PRÁCTICA CLASIFICACIÓN SUPERVISADA - Pima Indians Diabetes Database-

Alejandro Ayuso Exposito - b190238

Introducción:

En esta práctica se se va a proceder a la predicción de si un ciudadano de india tendrá o no diabetes teniendo en cuenta una serie de factores como la Edad, el IMC, el Grosor de la Piel, la Glucosa en sangre, en número de Embarazos, la presión Arterial, el uso de Insulina y la predisposición a tener diabetes o función del pedigrí de diabetes de a partir de la historia familiar.

Descripción del problema:

El conjunto de datos en cuestión proviene de un estudio de salud realizado entre la población femenina de ascendencia Pima, de 21 años o más.

La salud de las pacientes se caracteriza a través de distintas variables las cuales incluyen el número de embarazos que ha tenido, los niveles de glucosa en sangre , la presión arterial, el índice de masa corporal (IMC), la función del pedigrí de diabetes, la edad y el uso de insulina. Teniendo en cuenta estas variables, se ha etiquetado a cada paciente con respecto a si ha desarrollado o no diabetes, reflejado en la variable 'Outcome'.

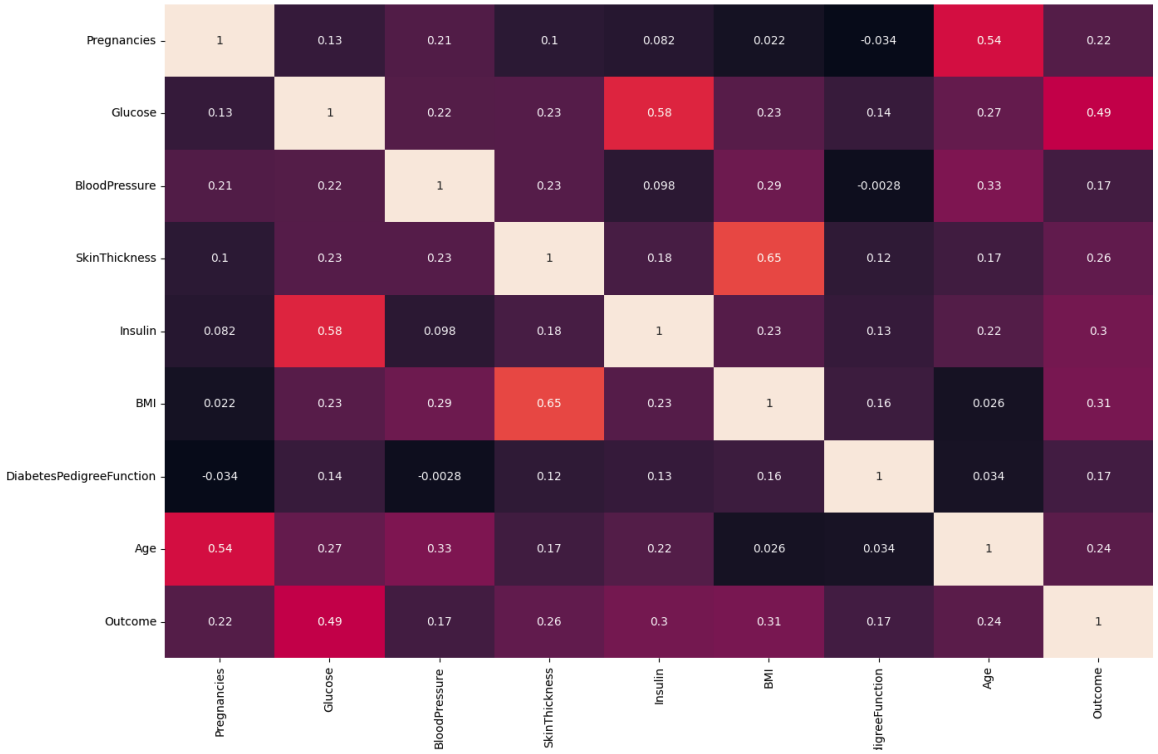
El desafío de este estudio es predecir con precisión la presencia o ausencia de diabetes (la variable 'Outcome') en base a las demás características de salud disponibles. En este contexto, la selección de un subconjunto apropiado de características para utilizar en el modelo de predicción es un paso esencial para asegurar una buena precisión y un modelo interpretativo. Esto requiere de una cuidadosa evaluación y selección de características, así como un riguroso proceso de validación del modelo para asegurar que las predicciones sean robustas y fiables. La elección de las métricas de rendimiento y su interpretación también juegan un papel crucial en la evaluación del rendimiento del modelo.

