Módulo 2 Análisis y Reporte sobre el desempeño del modelo.

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Estado de México

Inteligencia artificial avanzada para la ciencia de datos I

Alumno:

Joel Erick Martínez Espinosa
A01377945

Tecnológico de Monterrey

Introducción

La diabetes es una enfermedad crónica que afecta la forma en que el cuerpo humano utiliza la energía de los alimentos. En los últimos 20 años, el número de personas diagnosticadas con diabetes se ha duplicado.

La diabetes descompone la mayoría de los alimentos en glucosa, que luego se libera al torrente sanguíneo. En las personas con diabetes, el cuerpo no produce suficiente insulina, la hormona que convierte la glucosa en energía. Esto hace que los niveles de azúcar en sangre aumenten, lo que con el tiempo puede provocar problemas de salud graves, como enfermedades cardíacas, renales y pérdida de la visión. Hay diferentes tipos de diabetes:

Diabetes tipo 1:
Este tipo de diabetes es causada por una reacción autoinmune que impide que el cuerpo produzca insulina. Esto representa aproximadamente del 5 al 10% de todos los casos diagnosticados.

Diabetes tipo 2:
En este caso, el cuerpo no utiliza la insulina de forma eficaz y los niveles de azúcar en sangre aumentan de forma anormal. Esta es la forma más común de diabetes y afecta entre el 90 y el 95% de las personas diagnosticadas.

Diabetes gestacional:
Esta forma de diabetes ocurre en mujeres embarazadas debido a cambios hormonales en el cuerpo. Suele desaparecer después del nacimiento.

Prediabetes:
En esta afección, los niveles de azúcar en sangre son más altos de lo normal, pero no lo suficientemente altos como para diagnosticarlos como diabetes. Se estima que 96 millones de personas padecen diabetes y el 80% de ellas desconoce su afección.

Aunque actualmente no existe cura para la diabetes, existen varios métodos de diagnóstico que evalúan varias variables para determinar el tipo de diabetes que puede estar afectando al paciente. Estas variables incluyen, entre otras, obesidad, edad, antecedentes familiares, actividad física, antecedentes de diabetes gestacional y origen étnico como afroamericano, latino o asiático. Además, los síntomas comunes de la diabetes incluyen pérdida de peso involuntaria, visión borrosa, micción frecuente, hormigueo en manos y pies, cicatrización lenta de heridas, aumento del apetito, sed

constante, fatiga y piel. Esto incluye la sequía. Estos síntomas varían según el tipo de diabetes que tenga.

Si bien ningún método de diagnóstico es 100% preciso, los enfoques basados en datos que tienen en cuenta factores antropométricos y demográficos pueden ayudarnos a comprender mejor por qué las personas desarrollan esta afección.

Justificación del dataset y los modelos de ML

Al usar un dataset de una enfermedad tan común contamos con una buena cantidad de datos así mismo como con variables a analizar muy interesantes que nos ayudarán a entender mejor el comportamiento de esta misma y saber el porqué las personas pueden llegar a padecerlas, por lo que enlista las razones del uso de este dataset.

Diferentes variables: El conjunto de datos contiene diferentes variables considerando diferentes aspectos relacionados con la diabetes, tales como: B. Azúcar en sangre (glucosa), presión arterial (pb), grosor del pliegue cutáneo (piel), insulina (insulina) e índice de masa corporal (``IMC''), predisposición genética (``pedigree'') y edad (` `edad''). Esta variedad de variables proporciona una visión general completa de los factores que pueden influir en la diabetes.

Objetivo de clasificación: la variable Etiqueta indica que el objetivo del análisis es la clasificación. H. Predecir si una persona tiene diabetes. Esto es adecuado para aplicar técnicas como árboles de decisión y ANN, que son comunes en problemas de clasificación.

Posibilidad de descubrir relaciones: las variables de un conjunto de datos se pueden relacionar entre sí de formas complejas. Por ejemplo, se sabe que el azúcar en sangre, el IMC y la predisposición genética pueden desempeñar un papel en el desarrollo de la diabetes. Estas relaciones pueden explorarse y descubrirse mediante un análisis adecuado.

Buena dimensionalidad: Nueve variables le dan al conjunto de datos una buena cantidad de dimensiones. Esto facilita la aplicación de técnicas de reducción de dimensionalidad como el uso de la entropía para simplificar el análisis si es necesario.

