Experimento Propio

Patricio Guledjuan Azul Noguera Rocio Gonzalez

2023-08-22

Introducción

El objetivo de nuestro experimento consistió en explorar y comprender el impacto de las variables en el rendimiento del modelo. A través del análisis de las variables más influyentes, buscamos obtener una visión más profunda de cómo se toman decisiones en el contexto de cada conjunto de datos. Además, evaluamos si la selección de características relevantes podría potencialmente mejorar el desempeño del modelo.

Nuestro enfoque implicó identificar las variables más significativas en función del comportamiento del modelo original y, posteriormente, elegir las variables más influyentes para construir un nuevo modelo. Esta estrategia nos permitió comparar el rendimiento del nuevo modelo con el del original y analizar cualquier diferencia resultante.

En una primera etapa, evaluamos la importancia de los atributos en el modelo original utilizando rpart. En este contexto, una alta puntuación de importancia sugiere que la variable ejerce una mayor influencia en las decisiones tomadas por el árbol, y por ende, en las predicciones resultantes.

```
##                          Churn_Importance
## complains                     236.2103361
## seconds_of_use                204.9099906
## frequency_of_use              171.5954940
## customer_value                149.2826469
## distinct_called_numbers       132.2503734
## status                        112.0612624
## frequency_of_sms               97.1680251
## call_failure                   80.4796559
## subscription_length            62.5382885
## age_group                      49.6025298
## age                            41.0622159
## charge_amount                  33.6214038
## tariff_plan                     0.3453161
```

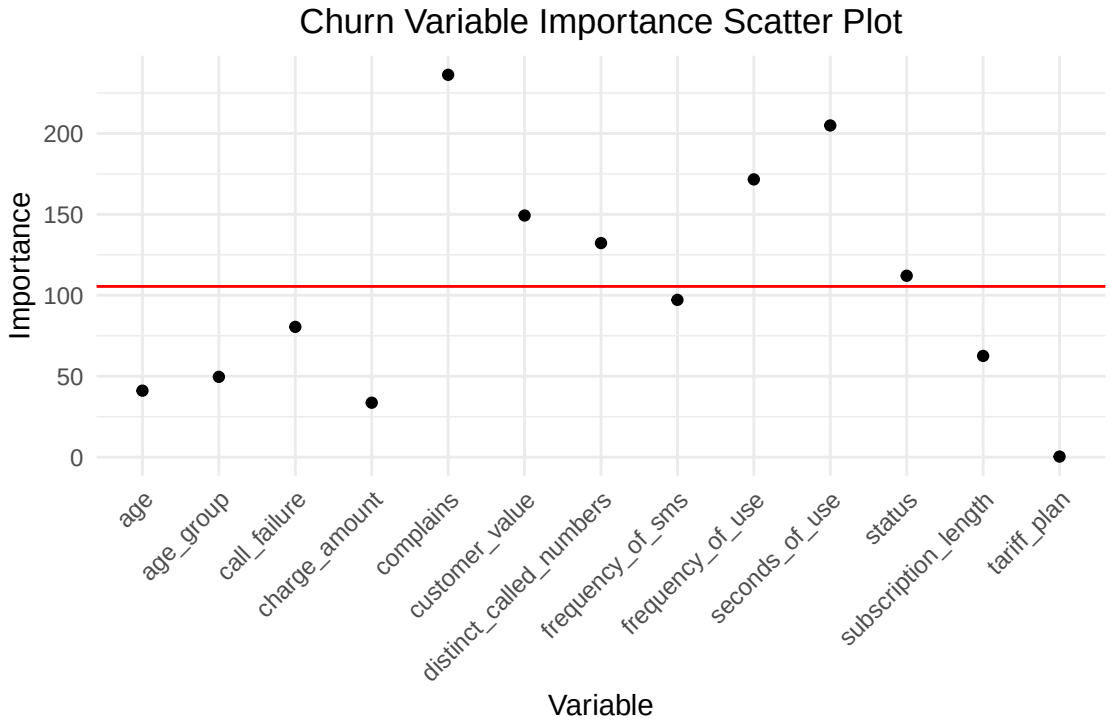
```
##                Heart_Importance
## ST_Slope               90.452387
## Oldpeak                57.272199
## MaxHR                  46.758755
## ChestPainType          43.218971
## Age                    36.473852
## ExerciseAngina         36.327523
## Cholesterol            30.218954
## RestingBP              24.524261
## Sex                     8.631311
## FastingBS               5.295776
## RestingECG              4.607943
```

```
##                      Wine_Importance
## alcohol                    213.69668
## density                    177.25551
## free.sulfur.dioxide        138.41699
## residual.sugar             136.51164
## volatile.acidity           134.56902
## total.sulfur.dioxide       129.91177
## chlorides                  114.29425
## fixed.acidity               90.97684
## citric.acid                 80.10864
## pH                          72.57032
## sulphates                   54.38498
```