Metodología:

1. **Preprocesado de los datos:**

   Para poder hacer buen uso de los datos y que se genere un modelo predictivo que tenga una precisión razonable antes de aplicar algún algoritmo hay que realizar un preprocesado de los datos, habría que fijarse en los atributos y seleccionar aquellos que tengan valores nulos o valores que carezcan de sentido, por ejemplo: una persona no puede tener una presión arterial o un nivel de glucosa en sangre de 0. Para solucionar este problema se realiza un script de python el cual revisa las columnas y su contenido y los datos erróneos y los sustituye por un elemento vacío.\

   Se genera una matriz de correlación para ver la multicolinealidad de los datos para ver donde nos podemos esperar que haya alguna relación.



Usaremos el nuevo csv en weka para cambiar todos los elementos vacíos de las celdas por la media en caso de ser atributos numerales o por la moda en caso de ser un atributo nominal.

Finalmente exportamos el nuevo csv para poder tenerlo guardado para poder aplicarle todas las operaciones y algoritmos posteriores.

2. **Predicciones Originales:**

   Como podemos observar aquí no se tiene en cuenta el orden de las variables para predecir el Outcome, no se realiza ningún filtrado, se introducen los datos tal y como se obtienen después del preprocesado y se les aplican los algoritmos de predicción para ver cual de ellos obtiene una mayor precisión.

3. **Selección de Variables:**
   a. **Univariante:**

      El método “InfoGainAttributeEval” nos devuelve en orden las variables que considera mejor para predecir, en este caso:
      [Glucose, Bmi, Age, Insulin, Pregnancies, SkinThickness, DiabetesPedigreeFunction, BloodPressure].

      Este tipo de filtrado tiene una gran cantidad de información en los atributos repetida, y al usar los mismos atributos que en las predicciones originales, aunque en este caso se encuentran ordenados de mayor a menor influencia en el resultado, y tampoco se tiene en cuenta la relación entre las variables, la precisión de los algoritmos de predicción es prácticamente igual, aunque

se puede notar una leve mejora en la precisión de los algoritmos JRIP, ID# J$* y Logistic.

b. Multivariante:

El método "CfsSubsetEval" considera el conjunto completo de atributos, evaluando no solo la importancia de cada atributo individualmente, sino también considerando su relación con los demás, en este caso las variables que se devuelven son:
[Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age].

En este caso el filtrado multivariante se hace para eliminar las características menos importantes y conservar las que tienen más influencia en la predicción. Estas características se usan para predecir si una persona tiene diabetes o no. Por ejemplo,con una persona con alta Glucosa en sangre, bajos niveles de insulina, un BMI elevado, un historial familiar con casos de diabetes y una edad elevada, son indicadores de que esa persona puede desarrollar diabetes.

c. Wrapper:

Con este tipo de filtrado los atributos varían según el algoritmo que se vaya a usar.

- IB1 => [Glucose, DiabetesPedigreeFunction]
- IBK => [Glucose, Insulin, BMI, DiabetesPedigreeFunction, Age]
- Naive Bayes => [Glucose, BMI, DiabetesPedigreeFunction, Age]
- TAN => [Pregnancies, Glucose, BloodPressure, BMI, Age]
- JRip =>[Pregnancies, Glucose, BloodPressure, Insulin, BMI, Age]
- ID3 => [Glucose, BMI, DiabetesPedigreeFunction, Age]
- J48 => [Glucose, BMI, DiabetesPedigreeFunction, Age]
- Logistic = > [Pregnancies, Glucose, Insulin, BMI, DiabetesPedigreeFunction,Age]

