Diego Alberto Morales Ibáñez

Director: Dra. Patricia Rayón Villela

Imputación de Datos mediante Métodos Estadísticos, Aprendizaje Automático y Aprendizaje Profundo

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Índice

1. Introducción

- Transformación digital liderada por los datos
- Fuentes de información con datos faltantes o erróneos
- Corrección desde el origen complicada
- Recuperación de información no viable
- Incompatible para reportería o modelamiento
- Estudio comparativo de técnicas de imputación
- Uso de métodos estadístico, aprendizaje automático y aprendizaje profundo

2.1 Objetivo General

- Realizar un estudio comparativo para la imputación de datos mediante métodos estadísticos, de aprendizaje automático, aprendizaje profundo para tareas de clasificación y regresión.

2.1 Objetivos Específicos

- Evaluar el desempeño de distintos algoritmos estadísticos, de aprendizaje automático y aprendizaje profundo sobre conjuntos de datos con información faltante para la imputación mediante regresión y clasificación.
- Medir el tiempo de ejecución de distintos algoritmos estadísticos, de aprendizaje automático y aprendizaje profundo sobre conjuntos de datos con información faltante para la imputación de datos.
- Determinar el método que ofrezca los mejores resultados para tareas de clasificación y regresión a partir de métricas de desempeño y tiempo de ejecución.

3. Metodología

Se define un iterador para el rango de nulos [0.05, 0.45]
Por cada valor de porcentaje de nulidad:
    Se divide el conjunto de datos en entrenamiento y prueba
    Por cada método de imputación, se realiza lo siguiente:
        Se entrena sobre el conjunto de entrenamiento
        Se inicia el temporizador
        Se realiza la imputación sobre el conjunto de nulos
        Se detiene el temporizador
        Se registran los resultados
Se generan los gráficos y tablas de resultados ordenados por desempeño

3.1 Regresión

- **California Housing**: Registros, 20640; Atributos, numéricos; Tarea, regresión.
- `X = data[['MedInc', 'HouseAge', 'AveRooms', 'AveBedrms', 'Population', 'AveOccup', 'Latitude', 'Longitude']]`
- `y = data['MedHouseVal']`

