Imputación de Datos

Objetivo

Llevar a cabo diferentes métodos de imputación de datos para el dataset “Student Success”, graficar las distribuciones antes y después de la imputación, analizar los resultados, y elegir las técnicas más adecuadas según el caso.

Introducción

Desde sus inicios el ser humano ha sabido manipular y aprovechar los datos del mundo que le rodea para su propio beneficio, esto le ha permitido acumular mucho conocimiento en distintas áreas. Actualmente, ha cobrado aún más relevancia el saber manipular los datos de manera correcta tarea que, en ocasiones, no es tan sencilla pues regularmente se tiene que lidiar con valores atípicos, baja calidad de datos o con valores faltantes, siendo este último en el que se enfocará esta práctica.

No importa el área de estudio a la que pertenezcan los datos, es muy probable que por algún motivo pueda haber valores faltantes dentro de los datasets manipulados, ya sea por errores humanos, problemas técnicos, censura, etc. Esto podría llegar a ser un gran problema, pues puede afectar de manera negativa a los modelos de machine learning que se vayan a utilizar, pues causan sesgos, aumentan el error y disminuyen la capacidad para realizar predicciones.[2] Para disminuir las consecuencias negativas que la falta de datos puede conllevar, se han desarrollado varias técnicas de “imputación”, las cuales permiten rellenar los valores faltantes con nuevos datos que permitan aprovechar toda la información disponible y evitar la pérdida de registros valiosos. [2]

En la presente práctica se aplican algunas técnicas de imputación para el dataset “Student Success”.

Marco Teórico

Cómo se mencionó anteriormente, los métodos de imputación permiten rellenar los valores faltantes de un dataset. Existen muchas técnicas de imputación, desde aquellas que se limitan a agregar valores al azar, hasta las que crean modelos que permiten recrear con mayor grado de verosimilitud los datos ausentes. A continuación, se mencionan algunos de los más relevantes:

a) Imputación por media: Reemplaza los valores faltantes por la media de los valores observados en el mismo atributo. [1]
b) Imputación por media de clases: Es una variante de la imputación anterior, con la diferencia de que si las instancias están etiquetadas el valor de la media se calcula por cada clase y se reemplaza el valor según la clase a la que pertenezca. [1]
c) Imputación por moda: Reemplaza los valores faltantes por la moda de los valores del atributo, se usa especialmente para datos categóricos. [2]
d) Imputación hot-deck: Copia los datos faltantes de una observación similar en el mismo conjunto de datos. [2]

e) Imputación por regresión: Estima los datos faltantes usando una regresión lineal, logística o polinomial, con las otras variables como predictores. [2]
f) Imputación aleatoria: Se reemplazan los datos con números aleatorios que estén dentro del rango del mínimo y máximo encontrados en el atributo. [2]

En la presente práctica se usarán la imputación por media, por media de clases, por moda, aleatoria y por regresión lineal.

Materiales y Métodos

En el presente trabajo se utilizó el dataset 'Student Success'. Por su parte, para el manejo de los datos e imputación, se recurrió a la librería Pandas y Numpy. Por último, para la creación de las gráficas se usó Plotly.

Resultados

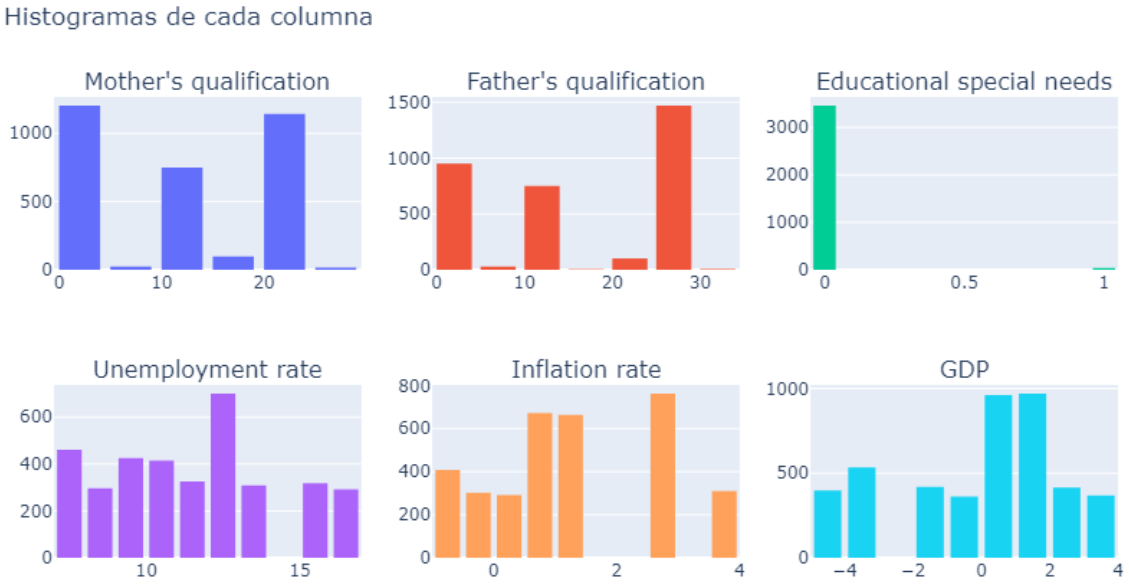
El dataset utilizado cuenta con 35 atributos sin datos faltantes, por lo que, para poder realizar la práctica de imputación, se procedió primero a establecer las columnas con las que se trabajarían, que fueron las siguientes:

Atributo	Tipo de valor
Mother's qualification	Discreto
Father's qualification	Discreto
Educational special needs	Binario
Unemployment rate	Continuo
Inflation rate	Continuo
GDP	Continuo
Target	Categórico

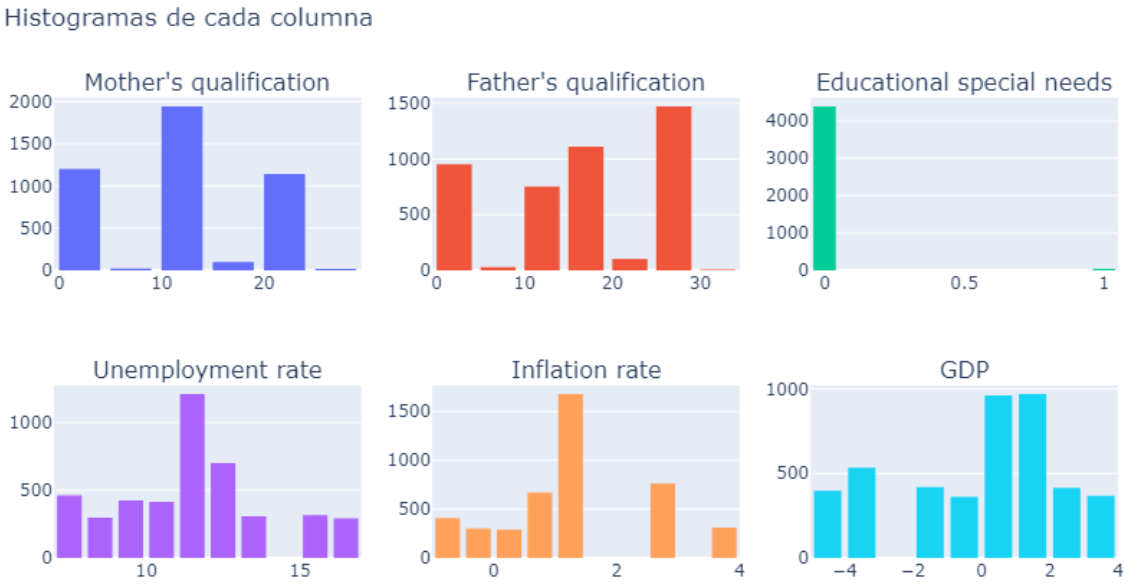
Posteriormente se creó una función para quitar un porcentaje específico de datos de manera aleatoria, esto sólo para las columnas *Mother's qualification*, *Father's qualification, Educational special needs, Unemployment rate, Inflation rate*. Por su parte, las columnas *GDP* y *Target*, se usaron como referencia para la imputación de regresión lineal y de media por clases respectivamente.

Hecho esto se procedió a graficar las distribuciones de los datos sin imputación, y finalmente a realizar las imputaciones e igualmente graficar las distribuciones obtenidas por los distintos métodos.