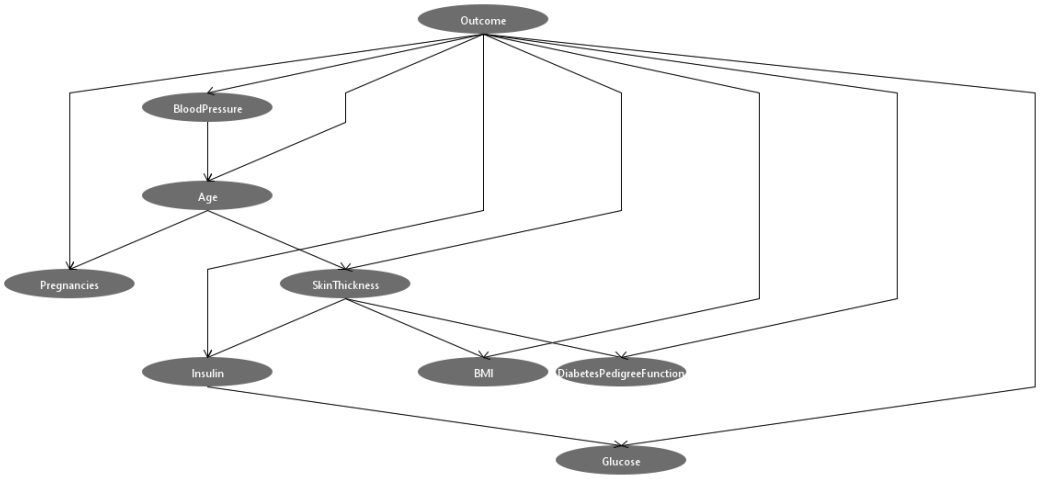
Resultados:

Algoritmo	Variables Originales	Filtrado Univariante	Filtrado Multivariante	Wrapper
IB1	67.71%	67.71%	76.20%	71.40%
IBK	74.74%	74.74%	77.34%	77.40%
Naive Bayes	75%	75%	76.43%	77%
TAN	73.83%	73.83%	74.74%	76.20%
JRip	75.39%	76.56%	74.74%	76.40%
ID3	57.42%	72.14%	78.91%	78.80%
J48	72.13%	75.39%	79.43%	79%
Logistic	77.08%	76.82%	77.34%	77.30%

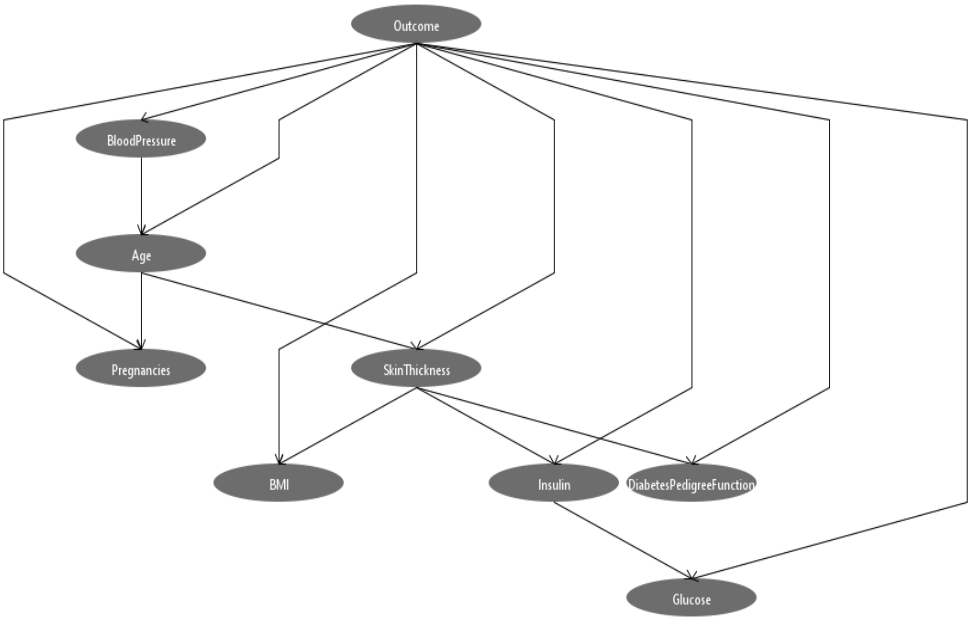
Como podemos observar en la tabla de resultados, los algoritmos que consiguen una mayor precisión son ID3 y J48 mediante el filtrado multivariante o con variables de tipo wrapper. Y el que menor precisión tiene es el ID3 pero usando las variables originales, sin ningun tipo de filtrado.

