UNIVERSIDAD TECNOLÓGICA METROPOLITANA
UTEM del Estado de Chile

Escuela de Informática
Facultad de Ingeniería
Minería de datos INFB8104

# “Modelo Predictivo de Retiros Escolar”

## Proyecto Semestral

**Integrantes:** Sebastián Alejandro Garrido Valenzuela
Javier Ignacio Gálvez González
Benjamín Ignacio Martínez Gárate

**Profesor(a):** Claudio Gerardo Collao Bahamondes

**Fecha de entrega:** 08/01/2021

# CONTENIDOS

# 1. Introducción

Un modelo predictivo es un sistema que emplea datos y estadísticas para predecir resultados a partir de unos modelos de datos. Estos modelos se utilizan para predicciones de todo tipo, como por ejemplo, resultados deportivos, audiencias televisivas, avances tecnológicos, ganancias empresariales, etc.

El modelado predictivo es útil porque proporciona información precisa sobre cualquier pregunta y permite a los usuarios crear previsiones de esta, así ir manteniendo una ventaja competitiva tras la obtención de información detallada de eventos y resultados futuros. En el análisis de minería de datos se suele extraer información esencial de diversas fuentes para complementar los modelos predictivos, estos son:

- Datos sobre transacciones
- Datos de servicio al cliente
- Datos de encuestas o sondeos
- Datos de marketing digital y publicidad
- Datos económicos
- Datos demográficos
- Datos geográficos
- Datos de tráfico web
- entre otros.

Predecir datos está relacionado con un análisis predictivo de estos, el cual se describen estos datos utilizados, posteriormente un análisis de diagnóstico para ver entre qué estadísticas se encuentra lo que se estudiará. Es por esto, que para el presente proyecto se requerirá una minería de datos, un análisis profundo y una comprensión de información para poder predecir finalmente el rendimiento escolar ligado al retiro de los estudiantes en establecimientos, esto depende de varias características por lo que predecir esto a futuro traerá una toma de decisiones más firme sobre el pensamiento y sobre qué se debe realizar para minimizar este retiro por parte de los estudiantes.

A lo largo del proyecto se darán a conocer distintos modelos predictivos que se encuentran hoy en día, estos se dividen en dos campos: paramétricos y no paramétricos. Si bien estos términos pueden parecer jerga técnica, la diferencia esencial es que los modelos paramétricos hacen más suposiciones, y más específicas, sobre las características de la población utilizada para crear el modelo. En concreto, algunos tipos de modelos predictivos son:

- Mínimos cuadrados ordinarios
- GLM (Modelos lineales generalizados)
- Regresión logística
- Bosques aleatorios
- Árboles de decisión
- Redes neuronales
- MARS (Ejes de regresión adaptativa multivariante)

Cada uno de estos tipos tiene un uso particular y responde a una pregunta específica o utiliza un determinado tipo de conjunto de datos. A pesar de las diferencias metodológicas y matemáticas entre los tipos de modelos, el objetivo general de todos ellos es similar: predecir resultados futuros o desconocidos basándose en datos pasados.

Finalmente lo más utilizado para predicción de este proyecto serán Regresión Logística Multivariable, Árboles de Decisión, Redes Neuronales, análisis PCA, aplicación de clustering, redes bayesianas y máquinas de vector de soporte, estos modelos nos traerán distintas visiones y acercamiento de precisión para el objetivo general, el cual se explicará más adelante.

# 2. Aspectos Generales del Proyecto

## 2.1 Delimitación del tema de proyecto:

El proyecto busca predecir el rendimiento escolar en el año 2021 para los colegios de la comuna de Santiago, Región Metropolitana, adquiriendo datos de la cantidad de alumnos retirados, con base a los datos recopilados del año 2018.

El problema principal es no saber qué sucederá con los alumnos de cada establecimiento para poder anteponerse al retiro de estos mismos, saber el porqué de ello, además de poder enfocarse en tener una mejor enseñanza no solamente para que sean aprobados los estudiantes, sino para que salgan de las instituciones educacionales bien preparadas y por otra para, que haya menos reprobados y si es que hay, llegar a la raíz de ello.

## 2.2 Formulación de la problemática del proyecto:

En el contexto que existe actualmente, es que en los colegios y liceos existe una cierta cantidad de alumnos que se retira del establecimiento. Para este caso, cabe mencionar que existen factores cualitativos los cuales influyen en los resultados del desempeño:

1) Estilo de vida.
2) Cultura.
3) Motivación.
4) Personalidad.

Cómo se logra ver, estos factores resultan muy difíciles de cuantificar. Sin embargo, existen otros factores que sí se podrán cuantificar, con el fin de adquirir un conjunto de datos que pueda predecir si el alumno abandona el año escolar.

Los datos a utilizar serán obtenidos del Centro de Estudios MINEDUC (formato de la base de datos: .CSV con codificación UTF-8), lo cual es un esquema de registros de Rendimiento escolar para el año 2018 de todos los establecimientos educacionales de toda la Región Metropolitana, teniendo registrados a estudiantes por nivel de enseñanza, situación final, sexo y grado. A

continuación, en la tabla N.° 1 se detalla el número de observaciones desde el año 2015 hasta el 2018, con el fin de visualizar el movimiento de años anteriores al 2018.

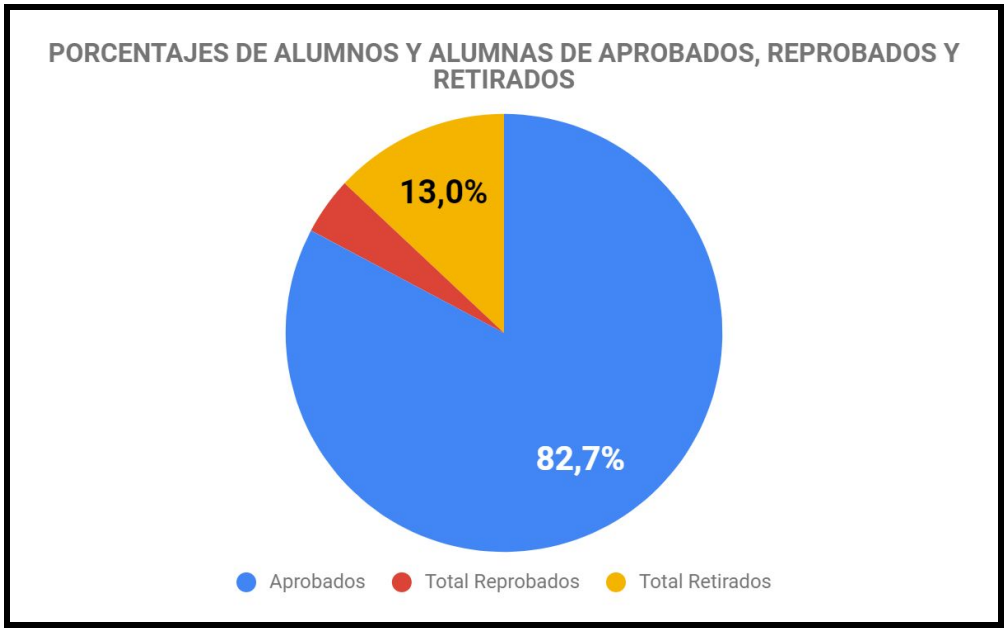
Tabla N.° 1, Número de Observaciones por año

| Año escolar | Número de Unidades Educativas | Número de Unidades Educativas en establecimientos en funcionamiento | Número de establecimientos | Número de establecimientos en funcionamiento |
|---|---|---|---|---|
| 2018 | 13.940 | 13925 | 9.192 | 9.179 |
| 2017 | 13.942 | 13936 | 9.273 | 9.268 |
| 2016 | 14.009 | 13.995 | 9.349 | 9.337 |
| 2015 | 14.069 | 14.041 | 9.458 | 9.437 |

Fuente: Elaboración propia, extracto de Registro de Estudiantes de Chile (RECH), Mineduc

A partir de las estadísticas del año 2018, según la página de ministerio de educación , donde se interpretan porcentajes de acuerdo la situación final de los alumnos, donde 82,7% de alumnos aprobaron y 4,23% de alumnos reprobaron en el desempeño académico, en cuanto a alumnos que se retiraron de la institución se interpreta un porcentaje de 13,03%. Todo esto se podrá visualizar en la siguiente figura:

Figura N.° 1, Porcentaje de alumnos y alumnas de acuerdo al resultado de desempeño académico 2018




Fuente: Elaboración propia en base a la estadística recabada en el sitio web de Mineduc

Una vez contextualizado en el ámbito académico, los problemas potenciales que se deben analizar son el caso de alumnos y alumnas que se hayan retirado del establecimiento, sin embargo no será posible saber por qué razón tomaron esa decisión, ya sea personal o por la situación que se encuentran en la familia.

## 2.3 Objetivos

### *2.3.1 Objetivo General:*

El objetivo general es formular un modelo predictivo capaz de identificar la cantidad de alumnos retirados de las instituciones educacionales de la comuna de santiago, para proporcionar ayuda al desempeño de dichas instituciones.

### *2.3.2 Objetivos Específicos:*

- Recopilar Información respecto a los alumnos retirados de colegios de santiago en el año 2018.
- Preparación de los datos para generar un modelo
- Generar un modelo predictivo con herramientas de minería de datos

## 2.4 Hipótesis:

***¿Es posible predecir el posible retiro de los alumnos para un determinado conjunto de instituciones de una comuna?***

Con este modelo se podrá predecir la posibilidad de que los alumnos se retiren desde sus respectivos establecimientos educacionales, de un determinado conjunto de instituciones de una comuna, para así en aspectos ideales tener para un plan de acción en el año 2021. Este modelo ayuda a prevenir a que los alumnos de un sector puedan seguir estudiando, para este caso es necesario tomar ciertas acciones para disminuir las posibilidades del abandono.

## 2.5 Definición de los datos a utilizar:

La extracción de datos para este proyecto estarán envueltos en:

- Ministerio de Educación (Mineduc).
- Biblioteca del Congreso Nacional de Chile (BCN)

Durante la recolección de datos, se ha extraído información que son necesarios para la creación del archivo CSV, los cuales estos datos son comunas de la región metropolitana, junto con sus características numéricas. A continuación se especificarán las columnas que contendrá este archivo:

- **Comuna:** Corresponde a los nombres de las comunas que se encuentran en la región metropolitana.
- **Cantidad total de retirados:** Corresponde a números de alumnos que se han retirado de sus establecimientos por comuna.
- **Superficie km^2:** Corresponde al área de superficie total que tiene cada comuna en kilometros al cuadrado.
- **Población 2020:** Corresponde a la cantidad de habitantes de personas que habitan en cada comuna, siendo que estas fueron registradas en el año 2020.
- **Tasa de pobreza por Ingresos (%):** Corresponde a la tasa de pobreza que tiene cada comuna.
- **Personas en hogares carentes de servicios básicos (%):** Corresponde al porcentaje de los hogares por familia que tienen problemas con los servicios básicos (agua, luz, alimentos, entre otros) en cada comuna por su carencia.
- **Hogares hacinados (%):** Corresponde al porcentaje de hogares hacinados, los cuales tienen problemas de amontonamiento o acumulación de individuos en un mismo lugar donde viven, teniendo también un problema serio de privacidad.
- **Cantidad de establecimientos municipales:** Pertenece a los números de establecimientos municipales que contiene cada comuna.
- **Cantidad de establecimientos particulares subvencionados:** Pertenece a los números de establecimientos particulares subvencionados que contiene cada comuna.

- **Cantidad de establecimientos particulares pagados:** Pertenece a los números de establecimientos particulares pagados que contiene cada comuna.
- **Cantidad total de matrículas en los establecimientos:** Corresponde al número de matrículas que se han realizado en los establecimientos municipales por comuna.
- **Ingresos Educación (M$)**: Pertenece a los ingresos que todas las instituciones de educación genera en millones de pesos por comuna.
- **Aporte Municipal al Sector Educación (M$):** Pertenece a los aportes que realizan las municipalidades para los establecimientos educacionales en millones de pesos por comuna.
- **Gastos Educación (M$):** Corresponde a los gastos que se producen en los establecimientos educacionales en millones de pesos por comuna.
- **Tasa de denuncias por delito de mayor connotación social:** Corresponde a la tasa de denuncias respecto a su cantidad, lo cual tiene mayor connotación social.
- **Tasa de denuncias por violencia intrafamiliar:** Corresponde a la tasa de denuncias respecto a su cantidad referente a la violencia intrafamiliar.

**NOTA:** No se utilizarán los datos académicos del año 2019, por el estallido social ocurrido en octubre, ya que esto repercute en una dispersión total de los datos, en cambio en el año 2018 se obtiene una nivelación “normal” de ellos. Para ver las fuentes donde se han extraídos estos datos, véase a [1] y [2] desde referencias.

A partir de la figura N°2, se visualiza los datos que contiene el archivo CSV:

Figura N.° 2, Estadísticas de cada comuna

| N° | Comuna | Cantidad Total de Retirados | Superficie km2 | Poblacion 2020 | Tasa de Pobreza % | Hogares Carentes de Servicios Básico % | Hogares hacinados % | Cantidad de Establecimientos Municipales | Cantidad de Establecimientos Particular Subvencionado | Cantidad de Establecimientos Particular Pagado | Cantidad de matriculas totales | Ingresos Educación (M$) | Aporte Municipal al Sector Educación (M$) | Gastos Educación (M$) | Tasa de Denuncias por Delito de Mayor Connotacion Social | Tasa de Denuncias por Violencia Intrafamiliar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Santiago | 2123 | 22 | 503.147 | 04.08 | 0.19 | 18.50 | 44 | 66 | 17 | 81529 | 68086543 | 8810430 | 68467876 | 21169.70 | 1283.00 |
| 2 | Recoleta | 713 | 16 | 190.075 | 6.89 | 13.1 | 19.1 | 19 | 43 | 2 | 31426 | 23125862 | 2767600 | 23880193 | 7335.60 | 859.90 |
| 3 | Independencia | 402 | 7 | 142.065 | 8.50 | 6 | 20 | 9 | 25 | 3 | 20149 | 11430455 | 162370 | 10658897 | 8697.20 | 1520.30 |
| 4 | Quinta Normal | 58 | 13 | 136.368 | 3.73 | 9.1 | 18.20 | 2 | 49 | 0 | 17584 | 21182891 | 1671736 | 20633866 | 6280.40 | 1107.30 |
| 5 | Maipú | 11 | 135.5 | 578.605 | 2.57 | 2.90 | 11.30 | 0 | 163 | 13 | 81309 | 44285876 | 3900000 | 44979840 | 868.70 | 197.40 |
| 6 | Estación Central | 432 | 15 | 206.792 | 5.81 | 14.70 | 17.1 | 15 | 40 | 3 | 25562 | 19921456 | 3913500 | 20474710 | 8510.80 | 821.20 |
| 7 | Cerrillos | 31 | 21 | 88.956 | 6.48 | 6 | 15.80 | 9 | 24 | 1 | 12914 | 8247485 | 892025 | 8213170 | 5336.80 | 973.60 |
| 8 | Quilicura | 4 | 58 | 254.694 | 5.68 | 2.70 | 15.80 | 12 | 46 | 3 | 44225 | 21117885 | 3500000 | 19690246 | 1775.40 | 406.70 |
| 9 | El Bosque | 292 | 14.20 | 172.000 | 9.58 | 5.50 | 17.30 | 21 | 71 | 1 | 38293 | 18899970 | 177500 | 18743472 | 2539.90 | 516.20 |
| 10 | La Cisterna | 119 | 10 | 100.434 | 6.60 | 20.1 | 16.60 | 8 | 54 | 2 | 28318 | 6930984 | 195551 | 7251887 | 6186.60 | 630.00 |
| 11 | La Florida | 885 | 70.20 | 402.433 | 4.50 | 4 | 15.80 | 0 | 147 | 16 | 66483 | 32886109 | 5717647 | 31563232 | 2693.80 | 474.10 |
| 12 | La Granja | 281 | 10 | 122.557 | 4.75 | 8.20 | 19.1 | 16 | 38 | 0 | 20910 | 13472253 | 346500 | 13929494 | 3119.70 | 637.90 |
| 13 | La Pintana | 265 | 30.60 | 189.335 | 14.14 | 4.80 | 23.30 | 13 | 59 | 0 | 36163 | 14404373 | 1054293 | 15597872 | 1994.80 | 661.00 |
| 14 | La Reina | 214 | 23 | 100.252 | 0.99 | 5.60 | 12.40 | 0 | 15 | 21 | 18838 | 11285046 | 2209852 | 10922664 | 3084.10 | 389.80 |
| 15 | Las Condes | 322 | 99 | 330.759 | 0.19 | 1.30 | 5.30 | 0 | 12 | 44 | 36448 | 21984167 | 10749110 | 21984167 | 2724.60 | 236.50 |
| 16 | Lo Barnechea | 157 | 1024 | 124.076 | 2.84 | 7.50 | 18.20 | 6 | 9 | 18 | 22494 | 8837181 | 4222782 | 7977522 | 1218.20 | 193.40 |
| 17 | Lo Espejo | 181 | 7 | 103.865 | 6.69 | 9.60 | 24.80 | 15 | 23 | 0 | 12446 | 12902004 | 1750000 | 14515456 | 3452.60 | 611.30 |
| 18 | Macul | 300 | 12.90 | 134.635 | 7.45 | 7.40 | 13.60 | 0 | 26 | 5 | 12558 | 10968716 | 870000 | 11904814 | 4226.10 | 709.20 |
| 19 | Ñuñoa | 584 | 16.90 | 250.192 | 0.90 | 6.20 | 7.60 | 0 | 27 | 30 | 23521 | 26333051 | 1651021 | 24502147 | 5899.60 | 614.30 |
| 20 | Pedro Aguirre Cerda | 270 | 10 | 107.803 | 6.21 | 9.80 | 18.1 | 15 | 33 | 0 | 16361 | 11550970 | 850000 | 12114241 | 3325.40 | 645.50 |
| 21 | Peñalolén | 531 | 54 | 266.798 | 4.37 | 9.20 | 17.90 | 0 | 40 | 15 | 30248 | 22423126 | 1716933 | 25176362 | 2078.20 | 467.30 |
| 22 | Providencia | 296 | 14.30 | 157.749 | 0.43 | 2.70 | 5.50 | 0 | 12 | 31 | 20889 | 23348743 | 4565191 | 23162400 | 8652.60 | 321.80 |
| 23 | San Joaquín | 267 | 9.70 | 103.485 | 5.24 | 12.80 | 18.20 | 0 | 30 | 0 | 7022 | 8813611 | 350953 | 10092617 | 4264.70 | 732.40 |
| 24 | San Miguel | 394 | 10 | 133.059 | 4.85 | 5.50 | 12.1 | 0 | 42 | 4 | 19478 | 9067408 | 1838042 | 8867165 | 7375.20 | 910.00 |
| 25 | San Ramón | 569 | 7 | 86.510 | 4.60 | 8.20 | 21 | 12 | 25 | 0 | 14547 | 10586286 | 1104999 | 10912572 | 2773.10 | 629.50 |
| 26 | Vitacura | 134 | 28.30 | 96.774 | 0.13 | 0.50 | 2.90 | 2 | 1 | 16 | 20437 | 7125426 | 2913000 | 6658703 | 4138.90 | 202.10 |
| | Total | 9835 | | | | | | | | | | | | | | |

