Tecnológico de Monterrey

Árboles de clasificación y regresión.

–

Minería de Datos

Mónica González - A01735626
Mariana Rico - A01735770
Alberto Muro - A01734046

Introducción

El conjunto de datos en cuestión fue proporcionada por el profesor, obtenida de la plataforma Kaggle, en donde se pudo encontrar un poco más de información acerca del origen de la misma. En un vistazo, este conjunto de datos proviene del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, y tiene como objetivo predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en ciertas mediciones de diagnóstico incluidas en el conjunto de datos.

Se impusieron varias restricciones en la selección de estas instancias, extraídas de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de edad de ascendencia india “Pima”.

La base de datos a continuación consta de varias variables predictivas médicas, tales como el número de embarazos que ha tenido la paciente, su IMC, nivel de insulina, edad, etc., mientras que consta de una sola variable objetivo, el resultado, en donde “1” es positivo a diabetes y “0” es negativo. La presente base de datos consta de 9 variables y 768 datos o filas.

	Embarazos	Glucosa	Presion	Piel	Prueba	BMI	DPF	Edad	Clase
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Explicación del método

En el análisis del modelo predictivo, abordaremos la metodología de árboles de decisión. En primer lugar, utilizaremos el modelo de clasificación, cuyo objetivo es realizar particiones binarias de la información. Este proceso se inicia con un nodo raíz, y a medida que avanzamos, la base de datos se divide mediante condicionales. Al final, evaluaremos el

nivel de homogeneidad de cada nodo para determinar en qué medida el análisis predictivo se ajusta a los datos analizados.

Por otro lado, también exploraremos la metodología del árbol de regresión. Este enfoque tiene como objetivo predecir el comportamiento de los datos con respecto a la variable respuesta, utilizando las variables independientes seleccionadas. El método de desarrollo es similar al mencionado anteriormente: se crean nodos y hojas que buscan determinar cuándo una persona podría tener diabetes y cuándo no.

Pasos Realizados y los Resultados Obtenidos

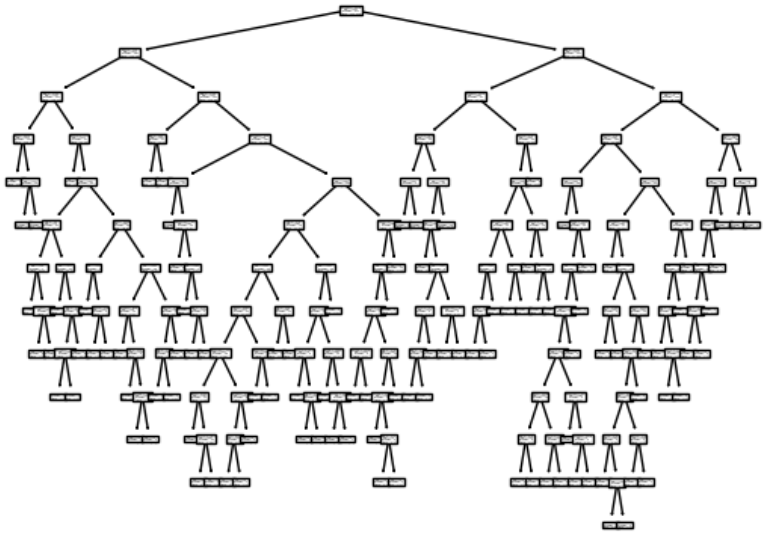
Árbol de Clasificación

Para la resolución del caso comenzamos elaborando la prueba de árboles de clasificación, en este tomamos como variable respuesta ***Clase*** y como independientes el resto de las variables dentro del conjunto de datos.

El primer paso será realizar el cálculo del índice Gini, el cual nos indicará la homogeneidad del modelo en partición binaria. El valor inicial obtenido fue de **0.45437282986111116**. Posterior a esto, partimos del modelo en entrenamiento y prueba (relación 80|20) y al calcularlo obtuvimos los siguientes resultados:

```
Exactitud: 0.7467532467532467
Precisión 0.7557102785013023
```