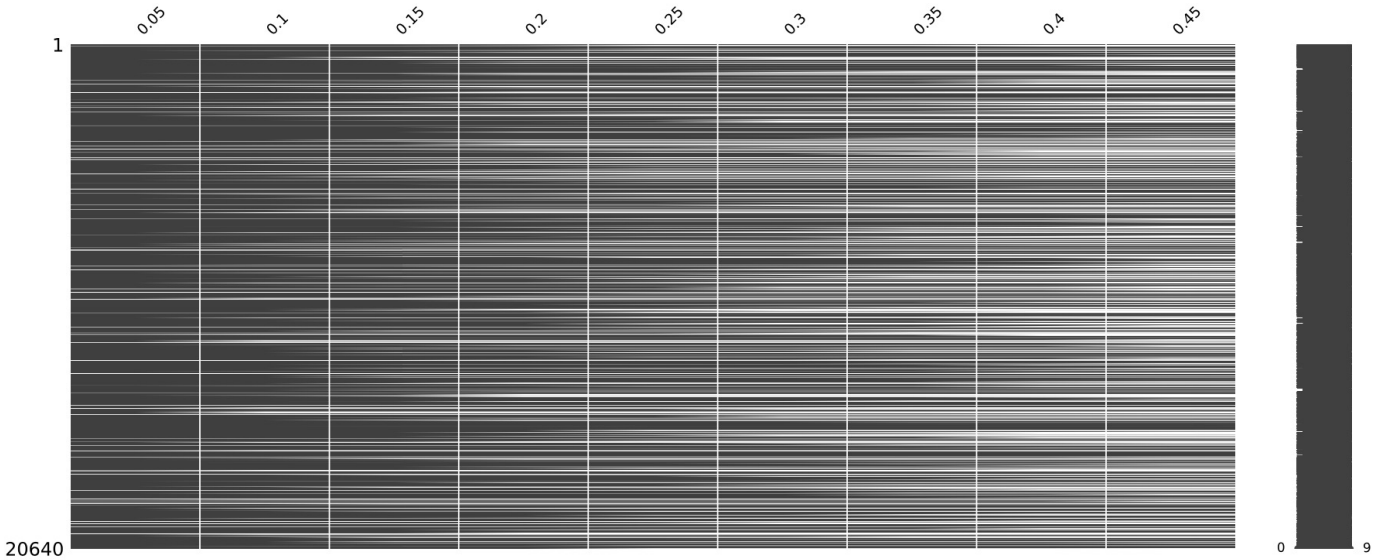

Figura 9.1: Nulos introducidos en MedHouseVal del California Housing

3.1 Regresión

- Medición de error cuadrático medio y tiempo de imputación

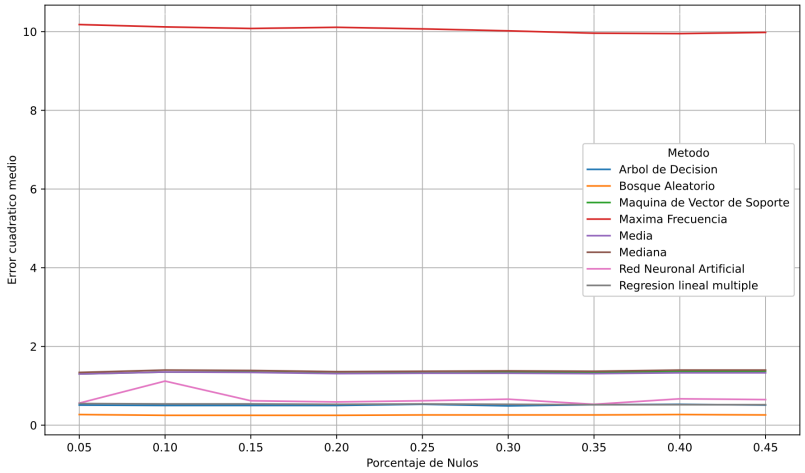

Figura 9.2: Error para métodos de imputación sobre MedHouseVal

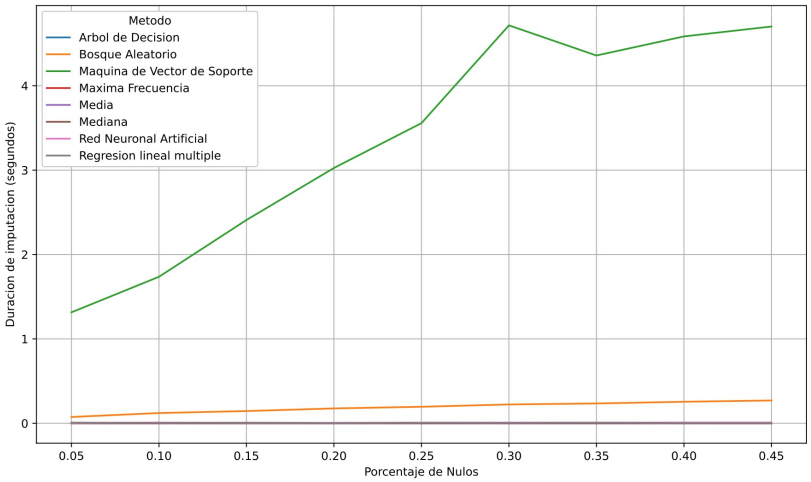

Figura 9.3: Duración para métodos de imputación sobre MedHouseVal

Metodo	RMSE	Duracion
Bosque Aleatorio	0.258889	0.205695
Arbol de Decision	0.510000	0.004361
Regresion lineal multiple	0.532222	0.004597
Red Neuronal Artificial	0.668889	0.007549
Media	1.323333	0.000265
Maquina de Vector de Soporte	1.340000	4.083446
Mediana	1.378889	0.000196
Maxima Frecuencia	10.052222	0.000143

Cuadro 9.4: Resultados promedio para métodos de imputación sobre MedHouseVal

3.2 Clasificación

- **Iris**: Registros, 150; Atributos, numéricos; Tarea, clasificación.
- `X = data[['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']]`
- `y = data['target']`