Tambien podemos observar los grafos que se producen con el algoritmo TAN:

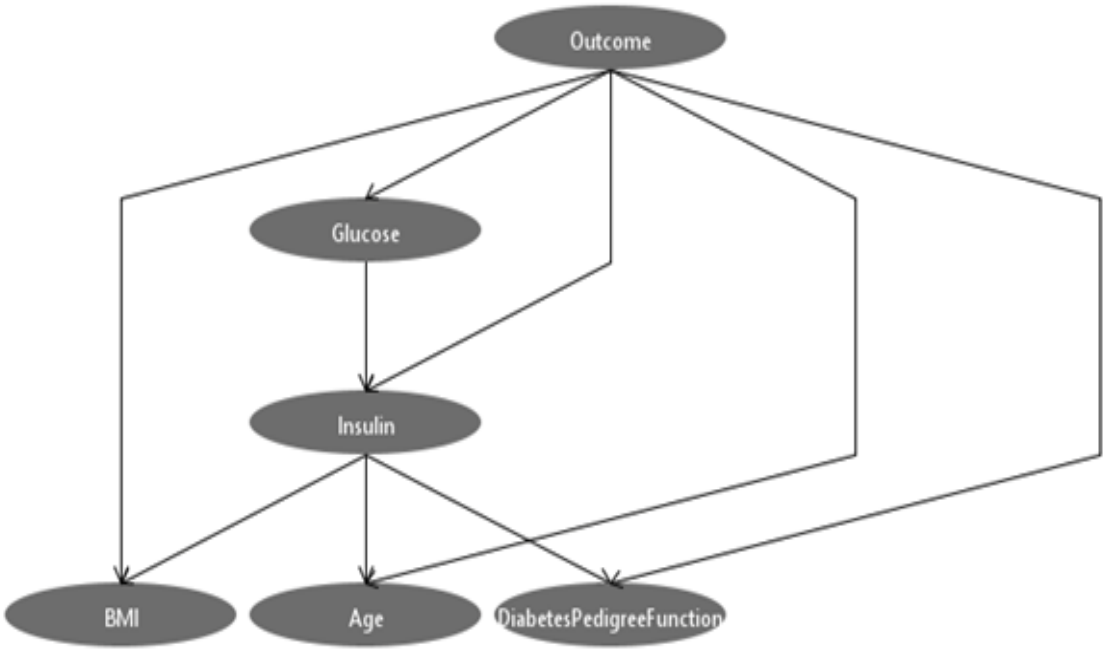
Distribución de los atributos según los datos originales:



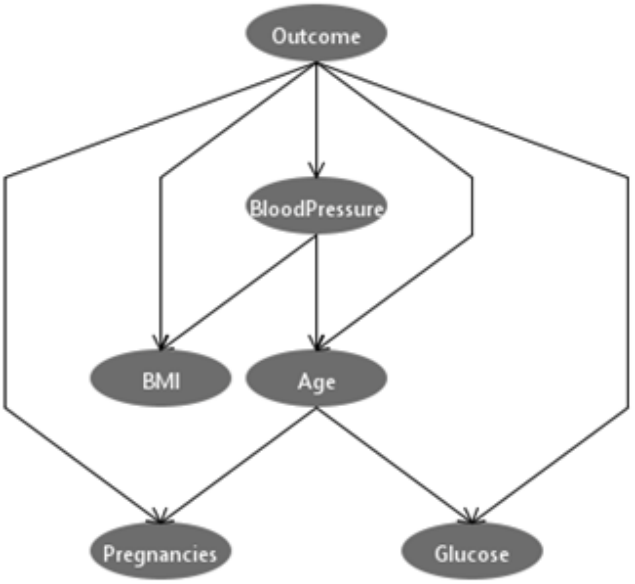
Distribución de los atributos según los el filtro univariante:



Distribución de los atributos según el filtrado multivariante:

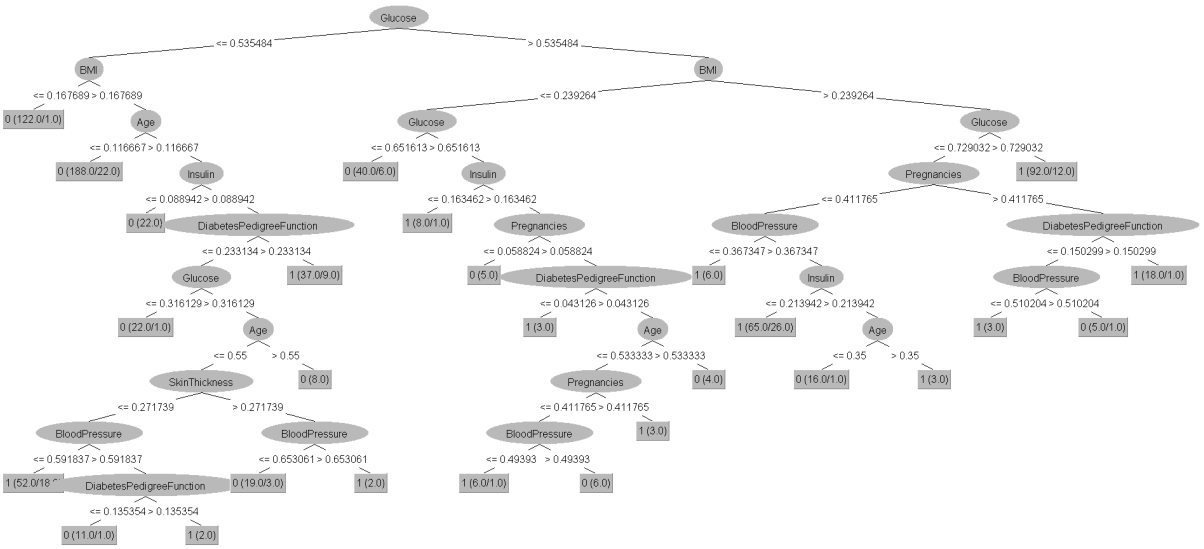


Distribución de los atributos con wrapper:

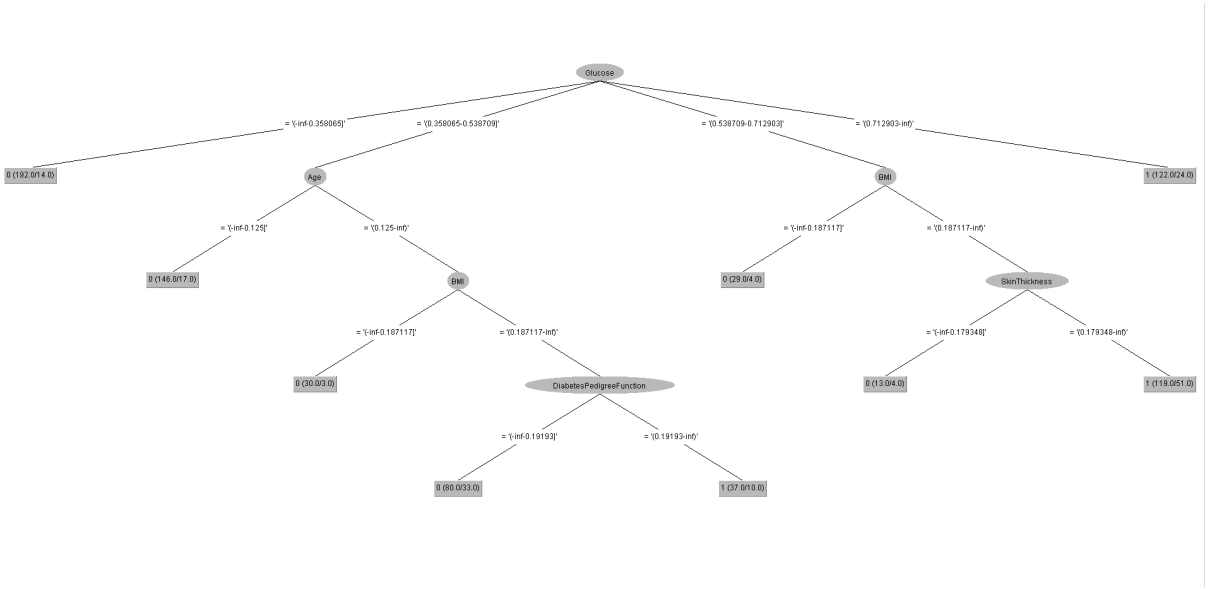


Árboles generados con J48:

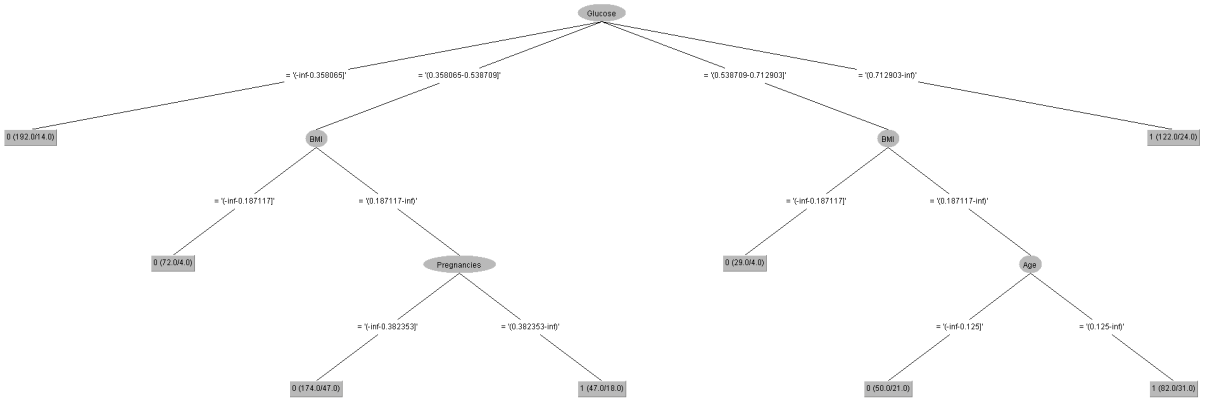
Árbol generado con datos originales:



Árbol con filtrado univariante:



Árbol generado con filtrado multivariante:



Reglas generadas con JRip:

Reglas generadas con los atributos originales:

```
(Glucose >= 0.541935) and (BMI >= 0.247444) => Outcome=1 (203.0/57.0)
(Age >= 0.166667) and (BMI >= 0.198364) and (Glucose >= 0.36129) and (Insulin >= 0.120192) => Outcome=1 (96.0/40.0)
 => Outcome=0 (469.0/66.0)
```

Reglas generadas con filtrado univariante:

```
(Glucose = '(0.712903-inf)') => Outcome=1 (122.0/24.0)
(Age = '(0.125-inf)') and (BMI = '(0.187117-inf)') and (DiabetesPedigreeFunction = '(0.19193-inf)') and (Insulin = '(0.088341-inf)') => Outcome=1 (76.0/22.0)
(Glucose = '(0.538709-0.712903]') and (BMI = '(0.187117-inf)') and (Age = '(0.125-inf)') => Outcome=1 (50.0/22.0)
 => Outcome=0 (520.0/88.0)
```

Reglas generadas con filtrado multivariante:

```
(Glucose = '(0.712903-inf)') => Outcome=1 (122.0/24.0)
(Age = '(0.125-inf)') and (BMI = '(0.187117-inf)') and (Glucose = '(0.538709-0.712903]') => Outcome=1 (82.0/31.0)
(Age = '(0.125-inf)') and (BMI = '(0.187117-inf)') and (Glucose = '(0.358065-0.538709]') and (Pregnancies = '(0.382353-inf)') => Outcome=1 (45.0/17.0)
 => Outcome=0 (519.0/91.0)
```