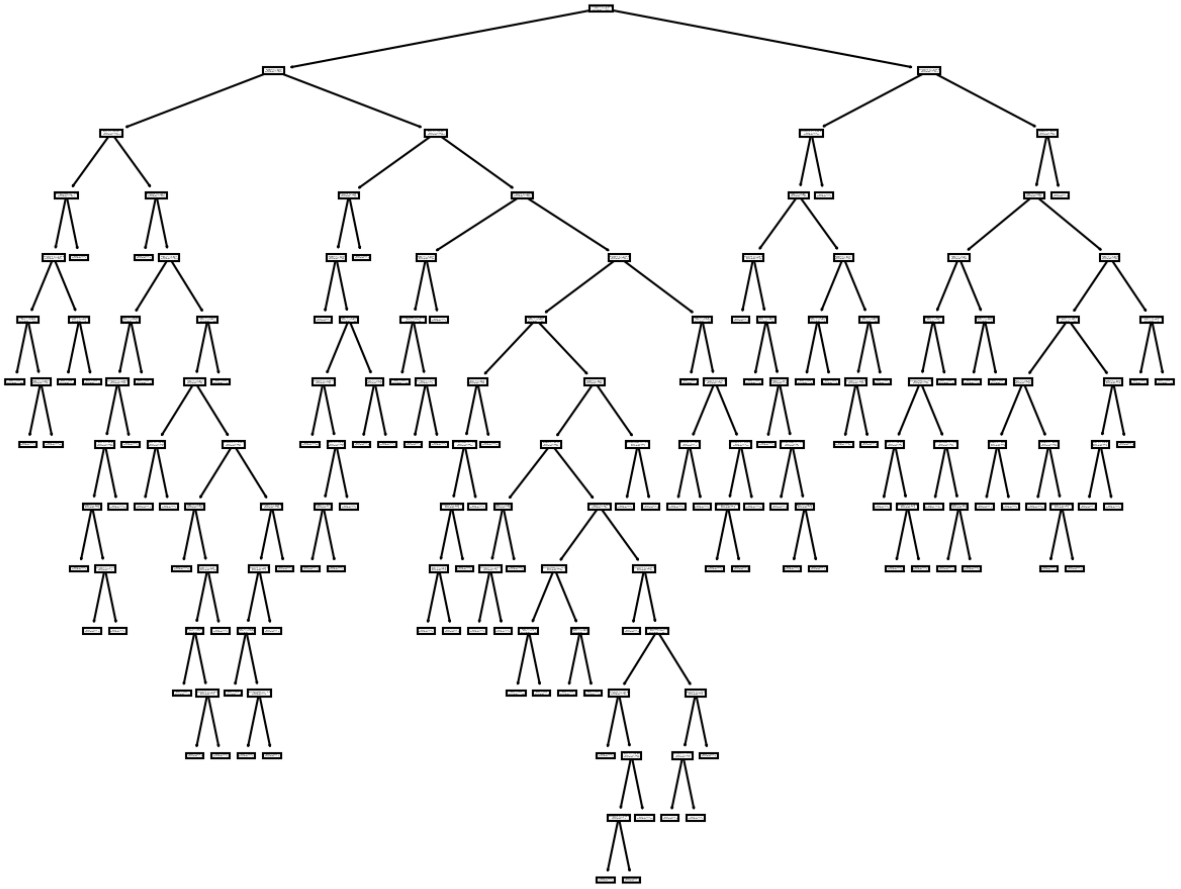
Este resultado podría proporcionarnos un indicio positivo de éxito. La exactitud nos ofrece la tasa de éxito con la que el árbol de decisión ha clasificado los casos positivos en el modelo de entrenamiento, mientras que el valor de precisión se refiere a la capacidad del modelo para detectar correctamente los casos en el modelo de prueba.



Al examinar la representación gráfica del modelo de árbol de clasificación, notamos que es complicado de explicar. No se aprecian claramente el índice Gini de los nodos, las variables más influyentes del modelo ni la profundidad del mismo (que suponemos podría ser de 12). Ante esta dificultad, hemos optado por llevar a cabo un segundo análisis con el objetivo de mejorar estos resultados.

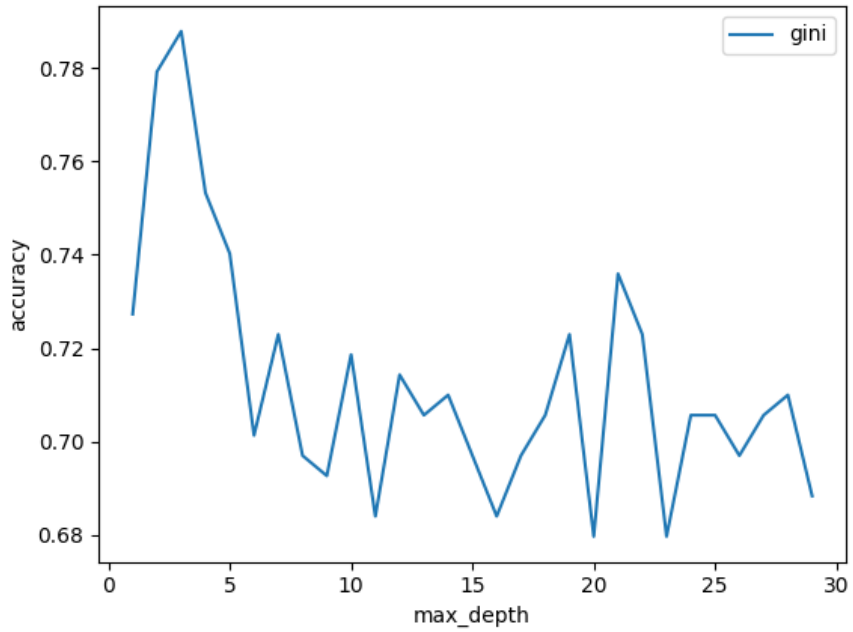
Árbol de regresión.

