Resultados obtenidos:

Primero que nada me base en el estudio de mi dataset con gráficos descriptivos para entender un poco el comportamiento y conocimientos básicos de mis datos.

Utilizando la función print(df.describe()) para ambas bases de datos tanto varones como mujeres , se puede notar que las defunciones de los hombres tiene su mayor cantidad de defunciones en el año 2020 , año de pandemia mientra que luego vuelve a disminuir sus valor años posteriores y en cambio el data set de las mujeres parece tener un comportamiento más lineal y aumentado la cantidad de defunciones año a año. Destacando que el numero de defunciones en varones es superior al numero de defunciones de mujeres.

Luego realice histogramas , de ambos data set para ver el comportamiento de las defunciones en los últimos 5 años de información que me brindaba el data set (2018 a 2022) pudiendo ver que hay valores bastantes separado.

Luego realice diagramas de dispersión de ambos data set en relacion a las cusas de muerte en los ultimos 5 años de info del data set para tener una idea de los cambios en las causas , notando que:

En el caso de los varones: las principales causas de muerte para los años 2018,2019 son en personas mayor a 45 y por lo general por enfermedades en el sistema circulatorio , respiratorios y tumores malignos. En el año 2020 se nota un aumento en defunciones por causas infecciosas en personas mayores a 65 que disminuye gradualmente sus valores en 2021 y 2022.

En el caso de las mujeres las causas son en su mayoría coincidentes con la de los varones a diferencia de que las mayores defunciones se centran en personas mayores de 65 años.

En cuanto a mi modelo de aprendizaje automático intente dos manera para ver cual predice con mayor exactitud. Lo primero que intente buscando e inspeccionando los modelos de aprendizaje automático que mayor se adaptaban a mi data set fue:

Modelos SARIMA: Una Herramienta Esencial para la Predicción de Series Temporales

¿Qué es un modelo SARIMA?

SARIMA, que significa *Seasonal AutoRegressive Integrated Moving Average*, es una extensión de los modelos ARIMA, específicamente diseñados para analizar y predecir series de tiempo que presentan estacionalidad. Esta estacionalidad puede ser diaria, semanal, mensual, anual o de cualquier otro período regular

Como resultado de mi modelo sarima para el data frame de varones obtuve:

MSE: 46.19060166666667

R2: 0.9299918623246577

esto quiere decir que el MSE bajo y un R^2 alto, como los indica el modelo SARIMA se ajusta muy bien a los datos. Esto quiere decir que las predicciones del modelo son precisas y confiables.Puedes tener una alta confianza en las predicciones generadas por este modelo para datos futuros que sigan patrones similares.

Como resultado de mi modelo sarima para el data frame de mujeres obtuve:

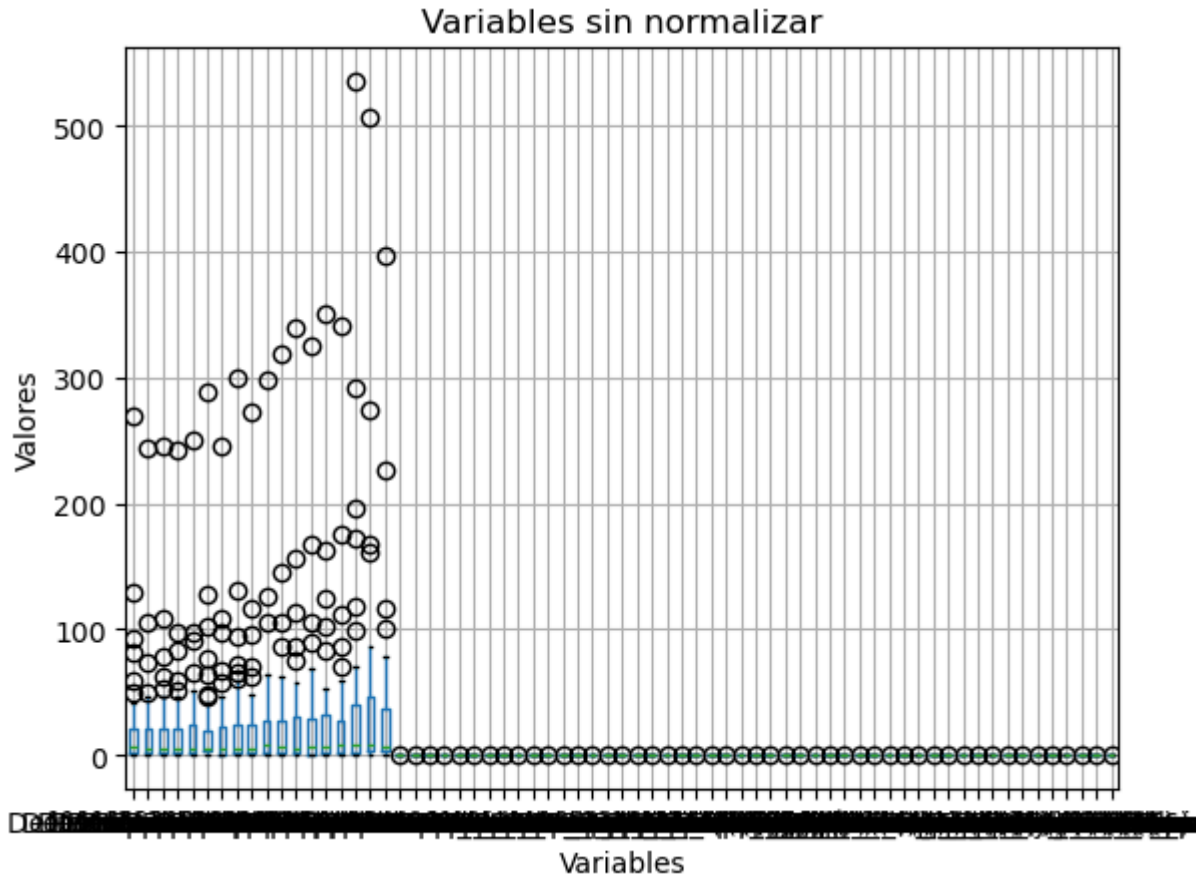
MSE: 12.972586666666666

R2: 0.9318461709864078

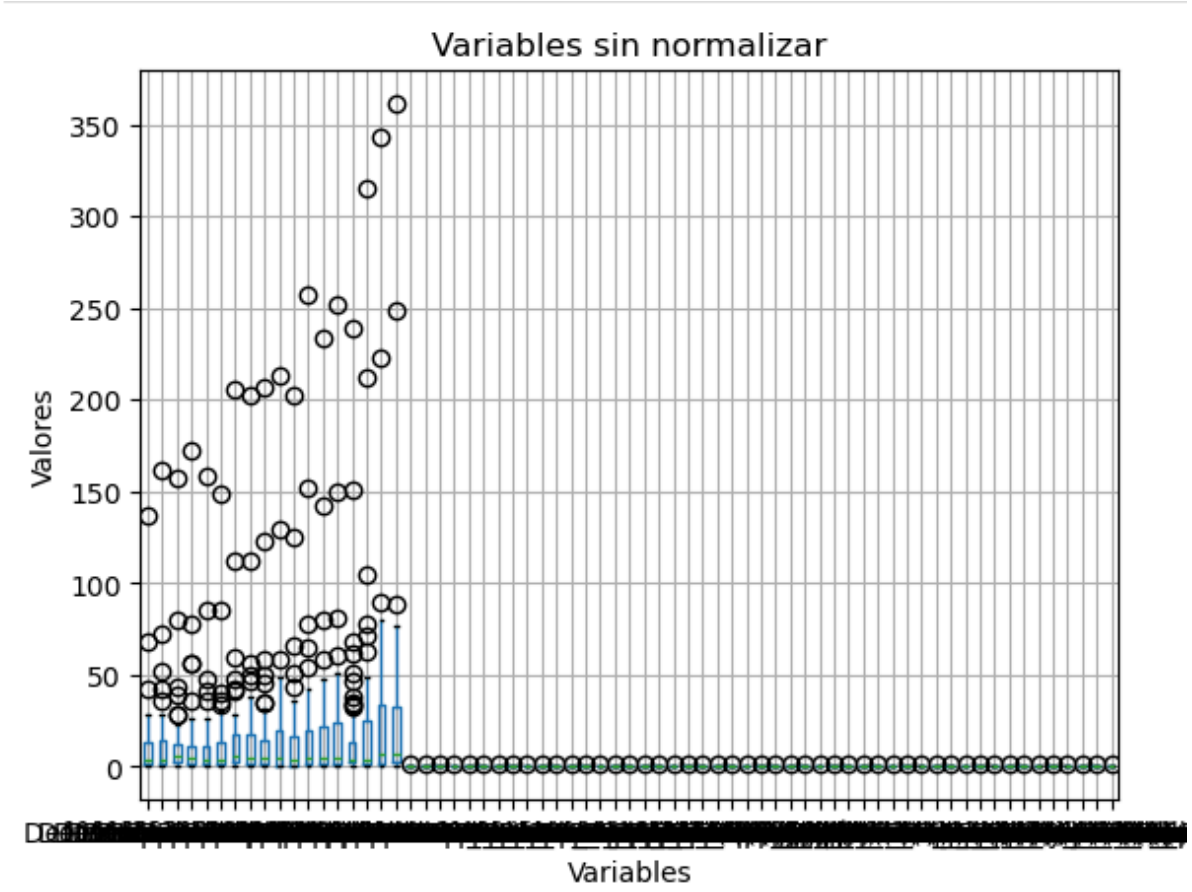
Es decir, mi MSE es aun menor que en comparación con con el data frame de los varones , este mse tan bajo quiere decir que mi modelo se ajusta aun menor y al aumentar R2 su valor y estas mas cercano a 1 , el modelo es mejor ya que explica el 93.18% de la variabilidad en los datos.

Para obtener otra opcion de modelo de aprendizaje automatico y comparar para ver que se ajusta mejor a mi dataframe probe con arboles de decisión. Para esto primero realice graficos de boxplot para ver la diversidad de datos en ambos data frame y obtuve lo siguiente:

boxplot varones:



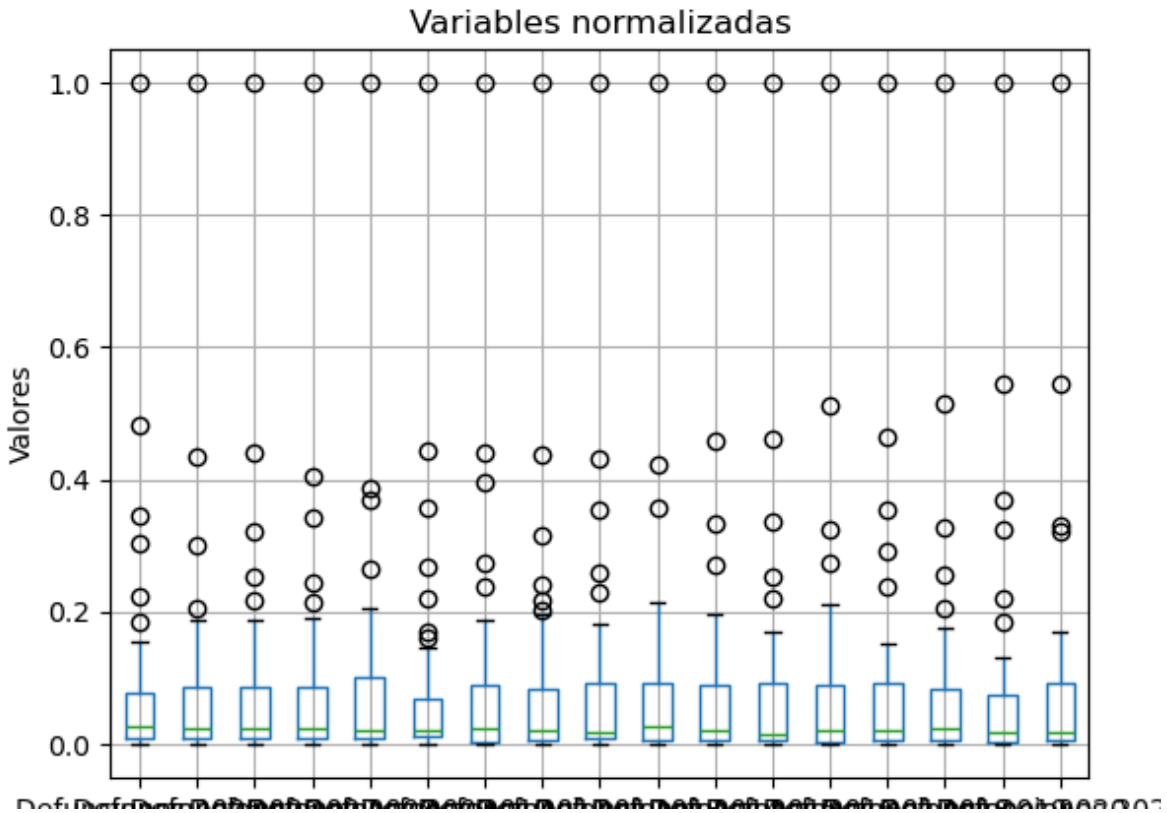
boxplot mujeres



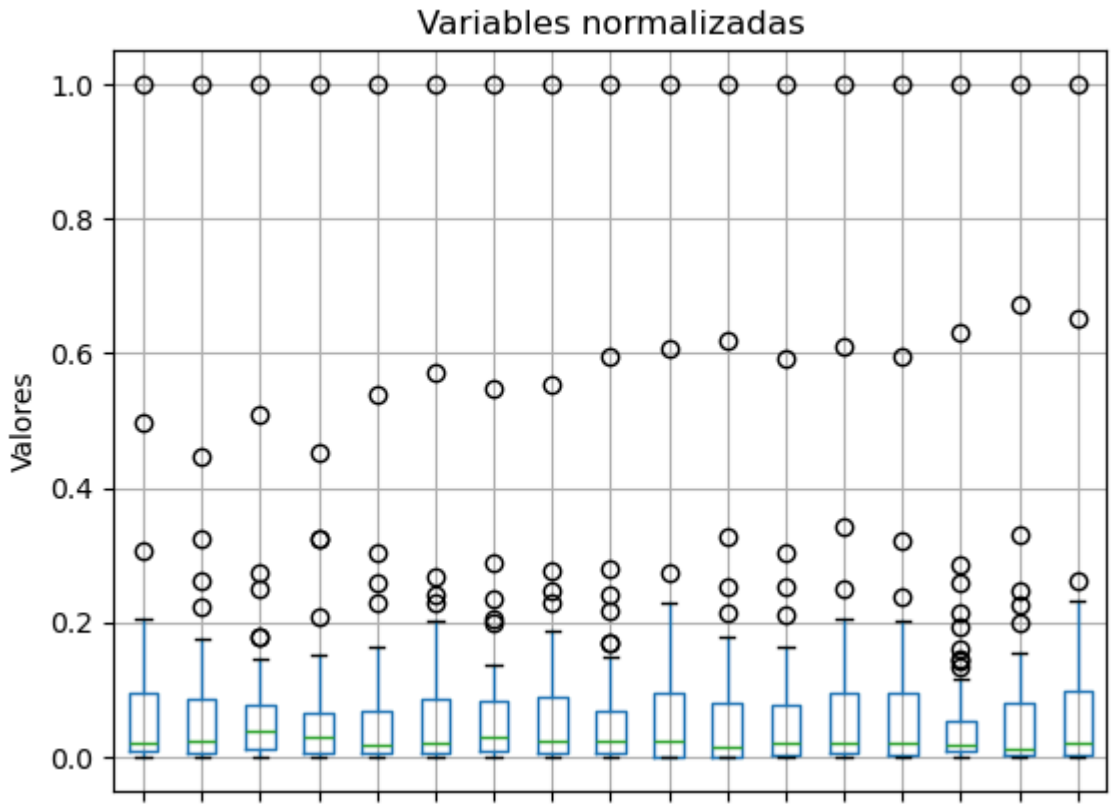
Se puede ver claramente que en ambos casos los valores estan muy distantes uno de otro y van desde cero a mas de 500 en el caso de los hombres y desde cero a mas de 350 en el caso de las mujeres por lo que dificulta poder generar un arbol de decisión asi.

Para mejorar esto decidi anonimizar los datos obteniendo unos boxplots muchas mas parejos como los que detallo a continuacion:

boxplto normalizado varones



boxplot normalizado mujeres



En este caso para el dataframe de varones el MSE me dio un valor de Error cuadrático medio: 1.791004469113144 , mucho menor al obtenido con el metodo SARIMA , y para ajustar este error cuadrático medio decidi tomar decidi no tomar todas las columnas como en el primer caso donde eran 18 columnas con informacion sino que decidi “entrenar menos el modelo “ y obtuve un error cuadratico de Error cuadrático medio: 0.0009299275603974688 ampliamente menor.

Para el dataframe de las mujeres obtuve un valor de mse de Error cuadrático medio: 0.12403007578625799 significativamente menor para mi modelo SARIMA y ajustando mi modelo de entrenamiento para evitar overfitting y reduciendo las variable de entrenamiento obtuve un valor de Error cuadrático medio: 0.0013968797872229324

En conclusión , el modelo de aprendizaje automático que más se ajusta a mi modelo es el de árbol de decisión superando ampliamente la diferencia con el de SARIMA y ajustándose mejor el modelo de predicción cuando no se le hace un overfitting con todas las columnas del dataset