Fuente: Elaboración propia

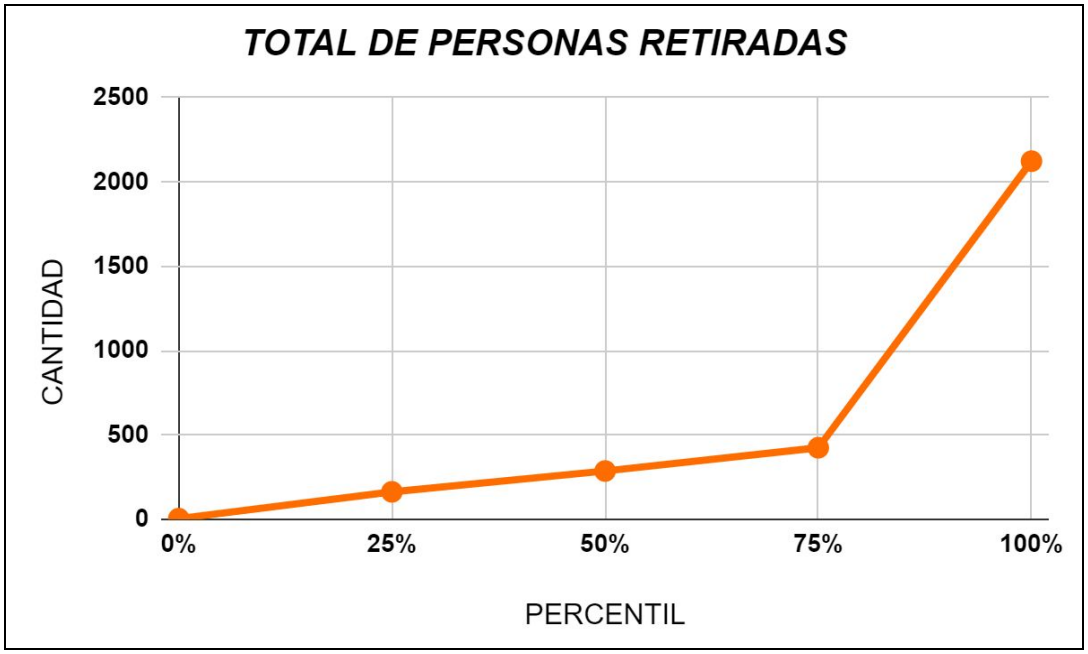
## 2.6 Normalización de Datos:

Para normalizar los datos desde el programa Python en Colaboratorio de Google (Colab), primero se ha extraído las columnas del archivo CSV para generar un Dataframe utilizando la librería Pandas, luego se eliminó la columna "*Comuna*" para no alterar los resultados por su tipo de datos String. Luego se realizó una división del dataset, quedando en 2 variables, "*variable_objetivo*" y "*variables_independientes*", donde la primera tiene almacenado la "cantidad de retirados totales" y la segunda tiene almacenado todas las columnas restantes. Después se realizaron distintas normalizaciones de "*variables_independientes*", de los cuales están incluidas dentro de la librería "*sklearn*", y se utilizaron las funciones como Scale, "*preprocessing.scale()*", Normalize, "*preprocessing.normalize()*", y MinMaxScale, "*preprocessing.minmax_scale( )*". Para este caso se utilizaron estas tres normalizaciones y también otros datos sin normalizar.

# 3. Análisis de Datos y Estadística Descriptiva

A partir de esta tabla, se podrá obtener los valores de promedio, desviación estándar, valores mínimo (percentil 0%), el primer cuartil (percentil de 25%), la mediana (percentil 50%), el tercer cuartil (percentil 75%) y valores máximos (percentil de 100%), para así poder describir las estadísticas presentes en la base de datos. Además se ha confeccionado un gráfico de percentil vs cantidad, para solamente casos de la cantidad total de personas que se retiraron. De los mismos gráficos, se logra visualizar que en el punto del valor máximo, fue lo que tiene mayor variación de datos en comparación a los puntos del valor mínimo, primer cuartil, mediana y tercer cuartil, lo cual implica la existencia de outliers. Todo lo anterior se visualiza en las figuras N°3, 4 y 5.

Figura N.° 3, Gráfico Percentil vs Cantidad respecto Total de Personas Retiradas




Fuente: Elaboración propia

Figura N.° 4, Descripción estadístico del datasets, parte 1

| | CANTIDAD TOTAL DE RETIRADOS | Superficie km2 | poblacion 2020 | tasa de pobreza por ingresos % | Personas en hogares carentes de servicios básico % | Hogares hacinados % | cantidad de establecimientos municipales | cantidad de establecimientos particular subvencionado | cantidad de establecimientos particular pagado | cantidad de matriculas totales | Ingresos Educación (M$) | Aporte Municipal al Sector Educación (M$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 2.600000e+01 | 2.600000e+01 |
| mean | 378.269231 | 66.869231 | 195516.076923 | 4.930769 | 7.061154 | 15.596154 | 8.384615 | 43.076923 | 9.423077 | 29236.615385 | 1.881607e+07 | 2.611578e+06 |
| std | 416.096677 | 197.689346 | 129042.959903 | 3.168674 | 4.589723 | 5.447236 | 10.135391 | 37.333013 | 11.876609 | 19675.390914 | 1.336517e+07 | 2.614291e+06 |
| min | 4.000000 | 7.000000 | 86510.000000 | 0.130000 | 0.190000 | 2.900000 | 0.000000 | 1.000000 | 0.000000 | 7022.000000 | 6.930984e+06 | 1.623700e+05 |
| 25% | 163.000000 | 10.000000 | 104849.500000 | 3.062500 | 4.200000 | 12.700000 | 0.000000 | 24.250000 | 0.250000 | 17897.500000 | 1.068189e+07 | 8.755062e+05 |
| 50% | 286.500000 | 15.500000 | 139216.500000 | 4.800000 | 6.100000 | 17.200000 | 7.000000 | 35.500000 | 3.000000 | 21702.000000 | 1.393831e+07 | 1.733466e+06 |
| 75% | 424.500000 | 30.025000 | 239342.000000 | 6.570000 | 9.175000 | 18.425000 | 14.500000 | 48.250000 | 16.000000 | 34978.750000 | 2.231339e+07 | 3.800000e+06 |
| max | 2123.000000 | 1024.000000 | 578605.000000 | 14.140000 | 20.100000 | 24.800000 | 44.000000 | 163.000000 | 44.000000 | 81529.000000 | 6.808654e+07 | 1.074911e+07 |

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 5, Descripción estadístico del datasets, parte 2

| | Gastos Educación (M$) | tasa de denuncias por delito de mayon connotacion social | tasa de denuncias por violencia intrafamiliar |
|---|---|---|---|
| count | 2.600000e+01 | 26.000000 | 26.000000 |
| mean | 1.895675e+07 | 4989.334615 | 644.296154 |
| std | 1.337046e+07 | 4054.570376 | 330.333837 |
| min | 6.658703e+06 | 868.700000 | 193.400000 |
| 25% | 1.072232e+07 | 2701.500000 | 421.850000 |
| 50% | 1.505666e+07 | 3795.750000 | 629.750000 |
| 75% | 2.286784e+07 | 6256.950000 | 799.000000 |
| max | 6.846788e+07 | 21169.700000 | 1520.300000 |