En la implementación de este análisis, nuevamente empleamos las mismas variables y dividimos la base de datos en dos conjuntos, uno para entrenamiento y otro para prueba, utilizando una proporción de 80|20. Como resultado inicial, obtenemos un valor de **0.7012987012987013**, que se aproxima significativamente al valor obtenido en nuestro árbol de clasificación anterior. Con base en este resultado cercano, procedemos a representar gráficamente el modelo para una mejor comprensión.

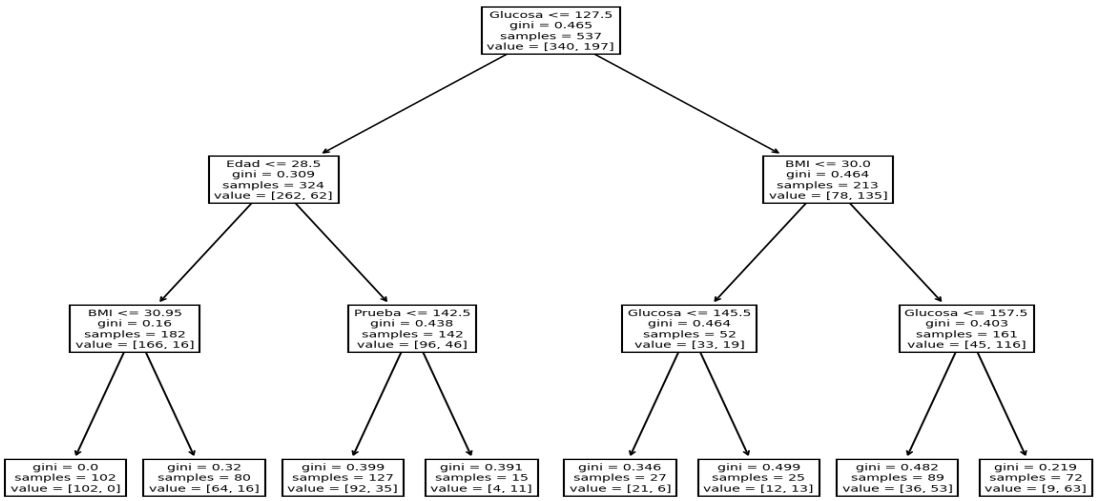


Similar al ejemplo anterior, el análisis genera un modelo de interpretación bastante compleja. En el caso del árbol de regresión, contamos con la técnica de "poda". Con esta técnica, podemos decidir si queremos definir el número de variables consideradas, los nodos que deberían ramificarse o la profundidad que debería tener el árbol. En este caso y basándonos en lo aprendido en clase, hemos optado por la metodología de poda de árbol, centrándonos en el nivel de profundidad. Para ello, ejecutaremos un modelo donde

variaremos constantemente el nivel de profundidad en una escala de 1 a 30. Este modelo registrará el índice de Gini de cada nivel y generará un gráfico con los resultados obtenidos, que se presentan a continuación: [Insertar resultados aquí].



En el caso del diagrama de árbol de regresión, el modelo busca la mayor homogeneidad entre cada partición, lo que lleva a alcanzar una profundidad muy elevada. En el gráfico anterior, se observa cómo el valor de precisión oscila entre 0.80 y 0.70, y el eje x indica el valor obtenido en función de la profundidad desarrollada por el modelo. De este modo, podemos determinar que el pico más alto del modelo se alcanza cuando la profundidad es de 3. Con esta información relevante, procedemos a presentar la representación gráfica correspondiente.



En este gráfico, se observa detalladamente que la partición en el nodo raíz utiliza la variable ***Glucosa.*** En el lado izquierdo, encontramos tres nodos que contienen las variables ***Edad***, ***BMI*** y ***Prueba***. En el lado derecho, hay otros tres nodos que comprenden las variables ***BMI*** y ***Glucosa***.

Conclusiones y Recomendaciones

Comparar dos modelos con la misma base de datos nos brinda la oportunidad de realizar un análisis más completo del caso. Por un lado, el árbol de clasificación exhibe valores más altos de exactitud y precisión, pero resulta difícil de explicar al observar las ramificaciones. No es sencillo distinguir las variables más relevantes, los nodos creados o incluso la profundidad del árbol.

Por otro lado, el árbol de regresión proporciona una exactitud menor, pero al aplicar la técnica de poda de árbol con una profundidad de 3, obtenemos un modelo deseado con una exactitud del 0.788. Este modelo destaca las particiones binarias de manera clara, permitiéndonos observar los componentes de los nodos y los niveles Gini en cada uno de ellos.

En conclusión, el árbol de regresión utilizando la técnica de "poda" con una profundidad de 3 es el modelo preferido. Presenta un nivel de precisión cercano a 1 y permite apreciar claramente los componentes con los que se ha formado.