Tecnológico de Monterrey

Datos de salud relevantes: las variables del conjunto de datos están relacionadas con la salud, por lo que este conjunto de datos es relevante para el tratamiento de problemas médicos importantes como la diabetes.

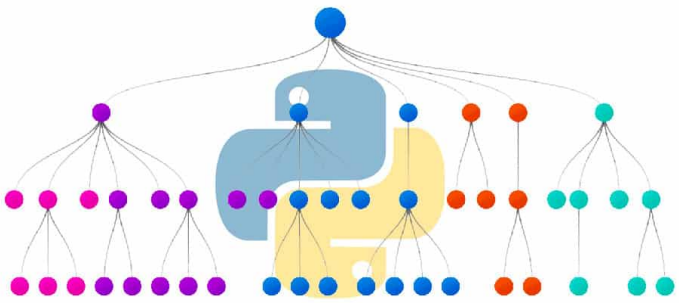
Tamaño del conjunto de datos: aunque no especificamos el tamaño exacto del conjunto de datos, la idoneidad también depende del tamaño. En general, al aplicar técnicas de aprendizaje automático, es deseable tener un conjunto de datos con suficientes observaciones para entrenar y evaluar el modelo de manera efectiva.

Árboles de Decisiones:

- Interpretación de Resultados: Los árboles de decisiones son especialmente útiles cuando deseas comprender cómo se toman decisiones en tu conjunto de datos. Cada nodo del árbol representa una característica y una decisión basada en esa característica, lo que proporciona una interpretación clara y transparente de cómo se realiza la clasificación.

- Selección de Características: Puedes utilizar árboles de decisiones para identificar las características más importantes en tu conjunto de datos. Esto es útil para el análisis de datos y la selección de características para modelos más avanzados.

- Clasificación: al ser el objetivo de la clasificación (la predicción de la presencia o ausencia de diabetes), los árboles de decisiones son adecuados para este propósito.



Desarrollo

Se usaron dos métodos diferentes; un árbol de decisión normal y otro con entropía para poder visualizar y analizar el dataset que nos daban

Antes de utilizar los métodos se realizó un análisis del dataset. Primero, se realizó un análisis univariante en el que obtuvimos la estadística descriptiva (medias, desviación típica, mínimo, máximo y cuartiles)

	pregnant	glucose	bp	skin	insulin	bmi	pedigree	age	diagnostic
count	768	768	768	768	768	768	768	768	768
unique	17	136	47	51	186	248	517	52	2
top	1	99	70	0	0	32	0.258	22	0
freq	135	17	57	227	374	13	6	72	500

Imagen 1.1 (Tabla del análisis descriptivo con python)

Se realizó un análisis multivariado, en el que creamos gráficos de barras para ver cómo las variables "interactúan" o se comportan en relación con otras variables. Esto nos muestra las relaciones entre las variables y determina qué variables son más importantes en este estudio de caso.
También realizamos una Matriz de confusión: La matriz de confusión proporciona una visión detallada de cómo se están clasificando las muestras en diferentes categorías. Puedes calcular el número de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos. Una alta proporción de falsos positivos o falsos negativos puede indicar sesgo en una dirección particular.

```
Confusion Matrix:
 [[117  29]
 [ 42  43]]
```

Imagen 1.2 (Matriz de confusión)

Así mismo sacamos la Precisión, exhaustividad y F1-score: Estas métricas proporcionan una medida más completa del rendimiento del modelo. Puedes calcular la precisión, la exhaustividad y el F1-score para cada clase y observar si hay un desequilibrio significativo en estas métricas.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.76      0.74       146
           1       0.54      0.48      0.51        85

    accuracy                           0.66       231
   macro avg       0.63      0.62      0.62       231
weighted avg       0.65      0.66      0.65       231
```

Imagen 1.3 (Classification report)

De igual manera sacamos la Curva ROC y AUC: La curva ROC (Receiver Operating Characteristic) y el área bajo la curva (AUC) pueden ayudar a evaluar el rendimiento del modelo en problemas de clasificación binaria. Un sesgo en una dirección particular puede afectar la posición de la curva ROC.

```
0.6213134568896053
```

Imagen 1.4 (Área bajo la curva)

Por último hicimos un cross validation que en conjunto con los demás análisis nos ayuda a comprender el bias. La validación cruzada utiliza técnicas de validación cruzada para evaluar el rendimiento del modelo en múltiples divisiones del conjunto de datos. Esto te dará una idea más robusta del sesgo del modelo.

