Normalización y Creación de Datos Sintéticos

Objetivo

Implementar en una base de datos los métodos de normalización que se consideren más adecuados y llevar a cabo un aumento de datos (datos sintéticos), para balancear las clases existentes.

Introducción

En el machine learning, los datos son el componente clave que permite que los modelos de inteligencia artificial aprendan los patrones necesarios para realizar una tarea específica; el correcto manejo ellos, puede hacer que el rendimiento de un modelo se incremente significativamente. Hay muchas prácticas y herramientas útiles que permiten incrementar la utilidad de los datos que se poseen como la "normalización", o incluso que son capaces de generar nuevos datos como lo es el *"data augmentation"*.

En lo que concierne a la normalización se puede decir, en términos resumidos, que es un proceso que consiste en transformar los valores de los atributos numéricos de un dataset con el fin de que tengan una escala común. Aunque parece algo muy sencillo, en realidad es un proceso clave para mejorar el desempeño de los modelos que se construyan, esto por dos motivos principales: en primer lugar, de esta manera se reduce la variabilidad y la dispersión de los datos, lo que facilita el aprendizaje de los algoritmos; y en segundo lugar, evita que los valores de un atributo dominen sobre otro, ya sea porque estén en diferentes escalas, o porque se estén manejando diferentes rangos entre los atributos. Por último, en este punto, cabe resaltar que existen muchas técnicas de normalización, la elección de la más adecuada dependerá del tipo de datos y del problema que se quiera resolver. [1]

Por otra parte, en lo que respecta al *"data augmentation"* (también llamado "datos sintéticos"), son un conjunto de técnicas que permiten aumentar la cantidad de datos a partir de los que ya se poseen. Esto es útil principalmente cuando los datos son escasos o costosos, incluso cuando se requiere mantener la confidencialidad. Si se utilizan los métodos correctos, es una herramienta muy poderosa que aumentará la cantidad de ejemplos para el entrenamiento del modelo y de esta manera mejorar su rendimiento.

Marco Teórico

Como se mencionó anteriormente, existen varios métodos tanto de normalización, como de creación de datos sintéticos, por lo que conviene dar un breve repaso de algunos de ellos. Algunas técnicas de normalización son:

a) Normalización min-max: Es uno de los métodos más comunes. La fórmula simple normaliza los datos en un rango entre 0 y 1, sin embargo, existe una fórmula general que permite escalar en cualquier rango deseado. La fórmula general se muestra a continuación [1]:

$$\dot{x} = a + \frac{x - \min(x)}{\max(x) - \min(x)} * (b - a) \qquad (1)$$

b) Normalización z-score: Este método escala los valores de tal forma que la media de todos los datos sea igual a 0 y la desviación estándar sea 1. Su ecuación es la siguiente [1]:

$$\dot{x} = \frac{x - \mu}{\sigma} \qquad (2)$$

c) Normalización por medias: Es similar a z-score, pero en este caso el denominador cambia como se muestra en seguida [1]:

$$\dot{x} = \frac{x - \mu}{\max(x) - \min(x)} \qquad (3)$$

d) Normalización por vector unitario: Esta técnica consiste en dividir cada entrada de un vector por su magnitud para crear un vector de longitud 1 conocido como vector unitario. Se usa la siguiente ecuación [1]:

$$\dot{x} = \frac{x}{[\![x]\!]} \qquad (4)$$

e) Normalización sigmoide: Consiste en aplicar una función sigmoide a cada valor de un conjunto de datos. Se sigue la siguiente fórmula:

$$\dot{x} = \frac{1}{1 + e^{-(\frac{x - \bar{x}}{\sigma})}} \qquad (4)$$

Por su parte, también existen muchas técnicas para la creación de datos sintéticos, entre ellas destacan Adasyn, SMOTE y el método de la ruleta. En la presente práctica sólo se implementa esta última.