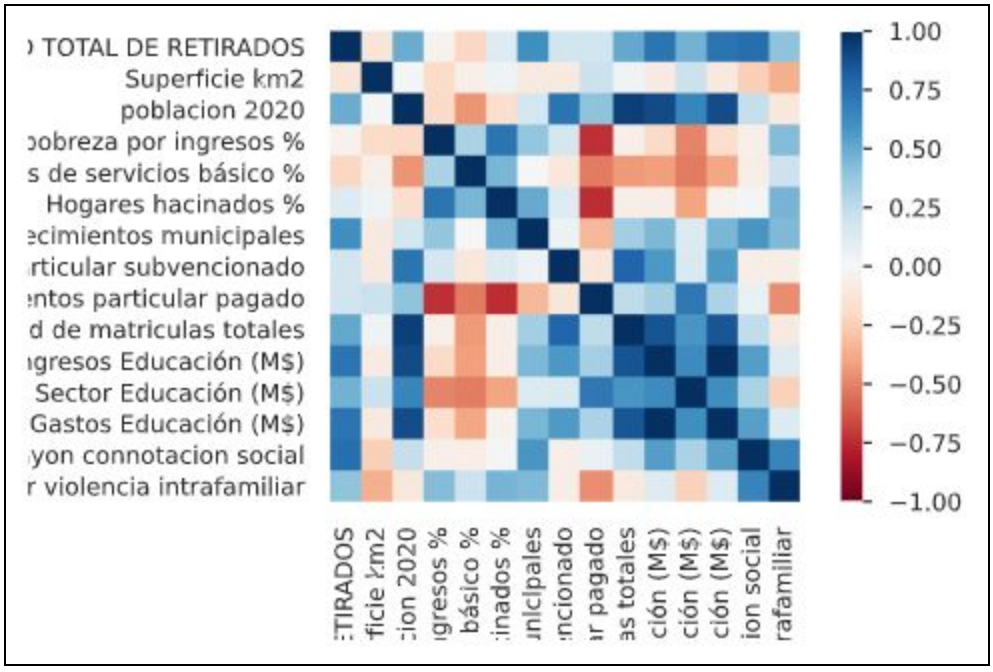
Fuente: Elaboración propia en base al código del proyecto

Por último, la correlación entre las variables se puede visualizar en la figura N°6, donde por orden se encuentran:

1) Cantidad total de retirados.
2) Superficie km^2.

3) Población 2020.
4) Tasa de pobreza por Ingresos (%).
5) Personas en hogares carentes de servicios básicos (%).
6) Hogares hacinados (%).
7) Cantidad de establecimientos municipales.
8) Cantidad de establecimientos particulares subvencionados.
9) Cantidad de establecimientos particulares pagados.
10) Cantidad total de matrículas en los establecimientos.
11) Ingresos Educación (M$).
12) Aporte Municipal al Sector Educación (M$).
13) Gastos Educación (M$).
14) Tasa de denuncias por delito de mayor connotación social.
15) Tasa de denuncias por violencia intrafamiliar.

Figura N.° 6, Correlación de Pearson entre variables




Fuente: Elaboración propia en base al código del proyecto

Por lo tanto, al comparar con la variable objetivo, que es la cantidad de personas que se retiraron, la fuerza de asociación positiva es fuerte para alguna correlación de Pearson como es el caso de la tasa de denuncias por delito de mayor connotación social, cantidad de establecimientos municipales, Gastos Educación (M$) e ingresos de Educación (M$). En cuanto al resto de otras variables, las correlaciones son débiles.

También se procedió con graficar los histogramas para cada campo, donde se tienen las frecuencias de los valores dentro de cada intervalo, los cuales se podrán ver en el anexo 13.3 estos detalles.

## 3.3 Detección de Outliers

Un Outlier se define como "aquella observación (o conjunto de observaciones) inconsistentes con el resto del conjunto de datos" (Barnett y Lewis, 1994). Es decir, aquella observación atípica y/o errónea que tiene un comportamiento muy diferente con respecto al resto de los datos analizados. Para esto conviene observar lo siguiente frente a la detección de Outliers:

- Las observaciones atípicas y erróneas exigen que los errores o variabilidades sean grandes.
- Los outliers no consideran todas las observaciones atípicas o erróneas, sino aquellas que tienen un comportamiento muy diferente respecto al resto de los datos.

Con esto, cabe mencionar que, existen métodos para corregir este tipo de observaciones atípicas y/o erróneas, pero para aquellas que no tienen un gran error o que se comportan como la mayoría, no van a afectar de forma determinante a las conclusiones que se realicen a partir de las mismas.

Por consiguiente, para la detección de Outliers se ocuparán los siguientes métodos:

- Método basado en el recorrido intercuartílico
- Diagrama Box-Plot

Con esto, se aplicarán los métodos para cada columna, con el fin de poder encontrar el/los Outliers de cada sección de datos.

Para el caso de método basado en el recorrido intercuartílico, se obtuvieron las cotas superior e inferior, el valor máximo y mínimo, cuartil inferior y superior, la mediana, rango inter cuartil y los Outliers ya sean severos o no, esto para cada sección de datos agrupados lo cual se podrá ver con mayor detalles en el anexo 13.1. A partir de la tabla N°2 se visualiza los outliers que se han detectado con el método basado en el recorrido intercuartílico.

