Análisis del Dataset ‘Student Succes’

Objetivo

Realizar un análisis básico del dataset ‘Student Success’ de tal forma que se identifiquen las medidas y características más importantes de los datos que contiene; e igualmente realizar 10 tipos de gráficas diferentes con dos librerías distintas de Python.

Introducción

Uno de los aspectos más importantes en una sociedad, es el nivel educativo de su gente, este es vital, pues un buen nivel educativo se traduce en una gran calidad de vida, trabajos justos y crecimiento económico. Por desgracia, esto no siempre es posible debido a, entre otros factores, problemas socioeconómicos, macroeconómicos, demográficos o simple dificultad en los asuntos académicos que la persona presente. El presente dataset con el que se trabajará tiene el objetivo, por tanto, de poder predecir el nivel de éxito que un estudiante puede tener según los factores que lo rodean. [3]

La base de datos tiene por nombre ‘Student Success’ y contiene información obtenida de diversas instituciones de educación superior de Portugal. El archivo está compuesto de 35 atributos y 4424 instancias correspondientes cada una a un estudiante. Entre los atributos se encuentran datos demográficos, socioeconómicos, macroeconómicos y de trayectoria académica. [3]

Por su parte, el análisis que se hizo tiene como objetivo conocer las siguientes características de los datos: número de atributos, número de instancias, valores mínimo y máximo de cada atributo, media, desviación estándar, correlación de Pearson, número de datos faltantes, número de *outliers*, balance de clases y distribución.

Por último, igualmente se hacen gráficas con dos librerías diferentes de Python para visualizar de mejor manera las características del dataset.

Marco Teórico

A continuación, se dará un breve repaso de cada una de las características y medidas que se calcularán para el dataset.

a) Atributos: Se refiere a cada una de las características o propiedades de los datos recopilados. En un dataframe es equivalente a las columnas. [1]
b) Instancias: Se refiere a cada uno de los registros individuales u observaciones hechas. En un dataframe es equivalente a las filas. [1]
c) Mínimo: Es el valor más pequeño registrado en un atributo.
d) Máximo: Es el valor más grande registrado en un atributo.
e) Media: También se conoce como promedio. Es una medida estadística que se utiliza para obtener el valor central de un conjunto de datos. Se obtiene mediante la siguiente ecuación: [2]

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

f) Desviación Estándar: Es una medida de dispersión que indica cuánto se desvían los valores de un conjunto de datos respecto a la media. Una desviación estándar alta indica que los valores están más dispersos, mientras que una desviación estándar baja indica que los valores están más cercanos a la media. [2]

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

g) Correlación de Pearson: Es una medida estadística que evalúa la relación lineal entre dos variables. Se utiliza para determinar si existe una relación entre dos variables y qué tan intensa es. Sus valores van de -1 a 1, siendo -1 para una correlación negativa, 1 para una correlación positiva y 0 para la ausencia de esta. [2]

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

h) *Outliers*: Este concepto se refiere a los valores que difieren significativamente del resto de los datos. También se les llama 'valores atípicos'. Se puede calcular su número estableciendo límites con base en el primer cuartil y el tercer cuartil, de tal forma que los datos fuera de este rango se considerarían outliers. Las fórmulas para establecer los límites serían las siguientes, siendo Q1 el cuartil uno, Q3 el cuartil tres e IQR que es la diferencia entre Q1 y Q3. [1]

$$L_{inf} = Q1 - 1.5 * IQR$$
$$L_{sup} = Q1 + 1.5 * IQR$$

i) Distribución: Es la forma en la que se distribuyen los valores para un atributo. Puede tener muchas formas, normal, unimodal, exponencial, etc. [1]

Materiales y Métodos

En el presente trabajo se utilizó el dataset 'Student Success'. Por su parte, para el manejo de los datos y cálculo de las medidas básicas de análisis, se recurrió a la librería Pandas y Numpy. Por último, para la creación de las gráficas se usaron las librerías Matplotlib y Plotly.

Resultados

1) Número de Instancias y Atributos del Dataset:

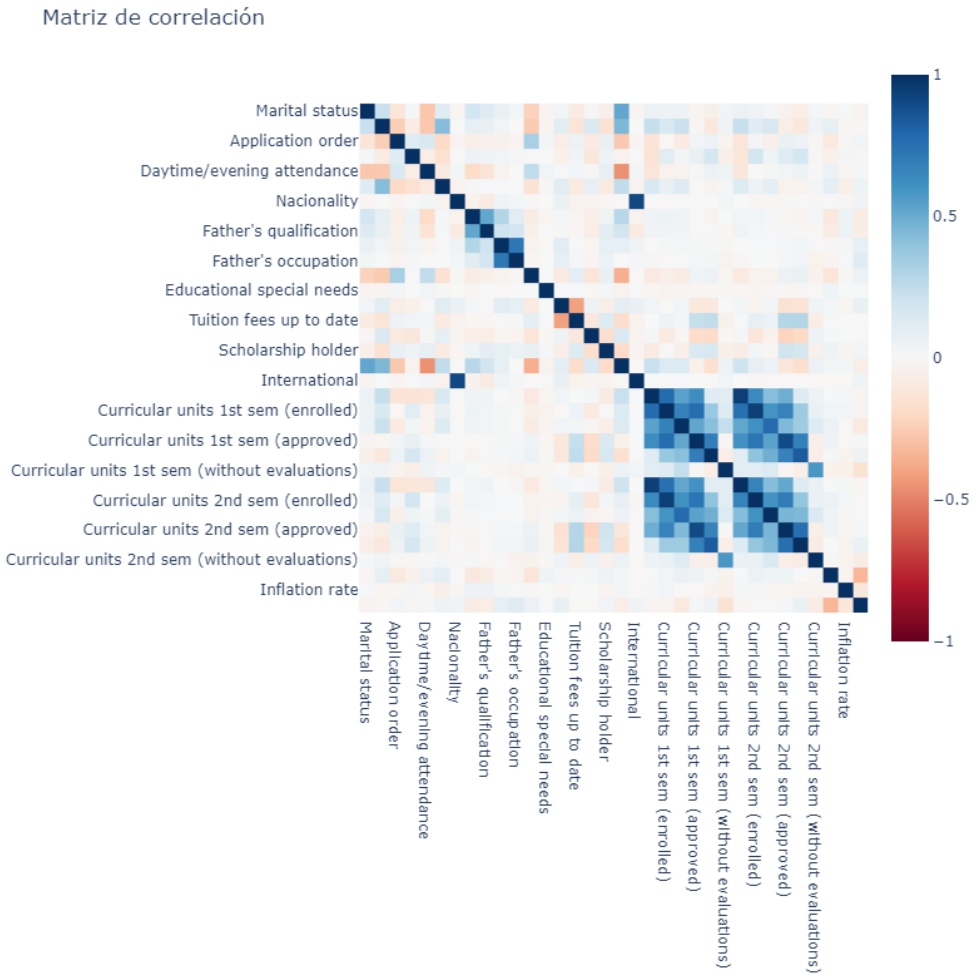
Número de Instancias	Número de Atributos
4424	35