Materiales y Métodos

Para el desarrollo de este trabajo se utilizó el dataset 'Student Success'. Por su parte, para el manejo de los datos y la creación de funciones de normalización y *'data augmentation'*, se recurrió a las librerías Pandas y Numpy. Por último, para la creación de las gráficas se usó Plotly.

Resultados

En el presente documento se muestran los resultados de implementar algunas técnicas de normalización y de creación de datos sintéticos. En primer lugar, se mostrará lo que se hizo en la normalización y posteriormente lo que se obtuvo en el *'data augmentation'*.

Normalización

El dataset que se usó cuenta con 35 atributos sin datos faltantes. La mayoría de los atributos poseen tipos de datos numéricos discretos, esto es, que cada número tiene un significado categórico, el cual cambia según el atributo. Por ejemplo, en una de las columnas llamada

“Marital Status”, se encuentran valores del 1 al 6, el significado de cada número se puede revisar en la siguiente tabla:

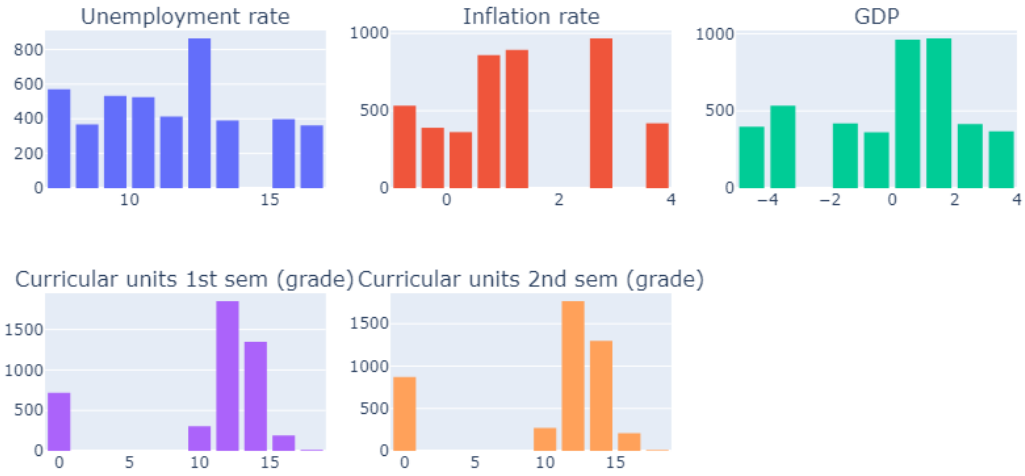
Codificación del atributo “Marital Status”	
Número	**Significado**
1	Soltero
2	Casado
3	Viudo
4	Divorciado
5	Unión Libre
6	Separado Legalmente

Tabla 1. Codificación del atributo “Marital Status”

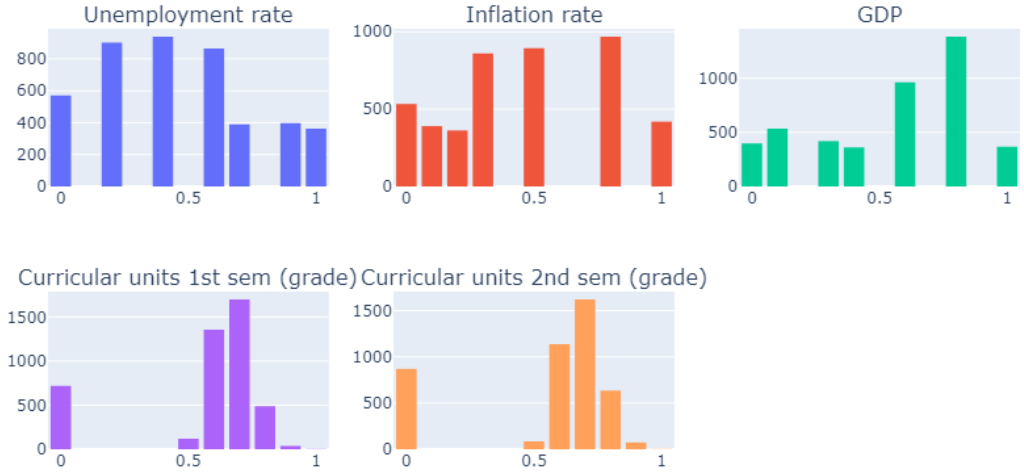
De esta manera se encuentran codificados muchos atributos del dataset, debido a esto, se consideró que no era adecuado aplicar normalización a todo el conjunto de datos, sino únicamente a aquellos atributos que poseían datos numéricos continuos, los cuales son: *‘Unemployment rate’, ‘Inflation rate’, ‘GDP’, ‘Curricular units 1st semester grade’, ‘Curricular units 2nd semester grade’*.

