Arboles de decisión para predecir académicamente exámenes ICFES.

Miguel Angel Echavarria Ocampo Eafit Colombia maechavaro@eafit.edu.co , Thomas Nieto Eafit Colombia , Miguel Correa Universidad Eafit Colombia [macorream@eafit.edu.co](mailto:macorream@eafit.edu.co) ,Mauricio Toro Universidad Eafit Colombia [mtorobe@eafit.edu.co](mailto:mtorobe@eafit.edu.co) .

El problema que queremos resolver en este proyecto es la predicción de exámenes de ICFES, para que los estudiantes mediante este puedan posiblemente saber su resultado, existen diferentes problemas relacionados con este son la predicción del clima para deportes y resultados de votaciones de diversos candidatos.

Palabras clave

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

1. INTRODUCCIÓN

En el presente semestre, tenemos como motivación intentar predecir mediante arboles de decisión el resultado de las pruebas ICFES, o de diferentes evoluciones que se realicen mediante nuestro código, al igual, que esto abre las puertas para la inteligencia artificial motiva mucho más para poder saber cómo se ejecuta estas y sus árboles.

1.1. En este semestre buscamos predecir los resultados de un estudiante en sus diversas evaluaciones, para poder lograr el éxito académico de este, utilizando diversos cálculos, métodos y estructuras para lograr esto en un árbol de decisión.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran aplicabilidad. Buscamos que el árbol viaje por los datos buscando la mayor probabilidad de éxito, que dato afecta más a la hora de efectuar la búsqueda y así separar a los aspirantes a éxito en los exámenes de ICFES.

1.3 Estructura del artículo En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

En este semestre, el trabajo relacionado es la investigación de árboles de decisión para predecir los resultados de los exámenes de los estudiantes, lo cual miramos en diversas fuentes, donde podemos tomar bases para nuestro proyecto.

Fuente 1 https://revistas.uptc.edu.co/index.php/investigacion\_duitama/article/view/9184

Fuente 2 http://scielo.sld.cu/scielo.php?script=sci\_arttext&pid=S0864-21412004000300002

Fuente 3 https://www.rcs.cic.ipn.mx/2018\_147\_5/Ensamble%20de%20clasificadores%20para%20determinar%20el%20perfil%20academico%20del%20estudiante%20usando%20arboles.pdf

Fuente 4 https://dialnet.unirioja.es/servlet/articulo?codigo=6976577

3. MATERIALES Y MÉTODOS En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

3.1 Recopilación y procesamiento de datos Obtuvimos datos del Instituto Colombiano de Fomento de la Educación Superior (ICFES), que están disponibles en línea en ftp.icfes.gov.co. Estos datos incluyen resultados anonimizados de Saber 11 y Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los gradudados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 como Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba. En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen? Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes. Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal, como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en <https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets> .

Conjunto de datos 1 Conjunto de datos 2 Conjunto de datos 3 Conjunto de datos 4 Conjunto de datos 5

Entrenamiento 15,000 45,000 75,000 105,000 135,000

Validación 5,000 15,000 25,000 35,000 45,000

Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión

Presentaremos diferentes algoritmos en los que tomar bases para crear un árbol de decisión y buscar diferentes alternativas para el proyecto.

3.2.1

El algoritmo ID3 es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

El conjunto de ejemplos deberá estar conformado por una serie de tuplas de valores, cada uno de ellos denominados atributos, en el que uno de ellos, es el objetivo, el cual es de tipo binario.

De esta forma el algoritmo trata de obtener las hipótesis que clasifiquen ante nuevas instancias, si dicho ejemplo va a ser positivo o negativo.

ID3 realiza esta labor mediante la construcción de un árbol de decisión.

Los elementos son:

Nodos: Los cuales contendrán atributos.

Arcos: Los cuales contienen valores posibles del nodo padre.

Hojas: Nodos que clasifican el ejemplo como positivo o negativo.

3.2.3

C4.5 es una extensión del algoritmo ID3 desarrollado anteriormente. Los árboles de decisión generados por C4.5 pueden ser usados para clasificación, y por esta razón, C4.5 está casi siempre referido como un clasificador estadístico.

3.2.4

El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión) fue diseñado por Breiman et al. ... Este modelo admite variables de entrada y de salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión.

4. DISEÑO DE LOS ALGORITMOS

Escogimos el algoritmo CART. Este nos permite predecir el éxito de las pruebas saber mas fácilmente por medio de variables, como es el puntaje de las diferentes materias que se han evaluado un ejemplo es: Matemáticas, fisica e inglés. A partir del resultado que el sujeto obtuvo se toman ciertas decisiones que nos retorna la probabilidad de éxito académico.

Diagrama

Descripción generada automáticamente

4.1 Estructura de los datos

Un árbol de decisión binario para predecir Saber Pro basado en los resultados de Saber 11. Los nodos los dividimos por materia a evaluar el cual según el puntaje se toma cierta decisión, luego se saca el Gini que nos ayuda a predecir el éxito académico.