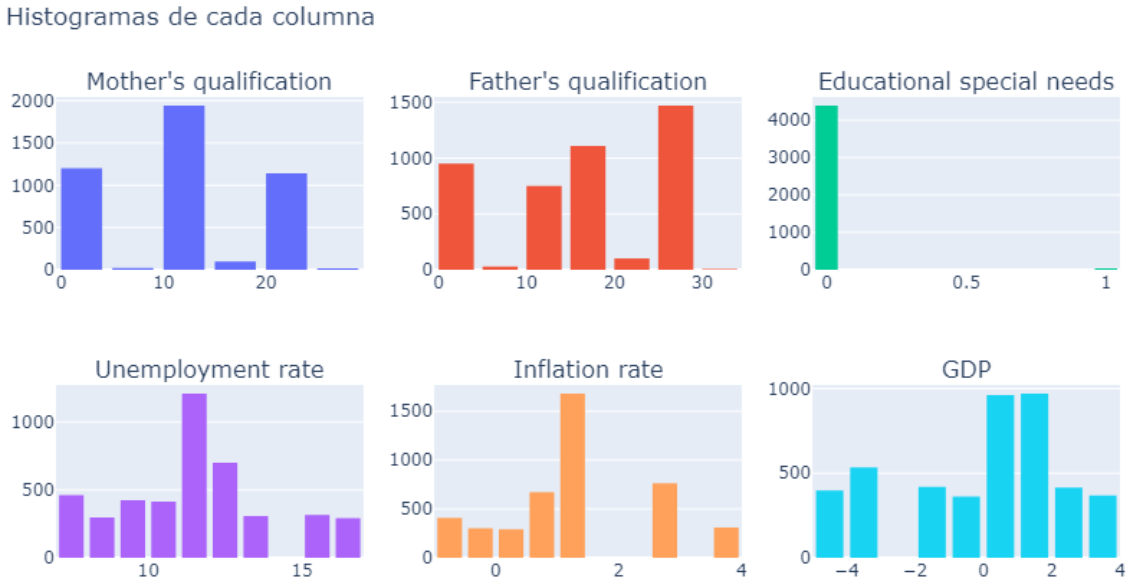
La distribución de los datos sin imputación se muestra en la siguiente página:



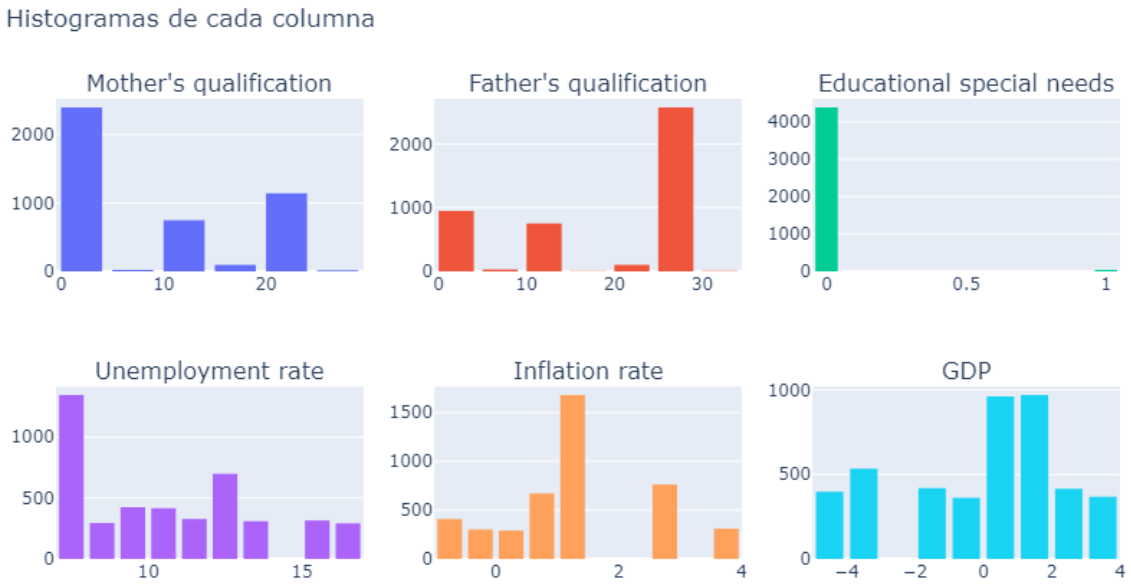
Por su parte al realizar la imputación por media se obtienen las siguientes distribuciones:



Para la imputación de media por clases los resultados son:

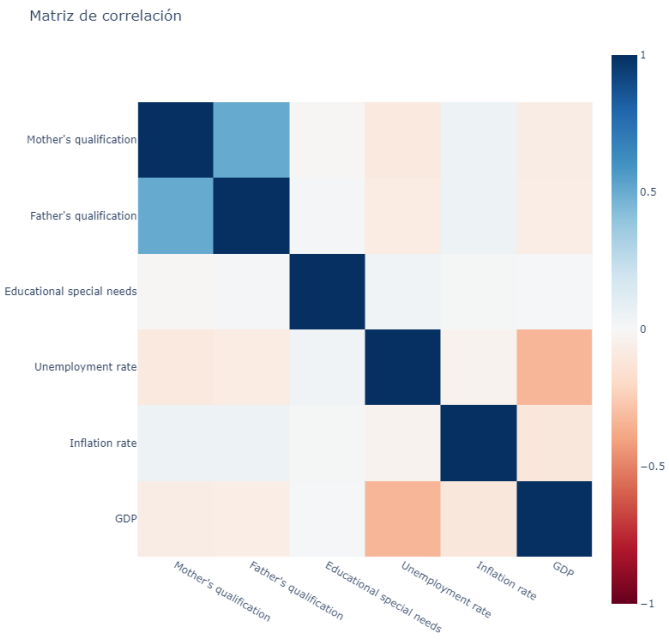


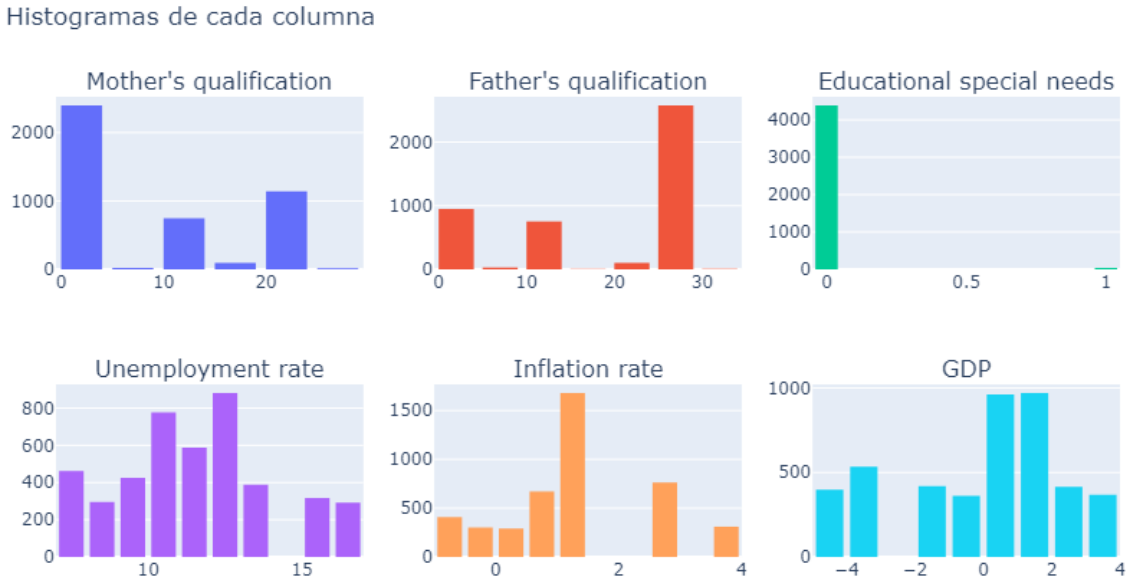
Las distribuciones al imputar por moda:



En la siguiente gráfica se muestran las distribuciones al imputar aleatoriamente (para las columnas Mother's qualification, Father's qualification, Educational special needs, Inflation rate) y al imputar por regresión lineal para la columna Unemployment rate. La razón de esta decisión es debido a que este último atributo parecía tener una correlación negativa con el atributo GDP, por lo que la regresión lineal parecía ser una mejor opción. A continuación se muestra el mapa

de calor para verificar lo dicho anteriormente, así como las distribuciones obtenidas al realizar estas imputaciones:





Por último, se realizó una imputación final, en donde los métodos que se usaron fueron aquellos que menores cambios hicieron en la distribución de los datos. Los métodos usados en este caso fueron imputación aleatoria para los atributos *Mother's qualification*, *Father's qualification, Educational special needs,* e imputación por regresión lineal para las columnas *Unemployment rate, Inflation rate*.

Conclusiones

Dados los resultados anteriores, se puede verificar que, al menos para este caso, el método que menos cambió las distribuciones para las columnas con valores discretos fue la imputación aleatoria, y para las columnas con valores continuos fue la imputación por regresión lineal. De este modo se puede observar que, si se realizan de manera correcta, las técnicas de imputación pueden ser de gran ayuda para poder aprovechar al máximo todos los datos que se tienen y crear modelos robustos de machine learning.

Bibliografía

[1] M. A. Fernandez, «Inteligencia Artificial para Programadores con Prisa,» 2021.

[2] C. Bits, «Guía completa para el Manejo de Datos Faltantes,» [En línea]. Available: https://www.codificandobits.com/blog/manejo-datos-faltantes/. [Último acceso: 02 marzo 2023].