ANÁLISIS PREDICTICO PARA LAS PRUEBAS SABER PRO

Luis Fernando Vargas Agudelo
Universidad Eafit
Colombia
lvarga12@eafit.edu.co

Tomás Bedoya Henao
Universidad Eafit
Colombia
Tbedoyah@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

RESUMEN
En el presente informe se presentan diferentes metodologías para el análisis predictivo que puedan dar respuesta a la problemática de poder conocer, analizar y predecir si un estudiante según sus resultados en la prueba de conocimiento saber 11 si será exitoso en los resultados de la prueba Saber Pro, esto con el fin de comprender cuales pueden ser las falencias y las variables que influyen directamente en estos resultados.

1. INTRODUCCIÓN

El considerable aumento de datos que se generan día a día con la interacción de las personas y los instrumentos tecnológicos, el análisis masivo de los datos se ha convertido en una herramienta cada vez más importante para las empresas y los países.

Gobiernos como China han comenzado a utilizar el análisis de datos para obtener puntajes de sus habitantes y de esta forma saber quiénes son aquellos de los cuales pueden tener más beneficios sociales, económicos, legales entre otros. Esta situación ha generado grandes escándalos éticos en los que se pone en duda la libertad de expresión y social.

En Colombia el análisis de los datos está comenzando a ser algo muy importante para las entidades académicas, estas mediante un análisis predictivo desean conocer cuál será el comportamiento de sus estudiantes en el transcurso del tiempo y de alguna manera predecir si sus resultados en pruebas futuras serán exitosos o no. Soluciones así impulsaría fuertemente a Colombia en la región, ya que actualmente los índices de educación media y superior no son los mejores, además estudios demuestran que el 45% de los estudiantes no están seguros de que la carrera seleccionada es realmente lo que les gusta, lo cual lleva a un alza en la deserción de la educación superior la cual se encuentra cerca del 43%.

Poder crear soluciones que contribuyan con estas situaciones tienen gran importancia para lo sociedad, de esta forma diferentes entidades podrán conocer en que tienen falencias las personas y más importante aún, en que son realmente buenas, permitiendo así que con herramientas de aprendizaje se pulan sus puntos fuertes y obtener el mayor beneficio para ellos mismos y la sociedad.

El objetivo de este informe es conocer, entender y analizar cuáles métodos existen para ayudar a darle una solución a esta necesidad y darle al lector una idea clara de su funcionamiento

2. PROBLEMA

La necesidad que se desea solucionar es poder conocer mediante un análisis predictivo si un estudiante de educación superior tendrá éxito o no en las pruebas Saber Pro. para calificar los resultados de la prueba como exitosos, el puntaje global obtenido debe estar por encima del promedio general.

Para poder predecir el éxito o no se utilizarán variables predictoras como el estado socioeconómico cuando el individuo presentó las pruebas saber 11 y el desempeño obtenido en las mismas.

Resolver esta problemática tiene diferentes objetivos, algunos pueden ser conocer las capacidades y falencias que tienen los estudiantes, identificar cuáles pueden ser gustos profesionales y permitir crear una educación un poco más personalizada lo que traería grandes beneficios para los diferentes entes interesados.

3. METODOLOGÍAS

Antes de conocer cuáles son las diferentes soluciones que existen para este tipo de problemáticas es conveniente conocer cuál es el funcionamiento general y entender el flujo del análisis predictivo.

Como su nombre lo indica el análisis predictivo se basa en realizar predicciones basadas en los datos, este proceso se realiza con técnicas estadísticas y aprendizaje automático con el fin de crear un modelo predictivo. Con esto modelos es posible predecir cosas como ¿Cuál es la posibilidad de que un cliente regrese por un nuevo producto financiero? o ¿En cuánto tiempo una máquina podría necesitar un cambio de pieza?

Los diferentes modelos predictivos cumplen un simple flujo general: recolección de los datos, preprocesamiento de los datos, desarrollo del modelo predictivo y finalmente la salida de estos modelos es integrada con sistemas analíticos para la lectura.

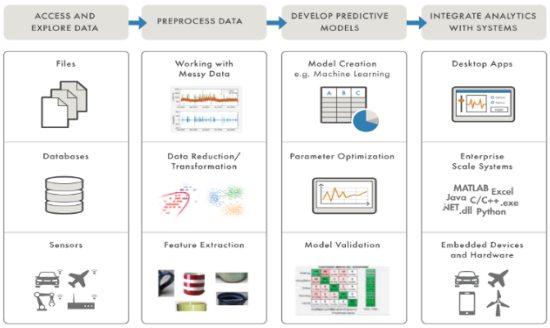

Ilustración 1 Flujo de trabajo de análisis predictivo. Tomado de:MathWorks, predictive-analytics

La manera más común y simple de realizar pronósticos para variables aleatorias, es la **regresión lineal**, que permite investigar las relaciones entre una variable dependiente y otras independientes mediante la formulación de ecuaciones matemáticas y procesos de optimización. De esta manera, se puede encontrar una línea (Función) que se ajuste lo mejor posible a los datos que se tienen o sobre los cuales se pretende trabajar en una predicción.