Puntaje Fisica >=50

4.2 En este semestre, usamos el algoritmo de árbol de decisión CART para entrenar y de igual forma usamos el mismo algoritmo para el conjunto de verificación, se uso este algoritmo ya que nos dio la impresión de ser mucho más rápido y eficaz que las demás opciones presentadas anteriormente.

4.2.1 Nosotros utilizamos el algoritmo dicho anteriormente, para entrenar este algoritmo utilizamos un archivo que le ingresaba datos de ciento treinta y cinco mil estudiantes, el algoritmo para crear el árbol empieza a mirar los puntajes de cada estudiante, utilizando como raíz la materia de ingles, si este tenia un puntaje mayor o igual a 60 seguía a la siguiente instrucción, la materia de Ciencias Sociales, acá miraba si tenia un puntaje mayor o igual a 50, luego nos retornaba el Gini que se calculaba por medio de la cantidad de estuante con éxito y los estudiante con poca probabilidad de éxito académico; si desde la raíz el estudiante no cumplía la primera condición se agregaba a la probabilidad baja de éxito.

Diagrama

Descripción generada automáticamente

4.3 Análisis de la complejidad de los algoritmos

Sabiendo que N es el número de filas y M el número de columnas de la matriz, sacamos la complejidad del algoritmo en base al código y de igual forma la velocidad en la que arrojaba los resultados con tal cantidad de datos ingresados, al sacarlo con el conjunto de entrenamiento y validación obtuvimos:

|  |  |  |
| --- | --- | --- |
|  | **Complejidad en tiempo** | **Complejidad en memoria** |
| Entrenamiento del modelo | O(N2\*M\*2M) | O(N\*M\*2M) |
| Validación del modelo | O(N\*M) | O(1) |

4.4 Criterios de diseño del algoritmo:

Como criterio queríamos que el algoritmo fuera eficaz, rápido y preciso, para ello probamos con los diferentes algoritmos que teníamos como opción pero el que nos convenció más fue el algoritmo CART, desde allí empezamos a crear el árbol de decisión y luego a implementar el Gini; viendo la cantidad tan abrumadora de datos que se nos presentaba teníamos pavor de que el consumo de memoria fue demasiado alto, pero afortunadamente no lo fue, esto lo verificamos gracias al código y al administrador de tareas de Windows, al final el algoritmo cumplió con lo que se esperaba a la hora de ingresar los datos y en el proyecto como general.

5. RESULTADOS

5.1 Evaluación del modelo En esta sección, presentamos algunas métricas para evaluar el modelo. La precisión es la relación entre el número de predicciones correctas y el número total de datos de entrada. Precisión. es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos identificados por el modelo. Por último, Sensibilidad es la proporción de estudiantes exitosos identificados correctamente por el modelo y estudiantes exitosos en el conjunto de datos.

5.1.1 Evaluación del modelo en entrenamiento A continuación presentamos las métricas de evaluación de los conjuntos de datos de entrenamiento y validación en la Tabla 3.

|  |  |  |
| --- | --- | --- |
|  | **Conjunto de entrenamiento** | **Conjunto de validación** |
| **Exactitud** | 0.7688 | 0.7489 |
| **Precisión** | 0.781833 | 0.7455 |
| **Sensibilidad** | 0.74950 | 0.7616 |

5.2 Tiempos de ejecución

Se calculo el tiempo de ejecución de cada conjunto de datos en Github. Se medio el tiempo de ejecución 100 veces, para cada conjunto de datos e informar del tiempo medio de ejecución para cada conjunto de datos al igual que el consumo de memoria.

6. DISCUSIÓN DE LOS RESULTADOS

Los resultados obtenidos por el algoritmo son precisos, pero como en todo se puede colar en los datos un falso positivo que incremente o disminuya el resultado que se le da al cliente, pero el algoritmo cumple con lo pedido para el proyecto y de igual forma, cumple con los criterios que teníamos sobre este en velocidad, precisión y eficacia, al ser un conjunto tan grande de datos se esperaba que el algoritmo fuera lento, pero gracias a la eficacia que nos permitió el algoritmo CART se pudo encontrar la rapidez que se buscaba. Al principio se tuvieron muchas dificultades a la hora de crear el código ya que nos arrojaba muchos errores, pero con ayuda de los monitores y profesor se pudieron solucionar y terminar el proyecto

AGRADECIMIENTOS

Agradecemos al profesor Mauricio Toro, que nos guio por el trayecto de este proyecto para poder finiquitarlo correctamente a lo largo de todo el curso, de igual forma agradecemos la ayuda del monitor Miguel Correa que nos soluciono muchas dudas a lo largo de los meses que nos iluminaron para la continuación del proyecto.

REFERENCIAS

1. Fischer, G. y Nakakoji, K. Amplificando la creatividad de los diseñadores con entornos de diseño orientados al dominio. en Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

2.Oracle:<http://docs.oracle.com/javase/7/docs/technotes/guides/visualvm/profiler.html>

3. “Dany”Script: https://es.scribd.com/document/422015007/Algoritmo-CART