributo* y sus *ejemplos*, así de esta manera se repetirá con todos y cada uno de los *atributos* empezando por los más importantes que son los que mayor *ganancia* proporcionan. El caso base de la función atiende a dos situaciones: 1) Que ya no queden atributos lo que significaría que le árbol está terminado, 2) Llegar a un nodo hoja ósea a una respuesta (Si o No).

Complejidad

O(mn log(n)) + O(n(log(n)) ^2)

Donde n es el número de datos y m el número de atributos- El primer término se refiere a construir un árbol de decisión sin considerar podado. El segundo término se refiere cuando realizamos podado. Esto es independiente de que los atributos sean continuos o no.





https://4.bp.blogspot.com/-DfqMoKFhmjI/U1QrCrt_omI/AAAAAAAAAf0/mJrrx0drM0U/s1600/arbol_desicion1.png

3.2.2 Algoritmo C4.5

Este algoritmo es una extensión del algoritmo ID3, con algunas modificaciones en sus características:

En C4.5 se hicieron un número de mejoras a ID3. Algunas de ellas son:

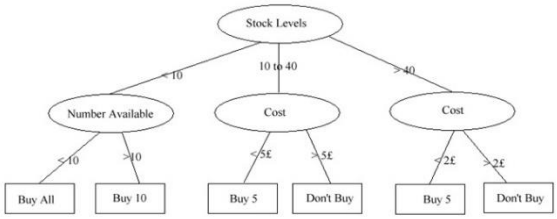
- Manejo de ambos atributos continuos y discretos - A fin de manejar atributos continuos, C4.5 crea un umbral y luego se divide la lista en aquellos cuyo valor de atributo es superior al umbral y los que son menores o iguales a él. (Wikipedia 2020).
- Manejo de atributos con costos diferentes. (Wikipedia 2020)
- Podando árboles después de la creación - C4.5 se remonta a través del árbol una vez que ha sido creado e intenta eliminar las ramas que no ayudan, reemplazándolos con los nodos de hoja. (Wikipedia 2020)

Este tiene la misma explicación del algoritmo ID3.

En resumidas cuentas, C4.5 es un algoritmo mucho más versátil y en muchos casos eficiente, idóneo para cuando la toma de decisiones no es binaria, ya que este puede devolver Clases a diferencia del ID3.

Su complejidad crece linealmente con el número de instancias de entrenamiento, pero exponencialmente con el número de atributos.

Ejemplo:



https://octaviansima.files.wordpress.com/2011/03/c45-tree.jpg

3.2.3 Algoritmo CART

El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión).

Este modelo admite variables generadas por el entorno, así como también variables de salida que sean, nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.

Este mismo hace uso principalmente del Índice GINI para calcular la homogeneidad, es decir, la pureza de un nodo en específico:

Índice GINIdel nodo = Peso relativo*Índice GINI Sub-nodo

Índice GINI de los Sub-nodos = p^2+q^2

Donde tenemos que p son aquellos datos que, si cumplen una dicha condición, mientras que q, son aquellos que no.

Ejemplo:

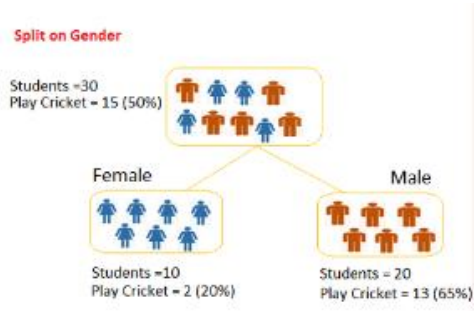
- 30 estudiantes

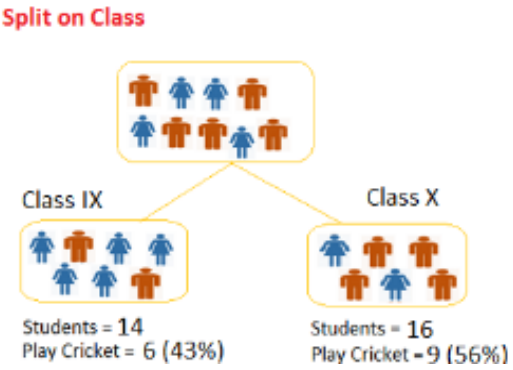
- 2 Variables: Genero (Hombre/Mujer) y Clase (IX/X)

- 15 estudiantes juegan cricket

- Crear un modelo para predecir quien jugará cricket

- Segregar estudiantes basados en todos los valores de las 2 variables e identificar aquella variable que crea los conjuntos más homogéneos de estudiantes y que a su vez son heterogéneos entre ellos.





Género	
Mujer	(0.2)^2 + (0.8)^2 = 0.68
Hombre	(0.65)^2 + (0.35)^2 = 0.55
Pond.	(10/30)*0.68* + *(20/30)*0.55 = **0.59**

Clase	
IX	(0.43)^2 +(0.57)^2 = 0.51
X	(0.56)^2 +(0.44)^2 = 0.51
Pond.	(14/30)*0.51* + *(16/30)*0.51 = **0.51**

Dando como mejor variable para segregar los datos, el género, ya que esta es de mayor magnitud que la de Clase.

3.2.4 Algoritmo CHAID

En el algoritmo CHAID (Kass, 1980) se propone para una variable respuesta de tipo cualitativo y predictores cualitativos y se basa en el test chi-cuadrado para contrastar las independencias. El algoritmo Chaid es un algoritmo secuencial y su esquema es el siguiente:

Dicho algoritmo contiene 4 fases, las cuales son:

Fase 1: En esta etapa Kass propone realizar el test chi-cuadrado, cruzando la variable respuesta con cada predictor, y ver que categorías tienen un perfil similar con respecto a la variable respuesta (que no son significativas). En ese caso, dichas categorías se agruparán.