Los modelos **autorregresivos** o AR buscan generar pronósticos del proceso modelado basándose en el valor que haya tomado en momentos anteriores, ya que usualmente los datos están ordenados cronológicamente. En otras palabras, la variable dependiente y la variable explicativa son la misma, pero en diferentes momentos.

Por otro lado, **las redes neuronales** artificiales buscan imitar el proceso de aprendizaje neuronal biológico. Dichas neuronas artificiales se conectan entre sí y 'aprenden' con la ejecución del proceso, en esencia, dados unos parámetros o estimadores, se busca llegar a un resultado determinado mediante la combinación de estos (Condicionada por pesos o funciones), saber qué combinación produce x o y resultado deseado, es el problema que pretenden solucionar las redes neuronales.

Por su parte, los **árboles de decisión** permiten seleccionar matemáticamente la opción más beneficiosa, pues en general un árbol de decisión es un mapa que contiene las diferentes opciones junto con los posibles resultados generados al escoger entre las opciones mapeadas, una tras otra. Por lo general comienza con un único nodo que se bifurca o se ramifica, generando otros nodos adicionales que se vuelven a ramificar, todos atados a probabilidades, costos o beneficios y de esta manera, la metodología permite encontrar el camino óptimo. Existen tres tipos de nodos: de decisión, de probabilidades y nodos de finalización.

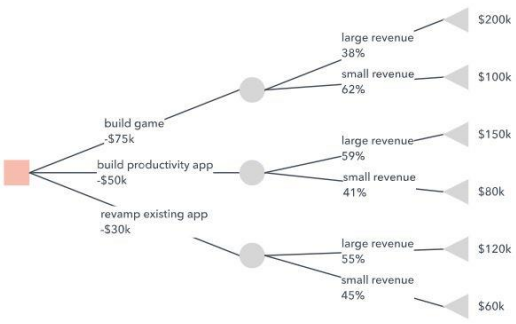

Ilustración 2 Mapa árbol de decisión. Tomado de: Árbol de decisión, lucidchart

3. TRABAJOS RELACIONADOS

Para desarrollar la solución al problema planteado en el inicio del informe se realizará un árbol de decisión, debido a esto es importante conocer cuáles son las diferentes formas que existen para desarrollarlos y los problemas similares que se han solucionado. A continuación, se profundizará en esto. Es importante recalcar que la diferencia entre los diferentes algoritmos radica, principalmente son las estrategias para 'podar' el árbol o reglas empleadas para dividir los nodos.

3.1 Algoritmo ID3

Las soluciones que se realizan con este algoritmo están basadas en nodos de decisión que a su vez están asociados a uno de los atributos, estos nodos cuentan con dos o más ramas que representan posibles valores del atributo. También está conformado por nodos-hojas, que representan al atributo objetivo que se quiere clasificar, pero todos al mismo tiempo, por lo que representa la decisión final del árbol. Este algoritmo utiliza el concepto de ganancia de información para seleccionar los atributos más útiles en cada iteración, mediante la entropía, pues a mayor homogeneidad de una muestra, más tiende a cero ésta. Si la muestra está igualmente distribuida, la entropía es igual a cero.

De esta manera, el algoritmo basa la ganancia e información en el decremento de la entropía. El atributo que crea las ramas más homogéneas (O con entropía más cercana a cero), se calcula de la siguiente manera:

i. Se calcula la entropía total.
ii. Se dividen los datos en función de los atributos.
iii. Se calcula la entropía de cada rama y se suma cada una de estas para obtener el total.
iv. Se resta el resultado del cálculo inicial (La entropía total).
v. El resultado es la ganancia de información o decremento de la entropía.
vi. El atributo que tenga mayor ganancia de información es un nodo de decisión del árbol planteado.

La fórmula para la entropía es:

$$E(S) = \sum_{i=1}^{c} -pi \log_2 pi$$

$$E(S) = -P \log_2 P - N \log_2 N$$

Donde P son los ejemplos positivos y N los negativos

Usar el algoritmo ID3 tiene algunas ventajas: se puede construir un árbol pequeño con reglas comprensibles; es un algoritmo rápido; dado que es recursivo no continúa si encuentra nodos hoja, por lo que se reduce el número de iteraciones o comprobaciones; utiliza todo el conjunto de datos que se pone a servicio.

