UNIVERSIDAD ICESI

Método de la Ingeniería
Aplicación para la Solución de un Problema de Machine Learning

-

Integrantes:
Julián Andrés Rivera
Juan Camilo Guerra
Diego Andrés Torres
Gustavo Adolfo Villada

Contexto Problemático

El ministerio de educación de Colombia se encuentra en un estado de preocupación debido a los bajos desempeños académicos que se presentan en los estudiantes de secundaria de ciertas poblaciones del país, por lo tanto, ha iniciado un proyecto de refuerzo en el cual se plantea subir el nivel educativo en las instituciones públicas para básica secundaria, para comenzar a realizar este proyecto se optó por la recolección de datos en diferentes instituciones, con la finalidad de crear una base de datos para que en la segunda etapa del proyecto se puedan analizar.

El ministerio de educación en este momento ya se encuentra en la segunda etapa del proyecto, por lo tanto se tiene la responsabilidad de lograr identificar las falencias y/o factores que afectan el desempeño educativo de los estudiantes.

Paso 1. Identificación del problema.

Para resolver la situación anterior se desarrollarán herramientas de analítica, inteligencia artificial y machine learning para lograr responder a las preguntas de interés que puedan surgir a raíz de cómo mejorar el desempeño de los estudiantes.

En esta solución del problema se va a utilizar un conjunto de datos obtenidos de una base de datos abierta (https://www.kaggle.com/spscientist/students-performance-in-exams). El cual está formado por 8 columnas y 100 filas. Las variables de estudio son las correspondientes a las notas de exámenes que se quieren estudiar tomando en cuentas diversos factores como género o institución educativa (factores que se encuentran en el dataset)

Debe realizarse un software en C# sobre .NET permita ofrecer una interfaz amigable con el usuario que pueda cumplir con los diferentes requerimientos.

Paso 2. Recopilación de Información.

Con el objetivo de tener total claridad en los conceptos involucrados se hace una búsqueda de las definiciones de los términos que se encuentran más relacionados con el problema a solucionar. Es importante lograr realizar esta búsqueda principalmente en fuentes

reconocidas y confiables para conocer cuáles elementos hacen parte del problema y cuáles no, teniendo cierta confianza de lo que se haya investigado.

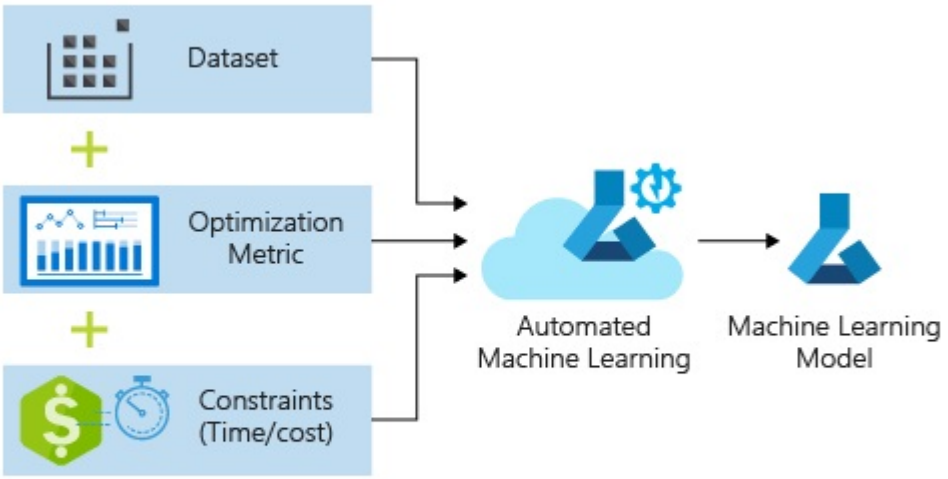
Fuentes:
https://es.wikipedia.org/wiki/Conjunto_de_datos
https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1
https://es.wikipedia.org/wiki/Red_neuronal_artificial

- **Conjunto de datos:** Es una colección de datos habitualmente tabulada, un conjunto de datos contiene los valores para cada una de las variables, como por ejemplo la altura y el peso de un objeto, que corresponden a cada miembro del conjunto de datos. Cada uno de estos valores se conoce con el nombre de dato. El conjunto de datos puede incluir datos para uno o más miembros en función de su número de filas.