```
Cross-Validation Scores: [0.74025974 0.69480519 0.66233766 0.77777778 0.76470588]
```

Imagen 1.5 (Resultados del cross-validation con 5 divisiones)

Antes de hacer nuestro primer modelo del árbol de decisión se hizo un análisis de nuestros datasets; tanto para el train como para el test con 30% de test, el cual nos permite observar si existe overfitting, underfitting o el modelo tiene un buen ajuste. esto con la finalidad de poder mejorar el modelo más adelante; en este caso lo hicimos con

4 niveles de profundidad; 1, 3, 5, 7 y 10. Los cuales en este caso se podría estar en overfitting.

```
Max Depth: 1
Training Accuracy: 0.76
Testing Accuracy: 0.74
El modelo podría estar overfitting.
Max Depth: 3
Training Accuracy: 0.78
Testing Accuracy: 0.74
El modelo podría estar overfitting.
Max Depth: 5
Training Accuracy: 0.83
Testing Accuracy: 0.68
El modelo podría estar overfitting.
Max Depth: 7
Training Accuracy: 0.90
Testing Accuracy: 0.69
El modelo podría estar overfitting.
Max Depth: 10
Training Accuracy: 0.97
Testing Accuracy: 0.70
El modelo podría estar overfitting.
```

Imagen 1.6 (Análisis de nuestros datasets para evaluar su desempeño con el modelo)

El primer método utilizado fue el Árbol de decisión que nos ayudó a simplificar la complejidad de los espacios muestrales con muchas dimensiones a la vez y conservar la información, esto fue de bastante utilidad ya que al ser una base de datos relativamente grande pudimos reducirla sin perder información, por ende nos dio resultados más precisos y con un procesamiento de los datos más veloz y menos carga de trabajo para nuestro entorno de programación, siendo un método muy útil para aplicarlo antes de hacer algún método estadístico. tambien nos ayudó a relacionar si se diagnosticaba diabetes y si dependia del sexo del paciente con los graficos generados.