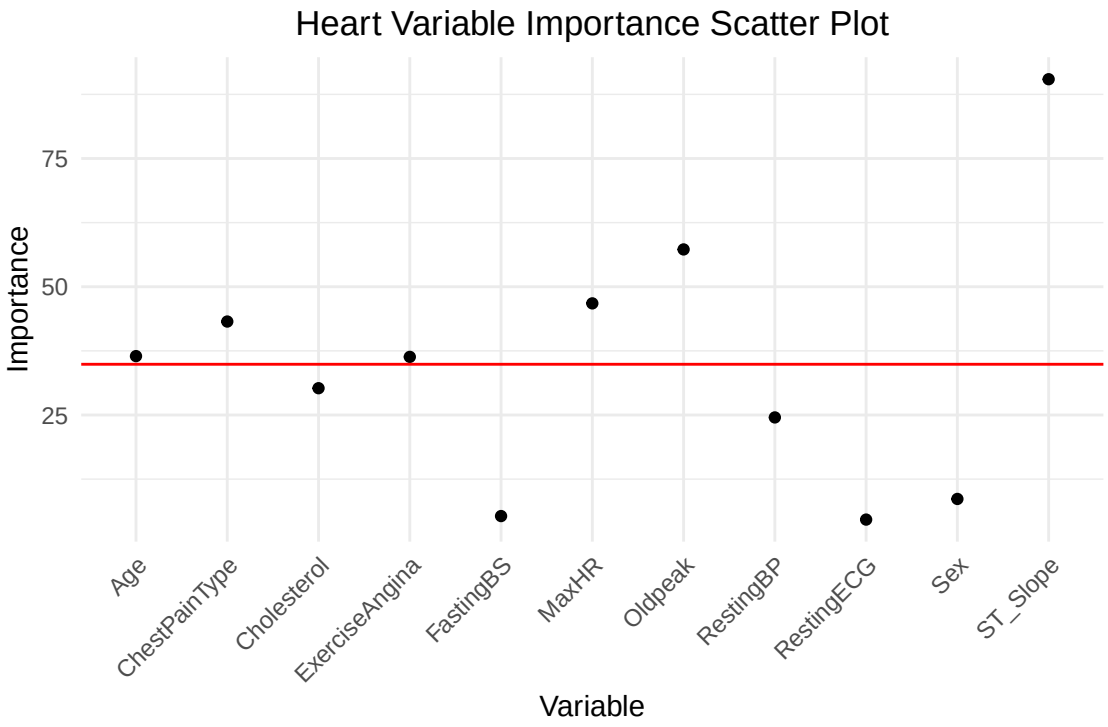
A partir de estos valores de importancia, establecimos una estrategia para determinar qué variables serían incluidas en el nuevo conjunto de datos. Para ello, calculamos el valor promedio de importancia de todas las variables presentes en el conjunto de datos original. Posteriormente, seleccionamos aquellas variables cuya importancia superara este promedio. Como resultado, cada uno de los conjuntos de datos originales se redujo a un conjunto de seis variables predictoras de mayor relevancia.

```
##        Dataset Importancia_Promedio
## 1    ChurnRate             105.47135
## 2 HeartDisease              34.88927
## 3  WineQuality             122.06333
```