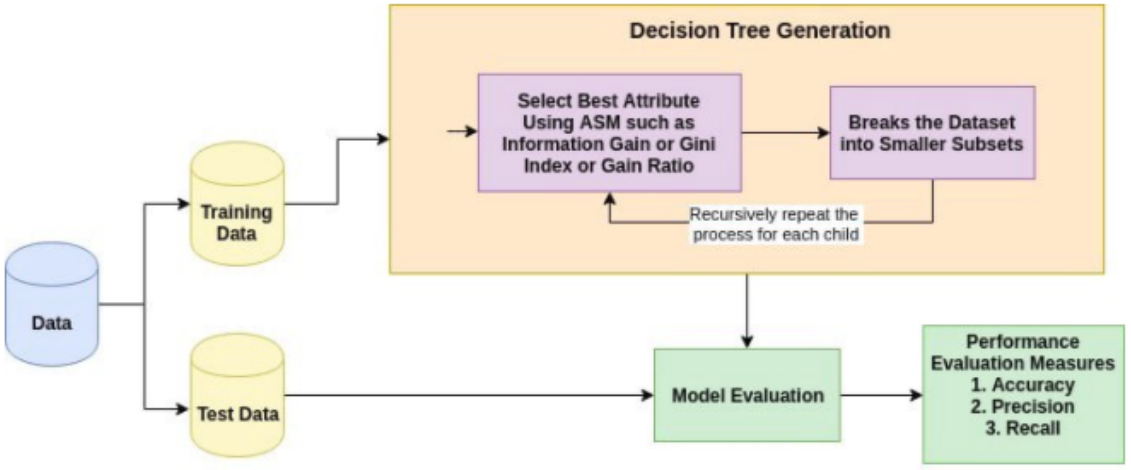
gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75
female	group B	associate's degree	standard	none	71	83	78
female	group B	some college	standard	completed	88	95	92
male	group B	some college	free/reduced	none	40	43	39
male	group D	high school	free/reduced	completed	64	64	67

Dataset con información de estudiantes.

- **Machine Learning**: Es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras puedan predecir ciertos elementos de acuerdo a información suministrada..



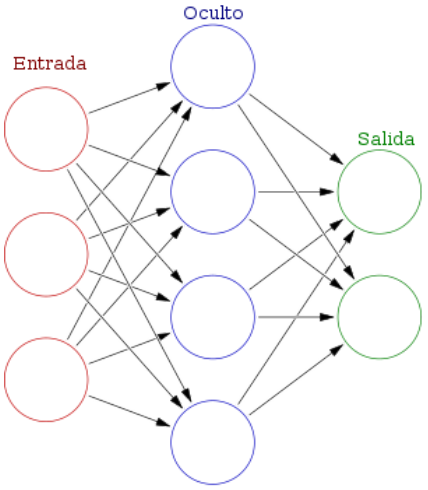
- **Árbol de desición:** Un árbol de decisión es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.



Funcionamiento de un Arbol de decisión en machine learning.

- **Redes Neuronales Artificiales:** son un modelo computacional el que fue evolucionando a partir de diversas aportaciones científicas que están registradas en la historia.1 Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida.

  Una red neuronal artificial es un grupo interconectado de nodos similar a la vasta red de neuronas en un cerebro biológico. Cada nodo circular representa una neurona artificial y cada flecha representa una conexión desde la salida de una neurona a la entrada de otra.



- **Redes Bayesianas:** Formalmente, las redes bayesianas son grafos dirigidos acíclicos cuyos nodos representan variables aleatorias en el sentido de Bayes: las mismas pueden ser cantidades observables, variables latentes, parámetros desconocidos o hipótesis. Las aristas representan dependencias condicionales; los nodos que no se encuentran conectados representan variables las cuales son condicionalmente independientes de las otras. Cada nodo tiene asociado una función de probabilidad que toma como entrada un conjunto particular de valores de las variables padres del nodo y devuelve la probabilidad de la variable representada por el nodo. Por ejemplo, si por padres son m variables booleanas entonces la función de probabilidad puede ser representada por una tabla de 2^m entradas, una entrada para cada una de las 2^m posibles combinaciones de los padres siendo verdadero o falso. Ideas similares pueden ser aplicadas a grafos no dirigidos, y posiblemente cíclicos; como son las llamadas redes de Markov.

Por ejemplo, si tuviéramos que predecir el movimiento de un tigre devorador de hombres a través de algunas aldeas del Himalaya que casualmente se encuentran en el límite de alguna reserva de tigres, podríamos modelarlo con cualquiera de los dos enfoques siguientes:

En un árbol de decisión, nos basaríamos en las estimaciones de los expertos para saber si un tigre, dada la posibilidad de elegir entre campos abiertos o ríos, se decantaría por estos últimos. En una **red bayesiana**, rastreamos al tigre por las marcas de carabina, pero razonamos de manera que se reconozca que estas marcas de carabina podrían haber sido las de algún otro tigre de tamaño similar que patrullara rutinariamente su territorio. Si utilizamos una **red neuronal**, tendríamos que entrenar el modelo repetidamente utilizando diversas peculiaridades de comportamiento del tigre en general, como su preferencia por nadar, su preferencia por las zonas cubiertas frente a las abiertas, su evitación de las viviendas humanas, para que la red pueda razonar de forma general sobre el rumbo que podría seguir el tigre.