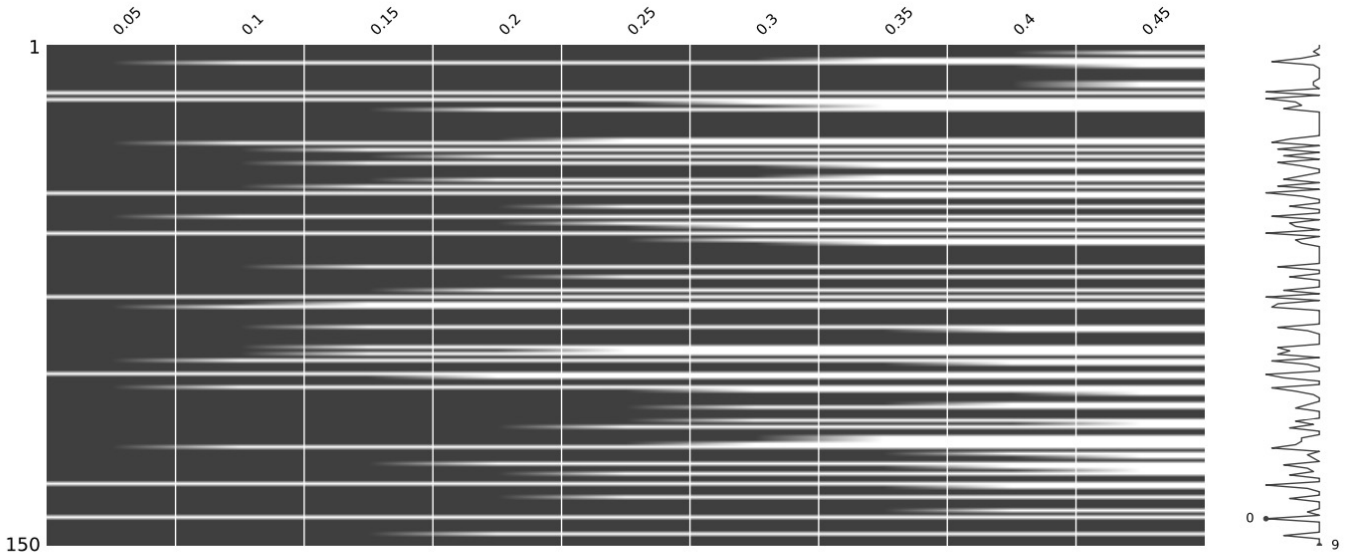

Figura 9.4: Nulos introducidos en target del Iris dataset

3.2 Clasificación

- Medición de precisión y tiempo de imputación

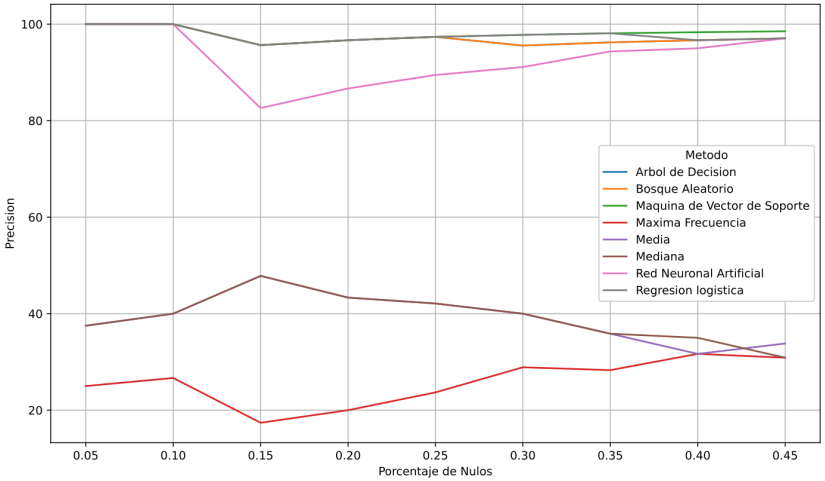

Figura 9.5: Precisión para métodos de imputación sobre target

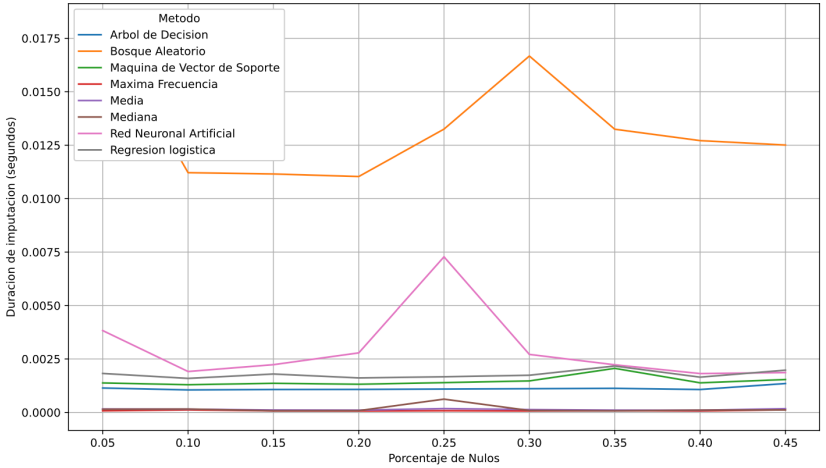

Figura 9.6: Duración para métodos de imputación sobre target

Metodo	Precision	Duracion
Regresion logistica	81.467778	0.002448
Red Neuronal Artificial	80.047778	0.003697
Bosque Aleatorio	79.481111	0.022638
Arbol de Decision	77.215556	0.001577
Maquina de Vector de Soporte	66.314444	0.017495
Maxima Frecuencia	60.874444	0.000146
Media	60.874444	0.000194
Mediana	60.874444	0.000129