2) Mínimo, máximo, media y desviación estándar de cada atributo:

	min	max	media	Desviación Estándar
Marital status	1	6	1.17857143	0.60574695
Application mode	1	18	6.88698011	5.29896372
Application order	0	9	1.7278481	1.31379308
Course	1	17	9.89918626	4.33179197
Daytime/evening attendance	0	1	0.89082278	0.31189668
Previous qualification	1	17	2.53141953	3.96370695
Nacionality	1	21	1.2545208	1.74844717
Mother's qualification	1	29	12.3221067	9.02625104
Father's qualification	1	34	16.4552441	11.0447995
Mother's occupation	1	32	7.31781193	3.99782771
Father's occupation	1	46	7.81916817	4.85669227
Displaced	0	1	0.54837251	0.49771085
Educational special needs	0	1	0.01152803	0.10676006
Debtor	0	1	0.11369801	0.31748001
Tuition fees up to date	0	1	0.88065099	0.32423538
Gender	0	1	0.3517179	0.47756044
Scholarship holder	0	1	0.24841772	0.43214415
Age at enrollment	17	70	23.2651447	7.58781562
International	0	1	0.02486438	0.15572932
Curricular units 1st sem (credited)	0	20	0.70999096	2.36050662
Curricular units 1st sem (enrolled)	0	26	6.27056962	2.48017818
Curricular units 1st sem (evaluations)	0	45	8.29905063	4.17910557
Curricular units 1st sem (approved)	0	26	4.70660036	3.09423798
Curricular units 1st sem (grade)	0	18.875	10.6408216	4.84366338
Curricular units 1st sem (without evaluations)	0	12	0.13765823	0.69088018
Curricular units 2nd sem (credited)	0	19	0.54181736	1.91854614
Curricular units 2nd sem (enrolled)	0	23	6.23214286	2.19595075
Curricular units 2nd sem (evaluations)	0	33	8.06329114	3.94795094
Curricular units 2nd sem (approved)	0	20	4.4358047	3.0147639
Curricular units 2nd sem (grade)	0	18.5714286	10.2302057	5.21080795
Curricular units 2nd sem (without evaluations)	0	12	0.15031646	0.75377407
Unemployment rate	7.6	16.2	11.5661392	2.66385048

Inflation rate	-0.8	3.7	1.22802893	1.38271069
GDP	-4.06	3.51	0.00196881	2.26993544

3) Correlación de Pearson:
Debido a la gran cantidad de atributos, se muestran los valores de la correlación de Pearson codificados en un mapa de calor.



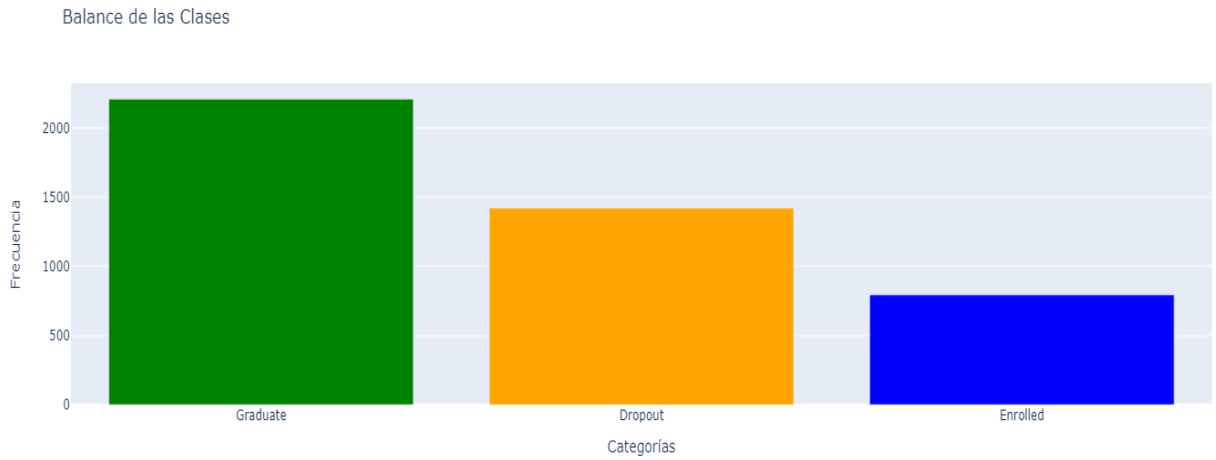
4) Número de datos faltantes: No hubo datos faltantes para ningún atributo.

5) Número de '*outliers*' por atributo:

	No. *Outliers*
Marital status	505
Application mode	0
Application order	541
Course	0
Daytime/evening attendance	483
Previous qualification	707
Nacionality	110

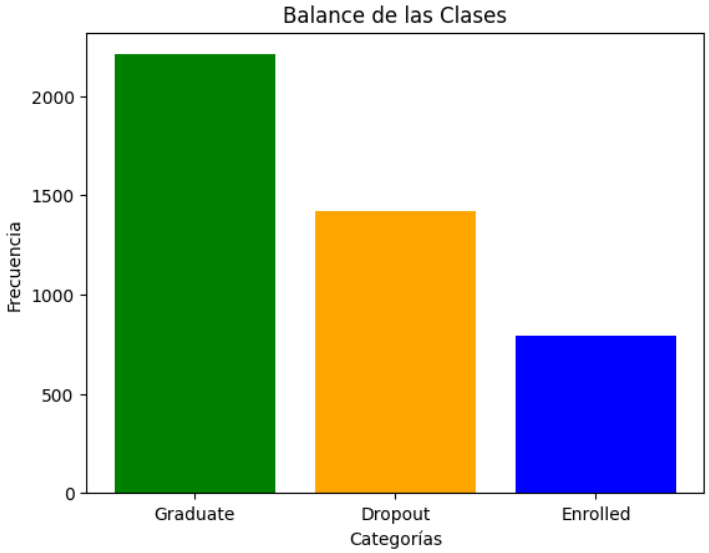
Mother's qualification	0
Father's qualification	0
Mother's occupation	84
Father's occupation	84
Displaced	0
Educational special needs	51
Debtor	503
Tuition fees up to date	528
Gender	0
Scholarship holder	1099
Age at enrollment	441
International	110
Curricular units 1st sem (credited)	577
Curricular units 1st sem (enrolled)	424
Curricular units 1st sem (evaluations)	158
Curricular units 1st sem (approved)	180
Curricular units 1st sem (grade)	726
Curricular units 1st sem (without evaluations)	294
Curricular units 2nd sem (credited)	530
Curricular units 2nd sem (enrolled)	369
Curricular units 2nd sem (evaluations)	109
Curricular units 2nd sem (approved)	44
Curricular units 2nd sem (grade)	877
Curricular units 2nd sem (without evaluations)	282
Unemployment rate	0
Inflation rate	0
GDP	0

6) Balance de clases:
Para visualizar el nivel de balance de las clases se recurrió a una gráfica de barras.