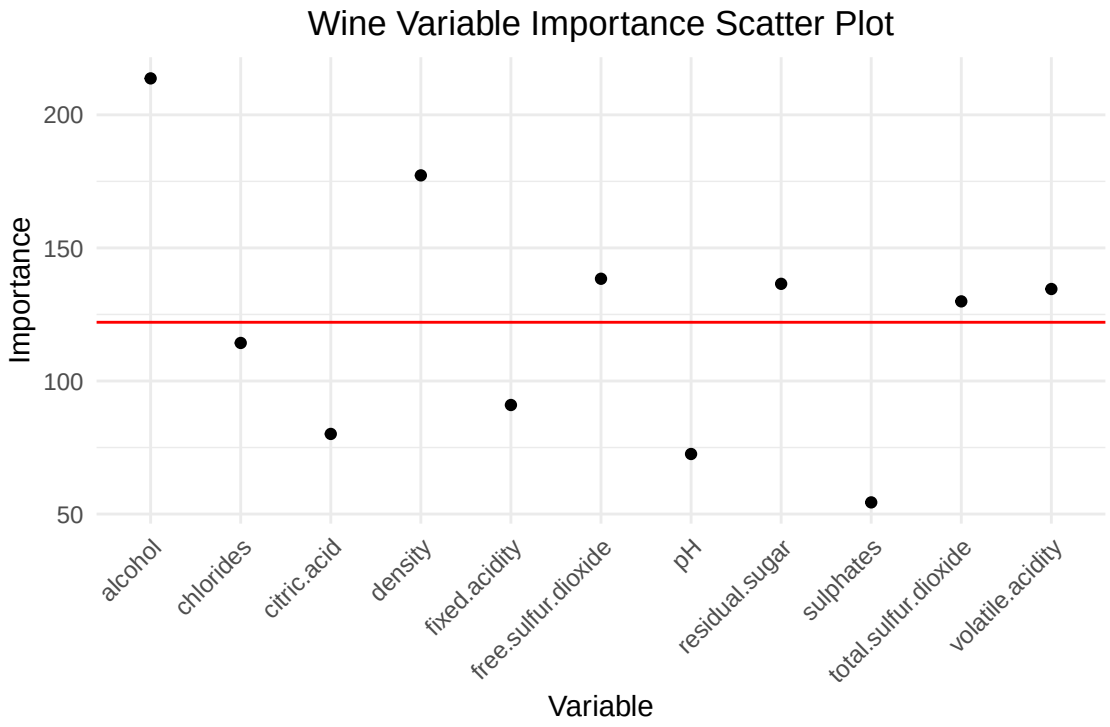
Variables importantes del dataset Churn



Variables importantes del dataset Heart



Variables importantes del dataset Wine

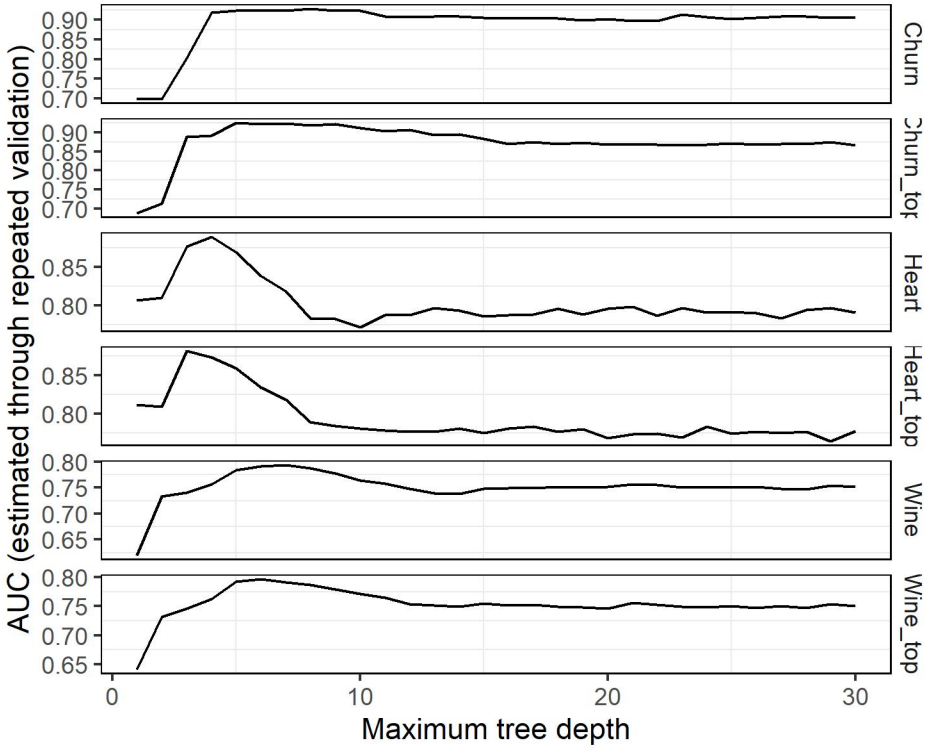


A partir de esto, obtuvimos los nuevos datasets que contienen las variables que cumplían con la condición planteada anteriormente.

Variables en datasets reducidos

```
##           Churn_variables Heart_variables       Wine_variables
## 1               complains        ST_Slope              alcohol
## 2          seconds_of_use         Oldpeak              density
## 3        frequency_of_use           MaxHR  free.sulfur.dioxide
## 4          customer_value   ChestPainType       residual.sugar
## 5 distinct_called_numbers             Age     volatile.acidity
## 6                  status  ExerciseAngina total.sulfur.dioxide
```

Resultados



Llevamos a cabo este experimento con la finalidad de indagar si incrementar el número de variables predictoras inevitablemente resulta en una mejora en el rendimiento del modelo, o si este resultado está condicionado por la importancia intrínseca de cada variable y las características particulares de los conjuntos de datos. Para efectuar este análisis, evaluamos el desempeño de los tres conjuntos de datos originales, junto con el de tres conjuntos de datos que fueron reducidos únicamente a las variables más significativas.

A grandes rasgos, podemos observar que las diferencias de rendimiento entre los conjuntos de datos originales y los conjuntos de datos reducidos son en su mayoría mínimas, siendo las performances prácticamente idénticas. Si profundizamos en los detalles, a medida que aumentamos la profundidad del árbol, se observa una ligera tendencia en la cual los conjuntos de datos originales tienen un rendimiento ligeramente superior en comparación con los conjuntos de datos reducidos.

Ahora, profundicemos en cada caso particular para determinar si este comportamiento tiene sentido en cada contexto.

Churn

Tanto el conjunto de datos original de churn como el conjunto reducido exhiben un comportamiento sorprendentemente parecido. Las discrepancias en el rendimiento a lo largo de las profundidades del árbol son mínimas, y se observa que los picos de rendimiento se alcanzan prácticamente al mismo tiempo. Además, los valores máximos de rendimiento se sitúan en torno a 0.9, señalando un nivel de predicción bastante sólido en ambos conjuntos de datos. Además, es interesante notar que el conjunto de datos original mantiene una ligera ventaja en términos de rendimiento en todas las iteraciones de entrenamiento de los árboles.

La aparente falta de impacto en el rendimiento podría sugerir que las variables que se eliminaron no desempeñaban un papel significativo en la capacidad del modelo para anticipar las tasas de abandono. Esto podría indicar que, dentro de este contexto particular, las variables que aún permanecen son capaces de capturar de manera adecuada la información relevante para realizar predicciones sólidas acerca del churn. En el marco de la tasa de abandono (churn rate), resulta lógico que las variables que estén directamente relacionadas con

los comportamientos y características de los clientes tengan un impacto sustancial en las predicciones. Las variables que mantuvieron una alta importancia se centran en aspectos como los patrones de llamadas, la frecuencia de uso y el valor del cliente. Estos factores pueden representar señales potentes acerca de si un cliente es más propenso a abandonar el servicio o a mantenerlo.