También se generó un modelo pero con entropía distinta, esto con el objetivo de realizar un análisis que ayuda a determinar cómo dividir los datos de manera óptima para lograr subconjuntos más puros y, en última instancia, construir un árbol de decisión que sea capaz de tomar decisiones de clasificación ya que hace más exhaustiva la clasificación por que el objetivo es dividir el conjunto de datos en subconjuntos más puros y homogéneos en términos de la variable objetivo que en este análisis fue el dar el diagnóstico.

Análisis y Resultados

Análisis univariable y gráficas de barras.

Se obtuvieron varías gráficas de barras con el fin de observar el número de pacientes que padecen o no diabetes, se observó que el grupo que no presenta la enfermedad es mayor, por lo que no se tiene una muestra homogénea. Además en esta gráfica se puede ver que la distribución entre las edades no es equitativa dentro de los casos que presentan o no la enfermedad. La desigualdad de casos con diabetes y sin diabetes dentro del dataset, causará mayor dificultad al diseñar un modelo capaz de identificar nuevos pacientes con diabetes.

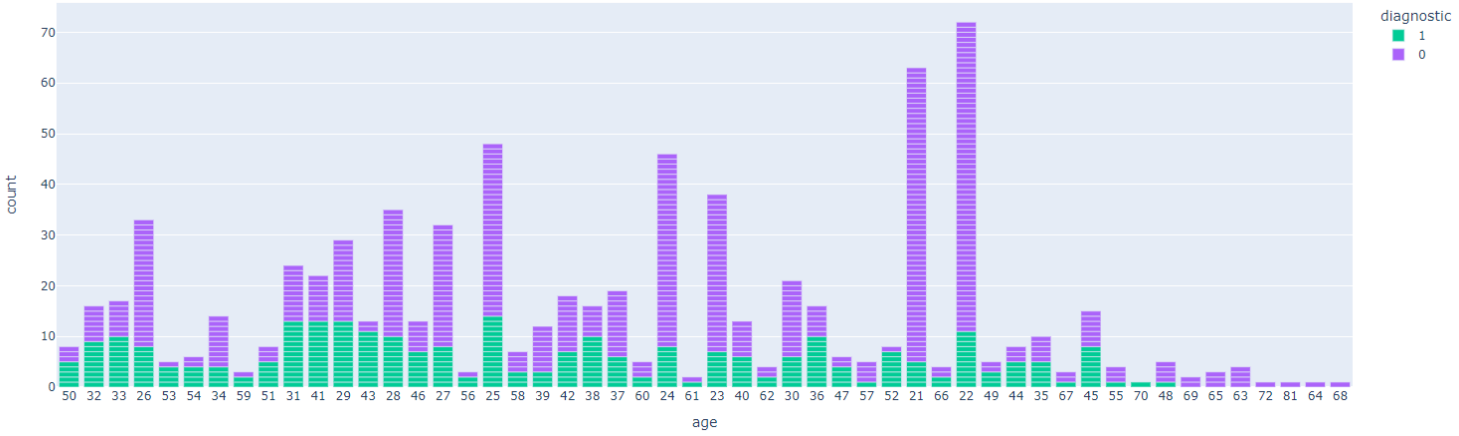


Imagen 1.7 (Gráfica de pacientes No_diabetes y Diabetes por edad)

De igual manera se realizó una gráfica de barras para saber el nivel de glucosa por edad, por lo que pudimos darnos cuenta que los niveles más altos están en el rango de edad de entre 20 a 30 años.

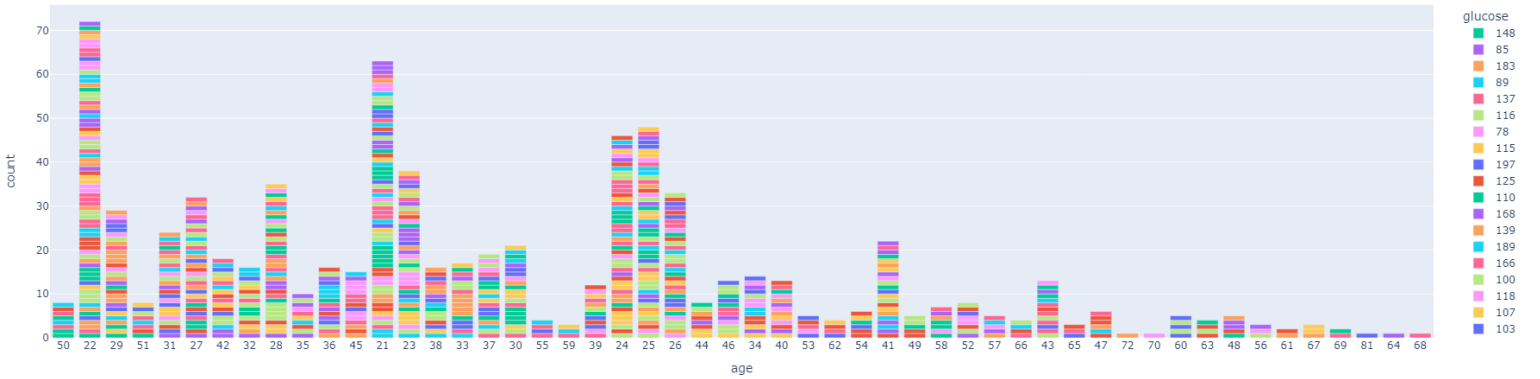


Imagen 1.8 (Gráfica de barras Edad y Nivel de glucosa)

Así mismo Realizamos la gráfica de barras donde podemos observar los casos donde tenían un nivel normal en color turquesa y casos con la insulina en un rango distinto al aceptable de distintos colores de acuerdo al valor de este. Aquí es más notorio el rango de edad en el cúal los niveles no son los idóneos siendo de nuevo el rango de los 20 a 30 años de edad.