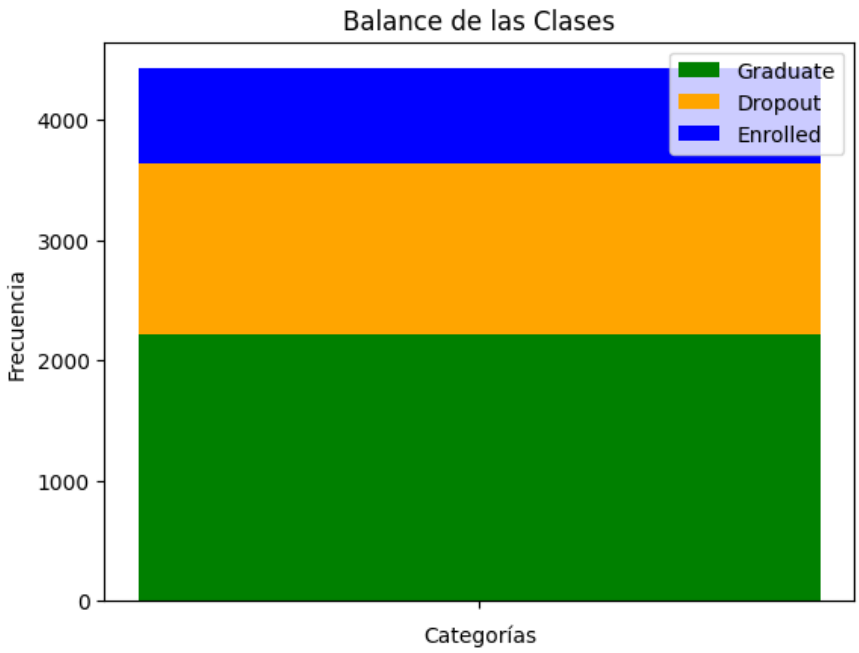


7) Distribución: Se usaron histogramas para graficar las distribuciones para cada atributo.

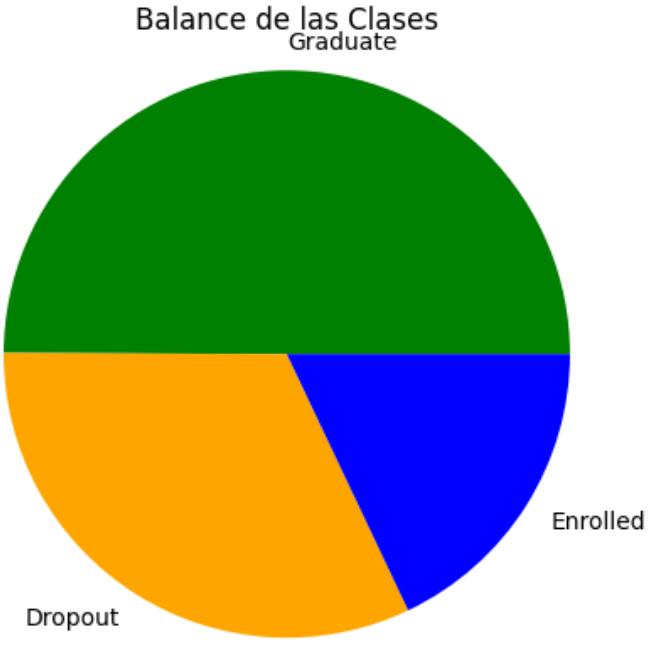
8) Gráficas con Matplotlib
   a. Gráfico de barras



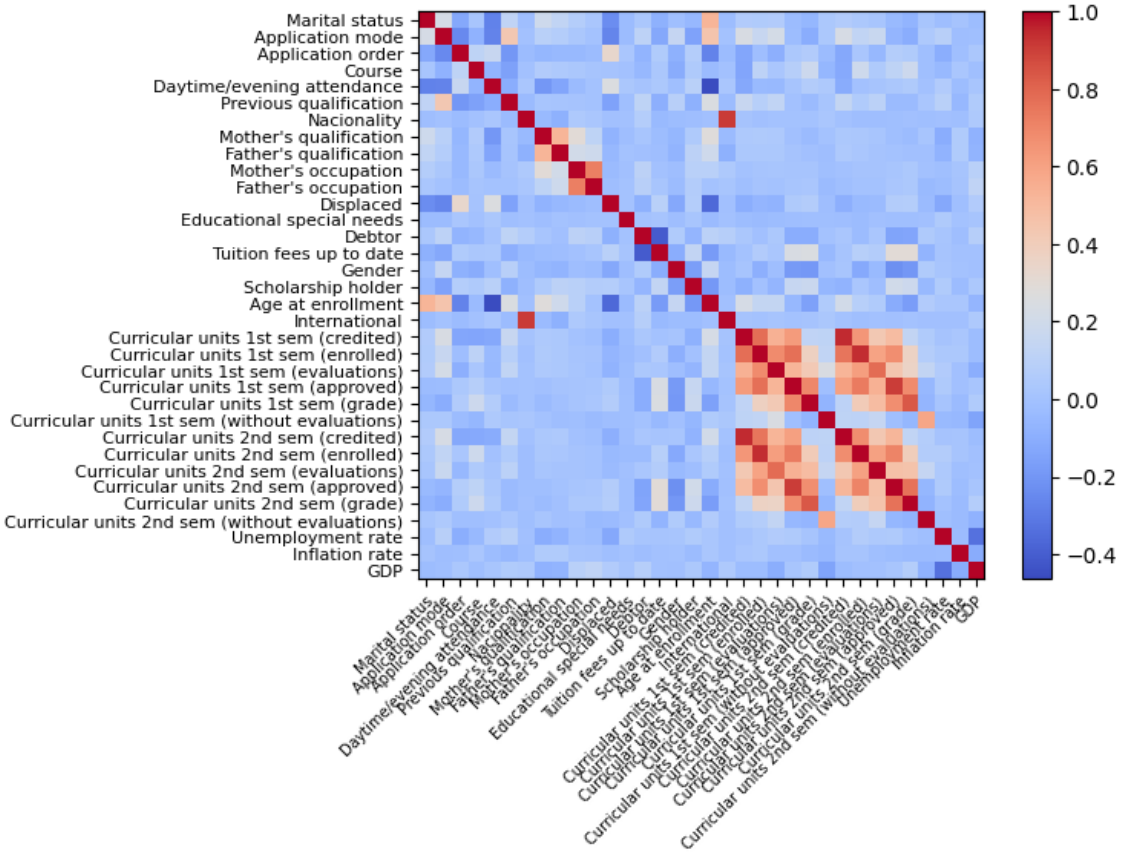
b. Gráfico de barras apiladas



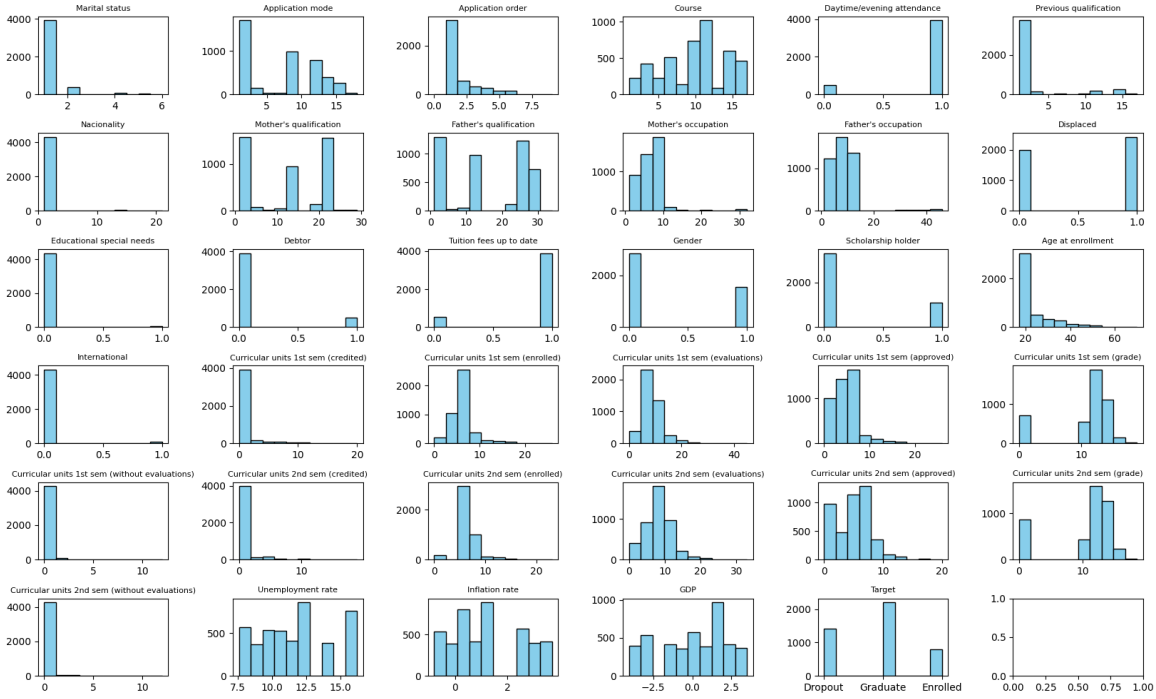
c. Gráfico de Pastel



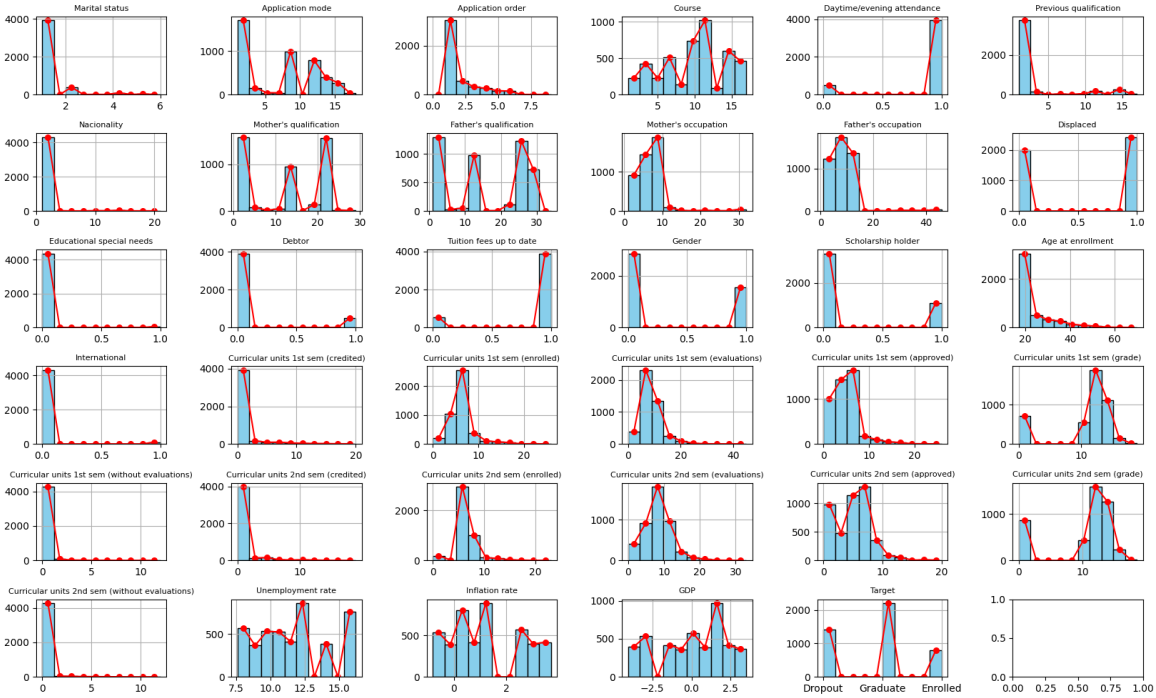
d. Mapa de calor



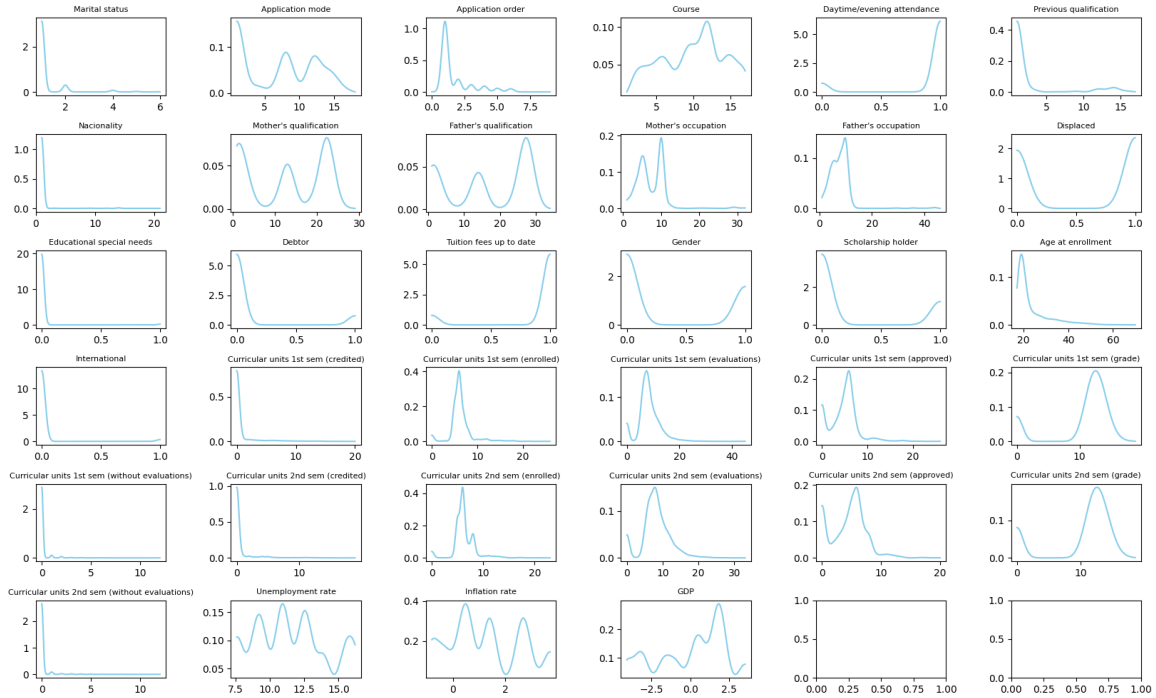
e. Histograma



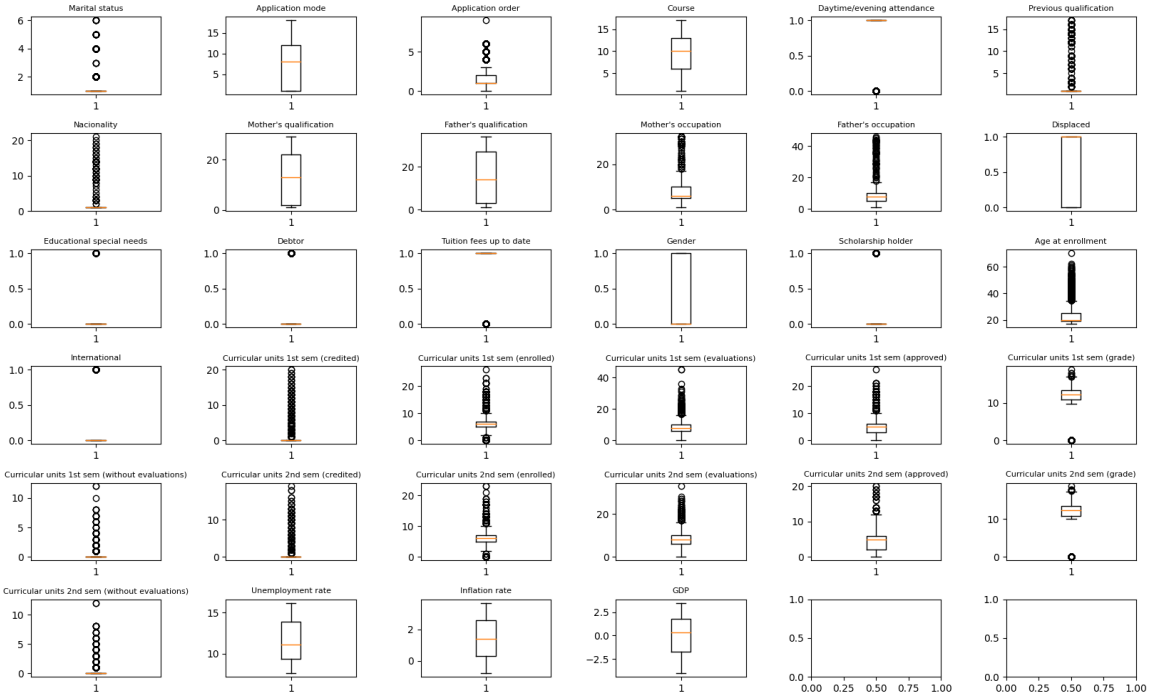
f. Histograma con líneas



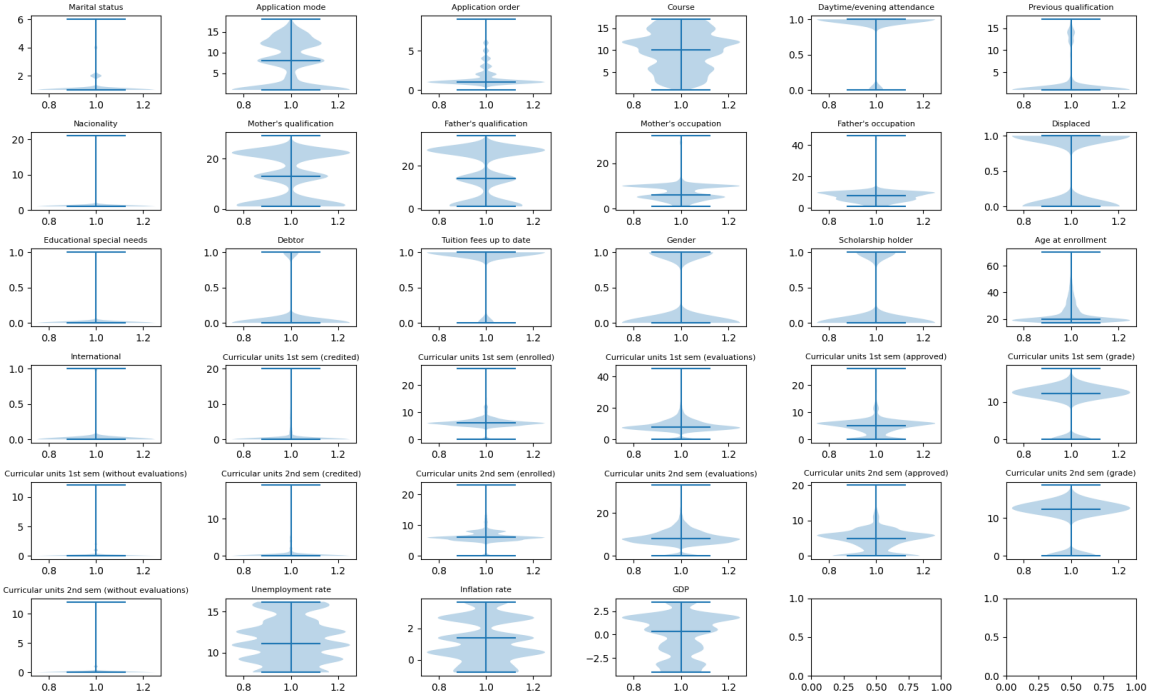
g. Gráfico de Densidad



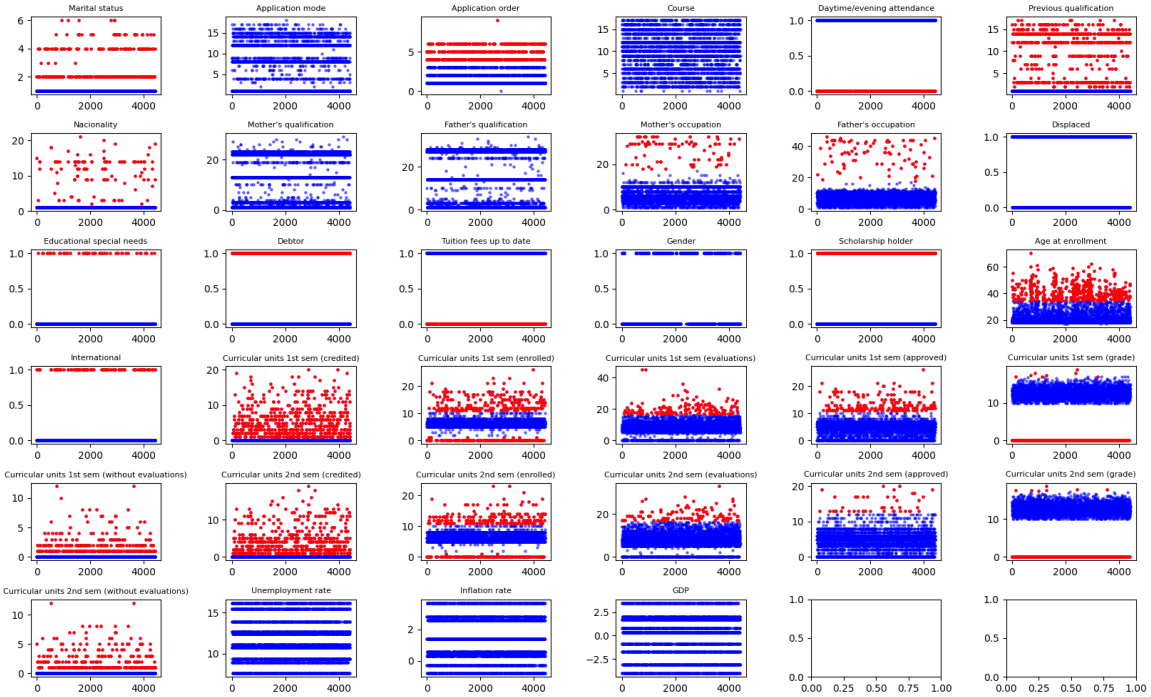
h. Boxplot



i. Gráfico de Violín

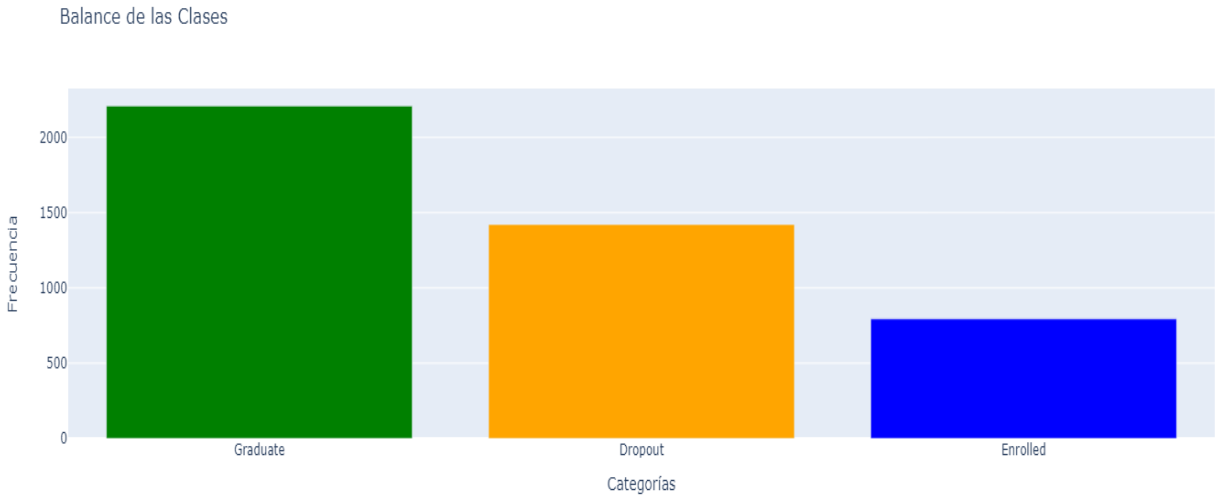


j. Gráfico de Dispersión

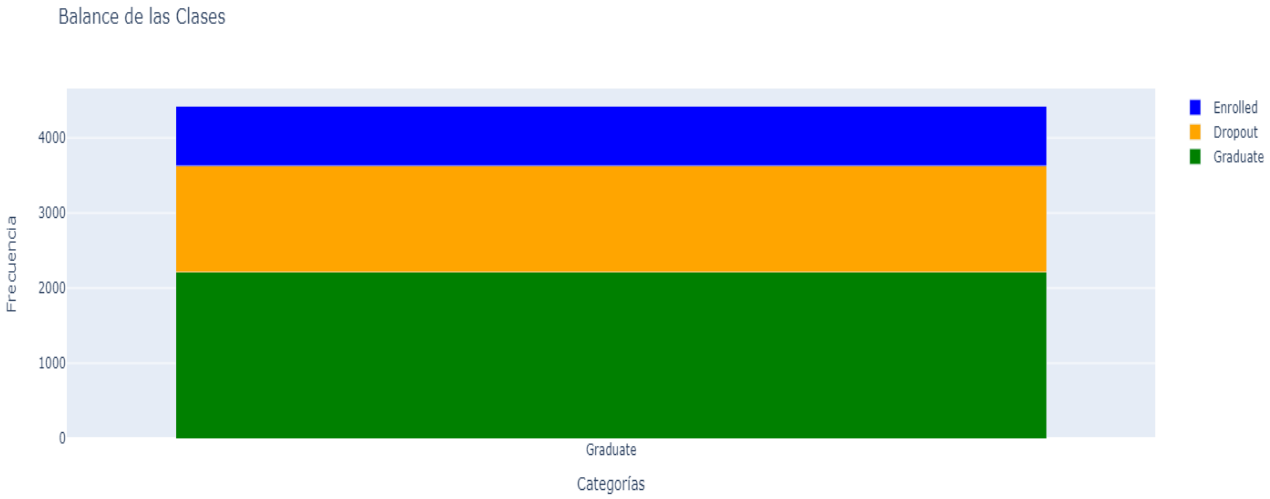


9) Gráficas con Plotly

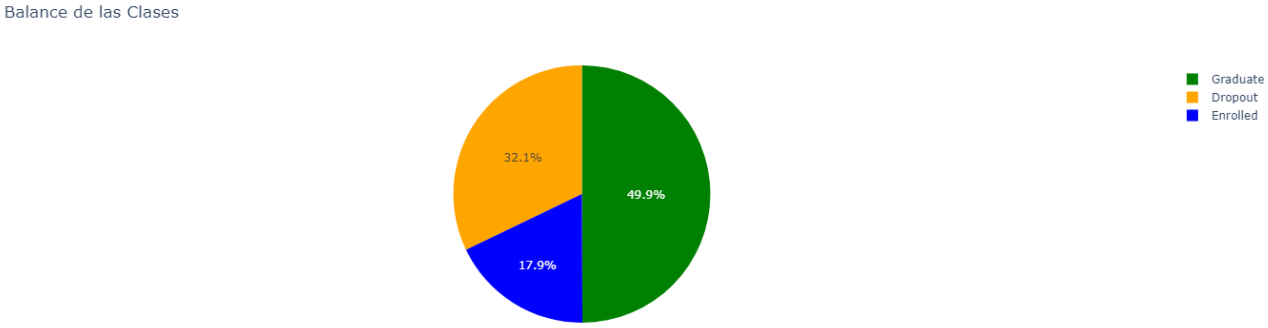
a. Gráfico de barras



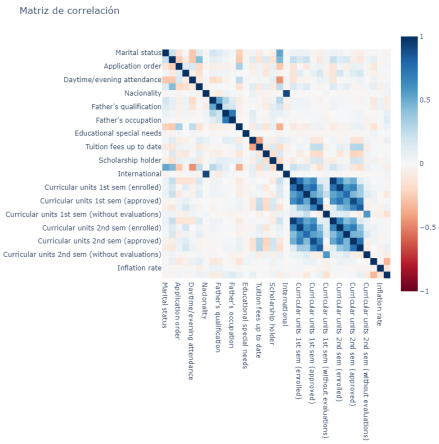
b. Gráfico de barras apiladas



c. Gráfico de Pastel

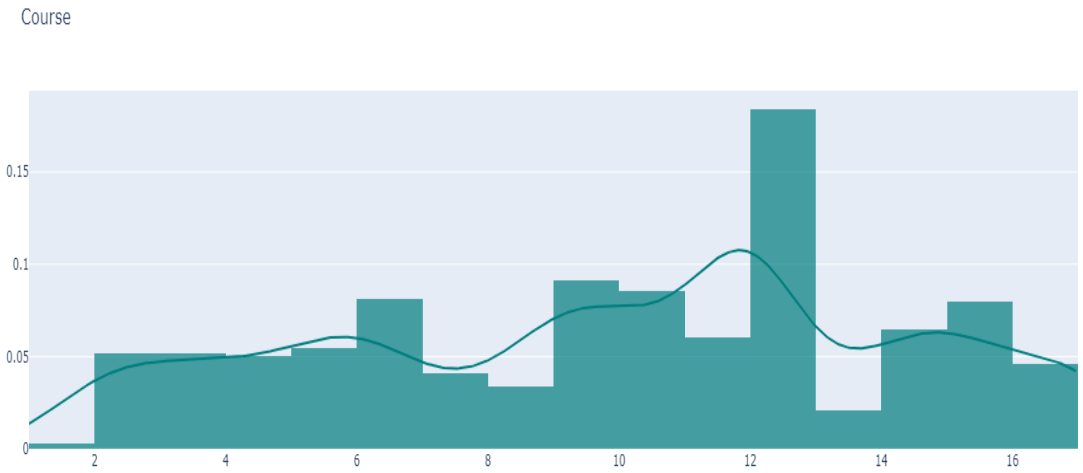


d. Mapa de calor

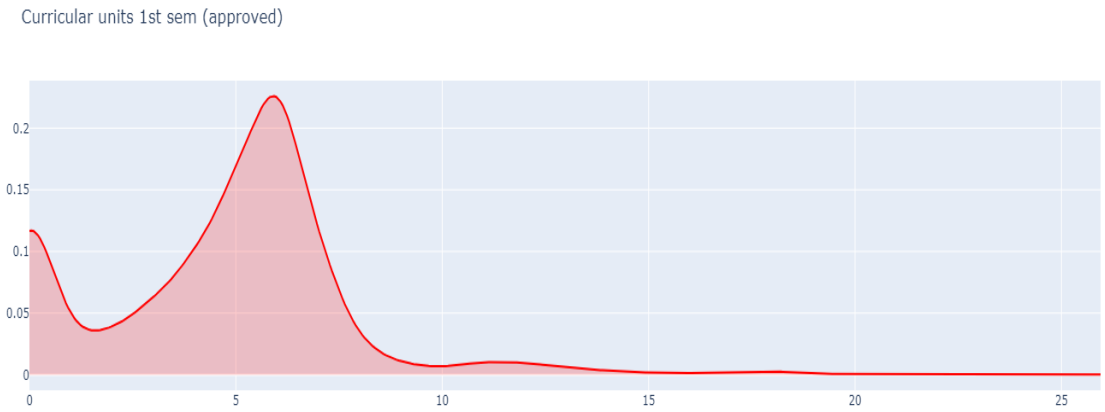