Reglas generadas con variables tipo wrapper:

```
(Glucose >= 0.741935) => Outcome=1 (104.0/19.0)
(Age >= 0.166667) and (BMI >= 0.198364) and (Glucose >= 0.387097) => Outcome=1 (152.0/57.0)
 => Outcome=0 (512.0/88.0)
```

Conclusiones:

Como se ha podido observar a lo largo de la realización de la práctica los algoritmos basados en arboles de decision (ID# y J48) demostraron un rendimiento bastante alto, especialmente cuando se usó la selección de características mediante filtrado multivariante y el método wrapper. Estos métodos de selección de variables permitieron a los árboles centrarse en las características más relevantes para la predicción, lo que llevó a un mejor rendimiento. En particular, J48 alcanzó la precisión más alta de todos los algoritmos, con un 79.43% utilizando el filtrado multivariante y un 79% utilizando el método wrapper.

En general, todos los algoritmos mejoran su precisión al introducir métodos de selección de variables. Particularmente los algoritmos IB1, IBK con k siendo 10, Naive Bayes y TAN obtienen la misma precisión al usar las variables originales y el filtrado univariante, pero al cambiar al filtrado multivariante y el método wrapper la precisión de estos algoritmos aumenta. Esto resalta la importancia de cómo una selección cuidadosa de los atributos puede variar el resultado final. Aunque el algoritmo Logistic fue el que obtuvo una mayor precisión al usar las variables originales, los resultados posteriores al usar otros métodos de selección de variables no mejoraron. Con esto podemos ver que no siempre hay un mejor algoritmo para el problema y el rendimiento puede variar en función de los atributos seleccionados.

Cada algoritmo seleccionó un conjunto diferente de características como las más importantes para la clasificación. Sin embargo, se pueden observar algunos patrones. En particular, 'Glucose' y 'BMI' fueron los atributos que se seleccionaron con mayor frecuencia, lo que sugiere que son factores clave en el diagnóstico de la diabetes. Otras características, como 'Insulin', 'DiabetesPedigreeFunction', y 'Age' también se seleccionaron en varios algoritmos, lo que destaca también su relevancia en el diagnóstico.Las diferencias en los atributos seleccionados por cada algoritmo pueden reflejar las interacciones entre ellos. Por ejemplo, 'Pregnancies' fue seleccionado por TAN y JRip pero no por el resto de los algoritmos. Esto puede indicar que la relevancia de 'Pregnancies' para la clasificación puede depender de otros factores, como 'BloodPressure' e 'Insulin', que también se seleccionaron en estos algoritmos. Este resultado subraya la complejidad del problema de diagnóstico de

la diabetes, que puede requerir tener en cuenta múltiples factores y sus interacciones. Aunque si se quiere realizar un diagnóstico más personal, habría que fijarse más en los perfiles y en las propiedades que puedan reflejar un factor de riesgo. Por ejemplo, un algoritmo que da importancia a 'Insulin', como pueden ser IBK o Logistic, pueden ser más adecuados para pacientes donde esto sea un factor de riesgo relevante.

Finalmente, a través de la aplicación de distintos métodos de selección de variables y técnicas de clasificación, hemos obtenido un valioso entendimiento sobre la complejidad que rodea el diagnóstico de la diabetes. Los datos y su análisis nos ofrecen un enfoque multifacético para un problema de salud que, a primera vista, puede parecer sencillo. La realización de este trabajo ha sido enriquecedora y creo firmemente que utilizar un conjunto de datos que resuene con el interés personal del investigador, en este caso la diabetes, potencia enormemente la motivación y el deseo de aprender. Esto no solo nos permite profundizar en los matices del aprendizaje automático, sino que también destaca la relevancia de su aplicación en la atención de la salud y el bienestar humano.

Bibliografía:

Pima Indians Diabetes Database

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Weka Manual

https://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf