ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos es una fase crítica en la ciencia de datos y el machine learning, y sin duda, es la que conlleva más tiempo. Utilizando las técnicas propuestas en [1] para resumir mediante gráficas los datos cualitativos y cuantitativos, se procede con el siguiente análisis.

1. Descripción del dataset inicial

El análisis exploratorio de lo realizó con el lenguaje Python en la plataforma Google Colab, utilizando el conjunto de datos almacenado en Google Drive. En la **Figura 1** se presenta el código implementado para dicho proceso.

```
df = pd.DataFrame()
def _loadCsv():
  global df
  id = '17_JBT4y9XoJYOAqui_tKxvICsbMOckMX'
  downloaded = drive.CreateFile({'id':id})
  downloaded.GetContentFile('dataset.csv')
  df = pd.read_csv('dataset.csv', converters={'cedula': lambda x: str(x)})
_loadCsv()
```

Figura 1 Script para cargar el conjunto de datos almacenado en Google Drive

Una vez cargado el dataset, se empezó con la exploración y comprensión inicial de los datos. La **Figura 2** muestra que el conjunto de datos tiene 36362 registros (filas) y 51 características (columnas).

```
print('El archivo tiene el siguiente número de filas y columnas:', df.shape)
print('Las columnas son:', df.columns)
```

```
El archivo tiene el siguiente número de filas y columnas: (36362, 51)

Las columnas son: Index(['facultad', 'carrera', 'periodo_lectivo', 'nivel', 'modalidad',
       'cedula', 'apellidos', 'nombres', 'fecha_nacimiento', 'genero',
       'estado_civil', 'etnia', 'sector_procedencia', 'nacionalidad',
       'pais_nacimiento', 'provincia_nacimiento', 'canton_nacimiento',
       'ciudad_nacimiento', 'direccion_nacimiento', 'pais_actual',
       'provincia_actual', 'canton_actual', 'parroquia_actual',
       'ciudad_actual', 'direccion_actual', 'trabaja', 'ingreso_estudiante',
       'numero_hijos', 'colegio', 'tipo_colegio', 'pais_colegio',
       'provincia_colegio', 'canton_colegio', 'oferta_academica', 'ciclo',
       'numero_matricula', 'malla_curricular', 'estado_matricula', 'paralelo',
       'jornada', 'asignatura', 'estado_asignatura', 'obligatoria',
       'arrastrable', 'asistencia_obligatoria', 'nota_ingresada',
       'nota_ponderada', 'porcentaje_asistencias', 'promedio_matricula',
       'homologada', 'observacion_homologacion'],
      dtype='object')
```

Figura 2 Información del dataset inicial

La **Figura 3** presenta los resultados luego de haber aplicado el comando ***describe().*** Este comando se aplicó para pesentar una descripción del dataset, y los resultados indican las propiedades como: la cantidad de registros, la media, la desviación estándar y otras correspondientes a las variables contínuas.

```
1 df.describe()
```

	ingreso_estudiante	numero_hijos	ciclo	numero_matricula	nota_ingresada	nota_ponderada	porcentaje_asistencias	promedio_matricula
count	36362.000000	36362.000000	36362.000000	36362.000000	30825.000000	30825.000000	33145.000000	34947.000000
mean	33.243331	0.163852	4.868324	307027.194186	7.716389	1.326403	87.211890	6.988508
std	113.357141	0.463211	3.068355	158763.528168	1.959020	0.991849	29.146071	2.844206
min	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	2.000000	187251.500000	7.100000	0.750000	91.890000	7.180000
50%	0.000000	0.000000	4.000000	343106.000000	8.030000	1.180000	98.330000	8.060000
75%	0.000000	0.000000	7.000000	443732.000000	8.960000	1.630000	100.000000	8.580000
max	1200.000000	10.000000	12.000000	526962.000000	10.000000	21.680000	400.000000	28.150000

Figura 3 Resultado del comando describe().

Para continuar con la comprensión del conjunto de datos, se procede describir a cada una de las 51 características, conforme se lo puede ver en la **Tabla 1**.

Tabla 1 Descripción de las características

CARACTERÍSTICA	TIPO DE DATO	DESCRIPCIÓN
facultad	TEXTO	Almacena las siglas de la facultad o área a la que pertenece la carrera que el alumno estudió o se encuentra estudiando. En este caso, contiene únicamente el texto "AEIRNNR".
carrera	TEXTO	Corresponde al nombre de la carrera o especialidad que estudió o se encuentra estudiando el alumno.
periodo_lectivo	TEXTO	Representa el año inicial y el año final en la que el alumno llevó a cabo un ciclo académico.
nivel	TEXTO	Simboliza el régimen al que pertenece la carrera que el alumno estudió o se encuentra estudiando.
modalidad	TEXTO	Contiene el nombre de la modalidad en que se desarrolla la carrera profesional.
cedula	TEXTO	Almacena el numero de la cedula de ciudadanía perteneciente a un alumno estudió o se encuentra estudiando.
apellidos	TEXTO	Este campo contiene los apellidos del estudiante.
nombres	TEXTO	Este campo contiene los nombres del estudiante.
fecha_nacimiento	TEXTO	Corresponde a la fecha en que nació el estudiante.
genero	TEXTO	Contiene el género masculino o femenino del alumno.
estado_civil	TEXTO	Almacena el estado civil que corresponde a cada estudiante.
etnia	TEXTO	Corresponde a la etnia con la que se identifica el estudiante.
sector_procedencia	TEXTO	Sector urbano o rural del que proviene el estudiante.
nacionalidad	TEXTO	Nacionalidad a la que pertenece el estudiante.
pais_nacimiento	TEXTO	País en donde nació el estudiante.
provincia_nacimiento	TEXTO	Provincia o departamento en donde nació el alumno.
canton_nacimiento	TEXTO	Cantón en donde nació el estudiante.
ciudad_nacimiento	TEXTO	Ciudad en donde nació el estudiante.
direccion_nacimiento	TEXTO	Dirección o ubicación en donde nació el estudiante.
pais_actual	TEXTO	País en donde vive actualmente el estudiante.
provincia_actual	TEXTO	Provincia en donde vive actualmente el estudiante.
canton_actual	TEXTO	Cantón en donde vive actualmente el estudiante.
parroquia_actual	TEXTO	Parroquia en donde vive actualmente el estudiante.
ciudad_actual	TEXTO	Ciudad en donde vive actualmente el estudiante.
direccion_actual	TEXTO	Dirección del domicilio en donde vive actualmente el estudiante.
trabaja	TEXTO	Condición de empleo actual del alumno, es decir si trabaja o no trabaja.

ingreso_estudiante	NUMÉRICO	Corresponde a la cantidad de dinero en dólares que el alumno tiene como ingresos mensuales.
numero_hijos	NUMÉRICO	Representa la cantidad de hijos o personas bajo la responsabilidad del estudiante.
colegio	TEXTO	Es el nombre del colegio del que proviene el alumno.
tipo_colegio	TEXTO	Corresponde al tipo (público, privado, etc.) de colegio del que proviene el estudiante.
pais_colegio	TEXTO	Nombre del país en que se encuentra el colegio del que proviene el estudiante.
provincia_colegio	TEXTO	Nombre de la provincia en que se encuentra el colegio del que proviene el estudiante.
canton_colegio	TEXTO	Nombre del cantón en que se encuentra el colegio del que proviene el estudiante.
oferta_academica	TEXTO	Corresponde al nombre la malla curricular que cumple el alumno.
ciclo	NUMÉRICO	Es el número que representa al ciclo académico que el estudiante cursó.
numero_matricula	NUMÉRICO	Es el número que identifica la matrícula en cada ciclo académico.
malla_curricular	TEXTO	Es el nombre del pensum académico al que pertenece el alumno.
estado_matricula	TEXTO	Corresponde al resultado final que obtuvo un estudiante al culminar el ciclo académico (aprobó o reprobó) o al estado actual de la matricula (matriculado si es que se encuentra actualmente cursando) o también puede tener el estado de reubicado, si es que el estudiante tuvo un cambio de malla.
paralelo	TEXTO	Es el nombre del paralelo al que perteneció el estudiante en cada ciclo académico.
jornada	TEXTO	Representa la jornada en que se desarrolla la carrera profesional del estudiante.
asignatura	TEXTO	Es el nombre de la asignatura que cursó el estudiante.
estado_asignatura	TEXTO	Corresponde al resultado final de la asignatura, que obtuvo un estudiante al culminar el ciclo (aprobó o reprobó) o al estado actual de la matricula (matriculado si es que se encuentra actualmente cursando).
obligatoria	TEXTO	Cualidad de las asignaturas, es decir, representa si una asignatura es obligatoria o no.
arrastrable	TEXTO	Cualidad de las asignaturas, es decir, representa si una asignatura es arrastrable o no.
asistencia_obligatoria	TEXTO	Cualidad de las asignaturas, es decir, representa si una asignatura es de asistencia obligatoria o no.
nota_ingresada	NUMÉRICO	Es el valor numérico de la calificación final que obtuvo el estudiante en cada asignatura.
nota_ponderada	NUMÉRICO	Corresponde al valor ponderado de las calificaciones ingresadas.
porcentaje_asistencias	NUMÉRICO	Corresponde al porcentaje de horas asistidas, en relación al total de horas planificadas.
promedio_matricula	NUMÉRICO	Corresponde al promedio final obtenido por el estudiante sobre las asignaturas desarrolladas en un ciclo específico.
homologada	TEXTO	Representa si un estudiante homologó alguna(s) materia(s).
observacion_homologacion	TEXTO	En este campo se encuentran las observaciones registradas cuando existe una homologación.

La variable ***estado_matricula*** es la variable que se va a predecir y también es la variable objetivo (target) que permitirá entrenar los modelos. Por tal motivo, se presenta una descripción de sus valores iniciales. Conforme a la **Figura 3**, se puede decir que, de 36362 registros, 35380 tienen un valor y es resto (982) son valores faltantes.

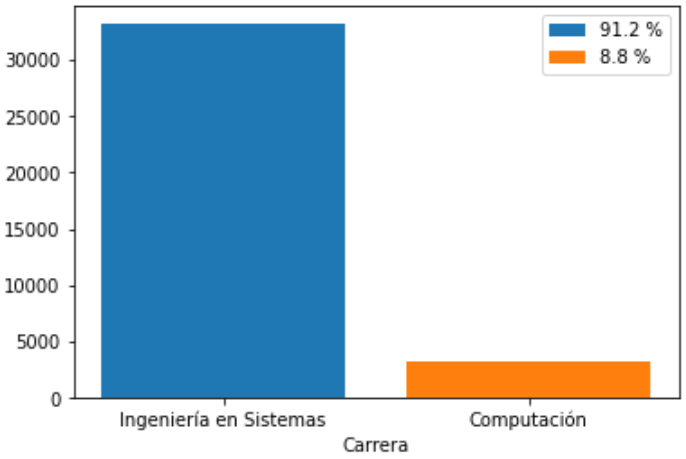
```
1 df['estado_matricula'].describe(include='all')
```

```
count        35380
unique           4
top       Aprobada
freq         24424
Name: estado_matricula, dtype: object
```

Figura 4 Descripción de la característica estado_matricula

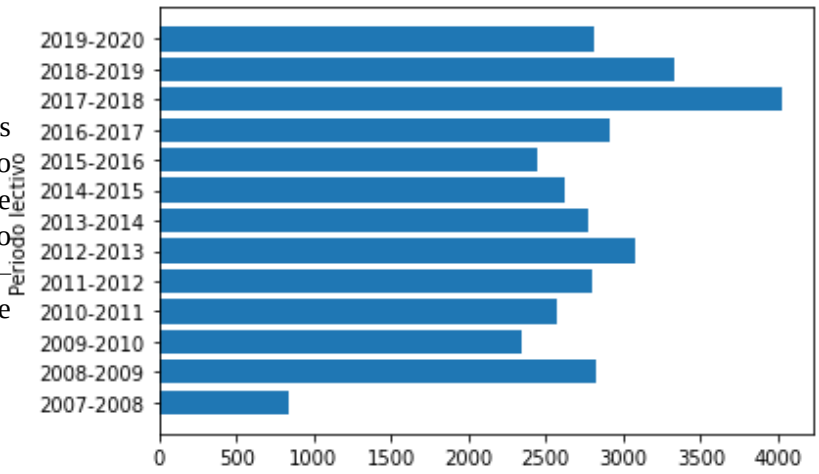
2. Resultados

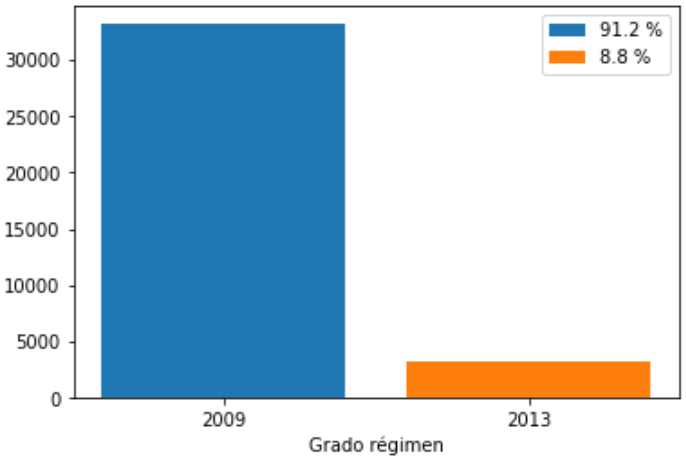
Para la presente exploración y visualización de datos se descarta las variables que contienen un valor único y también se ignora a las variables que se encuentran fueran del contexto de la investigación.



Cantidad de registros existentes de cada una de las carreras profesionalizantes utilizadas para el estudio. Donde se puede observar que la mayor cantidad de registros pertenecen a la carrera de Ingeniería en Sistemas.

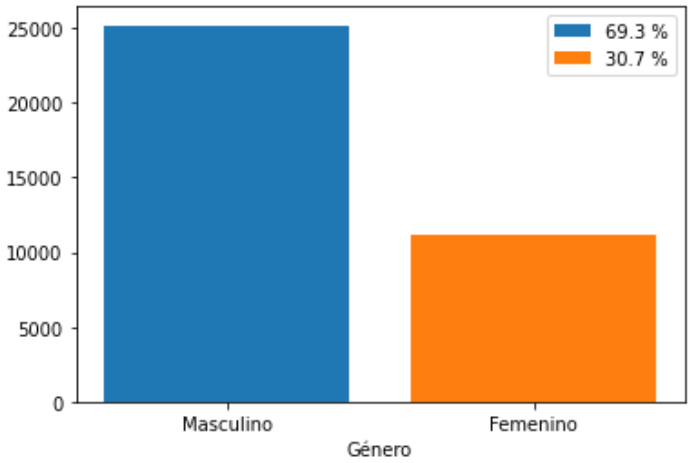
Cantidad de registros correspondientes a cada periodo lectivo desarrollado. Se puede observar que el primer periodo académico registrado es 2007 – 2008, y que es el periodo que menos registros contiene.

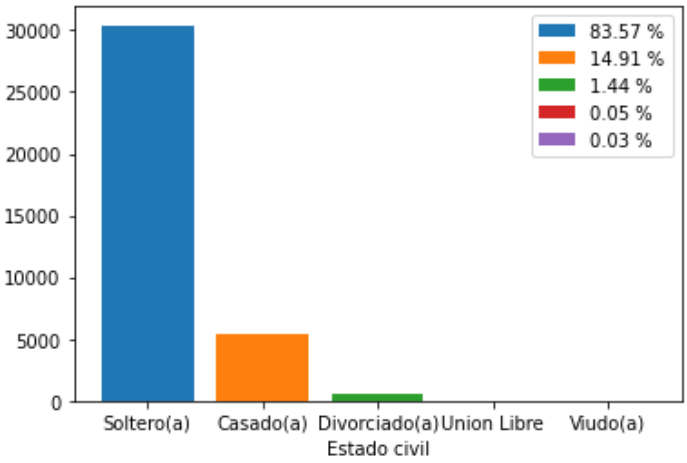




Cantidad de registros existentes de la característica grado régimen. Donde al régimen 2013 corresponde la carrera de Computación y al 2009 pertenece la carrera de Ingeniería en Sistemas.

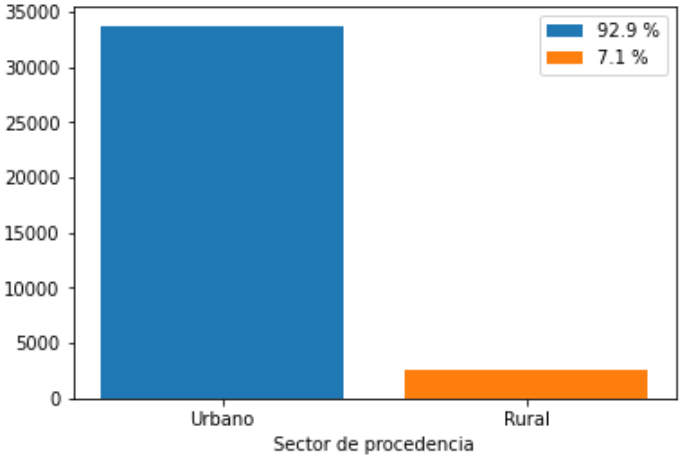
Registros correspondientes al género de los alumnos. Donde se visualiza claramente que el género masculino predomina en los registros.

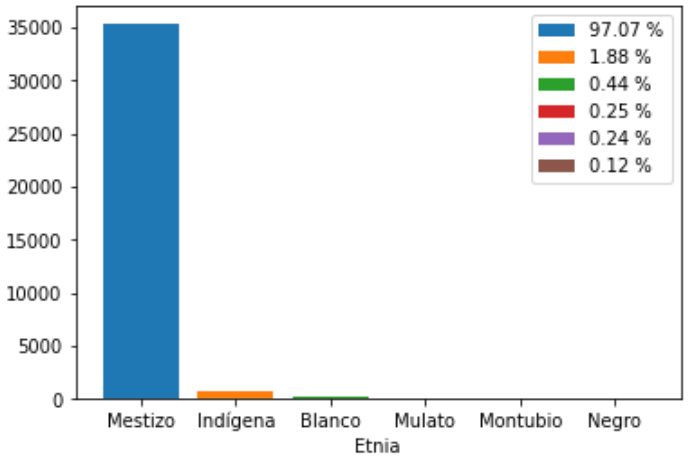




Gráfica que muestra la cantidad de registros correspondientes al estado civil de los alumnos. Donde se visualiza claramente que la mayoría son solteros.

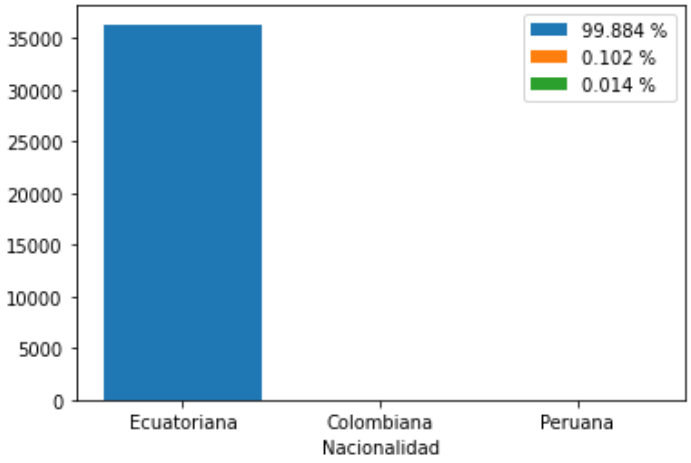
Gráfica que muestra la cantidad de registros correspondientes al sector de procedencia de los alumnos. Donde se visualiza que la mayoría pertenecen al sector urbano.

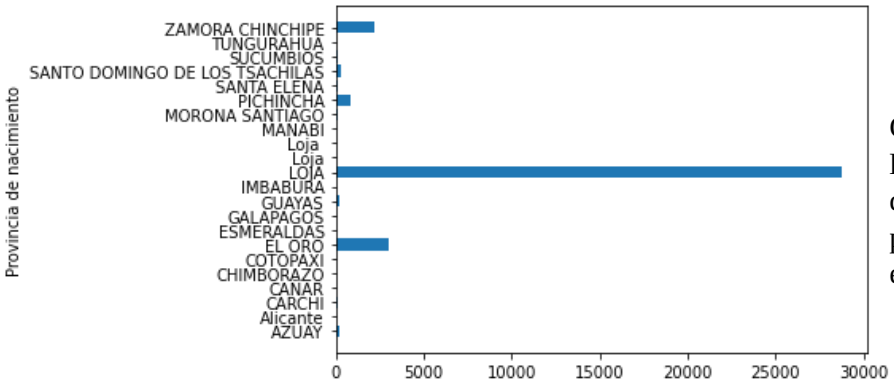




Gráfica que muestra la cantidad de registros correspondientes a la etnia de los alumnos. Donde se visualiza claramente que la mayoría son mestizos.

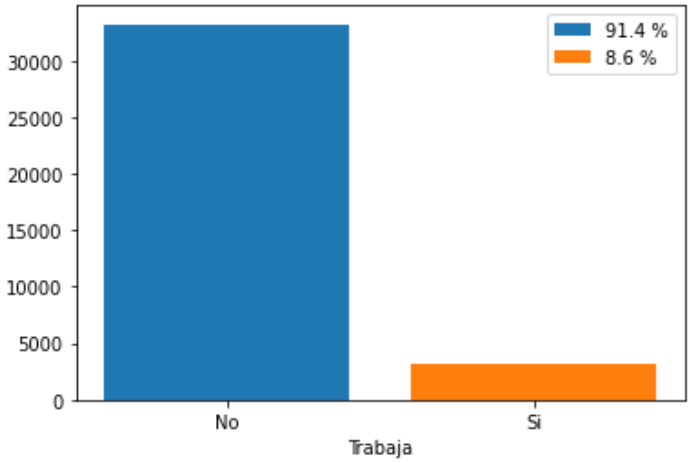
Gráfica correspondiente a los registros de la nacionalidad de los alumnos. Donde se visualiza que casi en la totalidad son ecuatorianos.

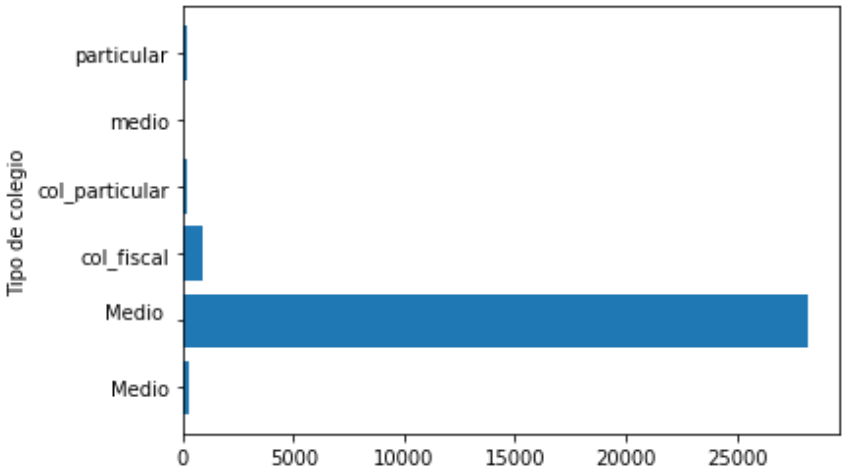




Gráfica correspondiente a la provincia de nacimiento de los alumnos. Donde la provincia más concurrente es Loja.

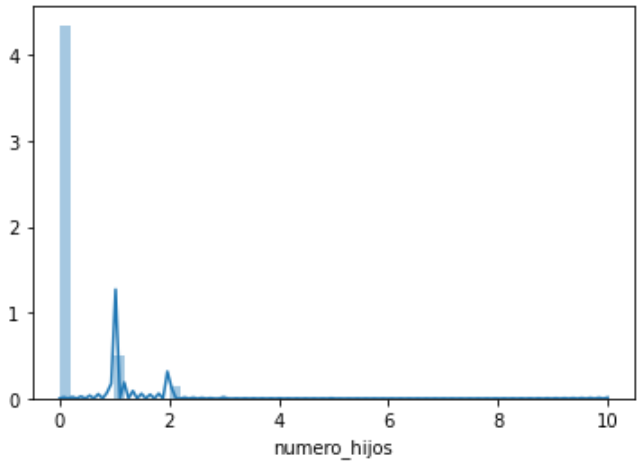
Gráfica que representa la cantidad de registros de los estudiantes que trabajan. Donde se visualiza de que la mayoría no trabaja.

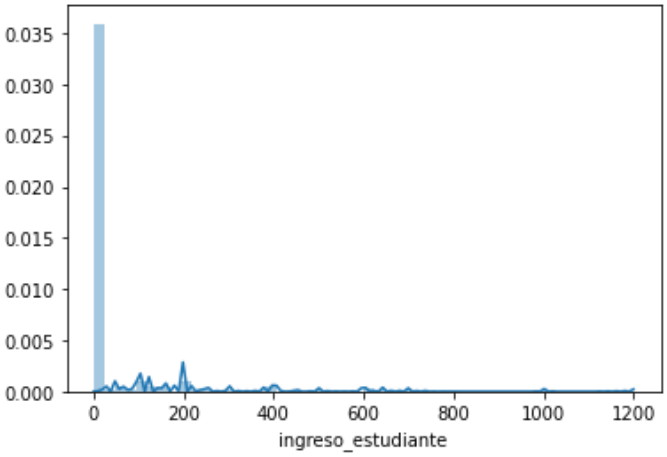




Registros correspondientes al tipo de colegio del que vienen los estudiantes. Donde se visualiza que la mayoría corresponden al tipo “medio”.

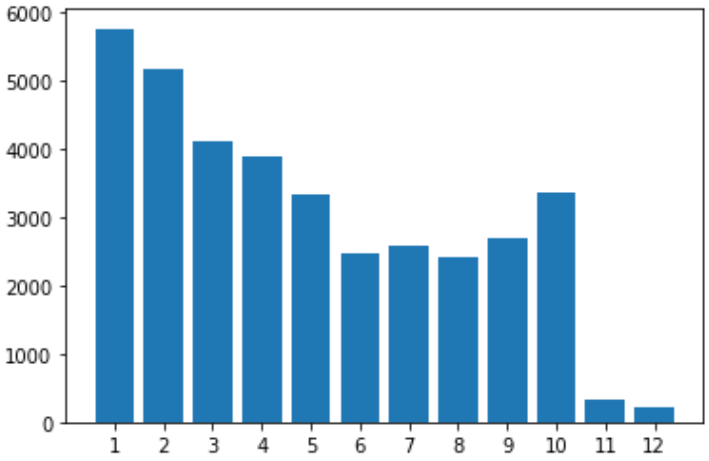
Gráfica que representa la cantidad de hijos que tienen los estudiantes. Donde se visualiza de que la mayoría no tiene hijos.

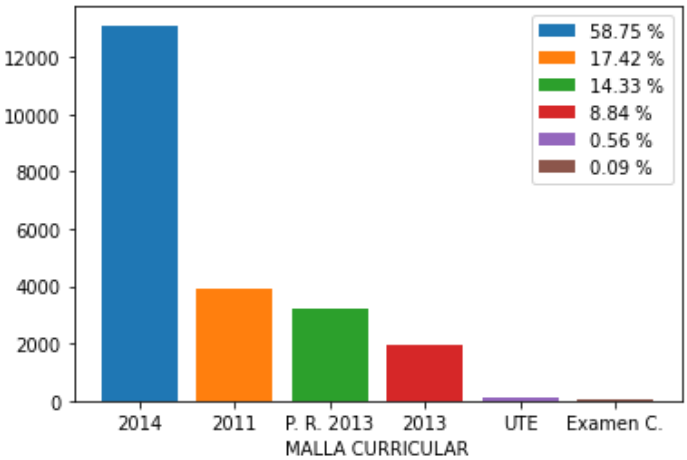




Gráfica que representa la cantidad de ingresos que existen en los registros académicos. Donde se visualiza de que la mayoría no tiene ingresos.