e. Histograma

f. Histograma con densidad

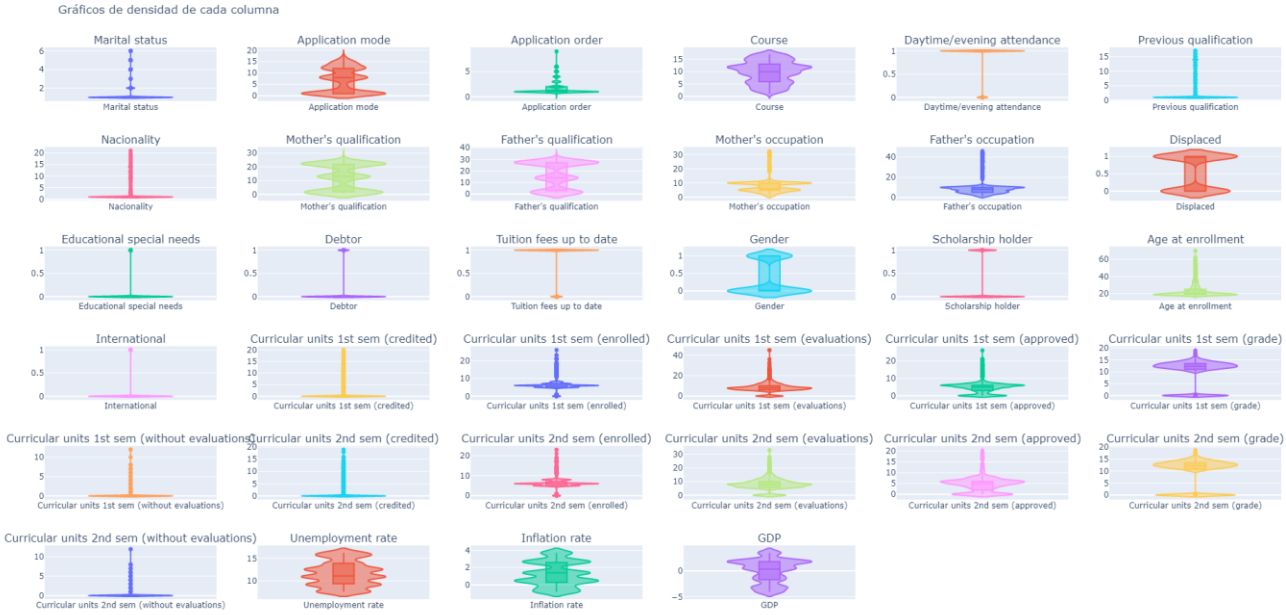


g. Gráfico de Densidad

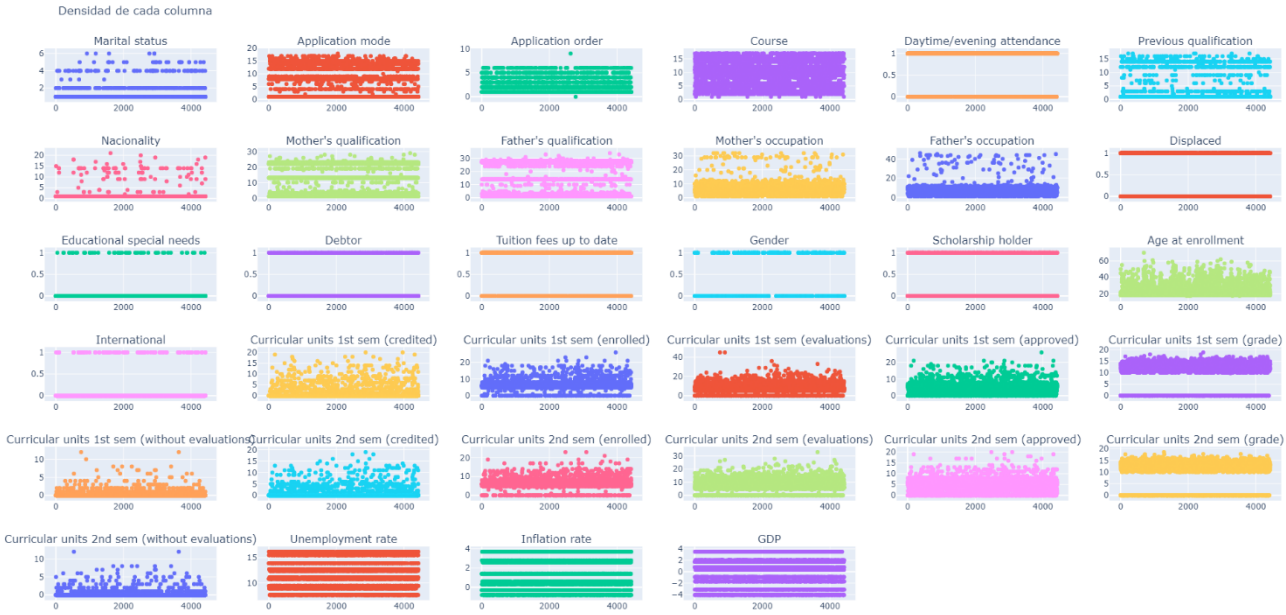


h. Boxplot

i. Gráfico de Violín



j. Gráfico de Dispersión



Conclusiones

Como se puede observar, analizar correctamente el conjunto de datos con el que se trabajará, siempre va a ser una parte importante antes de la creación de un modelo, principalmente porque ayudará a tomar mejores decisiones con respecto a la manera en que se manejarán los datos faltantes y las técnicas de machine learning que se usarán. El análisis puede revelar patrones, tendencias y relaciones entre los datos, lo cual contribuirá a llegar a conclusiones bien respaldadas. Es gracias a un buen manejo y análisis de los datos, que las empresas más exitosas logran identificar problemas y áreas de mejora.

Bibliografía

[1] M. A. Aceves Fernandez, Inteligencia Artificial para Programadores con Prisa. Independently Published, 2021.

[2] W. Mendenhall, R. J. Beaver, y B. M. Beaver, Introduccion a la Probabilidad y Estadistica, 12a ed. Valle de México: Cengage Learning Editores S.A. de C.V, 2007.

[3] V. Realinho, J. Machado, L. Baptista, y M. V. Martins, “Predicting student dropout and academic success”, Data (Basel), vol. 7, núm. 11, p. 146, 2022.