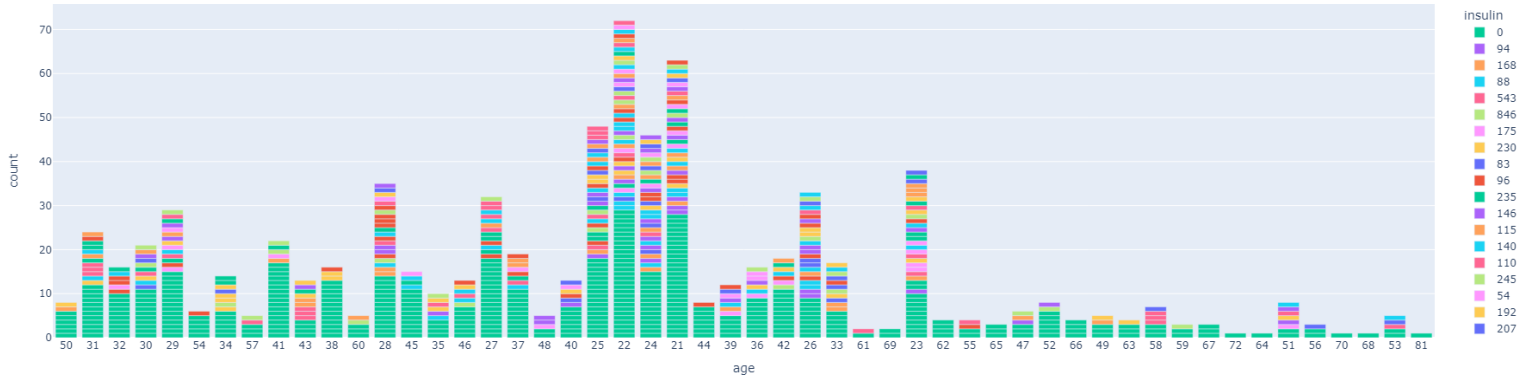


Imagen 1.9 (Gráfica de barras del Nivel de insulina por edad)

A continuación se realizó el dividir los datos para el train y el test para nuestros modelos de machine learning. el cual dividimos los datos en 30% de test y 70% en train para poder después entrenar el modelo de nuestros árboles de decisión.

Datos de entrenamiento:	pregnant	insulin	bmi	age	glucose	bp	pedigree
89	15	110	37.1	43	136	70	0.153
468	0	100	36.8	25	97	64	0.6
551	1	0	27.4	21	116	70	0.204
148	2	119	30.5	34	106	64	1.4
482	0	0	35.2	29	123	88	0.197
..	...	...	...	..	...	..	...
646	2	440	39.4	30	157	74	0.134
716	7	392	33.9	34	187	50	0.826
73	13	0	43.4	42	126	90	0.583
236	4	0	43.6	26	171	72	0.479
38	9	0	32.9	46	102	76	0.665

Imagen 1.10 (Muestra de datos de entrenamiento 70%, de cada variable)

Datos de Testeo:	pregnant	insulin	bmi	age	glucose	bp	pedigree
286	7	135	26	51	136	74	0.647
102	1	0	26.1	22	151	60	0.179
582	6	0	25	27	109	60	0.206
353	3	0	34.4	46	61	82	0.243
727	1	180	36.1	25	116	78	0.496
..	...	...	...	..	...	..	...
242	4	88	33.1	22	91	70	0.446
600	1	120	23.1	26	109	38	0.407
651	1	100	25.2	23	91	54	0.234
12	10	0	38	34	168	74	0.537
215	9	175	34.2	36	112	82	0.26

Imagen 1.11 (Muestra de datos de Testeo 30%, de cada variable)

Creación y entrenamiento del primer árbol de decisión

Para realizar el primer árbol de decisión creamos un objeto al cual le pusimos Decision Tree Classifier, el modelo lo entrenamos con el conjunto de datos de entrenamiento (X_train, y_train) y usamos el fit así mismo hacemos la predicción con el conjunto de datos de testeo Utilizas el modelo entrenado para hacer predicciones en el conjunto de prueba (X_test) y almacenas las predicciones en (y_pred) después calculamos métricas de evaluación, como la precisión, la recuperación y la puntuación F1, utilizando las funciones de metrics.