Para normalizar, se decidió aplicar 5 métodos de normalización distintos a estos atributos, y posteriormente comparar cuáles conservaban de mejor manera la distribución de estos. Las técnicas aplicadas fueron las vistas en el marco teórico, normalización min-max, z-score, por medias, por vector unitario y sigmoide. Las distribuciones resultantes se muestran en seguida:

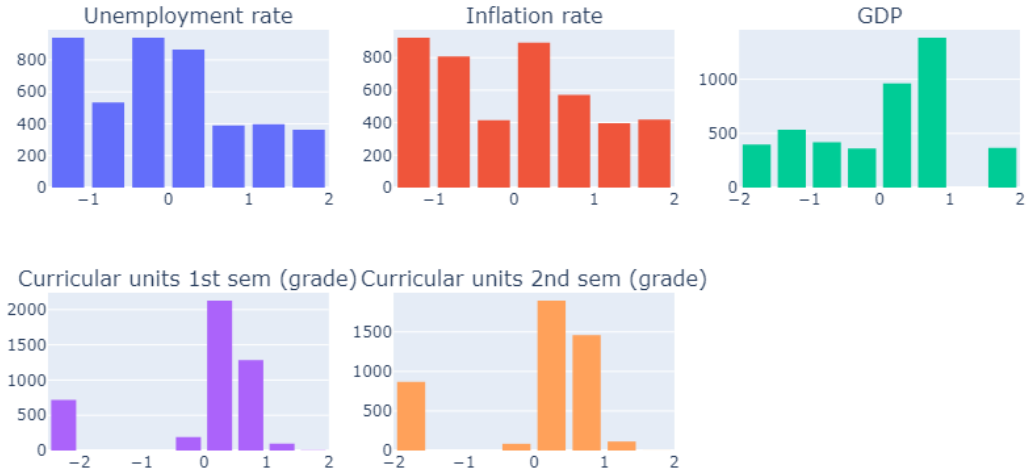
a) Distribución original de los atributos:



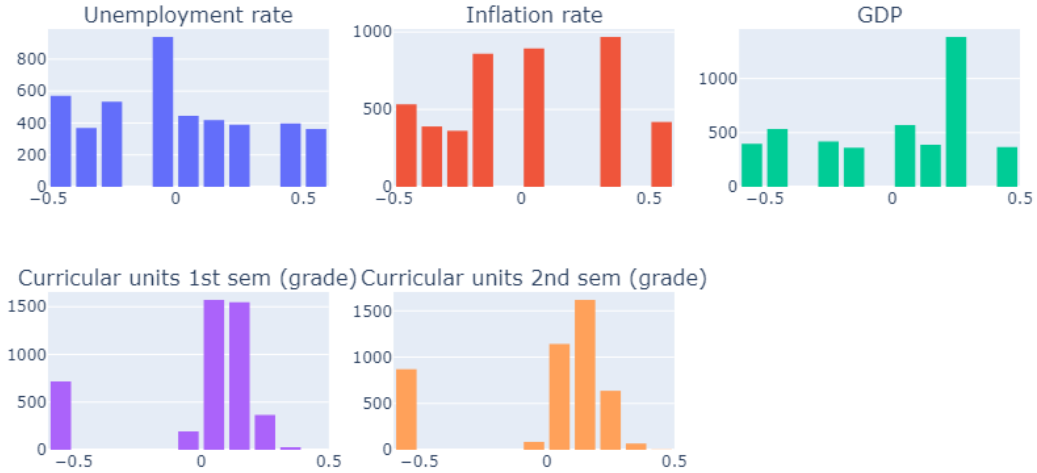
b) Distribución al aplicar normalización min-max:



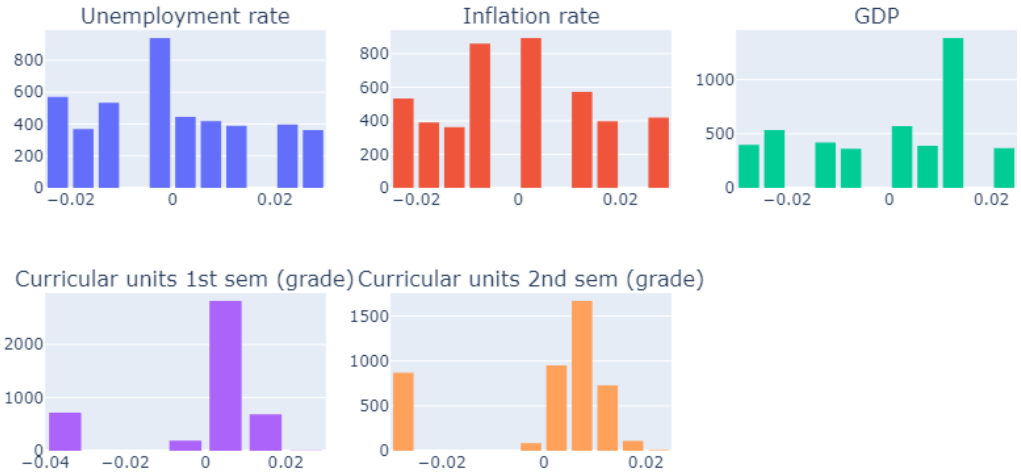
c) Distribución al aplicar normalización z-score:



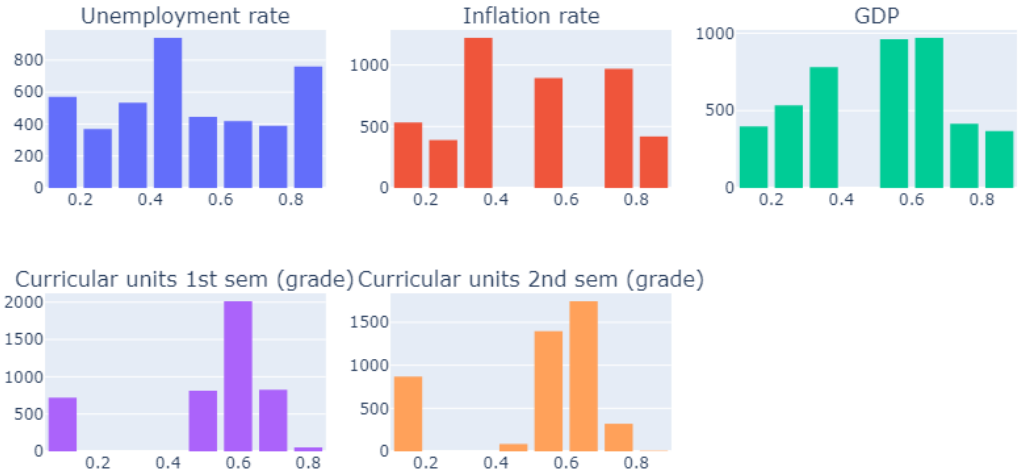
d) Distribución al aplicar normalización por medias:



e) Distribución al aplicar normalización por vector unitario:



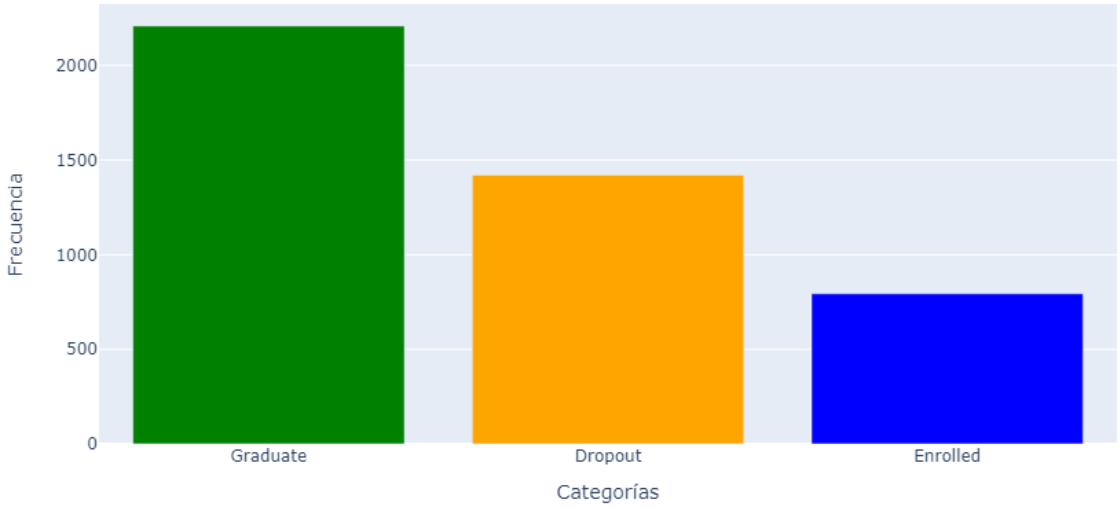
f) Distribución al aplicar normalización sigmoide:



Como se puede observar, cada método de normalización realiza algunos cambios en la distribución en mayor o menor medida. Analizando el primer atributo *'Unemployment rate'*, se observa que realmente presenta grandes cambios en su distribución al aplicar cualquier método, siendo z-score el que más le afecta y la normalización por medias y vector unitario los que menos. El atributo *'Inflation rate'*, por otra parte, conserva su distribución aproximada al aplicar normalización min-max y por medias, y se modifica bastante al aplicar las otras. De esta manera se puede ir analizando cada atributo con cada método. Al analizar los resultados se concluyó que las mejores normalizaciones para los tres atributos restantes *'GDP', 'Curricular units 1st'* y *'Curricular units 2nd'*, eran z-score, que es la que mejor conserva sus distribuciones.

Datos Sintéticos

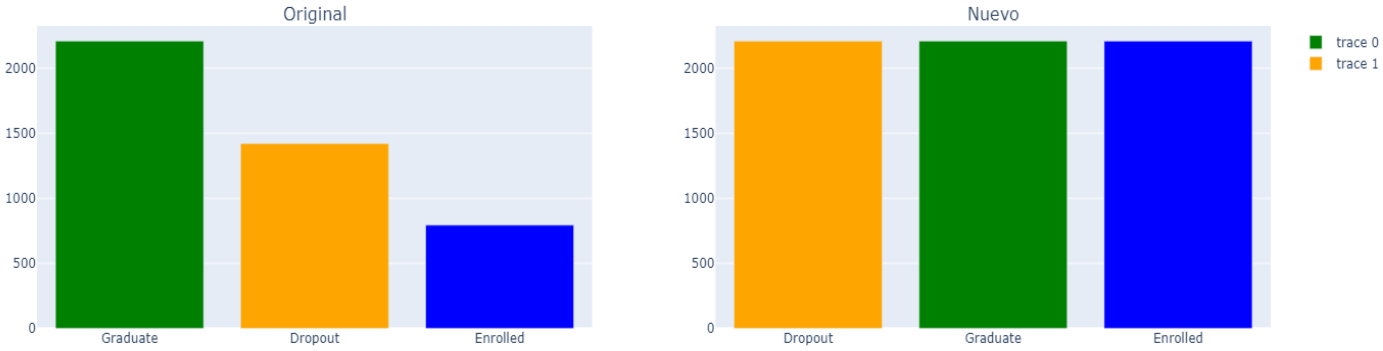
Por último, también se realizó una práctica para crear datos sintéticos en el mismo dataset; esto con el fin de la balancear las clases existentes, pues como se puede observar a continuación, existía un gran desbalance entre las clases:



Para esta sección se usó el método de la ruleta, esta técnica consiste básicamente en elegir valores al azar entre los que ya existen en ese atributo para una de las clases, sin embargo, la peculiaridad de este método es que los valores que aparecen en más ocasiones tienen mayor probabilidad de ser elegidos.

Para analizar los resultados, se procedió a comparar las distribuciones y desviaciones estándar que tenía cada atributo antes y después de aplicar la técnica.

a) Clases correctamente balanceadas:



b) Distribuciones originales:

c) Distribuciones con datos sintéticos:

d) Desviaciones estándar:

Columna	Dataset Original	Dataset Nuevo
Marital status	0.60574695	0.64703662
Application mode	5.29896372	5.31106675
Application order	1.31379308	1.27191823
Course	4.33179197	4.36509639
Daytime/evening attendance	0.31189668	0.31460437
Previous qualification	3.96370695	3.9700478
Nacionality	1.74844717	1.8124781
Mother's qualification	9.02625104	9.00129601
Father's qualification	11.0447995	11.0592583
Mother's occupation	3.99782771	4.33326913
Father's occupation	4.85669227	5.39189948
Displaced	0.49771085	0.49858377
Educational special needs	0.10676006	0.10785756
Debtor	0.31748001	0.33806227
Tuition fees up to date	0.32423538	0.33491045
Gender	0.47756044	0.48434881
Scholarship holder	0.43214415	0.40812429
Age at enrollment	7.58781562	7.57620428
International	0.15572932	0.15628587
Curricular units 1st sem (credited)	2.36050662	2.22920913
Curricular units 1st sem (enrolled)	2.48017818	2.40082422
Curricular units 1st sem (evaluations)	4.17910557	4.17379727
Curricular units 1st sem (approved)	3.09423798	3.00317173
Curricular units 1st sem (grade)	4.84366338	4.89106051
Curricular units 1st sem (without evaluations)	0.69088018	0.70861012
Curricular units 2nd sem (credited)	1.91854614	1.80930407
Curricular units 2nd sem (enrolled)	2.19595075	2.11790867
Curricular units 2nd sem (evaluations)	3.94795094	4.02991543
Curricular units 2nd sem (approved)	3.0147639	2.92578075
Curricular units 2nd sem (grade)	5.21080795	5.27056703
Curricular units 2nd sem (without evaluations)	0.75377407	0.77787666
Unemployment rate	2.66385048	2.65630466
Inflation rate	1.38271069	1.38053315
GDP	2.26993544	2.28188404

Hecho esto, se puede observar claramente, que la creación de datos sintéticos se ha hecho de manera correcta, pues la distribución de los datos no ha cambiado y sus desviaciones estándar no presentan grandes variaciones que puedan afectar.

Conclusiones

En esta práctica se aplicaron técnicas de normalización de datos y de creación de datos sintéticos. En los resultados de los ejercicios de normalización, se puede observar que ciertas técnicas con ciertos atributos logran cambiar la distribución de los datos significativamente, para la mayoría de los atributos (tres de ellos) la mejor normalización fue el z-score, pero en otros dos resultaron de mayor ayuda la normalización por medias, vector unitario y min-max. Por su parte, en la creación de datos sintéticos, la técnica de la ruleta funcionó muy bien para este dataset, conservando casi de la misma forma las distribuciones de todos los atributos y las desviaciones estándar originales.

Bibliografía

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