Tabla N.° 2, Outliers detectados en el dataset del proyecto

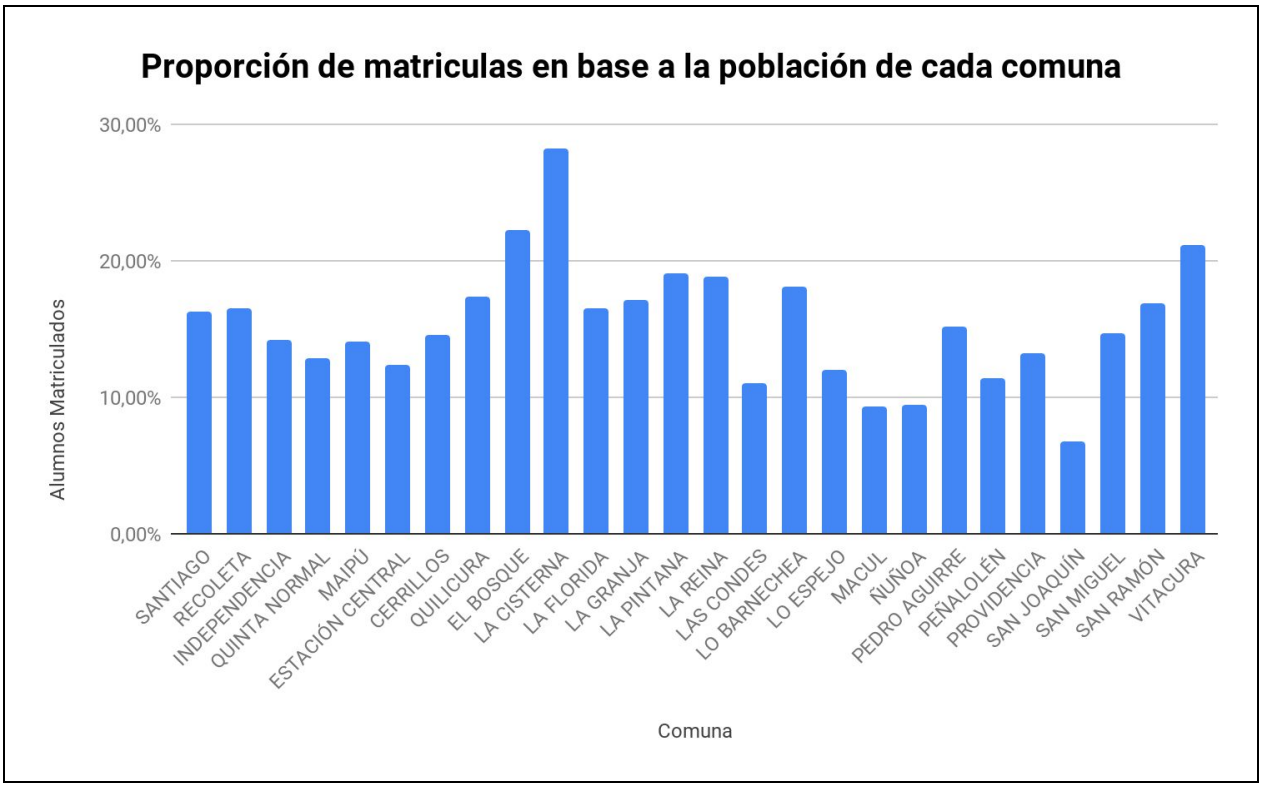
| N° | VARIABLES | OUTLIERS | OUTLIERS SEVERO |
|---|---|---|---|
| 1 | Cantidad total de retirados | - | 2123 |
| 2 | Superficie km2 | 99 y 135 | 1024 |
| 3 | Población 2020 | 503147 y 578605 | - |
| 4 | Tasa de pobreza por ingresos % | 14,14 | - |
| 5 | Personas en hogares carentes de servicios básico % | 20,1 | - |
| 6 | Hogares hacinados % | - | - |
| 7 | Cantidad de establecimientos municipales | 44 | - |
| 8 | Cantidad de establecimientos particular subvencionado | - | 147 y 163 |
| 9 | Cantidad de establecimientos particular pagado | 44 | - |
| 10 | Cantidad total de matrículas en los establecimientos | 81.309 y 81.529 | - |
| 11 | Ingresos Educación (M$) | 44285876 | 68086543 |
| 12 | Aporte Municipal al Sector Educación (M$) | 8810430 y 10749110 | - |
| 13 | Gastos Educación (M$) | 44979840 | 68467876 |
| 14 | Tasa de denuncias por delito de mayor connotación social | - | 21169,7 |
| 15 | Tasa de denuncias por violencia intrafamiliar | 1520,3 | - |

Fuente: Elaboración propia

Para el caso de diagramas de Box-Plot, véase anexo 13.2.1 que contiene detalles respecto la presencia de los outliers, con la diferencia que su rango, el cual los datos puedan estar dentro de lo normal, son más pequeñas.

Cabe mencionar que no todas las comunas tienen las mismas cantidades respecto a los números de alumnos que se han matriculado y también sobre cuántos de ellos se han retirado, esto se debe a la población de habitantes que tiene en cada comuna. Para este caso se determinó los porcentajes de alumnos matriculados para cada comuna, y como resultado la comuna con mayor porcentaje es "La Cisterna", mientras que otra comuna "San Joaquín" es la que tiene menor porcentaje de ellos, todo esto en base a la proporción sobre la población registrada en 2020 lo cual se podrá visualizar en la figura N°7, destacando que los porcentajes de varios son similares:

Figura N.° 7, Proporción de matriculados en base a la población total de cada comuna

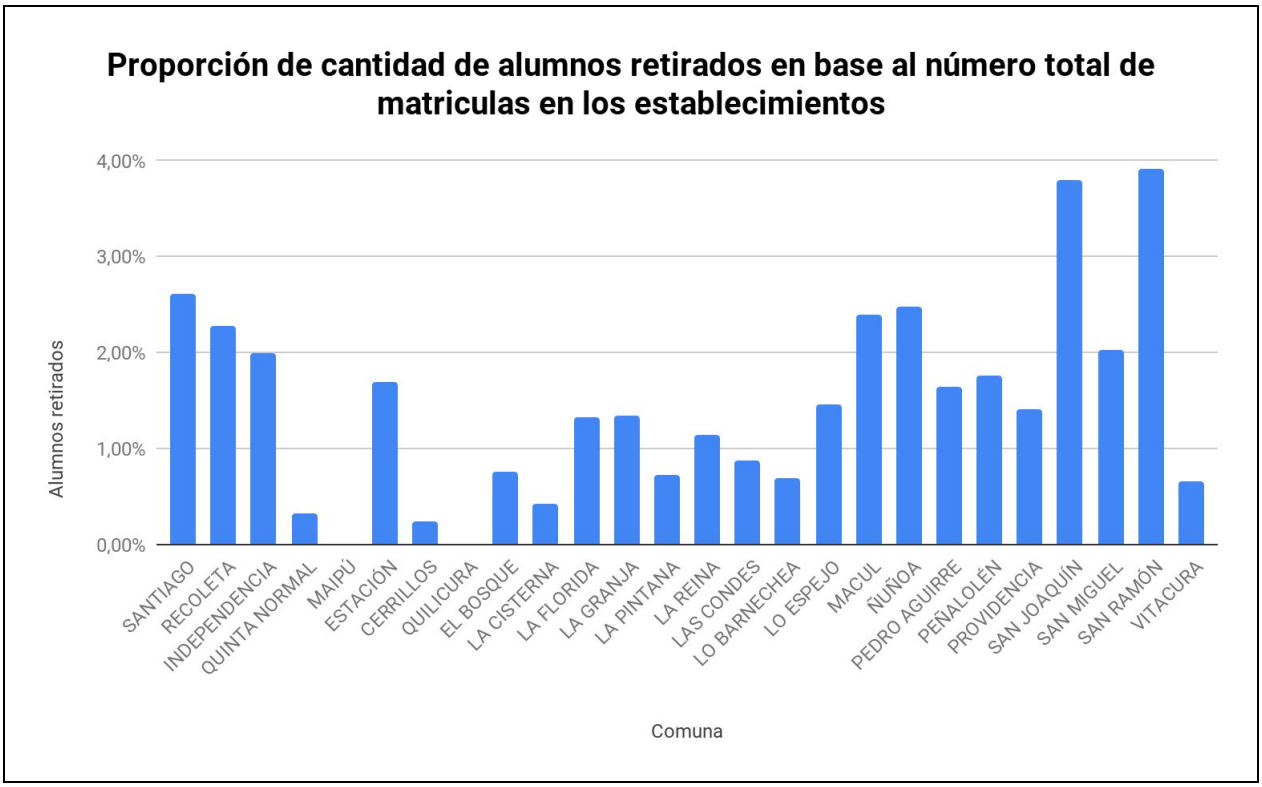



Fuente: Elaboración propia

En cuanto a la proporción de cantidad de retirados en base a número de matrículas registradas en las comunas, se obtuvieron porcentajes los cuales indican una contradicción a la detección de outliers con el método basado en el recorrido intercuartílico y boxplot, ya que Santiago, en base a esta proporción, no se presenta como outliers en comparación con el resto, y la comuna "San Ramón" es el que ha tenido mayor porcentaje sobre el retiro de los alumnos, destacando por otra

parte que todos los porcentajes indica un valor pequeño sobre estos retiros. Todo lo anterior se podrá visualizar en la figura N°8.