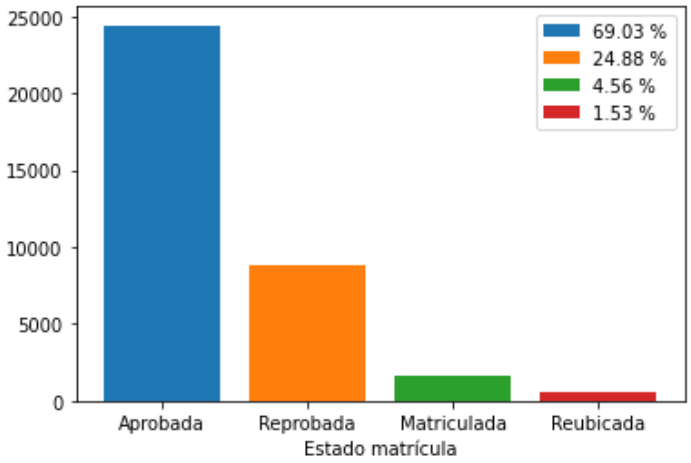
Gráfica que representa la cantidad de registros correspondientes al ciclo académico. Actualmente son 10 ciclos académicos, pero anteriormente habían sido doce.

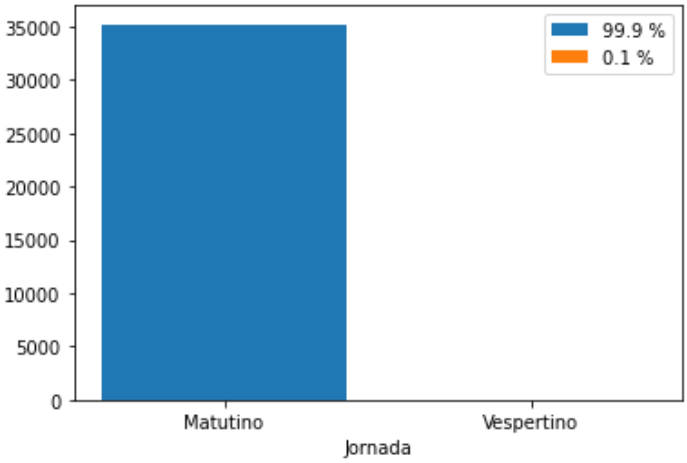




Gráfica que representa la cantidad de registros correspondientes a la malla curricular.

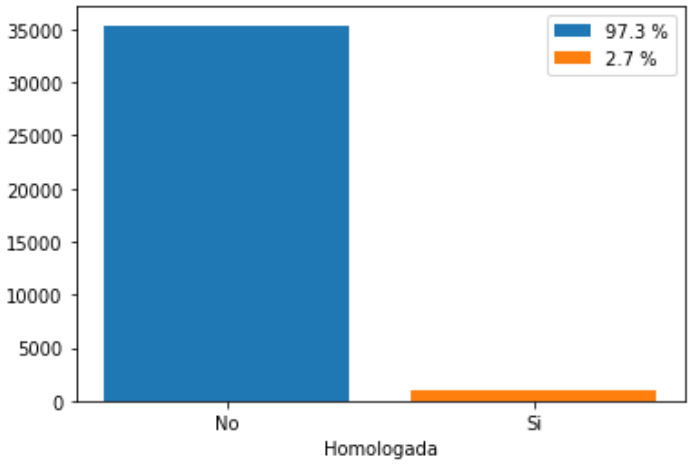
Cantidad de registros correspondientes al estado de matrícula. El estado “Matriculada” significa que esos registros se encuentran cursando el ciclo.





Gráfica que representa la cantidad de registros correspondientes al estado de matrícula. El estado “Matriculada” significa que esos registros se encuentran cursando el ciclo.

Gráfica que representa la cantidad de registros correspondientes a la característica “Homologada”. Donde el valor “NO”, significa que el estudiante no ha tenido homologaciones.



3. Conclusiones

- Luego de haber realizado el presente análisis exploratorio, se comprendió que existen características totalmente irrelevantes, y que se deberían descartar para entrenar los modelos.
- Existen registros que se pueden unificar, con la finalidad de crear menos clases o categorías, y así generar modelos de mejor calidad.
- Existen registros (filas) de los estudiantes por cada asignatura cursada, por lo que se debería presentar un solo registro por estudiante, en donde las asignaturas y su estado, se las ubique como columnas.

4. Bibliografía

[1] L. González-Támara, "Una introducción a la estadística descriptiva y probabilidad Análisis exploratorio de datos," *Univ. Bogotá Jorge Tadeo Lozano.*, 2017.