PREDICCIÓN ACADÉMICA: Una mirada al futuro de los estudiantes universitarios

Juan Jose Gabriel Cañon Diaz
Universidad Eafit
Colombia
jjcanond@eafit.edu.co

Bryant Samuel Prada García
Universidad Eafit
Colombia
bspradag@eafit.edu.co

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

RESUMEN
El objetivo de esta entrega es encontrar las posibles problemáticas del proyecto el cual es poder predecir si la persona tendrá o no tendrá éxito académicamente a nivel superior; una vez encontradas las problemáticas buscar las posibles soluciones más eficientes o más factibles, ya que la idea principal del proyecto es predecir con mayor precisión posible mediante una serie de datos que recibirá la máquina y así, en base al historial, descubra cual es el factor más grande que influye si la persona tendrá éxito o no.

Palabras clave
Información

Palabras clave de la clasificación de la ACM

1. INTRODUCCIÓN

En Colombia las pruebas Icfes miden el nivel académico de cada estudiante de bachillerato que llega al grado 11, recolectando así información social, económica, demográfica y familiar de todos los estudiantes del país que las presentan, de esta forma con la basta información recolectada se puede hacer infinidad de predicciones respecto al futuro académico de cada estudiante.

En este proyecto pretendemos predecir si el resultado que el estudiante obtendrá en las pruebas Saber Pro aplicada en los estudiantes universitarios del país con el mismo fin de evaluar conocimientos, estará por encima o por debajo del promedio.

Cabe resaltar que los estudiantes que están por encima del promedio (en ambas pruebas) ocasionalmente reciben becas o ayudas de diferentes entidades del estado o privadas, asegurando una mejor calidad de vida.

2. PROBLEMA

Lograr predecir cuál es la probabilidad de éxito de una persona en base a sus datos y descubrir cuál es el factor que más afecta este resultado de si tendrá éxito o no tendrá éxito conforme aprende la máquina

3. TRABAJOS RELACIONADOS

3.1 Predicción de la estructura secundaria de la proteína

Pespad es una herramienta algorítmica publicado por Claudia X. Mazo y Oscar F. Bedoya en el año 2010, que con ayuda de diferentes modelos de arboles de decision logra realizar una predicción mucho más contundente de la estructura secundaria de la proteína que los arrojados por los métodos individuales generalmente usados en el campo de la bioinformática. [1]

3.2 Predicción de resultados deportivos con técnicas de Machine Learning

Sergio Calderón Pérez en el año 2017 publicó un algoritmo que con ayuda de diferentes herramientas inherentes al Machine Learning, logra realizar una predicción “más acertada” que las de las casas de apuesta, teniendo en cuenta la calidad o momento por el cual está pasando cada jugador debido a sus actuaciones recientes y diversos datos recolectados por diferentes medios. [2]

3.3 Problema overfitting y underfitting:

Este problema consiste que el programa es entrenado con un número de datos limitado, el cual le ayuda a aprender conforme va leyendo los datos, sin embargo, la máquina solo aprende en base a los datos con los que ya ha interactuado y al recibir un dato diferente, no lo reconoce, por ejemplo,

- Underfitting: se entrena la máquina con un solo tipo de planta, al recibir otro tipo de planta, no la reconocerá como planta.
- Overfitting: se entrena con varios tipos de planta con cualidades semejantes, al recibir un tipo de planta que no tenga ningún atributo semejante a las que ya reconoce, no la reconocerá como una planta [3]

3.4 problema de predicción de series temporales:
este problema se basa en que las respuestas se basan en los datos "x" sin embargo, puede haber datos independientes que no tengan relación con el resultado, un ejemplo claro de este sería el salario de un trabajador y el peso de una persona.
sin embargo, el programa lee toda la información teniendo en cuenta estos datos independientes, generando un error de predicción conforme a la relación de los datos ya que este tiene en cuenta todos los datos entregados
Solución:
usar el método de clasificación, el cual consiste en leer todos los datos y seleccionar los datos no independientes enviarlos al programa de predicción para obtener un resultado más cercano al deseado.[4]

4. Lista de listas

La estructura de datos, es una lista de listas, es decir, que en cada espacio de la lista hay una persona y por consiguiente todos sus datos, de esta forma se puede acceder de manera sencilla a la información, pudiendo traspasar dicha información a arreglos o strings simples en donde se puede desglosar cada "sub-dato" para agregarlo a el algoritmo de predicción.

Figura 1: Representacion grafica de lista simple.

4.1

4.2 Criterios de Diseño

Decidimos elegir esta estructura de datos por que a nuestro parecer es la forma menos complicada de solucionar el problema que estamos tratando, y no sólo "sencillo" en el código sino que a la hora de la complejidad algorítmica ayuda bastante, claro está que para algunas operaciones era mejor utilizar una estructura distinta pero esta no está del todo mal y además ya nos habíamos familiarizado con una parte de esta estructura en el curso de Fundamentos de Programación lo que a la hora del entendimiento y elaboración de dicha estructura nos facilitó en gran manera las cosas.

4.3Complejidad

Operación	Complejidad
Acceder	O(n)
Buscar	O(n)
Leer	O(n)

Figura 2: Complejidad algorítmica hallada para el peor de los casos en una lista simplemente enlazada.

4.4 Tiempo

operación	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3
lectura (S)	281	415	550
busqueda (S)	268	460	467
acceder (S)	268	460	467

Figura 3: Tiempo que tarda en realizar cada operación respectivamente considerando el peor de los casos

4.5 Memoria

```
iniciando lectura de datos
los datos han sido guardados con exito
memoria consumida: 33554432
iniciando lectura de datos
los datos han sido guardados con exito
memoria consumida: 52953088
iniciando lectura de datos
los datos han sido guardados con exito
memoria consumida: 83361792
```

Figura 4: Memoria consumida por el programa para los 3 conjuntos de datos respectivamente.

numero de datos	5000	15000	25000
consumo de memoria (Bytes)	33554432	51904512	83886080

Figura 5: Tabla de la memoria que ocupa el programa según el número de datos que recibe o lee.

4.6 Análisis de Resultados

REFERENCIAS

1. Claudia X. Mazo, Oscar F. Bedoya. PESPAD: una nueva herramienta para la predicción de la estructura secundaria de la proteína basada en árboles de decisión. Ingeniería y Competitividad, 12 (2). 9-22.
2. Sergio Calderón Pérez-Lozao. 2017. Predicción de resultados deportivos con técnicas de Machine Learning aplicado al fútbol. Reporte Final. Universidad Carlos III de Madrid, Madrid.
3. Na8, Que es overfitting y underfitting y cómo solucionarlo, diciembre 12 de 2017, Aprende

Machine Learning antes de que sea demasiado tarde, https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

4. Álvaro G., Problemas Comunes En Aprendizaje automático, 10 de Mayo de 2018, https://machinelearningparatodos.com/problemas-comunes-en-aprendizaje-automatico/