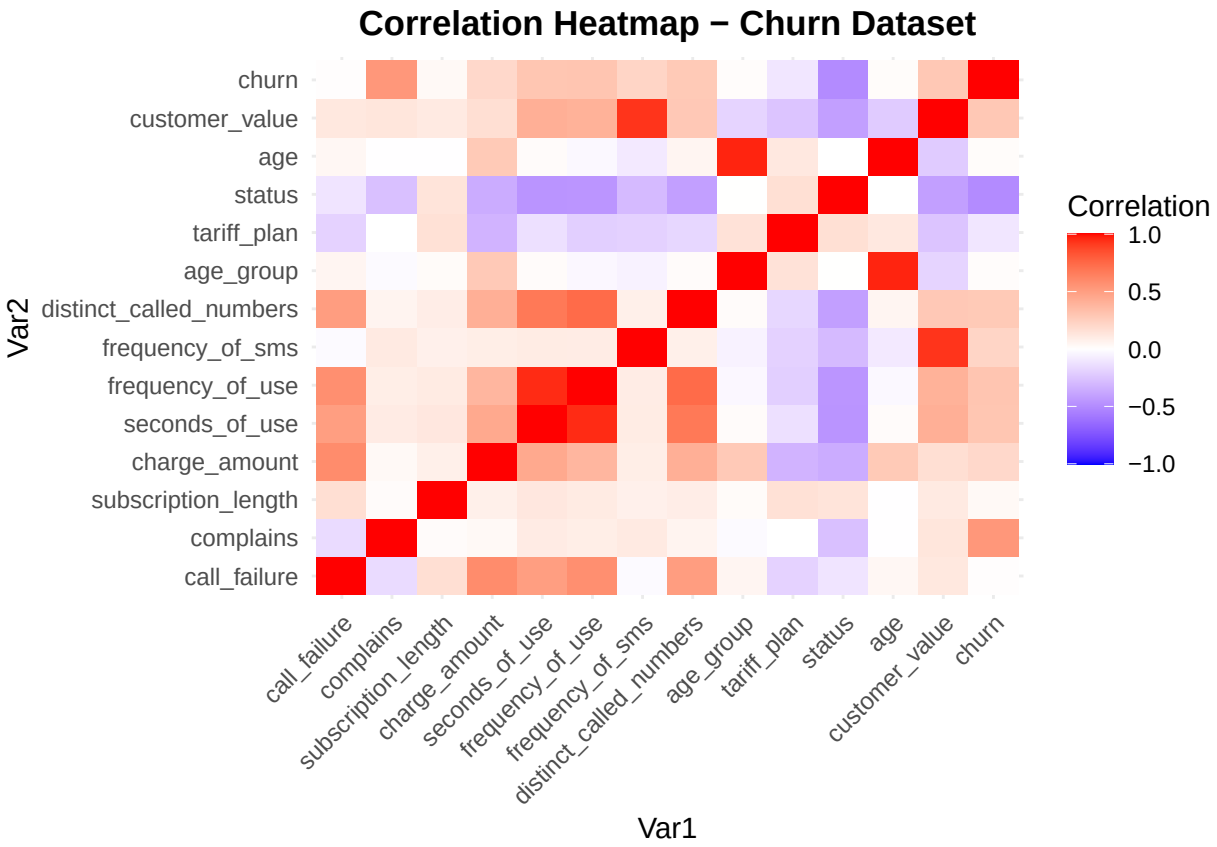
Una observación importante es que las variables eliminadas podrían haber carecido de aporte informativo único o estar altamente correlacionadas con las variables restantes. Esta correlación podría haber disminuido su valor predictivo cuando ya se consideran las otras características relevantes.

Veamos la corrrelación...

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```



Al analizar los valores de correlación entre las variables eliminadas y las restantes, notamos que en su mayoría son bajos. Esto sugiere que no existía una correlación fuerte entre las variables descartadas y las que se mantuvieron. Como resultado, es plausible pensar que las variables eliminadas posiblemente no añadían información distintiva que permitiera diferenciar entre las clases de predicción.

En conjunto, estos resultados sugieren que las variables más influyentes para predecir el churn están siendo capturadas por las variables que se mantienen, y que la inclusión o exclusión de las variables menos relevantes no afecta de manera significativa la capacidad del modelo para realizar predicciones precisas en este contexto particular.

Heart

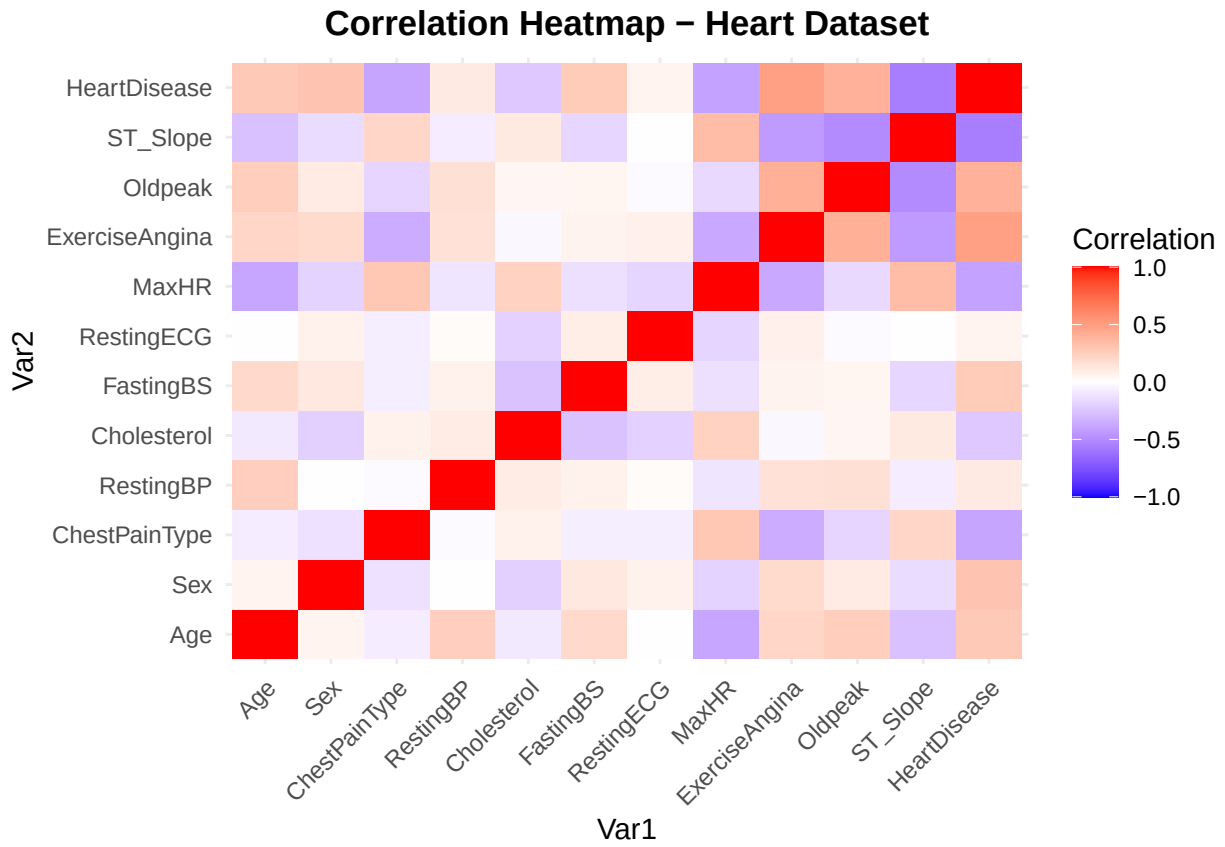
Al observar detenidamente los gráficos de rendimiento comparativo entre el dataset original y el dataset reducido de "Heart Disease", no se evidencian diferencias significativas que merezcan ser destacadas. Las tendencias predominantes en la evolución del rendimiento son sorprendentemente similares. Los picos de rendimiento máximo, se alcanzan prácticamente con la misma profundidad. Asimismo, la evolución del rendimiento a lo largo de las diferentes profundidades es casi idéntico y ambos datasets se comportan de la No obstante, es interesante resaltar que cuando se alcanza una profundidad más allá de un nivel 12, el dataset original muestra un rendimiento ligeramente superior en comparación con el dataset reducido.