Algunas de las principales desventajas son: el sobre entrenamiento o sobre clasificación que puede generar el algoritmo, sólo se comprueba o evalúa un atributo en cada paso; clasificar datos continuos es muy costoso y requiere una gran capacidad computacional, pues se deben crear

varios árboles para conocer dónde se rompe esa continuidad.

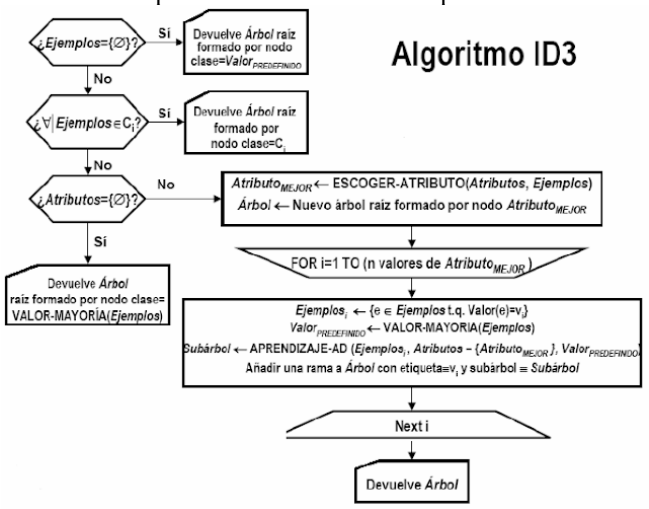

Ilustración 3 Pseudocódigo ID3. Tomado de: López, Bruno. Algoritmo ID3.

EL algoritmo ID3 es frecuentemente utilizado por diferentes empresas destinadas al marketing, ya que con estos pueden tomar decisiones de gran importancia en aspectos de lanzamiento de un nuevo producto al mercado, una problemática muy común es saber si un producto debe ser lanzado a escala nacional o regional y si se debe distribuir a nivel nacional o solo en algunas regiones. Problemáticas como estas son muy comunes y una gran herramienta para su solución son los árboles de decisión, para darle una solución a esta problemática se aconseja realizar el dibujo desarrolla del árbol de decisión como el siguiente:

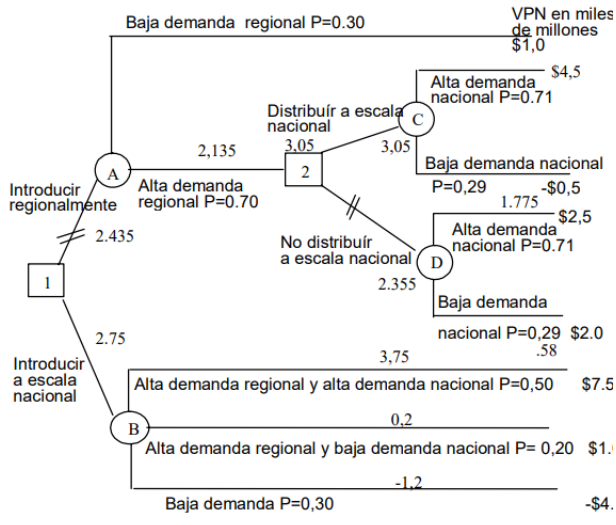

Ilustración 4 Árbol de decisión desarrollado. Tomado de: cashflow88.com/decisiones/riesgo/capitulo8

En el ejemplo se puede observar que la probabilidad por una demanda alta a nivel regional es del 70% Cada combinación puede generar una decisión diferente y así tomar alternativas.

3.2 Algoritmo CART

Esta metodología básicamente construye un árbol máximo que sobreajuste la información suministrada y posteriormente hace el árbol más sencillo, es decir, deja los nodos más importantes para finalmente seleccionar el árbol óptimo.

En primera instancia, se tiene el nodo raíz, por lo que el algoritmo busca partirlo en dos nodos hijos tomando como criterio la variable más adecuada para ello. Para elegir la mejor variable se implementa una medida de 'pureza' en la valoración de los dos nodos hijos. Adicionalmente, se asegura que la pureza de ambos nodos sea la máxima, y para ellos se utiliza regularmente la función Gini. También se puede evaluar la pureza de forma conjunta para todo el árbol.

Desde la raíz del árbol (Nodo inicial) hasta los nodos hojas se deben asignar una clase o una etiqueta a los datos. Esta asignación se realiza mediante una función que tiene en cuenta el número de apariciones de esta, las probabilidades y la pureza de la partición.