```
Accuracy: 0.670995670995671
Precision: 0.5569620253164557
Recall: 0.5176470588235295
F1: 0.5365853658536586
```

Imagen 1.12 (Métricas correspondientes al primer modelo del árbol de decisiones)

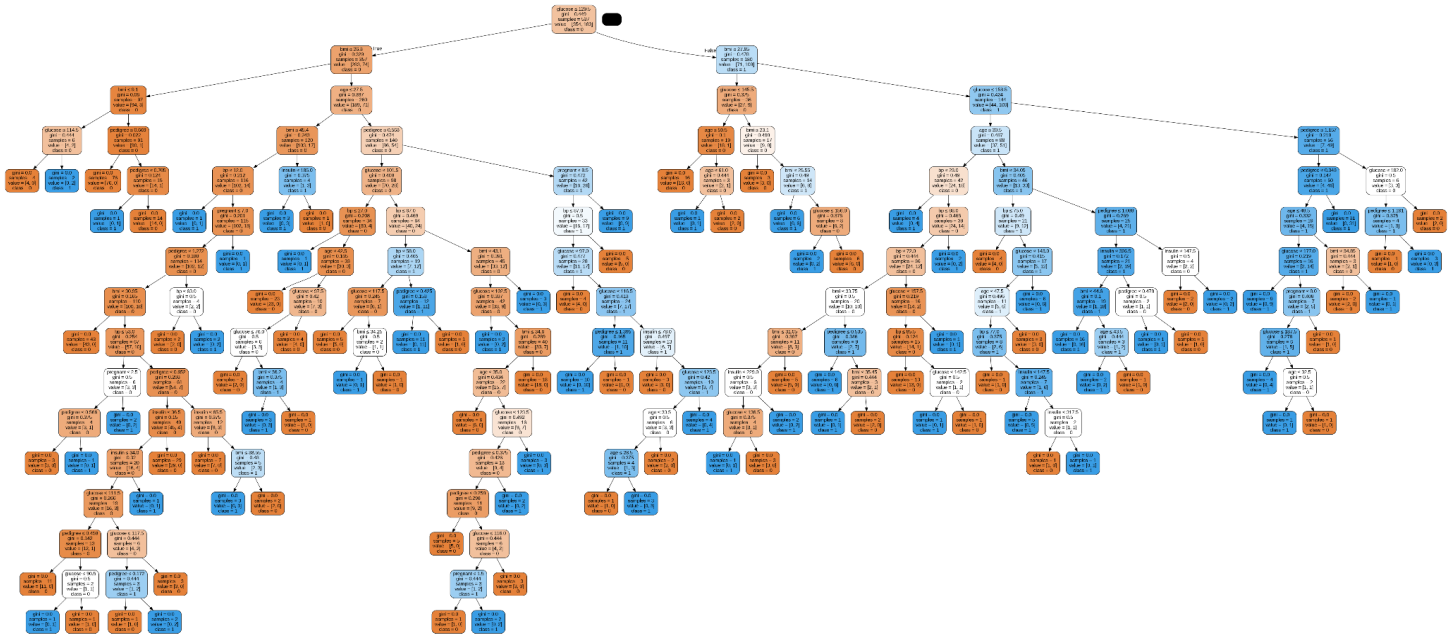


Imagen 1.8 (Imagen del árbol de decisiones del primer modelo.)

Tecnológico de Monterrey

Conclusión del primer modelo

La precisión del modelo es de 0,670995670995671, lo que significa que el modelo predijo correctamente el 67,09% de los datos. La precisión es de 0,5569620253164557, lo que significa que el 55,69% de los pacientes que el modelo predijo que tenían diabetes en realidad la tenían.
En general, una precisión de 0,67 se considera buena. Sin embargo, la precisión y la recuperación son un poco más bajas, lo que significa que el modelo está dando algunos falsos positivos y falsos negativos. Se puede intentar mejorar la precisión y la recuperación del modelo ajustando los hiper parámetros del algoritmo del árbol de decisión.

Se decidió seguir el análisis utilizando de nuevo un árbol de decisión pero en este caso vamos a modificar el valor de la entropía ayudando a mejorar el nivel de impureza de los datos; La entropía mide la impureza en un conjunto de datos. Un conjunto de datos con entropía baja significa que es bastante homogéneo en términos de la variable objetivo, mientras que una entropía alta indica que es heterogéneo. El árbol de decisión busca reducir la entropía en cada división para obtener subconjuntos más puros, así mismo nos ayuda a simplificar los nodos de nuestro árbol, por lo que al buscar la característica y el umbral de división que minimizan la entropía o maximizan la ganancia de información. La idea es encontrar la división que reduzca la impureza o la incertidumbre en la clasificación de los datos.

Creación y entrenamiento del segundo árbol de decisión con distinta entropía..

Se creó un segundo objeto Decision Tree Classifier llamado clf con el criterio de "entropía" y una profundidad máxima de 3. Estrenamos el nuevo modelo de árbol de decisión en el conjunto de entrenamiento (X_train, y_train) utilizando fit.

Claramente mejoraron los resultados de nuestras métricas ya que como lo dedujimos anteriormente al aumentar la entropía y la profundidad mejorando el nivel de impureza y con eso dando mejores métricas,