Figura N.° 8, Proporción de cantidad de alumnos retirados en base al número total de matrículas de cada comuna




Fuente: Elaboración propia

Por lo tanto, para estos dos casos particulares se recomienda no realizar algún tratamiento sobre la presencia de los outliers.

## 3.4 Estrategias de Segmentación y recuperación de Outliers

Los datos atípicos (outliers) pueden ser un problema para precisar mejor los resultados de predicción, y como se busca predecir la cantidad de alumnos retirados a de 26 instituciones educacionales de colegios de la Región Metropolitana. Para este caso es importante tener definido una estrategia para reducir su impacto, ya que si no se considera un dato con valor muy extremo, tiene mayores consecuencias en la estimación de la media que eliminar un dato de la región con mayor densidad. En el ámbito de inferencia, las pruebas de hipótesis son sensibles al

incumplimiento de supuestos en los modelos y a la presencia de outliers, pero no significa que todos estos datos sean erróneos. Cuando los datos no cumplen con estos supuestos disminuye la capacidad de detectar efectos reales, por lo que cualquier interpretación de los datos pueden ser erróneas.

Para el tratamiento de estos outliers los cuales perjudican los resultados de predicción respecto a la variable objetivo (cantidad de alumnos retirados), se procedió con la determinación de promedios para cada columna que contengan presencias de estos outliers y que son severos, para después reemplazar su valor y así poder equilibrar el conjunto de datos. Cabe destacar que se aplicará a todos los campos, con excepción a la variable objetivo (cantidad de retirados) y la cantidad total de matrículas. A partir de las figuras N° 9 y 10, se visualiza los nuevos valores de promedios, desviación estándar, la mediana, el valor máximo, mínimo, primer y tercer cuartil para cada columna tratada. Además, se muestra un cambio de fuerza en el correlación de Pearson visualizada en la figura N°11, donde la asociación de fuerza ha disminuido en las variables ya mencionadas anteriormente. Para mayor detalles respecto a los cambios producidos mediante diagramas de Box-Plot, véase anexo 13.2.2.

Figura N.° 9, Descripción estadístico del datasets con outliers tratados, parte 1

| | CANTIDAD TOTAL DE RETIRADOS | Superficie km2 | poblacion 2020 | tasa de pobreza por ingresos % | Personas en hogares carentes de servicios básico % | Hogares hacinados % | cantidad de establecimientos municipales | cantidad de establecimientos particular subvencionado | cantidad de establecimientos particular pagado | cantidad de matriculas totales | Ingresos Educación (M$) | Aporte Municipal al Sector Educación (M$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 26.000000 | 2.600000e+01 | 2.600000e+01 |
| mean | 378.269231 | 26.184615 | 168441.538462 | 4.576538 | 6.561154 | 15.596154 | 7.015385 | 34.461538 | 8.076923 | 29236.615385 | 1.594145e+07 | 2.060179e+06 |
| std | 416.096677 | 22.044386 | 78977.665966 | 2.552959 | 3.742346 | 5.447236 | 7.073765 | 17.586883 | 9.557921 | 19675.390914 | 6.854353e+06 | 1.527255e+06 |
| min | 4.000000 | 7.000000 | 86510.000000 | 0.130000 | 0.190000 | 2.900000 | 0.000000 | 1.000000 | 0.000000 | 7022.000000 | 6.930984e+06 | 1.623700e+05 |
| 25% | 163.000000 | 10.000000 | 104849.500000 | 3.062500 | 4.200000 | 12.700000 | 0.000000 | 24.250000 | 0.250000 | 17897.500000 | 1.068189e+07 | 8.755062e+05 |
| 50% | 286.500000 | 15.500000 | 139216.500000 | 4.800000 | 6.100000 | 17.200000 | 7.000000 | 35.500000 | 3.000000 | 21702.000000 | 1.393831e+07 | 1.733466e+06 |
| 75% | 424.500000 | 30.025000 | 189890.000000 | 6.412500 | 8.875000 | 18.425000 | 12.750000 | 43.000000 | 15.750000 | 34978.750000 | 2.116664e+07 | 2.876650e+06 |
| max | 2123.000000 | 70.200000 | 402433.000000 | 9.580000 | 14.700000 | 24.800000 | 21.000000 | 71.000000 | 31.000000 | 81529.000000 | 3.288611e+07 | 5.717647e+06 |

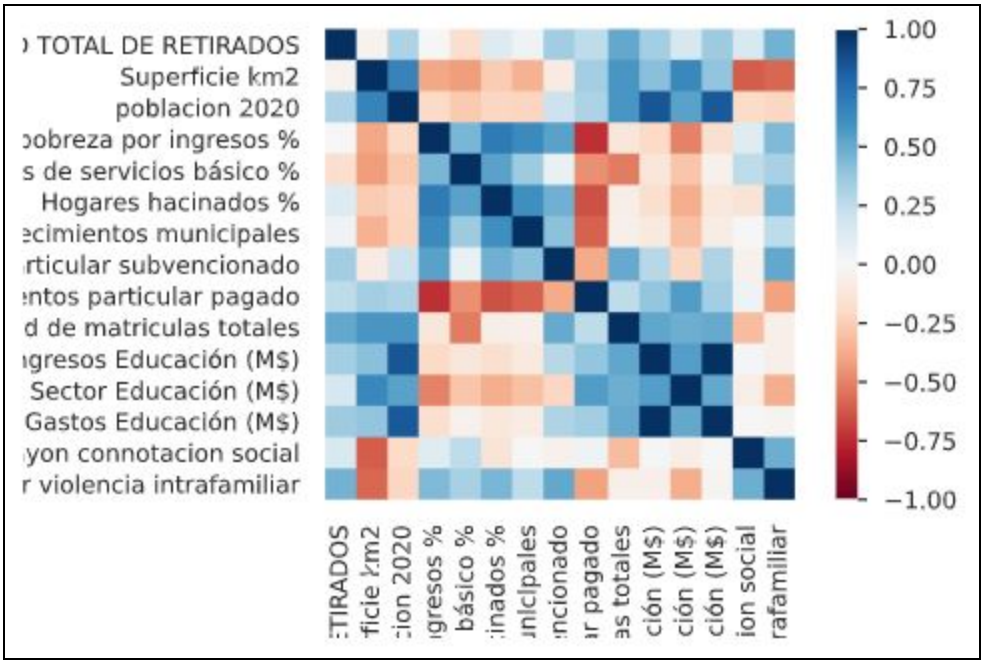
Fuente: Elaboración propia en base al código