Fuente:
https://www.i-ciencias.com/pregunta/56589/diferencia-entre-la-red-de-bayes-la-red-neuronal-las-redes-de-petri-y-el-arbol-de-decision

Paso 3. Búsqueda de Soluciones Creativas

Mediante una lluvia de ideas, se encontraron diferentes alternativas para satisfacer ciertas necesidades al momento de realizar las funcionalidades del software.

Para lograr dar una respuesta confiable mediante machine learning se obtuvieron las siguientes técnicas para clasificar la información:

- ***Alternativa 1 -Árboles de decisión:***

  La primera alternativa es utilizando árboles de decisión, tenemos una muestra de 1000 estudiantes y queremos identificar factores que afecten las notas de dichos individuos.

  Ahora, podríamos crear un modelo para predecir cuáles estudiantes tendrán bajas notas, de acuerdo al nivel de estudio de los padres, si tienen una buena alimentación o si realizaron el test de prueba previo al curso en el que se encuentran.

- ***Alternativa 2 – Redes neuronales artificiales:***

  Otra de las alternativas de técnicas de machine learning es utilizando redes neuronales artificiales para lograr crear un modelo para identificar y predecir qué nuevos estudiantes pueden estar en riesgo de sacar malas calificaciones, luego de utilizar una muestra de 1000 estudiantes.

- ***Alternativa 3 – Redes Bayesianas:***

  Siguiendo con las alternativas para la solución del problema podemos utilizar la técnica de redes bayesianas, para crear una red bayesiana donde se logren identificar en el grafo aquellas variables que tengan cierta relación entre sí, y así poder determinar los factores que puedan ayudar a predecir los estudiantes que puedan sacar malas calificaciones.

Paso 4. Transición de las Ideas a los Diseños Preliminares

Lo primero que hacemos en este paso es descartar las ideas que no son factibles. En este sentido ***descartamos la Alternativa 3 (Redes Bayesianas)*** debido a que en las redes bayesianas, la decisión se basa en la distribución de "pruebas" que apuntan a que un suceso ha ocurrido, en lugar de la observación directa del propio suceso, algo super importante para la optimización de las predicciones.

La revisión cuidadosa de las otras alternativas nos conduce a lo siguiente:

Alternativa 1.Árboles de decisión.

El árbol de decisión es de nuevo una red, Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

Cada nodo tiene más inteligencia que la red neuronal y la ramificación puede decidirse mediante evaluaciones matemáticas o probabilísticas.

Las decisiones son evaluaciones directas basadas en distribuciones de frecuencia de eventos probables, donde la decisión es probabilística.

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

Primer nodo o nodo raíz: en él se produce la primera división en función de la variable más importante.
Nodos internos o intermedios: tras la primera división encontramos estos nodos, que vuelven a dividir el conjunto de datos en función de las variables.
Nodos terminales u hojas: se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.

Alternativa 2.Redes neuronales artificiales

Las redes bayesianas modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas

Paso 5. Evaluación y Selección de la Mejor Solución

Criterios
Deben definirse los criterios que permitirán evaluar las alternativas de solución y con base en este resultado elegir la solución que mejor satisface las necesidades del problema planteado. Los criterios que escogimos en este caso son los que enumeramos a continuación. Al lado de cada uno se ha establecido un valor numérico con el objetivo de establecer un peso que indique cuáles de los valores posibles de cada criterio tienen más peso.

Criterio A. Precisión de la solución. La alternativa entrega una solución:
[2] Exacta (se prefiere una solución exacta)
[1] Aproximada

Criterio B. Eficiencia. Se prefiere una solución con mejor eficiencia que las otras consideradas. La eficiencia puede ser:

[4] Excelente
[3] Buena
[2] Regular
[1] Mala

Criterio C. Completitud. Se prefiere una solución que encuentre todas las soluciones. Cuántas soluciones entrega:

[3] Todas
[2] Más de una si las hay, aunque no todas
[1] Sólo una o ninguna

Criterio D. Facilidad en implementación:

[2] Fácil
[1] Difícil

Evaluación

Evaluando los criterios anteriores en las alternativas que se mantienen, obtenemos la siguiente tabla:

	Criterio A	**Criterio B**	**Criterio C**	**Criterio D**	**Total**
Alternativa 1. Árboles de Decisión	2	4	2	4	12
Alternativa 2. Redes Neuronales Artificiales	2	3	2	1	8

Selección

De acuerdo con la evaluación anterior se debe seleccionar la Alternativa 1, ya que obtuvo la mayor puntuación de acuerdo con los criterios definidos. Se debe tener en cuenta que hay que hacer un manejo adecuado del criterio en el cual la alternativa fue peor evaluada que la otra alternativa.