Propone cruzar cada par de categorías y fusionar el par con mayor p-valor no significativo. Este proceso se repetirá hasta que no pudiesen agruparse más categorías.

Fase 2: Finalizada la fase anterior, hay que seleccionar el mejor predictor. De entre los que se tiene se elige para segmentar con un menor p-valor tras realizar el test chi-cuadrado. Entonces, el mejor predictor será aquel que discrimine mejor a los individuos según la variable respuesta.

Fase 3: Fijado el p-valor si el predictor seleccionado es significativo segmentamos en tantas ramificaciones como categorías tenga este.

Fase 4: Finalizada la segmentación se realiza nuevamente el proceso completo en cada una de las ramificaciones hasta que los nodos sean terminales, es decir, que no haya predictores significativos para dicho nodo.

Estadística X^2 (Chi cuadrado)

$$X^2_{(P\alpha;gl)}\ (H_0) = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

$P\alpha$ = Valor de probabilidad que se observa en la relación entre el Predictor y la VD que se presentará si el Predictor y la VD fueran estadísticamente independientes.

gl = Grados de libertad: (I-1)*(J-1), es decir, (columnas-1)*(filas-1)

f_{ij} = Frecuencia condicional o frecuencia empírica, valor que asume el Predictor y el Criterio en la celda ij.

F_{ij} = Frecuencia esperada o teórica bajo la hipótesis de independencia

P-Value (Valor de probabilidad asociada a Chi cuadrado) (a) REGLA DE DECISIÓN

P-value: Probabilidad asociada con la obtención de los resultados del estadístico X^2 meramente por azar o casualidad. El P-value mide la credibilidad de que la H_0 sea cierta. La posibilidad de que X^2 sea debido al azar es igual al P-value.

Cuando esta probabilidad es inferior a 0,05, (P α: Nivel de Significación=5%) se suele rechazar la hipótesis de independencia H 0 de no relación entre las variables, para aceptar la hipótesis alternativa H_1, que indica que la relación entre las variables existe y es estadísticamente significativa y no se debe al azar.

$$\textbf{Si P-value} < P\,\alpha \Rightarrow \not{H}_0 \text{ y } H_1$$

PROCEDIMIENTO CHAID

En cada etapa del análisis, CHAID divide el árbol en la variable del predictor que tenga el valor de probabilidad o p-value más bajo, siempre y cuando el valor p sea menor que el valor del nivel de significación (Pα = 0,05). Un valor p de 0,05 significa que la relación observada entre el predictor y la variable dependiente ocurrirá meramente por azar, y por lo tanto es poco probable que estén relacionadas. Si el predictor obtiene un valor más bajo, es menos probable que se deba al azar, por lo tanto, cabe suponer que existe relación con la VD.

REFERENCIAS

Álvarez Blanco, Jorge, Lau Fernández, Rogelio, Pérez Lovelle, Sonia, & Leyva Pérez, Exiquio C. Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales. *Ingeniare. Revista chilena de ingeniería*, Facultad de Ingeniería Informática. Instituto Superior Politécnico José Antonio Echeverría, La Habana- Cuba, 2016, *24*(4), 715-727.

Jose G.G, Paola S.S, Manuel O.,Sergio O. Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. Formación universitaria, Facultad de Ingenierías, Universidad Simón Bolívar, Barranquilla - Colombia. 2019, 12(4), 55-62.

A. Oviedo, J. Jiménez. "Minería de datos educativos: análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO", Revista Politécnica, vol. 15, no.29 pp.128-140

Gutiérrez Fernández, A., Guerrero Higueras, A. M., Conde González, M. A., y Fernández Llamas. "Evaluación del resultado académico de los estudiantes a partir del análisis del uso de los Sistemas de Control de Versiones". RIED(Revista Iberoamericana de Educación a Distancia), vol. 23 no.2, pp. 127-145.

Wikipedia, La enciclopedia libre. C4.5. (2020, 8 de febrero). Fecha de consulta: 13:44, agosto 16, 2020 desde

https://es.wikipedia.org/w/index.php?title=C4.5&oldid=123402893.

Johanna Orellana Alvear. Correo: johanna.orellana@ucuenca.edu.ec . 12-16 noviembre 2018. Arboles de decisión y Random Forest – Bookdown. Fecha de consulta: 14:13 el 15 de agosto de 2020 desde

https://bookdown.org/content/2031/arboles-de-decision-parte-i.html#indice-gini

Prof. Rubén J. Rodríguez. Estadística II Licenciatura en Sociología, UCES. INTRODUCCIÓN A LA SEGMENTACIÓN JERÁRQUICA y ANÁLISIS CHAID. Fecha de consulta: 3:23 el 16 de agosto de 2020 desde

http://www.rubenjoserodriguez.com.ar/wp-content/uploads/2011/11/Introduccion_a_la_Segmentacion_Jerarquica-y-Analisis-CHAID.pdf

Sergio Ramos Hernández. Tutoras: María Purificación Galindo Villadon y María del Carmen Patino Alonso. Universidad de Salamanca. TAID versus CHAID. Fecha de consulta: 2:42 el 16 de agosto de 2020 desde

https://gredos.usal.es/bitstream/handle/10366/128233/TFM_MAADM_Ramos_Hernandez_Sergio.pdf?sequence=4&isAllowed=y