ALUMNO: María Celeste Moreno
PROFESOR: Martín Mirabete
MATERIA : Aprendizaje Automático
ENTREGA 3: Presentación del Modelo y Análisis de Resultados

Introduccion y origenes de datos

- **Fuente de datos:** Los datos utilizados son de IPIEC, con base en datos del Ministerio de Salud y Desarrollo Social. Secretaria de Gobierno de Salud. Dirección de Estadísticas e Información en Salud (DEIS).
- **Objetivo del proyecto:** El objetivo principal de este proyecto es desarrollar un modelo de aprendizaje automático capaz de predecir la causa de muerte en la provincia de Tierra del Fuego. A través de este modelo, se busca identificar patrones y tendencias en la mortalidad que permitan comprender mejor los factores de riesgo asociados a cada causa de muerte.

Análisis exploratorio de datos

- **Resumen estadístico:** Realice un análisis preliminar del dataset para comprender la distribución de las variables e identificar valores atípicos.

- **Gráficos y visualizaciones:** Inclui gráfico de dispersión, gráficos de barra apiladas y boxplot para visualizar las relaciones entre las variables.

Conclusiones del análisis exploratorio:

Durante el análisis se observó que los hombres tienen su máximo pico de defunciones en el año 2020 , año de la pandemia , mientras que años posteriores disminuye gradualmente. Por el contrario el dataset de las mujeres se puede notar un comportamiento más lineal aumentando la cantidad de defunciones año a año. Destacando que el total de defunciones de los hombres es mayor a las defunciones de las mujeres.

Luego realice histogramas , de ambos dataset para ver el comportamiento de las defunciones en los últimos 5 años de información que me brindaba el dataset (2018 a 2022) pudiendo ver que hay valores bastantes separados, es decir, causas variadas en las defunciones totales.

Luego realice diagramas de dispersión de ambos dataset en relación a las causas de muerte en los últimos 5 años de info del dataset 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.

Evaluacion y rendicion de los modelos

En cuanto a los modelo de aprendizaje automático seleccionamos para mi tipo de dataset utilice:

Modelo 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 hombres obtuve:

MSE: 6.270923333333335
R2: 0.9484351047146039

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. Es decir que las predicciones del modelo son precisas y confiables.

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

MSE: 56.39498166666667
R2: 0.9045467524696017

Para poder evaluar otra opción de modelo de aprendizaje automático y comparar para ver que se ajusta mejor a mi dataframe probé con modelo de árboles de decisión. Para esto primero realice gráficas de boxplot para ver la diversidad de datos en ambos data frame y al ver que los valores eran muy dispersos normalice las variables a fin de obtener un modelo más ajustado.

Como resultados de mi modelo arbol decision obtuve los siguientes resultados:

dataset hombres:

arbol de decision Error MSE: 0.0014035157874466744

arbol de decision r2: 0.8067498972775866

dataset mujeres:

árbol de decisión MSE: 0.0007818899355115524

arbol de decision r2: 0.844321178471499

Los valores de MSE extremadamente bajos y los valores de R^2 razonablemente altos sugieren que los árboles de decisión también se ajustan bien a los datos, especialmente después de la normalización.

Conclusiones y recomendaciones

Rendimiento General: Ambos modelos, SARIMA y árbol de decisión, parecen ajustarse bien a los datos. Sin embargo, el SARIMA parece ser ligeramente mejor en el caso de los hombres, mientras que el árbol de decisión sobresale en el caso de las mujeres, especialmente en términos de MSE.

Normalización: La normalización de los datos antes de aplicar el árbol de decisión fue una decisión que ayudó a mejorar el rendimiento del modelo.

Referencias y acceso al proyecto

Repositorio GitHub: el código, el dataset y el notebooks están disponibles en el siguiente repositorio GitHub: Acceso a repositorio