Cuadro 9.8: Resultados promedio para métodos de imputación sobre target

3.3 Mixto

- **Titanic**: Registros, 1309; Atributos numéricos y categóricos; Tarea, regresión y clasificación.
- Resumen de resultados para imputación sobre age y survived

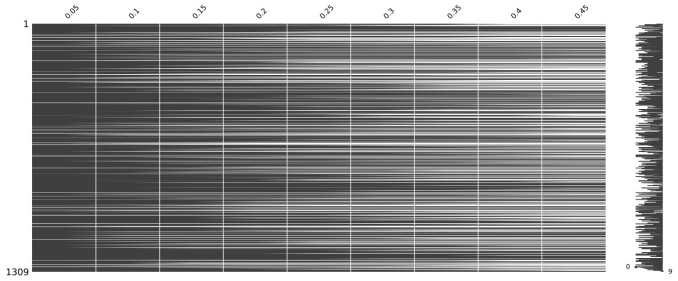

Figura 9.7: Nulos introducidos en age del Titanic Dataset

Metodo	RMSE	Duracion
Bosque Aleatorio	128.075556	0.041763
Regresion lineal multiple	129.690000	0.005684
Red Neuronal Artificial	157.654444	0.008528
Maquina de Vector de Soporte	164.591111	0.034454
Maxima Frecuencia	167.775556	0.000147
Mediana	167.775556	0.000727
Media	167.833333	0.000188
Arbol de Decision	173.480000	0.003430

Cuadro 9.12: Resultados promedio para métodos de imputación sobre age

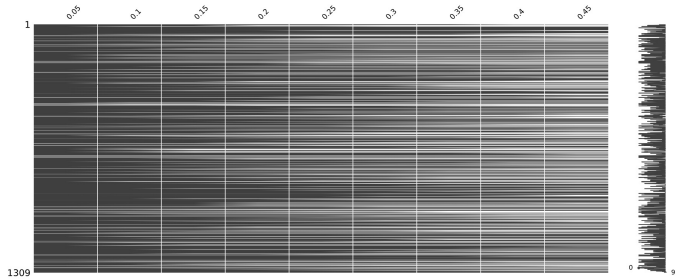

Figura 9.10: Nulos introducidos en survived del Titanic Dataset

Metodo	Precision	Duracion
Regresion logistica	81.467778	0.002448
Red Neuronal Artificial	80.047778	0.003697
Bosque Aleatorio	79.481111	0.022638
Arbol de Decision	77.215556	0.001577
Maquina de Vector de Soporte	66.314444	0.017495
Maxima Frecuencia	60.874444	0.000146
Media	60.874444	0.000194
Mediana	60.874444	0.000129

Cuadro 9.15: Resultados promedio para métodos de imputación sobre survived

4. Discusión de Resultados

- Los algoritmos de aprendizaje de máquina obtuvieron mejores resultados para imputación de valores numéricos.
- Los algoritmos de aprendizaje de máquina obtuvieron mejores resultados para imputación de valores categóricos.
- El bosque aleatorio y la regresión lineal múltiple obtuvieron errores similares en regresión, pero el tiempo de imputación del bosque aleatorio fue mayor.
- La regresión logística fue el método más eficiente para tareas de clasificación.

5. Conclusiones

- La imputación mediante métodos de aprendizaje automático y aprendizaje profundo obtuvieron el **menor error cuadrático medio** y la **mayor precisión** comparado con los métodos estadísticos convencionales.
- Los métodos de **bosque aleatorio y regresión lineal** múltiple obtuvieron los menores errores para la **imputación** de datos **numéricos**. El tiempo de ejecución fue menor para la regresión lineal múltiple.
- Los métodos de **regresión logística, red neuronal artificial y bosque aleatorio** obtuvieron las mayores precisiones para la **imputación** de datos **categóricos**. El tiempo de ejecución de la regresión logística y la red neuronal fueron inferiores al del bosque aleatorio.

6. Recomendaciones

- Evaluar metodología propuesta en diferentes conjuntos de datos (tamaño, tipología y distribución).
- Modificar parámetros de configuración de los algoritmos para minimizar el error y maximizar la precisión.
- Replicar metodología en otros métodos de imputación.