A partir de esto, podríamos inferir que la reducción del dataset no parece haber ejercido una influencia notoria ni positiva ni negativa en el rendimiento del modelo. En consecuencia, se podría concluir que las variables eliminadas no aportaban un volumen significativo de información valiosa al modelo.

A continuación, es importante examinar si este resultado tiene fundamento en el contexto de las enfermedades cardíacas. Es decir, si las variables que fueron eliminadas intuitivamente carecen de información sustancial o relevante para la detección de patrones relacionados con las enfermedades cardíacas.

Dentro del ámbito de las enfermedades cardíacas, las variables que se conservan abordan elementos cruciales como la reacción del corazón al ejercicio, la manifestación de angina inducida por el ejercicio, y características relacionadas con la edad y la sensación de dolor en el pecho. Estas variables actúan como indicadores fundamentales para evaluar la salud cardíaca y podrían desempeñar un papel esencial en la predicción de enfermedades cardíacas. Es lógico suponer que ciertas variables, como el género, posiblemente carezcan de información única suficiente para diferenciar de manera relevante entre las distintas clases de predicción.

Examinemos si las variables que fueron eliminadas podrían haber estado relacionadas con aquellas que demostraron una mayor importancia, lo que podría haber motivado su exclusión del modelo.



Observando la matriz de correlación entre las variables, notamos que las correlaciones entre las variables eliminadas y las variables restantes son en su mayoría bastante bajas. Esto sugiere que no había una correlación fuerte entre las variables que fueron eliminadas y las que se mantuvieron en el modelo. Es decir, estas variables eliminadas no estaban altamente correlacionadas con ninguna de las variables que se consideraron importantes.

Por lo tanto, la eliminación de estas variables no parece haber sido influenciada por la correlación directa con las variables restantes. Es posible que estas variables no aportaran suficiente información única o no fueran consideradas relevantes por otras razones, lo que podría haber llevado a su exclusión del modelo.

En resumen, parece que la eliminación de variables en el modelo de enfermedades cardíacas no estuvo fuertemente influenciada por la correlación con las variables restantes. Es probable que la selección se haya basado más en la relevancia clínica de las variables y en la capacidad de aportar información distintiva al modelo de predicción de enfermedades cardíacas.

Wine

Al realizar un análisis de los gráficos de rendimiento del conjunto de datos de calidad de vinos, se observa que este dataset es el que parece presentar la menor influencia por la eliminación de variables. Al comparar detalladamente el rendimiento del conjunto de datos original con el conjunto de datos reducido, no se evidencia ninguna alteración significativa en su comportamiento general.