```
Accuracy: 0.7705627705627706
Precision: 0.7105263157894737
Recall: 0.6352941176470588
F1: 0.6708074534161491
```

Imagen 1.13 (Métricas del segundo Árbol de decisión con entropía y profundidad máxima de 3)

De igual forma el segundo árbol de decisión mejoró al momento de hacer la clasificación simplificando los nodos y siendo mucho más preciso.

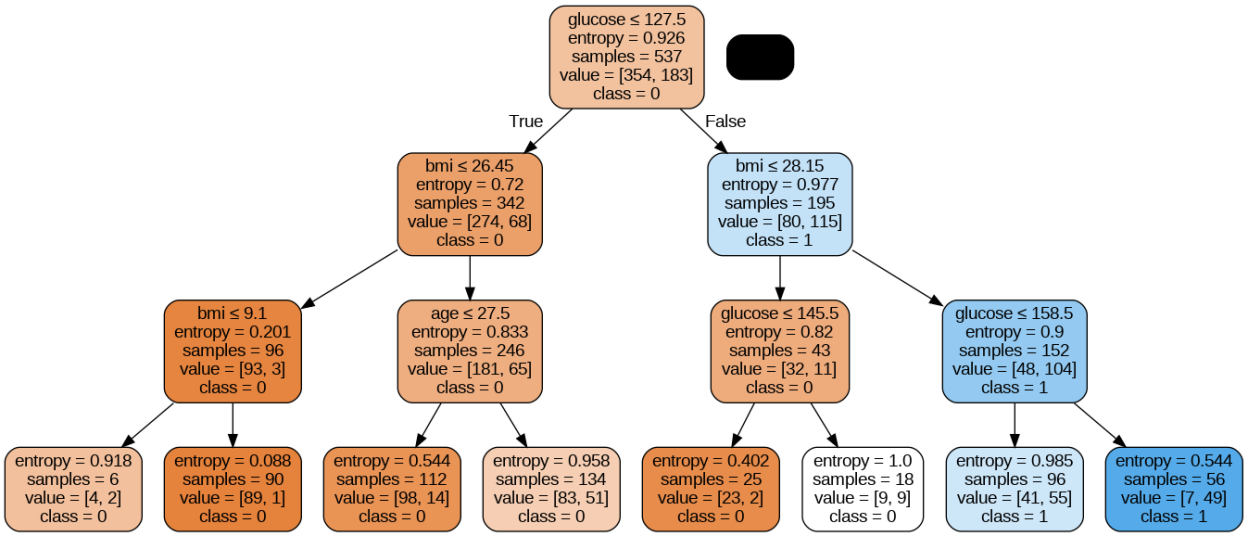


Imagen 1.14 (Gráfica del segundo árbol de decisión con los nodos simplificados y mejor clasificado)

Comparativa de ambos modelos

Primer modelo:

- **Accuracy**: 0.670995670995671: Esto significa que el modelo clasifica correctamente el 67.10% de las muestras.
- **Precision**: 0.5569620253164557: La precisión mide cuántas de las predicciones positivas son realmente positivas. En este caso, el 55.70% de las predicciones positivas son correctas.

- **Recall**: 0.5176470588235295: El recall (o sensibilidad) mide cuántos de los casos positivos reales se capturan correctamente. En este caso, el 51.76% de los casos positivos reales se capturan.
- **F1**: 0.5365853658536586: El valor F1 es una métrica que combina precisión y recall en un solo número. Cuanto más cercano a 1, mejor. En este caso, es 0.54.

Segundo modelo (con entropía y profundidad máxima ajustada):

- **Accuracy**: 0.7705627705627706: Este modelo clasifica correctamente el 77.06% de las muestras, lo que es un aumento significativo con respecto al primer modelo.
- **Precision**: 0.7105263157894737: La precisión mejoró al 71.05%, lo que significa que más de las predicciones positivas son correctas.
- **Recall**: 0.6352941176470588: El recall también mejoró al 63.53%, lo que indica que se capturan más casos positivos reales.
- **F1**: 0.6708074534161491: El valor F1 también aumentó considerablemente a 0.67, lo que indica un equilibrio más sólido entre precisión y recall.

El segundo modelo con entropía y profundidad máxima ajustada tiene un mejor rendimiento en todas las métricas (accuracy, precisión, recall y F1) en comparación con el primer modelo. Esto sugiere que el segundo modelo es una mejora significativa y podría ser la mejor elección para tu problema de clasificación. Sin embargo, siempre es importante evaluar modelos en función de los requisitos específicos de tu aplicación y realizar una validación cruzada adecuada para garantizar que el rendimiento sea robusto.

Por supuesto, aquí tienes una comparativa y una conclusión sobre cómo aumentar la entropía ayudó a mejorar tu modelo de árbol de decisión:

Comparativa

1. Precisión: Aumentar la entropía mejoró la precisión del modelo del 55.70% al 71.05%. Esto significa que en el segundo modelo, una mayor proporción de las predicciones positivas fueron correctas, lo que es crucial en problemas donde la precisión es fundamental.

2. Recall: El recall también aumentó del 51.76% al 63.53% con la mayor entropía. Esto indica que el segundo modelo logra capturar más de los casos positivos reales, lo que es especialmente importante cuando no quieres pasar por alto casos importantes.

3. F1-score: El valor F1 mejoró de 0.54 a 0.67 con la mayor entropía. El F1-score combina precisión y recall, y un valor más alto indica un equilibrio mejor entre ambas métricas. Por lo tanto, el segundo modelo es más equilibrado en términos de precisión y recall.