El proceso se repite hasta que se llega al tope máximo de niveles del árbol (Si se ha fijado); sólo hay una observación en cada nodo-hoja (por lo que estarían ya clasificados); o todas las observaciones tienen la misma probabilidad asignada en los nodos hoja, debido a esto el criterio de máxima pureza es imposible de determinar.

El árbol complejo debe simplificarse, para esto es usado un método para 'podar' el árbol. El procedimiento debe retirar los nodos que aportan muy poca precisión al modelo; se usa entonces una medida de costo-complejidad y se busca el árbol que obtiene menos valor en el parámetro.

De esta manera, los árboles sencillos pueden ser generados con menor dificultad.

Para seleccionar el árbol óptimo el algoritmo compara los resultados obtenidos con los del registro real (Con los que aprendió) para ver cuál es el que más se ajusta.

3.3 Algoritmo C 4.5

Este algoritmo, como el ID3, genera árboles de decisión con arreglos de datos de entrenamiento mediante el concepto de entropía (De la teoría de información), usándolo como criterio para la división eficaz; esto ya se explicó en el algoritmo ID3.

El algoritmo tiene algunos casos base: todas las muestras pertenecen a una misma clase o tienen una misma etiqueta, por lo que genera un nodo hoja eligiendo esa clase; ninguna de las características genera una ganancia de información, por lo que crea un nodo de decisión con el valor esperado de

la clase; instancia de una clase antes no vista es encontrada, creando un nodo con el valor esperado de la clase.

Sin embargo, tiene mejoras sustanciales con respecto al algoritmo ID3:

El algoritmo no sólo maneja atributos discretos sino también continuos, mediante el establecimiento de un umbral y la comparación de valores del atributo con éste para dividir la lista.

Permite marcar valores de atributos con (?) para indicar falta de los mismos y no utilizarlos posteriormente en los cálculos de ganancia y entropía.

Poda los árboles después de su creación, sustituyendo a aquellos que no aportan precisión por nodos hoja.

3.4 Algoritmo CHAID

Dado que las principales diferencias radican en la manera como el algoritmo clasifica o divide los datos, podemos decir que CHAID (Chi-are automatic interaction detector) usa como base la distribución x^2. El índice de probabilidad Chi-cuadrado es un método estadístico que pretende establecer la independencia entre la distribución de los datos observados u obtenidos empíricamente y una distribución teórica. Testea la hipótesis nula de que dos variables son independientes la una de la otra o, en otras palabras, que el comportamiento o valor de una de ellas no induce o condiciona de ninguna manera el comportamiento de la otra. Para el caso particular de árboles de decisión, se postula la hipótesis nula de que la división y la variable de la clase son independientes.

Para terminar el proceso, se usa un umbral para comparar los valores del atributo. Adicionalmente, usa la corrección de Bonferroni para solucionar el problema de comparaciones múltiples. Esta corrección, aplicada a árboles de decisión, lo que hace básicamente es mitigar el sesgo que pueda haber cuando se tienen entradas con muchos valores, es decir, esta corrección o ajuste al número de valores categóricos de la variable de entrada.

Finalmente, este algoritmo, como el C 4.5 (Y el c5, su evolución), también utiliza una metodología para marcar datos o valores faltantes en los atributos y así no incluirlos en los cálculos de división. Puede reconocer valores de clase que sesguen su funcionamiento o precisión marcándolos como faltantes y así seguir trabajando con normalidad.

REFERENCIAS

Analicaweb. (julio de 30 de 2015). *La predicción del dato: Redes Neuronales Artificiales.* Recuperado el 7 de febrero de 2020, de https://www.analiticaweb.es/la-prediccion-del-dato-redes-neuronales-artificiales/

Caparrini Sancho, F. (enero de 5 de 2013). *Árboles de desición: Algoritmo ID3.* Recuperado el 7 de febrero de 2020, de https://es.slideshare.net/FernandoCaparrini/arboles-decision-id3

Casas Mogollón, P. A. (6 de diciembre de 2018). *El problema no es solo plata: 42 % de los universitarios deserta.* Recuperado el 7 de febrero de 2020, de El Espectador: https://www.elespectador.com/noticias/educacion/el-problema-no-es-solo-plata-42-de-los-universitarios-deserta-articulo-827739

Helpes. (19 de febrero de 2019). *Algoritmo de C4.5.* Recuperado el 7 de febrero de 2020, de http://www3.helpes.eu/01096402/AlgoritmoDeC45

Numerentur. (29 de marzo de 2019). *Árboles de desición - DT.* Recuperado el 7 de febrero de 2020, de http://numerentur.org/arboles-de-decision/

Scielo. (febrero de 2018). *de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio.* Recuperado el 7 de febrero de 2020, de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-91112008000100013