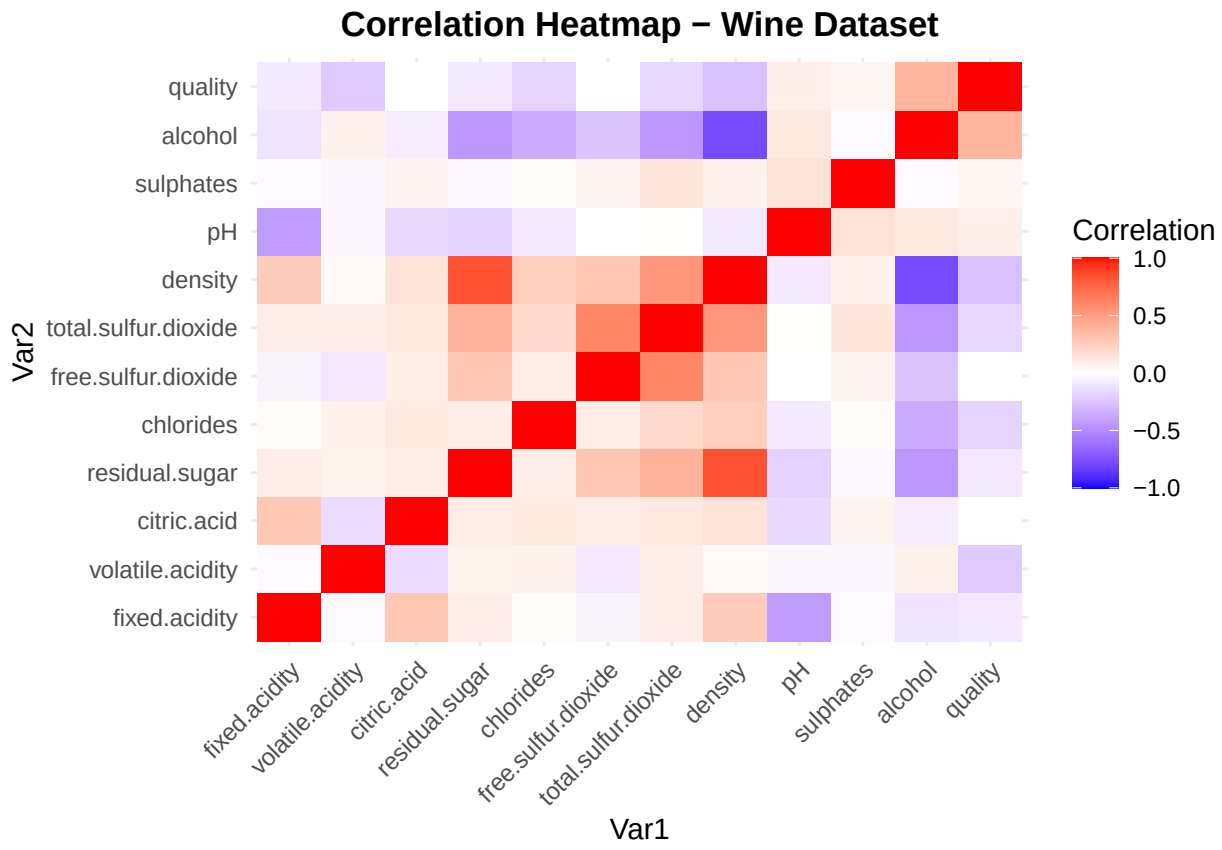
Esta estabilidad en el rendimiento podría sugerir que las variables que se eliminaron del conjunto de datos, en este caso “residual sugar” (azúcar residual) y “pH,” posiblemente no desempeñaban un papel crítico en la capacidad del modelo para predecir la calidad de los vinos. A pesar de la exclusión de estas variables, el modelo mantiene una consistencia en su rendimiento al predecir la calidad de los vinos.

Una hipótesis plausible para este fenómeno es que, en el contexto de la calidad del vino, las características

aportadas por las variables eliminadas podrían haber sido redundantes o estar altamente correlacionadas con otras propiedades sensoriales y químicas del vino que siguen presentes en el conjunto de datos. Por lo tanto, la eliminación de estas variables parece no afectar sustancialmente la capacidad del modelo para hacer predicciones precisas sobre la calidad del vino.

En el contexto de la calidad del vino, las variables restantes incluyen características químicas y físicas del vino. Estas variables pueden influir en la calidad percibida y real del vino. Dado que la calidad del vino se basa en componentes químicos y sensoriales, las variables seleccionadas son fundamentales para determinar cómo se percibirá el vino en términos de calidad.

Es importante analizar si las variables que se eliminaron del conjunto de datos de calidad de vinos podrían haber estado relacionadas con aquellas que se mantuvieron en el modelo. Para hacerlo, observemos el mapa de calor que muestra la correlación entre las variables.



En este caso, no se observan correlaciones particularmente altas o evidentes entre las variables eliminadas y las que se mantuvieron. Por ejemplo, las variables "fixed.acidity," "chlorides," y "pH" presentan correlaciones con las variables restantes que son bastante bajas, lo que sugiere que no hay una fuerte relación lineal entre estas variables.

Dado que las correlaciones no muestran una conexión clara entre las variables eliminadas y las variables que se mantuvieron en el modelo, es posible que las variables eliminadas no hayan sido excluidas debido a una correlación directa con las variables restantes. En cambio, otras consideraciones, como redundancia de información, relevancia contextual o impacto en la capacidad predictiva, parecen haber influido en su eliminación.