4. Exactitud (Accuracy): La exactitud también aumentó del 67.10% al 77.06%, lo que significa que el segundo modelo clasifica correctamente más muestras en general.

Conclusión sobre el mejoramiento del modelo

Aumentar la entropía en tu modelo de árbol de decisión claramente tuvo un impacto positivo en su rendimiento. Esto se debe a que una mayor entropía permite que el modelo considere un conjunto más amplio de divisiones de nodos durante la construcción del árbol, lo que puede resultar en un modelo más complejo y capaz de aprender patrones más sutiles en los datos.

En pocas palabras, el aumento de la entropía permitió al modelo ser más flexible y adaptarse mejor a los datos de entrenamiento. Esto se tradujo en una mejora en todas las métricas de evaluación, lo que indica que el segundo modelo es una elección más sólida y efectiva para el problema de clasificación en comparación con el primer modelo con una menor entropía. Sin embargo, es importante recordar que el ajuste de hiper parámetros como la profundidad máxima y la entropía debe realizarse con cuidado y validación cruzada para evitar el sobreajuste a los datos de entrenamiento.

Conclusiones

El análisis de información mediante herramientas de estadística multivariada nos permite identificar relaciones entre variables sin establecer causalidad. Al emplear algoritmos de clasificación, como árboles de decisión, podemos determinar qué variables pueden aumentar la probabilidad de que una persona sea propensa a desarrollar diabetes.

A pesar de que los factores de riesgo pueden variar según el tipo de diabetes y el IMC de la persona, en general, las variables antropométricas y las características sanguíneas requieren una atención especial. Estas variables, en conjunto, pueden revelar hábitos perjudiciales para la salud, como una mala nutrición o el sedentarismo. La glucosa, en particular, desempeña un papel crucial en la determinación de la diabetes.

En términos prácticos, un diagnóstico de diabetes puede realizarse cuando los niveles de glucosa superan los valores normales y están acompañados por otros factores de riesgo, ya sean antropomórficos o relacionados con índices sanguíneos, como el colesterol. Este enfoque combina el poder de los árboles de decisión con el ajuste de la entropía y el análisis univariado junto con gráficas de barras para identificar y comprender las relaciones y factores clave que contribuyen al riesgo de diabetes.

Referencias

¿Qué es la diabetes? | *Información Básica* | *Diabetes* | *CDC*. (s. f.), de https://www.cdc.gov/diabetes/spanish/basics/diabetes.html

Diabetes. (s. f.). OPS/OMS | Organización Panamericana de la Salud, de https://www.paho.org/es/temas/diabetes

Diabetes, F. P. (s. f.-b). *QuÃ© es la diabetes*, de https://www.fundaciondiabetes.org/prevencion/309/que-es-la-diabetes-2

Pennsylvania State University. (s. f.). *Lesson 12: Factor Analysis | STAT 505*. PennState: Statistics Online Courses, de https://online.stat.psu.edu/stat505/lesson/12

Plotly. (n.d.). *Plotly: Low-Code Data App Development*, from https://plotly.com

scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation. (n.d.), from https://scikit-learn.org/stable/

1.10. Decision Trees. (s. f.). scikit-learn, https://scikit-learn.org/stable/modules/tree.html