Figura N.° 10, Descripción estadístico del datasets, parte 2

| | Gastos Educación (M$) | tasa de denuncias por delito de mayon connotacion social | tasa de denuncias por violencia intrafamiliar |
|---|---|---|---|
| count | 2.600000e+01 | 26.000000 | 26.000000 |
| mean | 1.605159e+07 | 4367.011538 | 610.603846 |
| std | 6.696983e+06 | 2358.939278 | 277.929048 |
| min | 6.658703e+06 | 868.700000 | 193.400000 |
| 25% | 1.072232e+07 | 2701.500000 | 421.850000 |
| 50% | 1.505666e+07 | 3795.750000 | 629.750000 |
| 75% | 2.059408e+07 | 6114.850000 | 726.600000 |
| max | 3.156323e+07 | 8697.200000 | 1283.000000 |

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 11, Correlación de Pearson entre variables con Outliers tratados



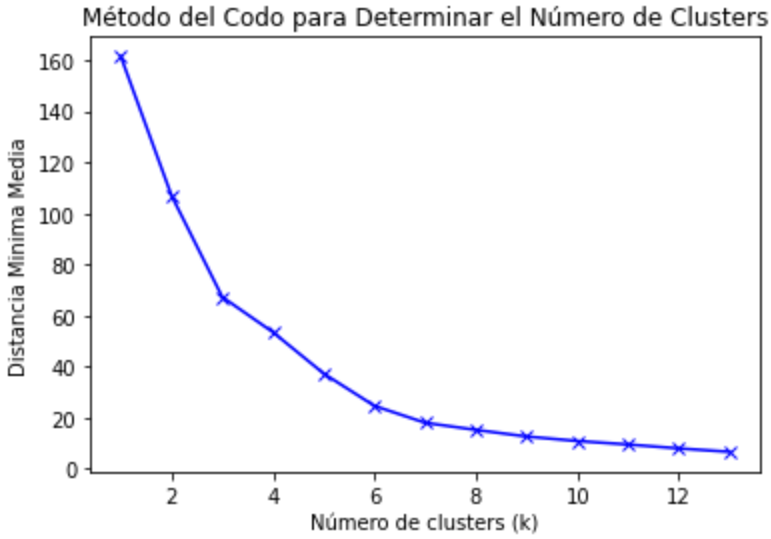

Fuente: Elaboración propia en base al código del proyecto

# 4. Aplicación de Clustering

## 4.1 Aplicación de KNN

En esta sesión, se procedió a aplicar el algoritmo K-means para graficar el agrupamiento particional, respecto la cantidad de alumnos que se retiraron y la población registrada del año 2020. Para este caso se logró determinar el número de clusters a través de método de codo, lo cual son entre 3 y 7, donde el primero implica que con este se produce un gran cambio en la curva, y este último donde se estabiliza la pendiente de la misma curva. A partir del gráfico número de clusters (k) vs distancia mínima media visualizada en la figura N°12.

Figura N.° 12, Gráfico Número de clusters vs Distancia Mínima Media



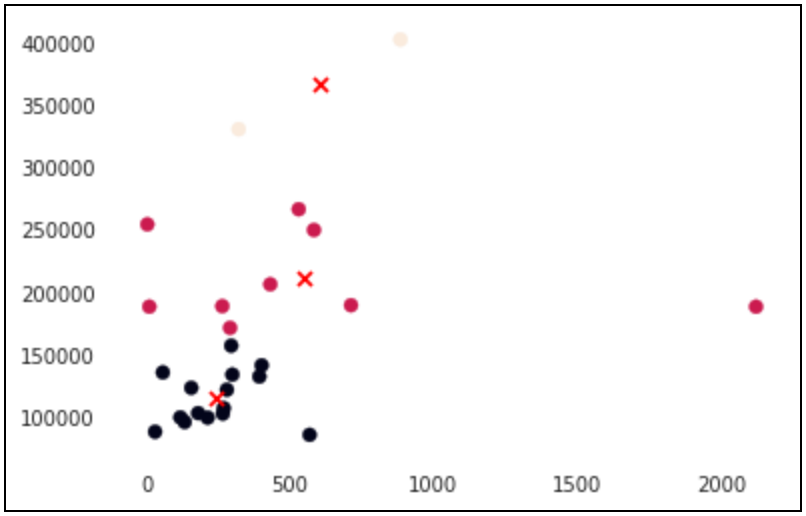

Fuente: Elaboración propia en base al código del proyecto

Por lo tanto, el número de clusters donde se produce un gran cambio en la curva es 3. A continuación se detallarán los gráficos para grupos KNN de 3,4,5, 6 y 7.

### 4.1.1 KNN para 3 Grupos:

A partir de la figura N°13, se visualiza el gráfico respecto al clustering con 3 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides, destacando que mayor distancia implica mayor diferencia de estos.

Figura N.° 13, Clustering basados en alumnos retirados y población 2020 con 3 grupos

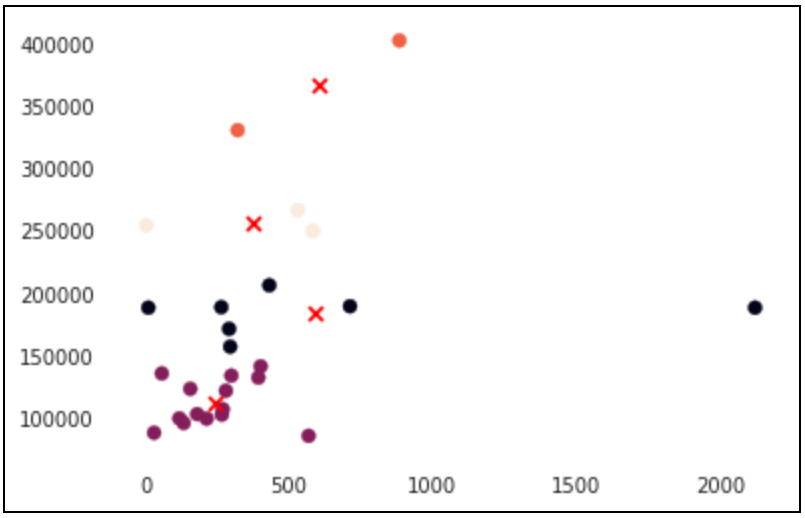



Fuente: Elaboración propia en base al código del proyecto

### 4.1.2 KNN para 4 Grupos:

A partir de la figura N°14, se visualiza el gráfico respecto al clustering con 4 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides, destacando que mayor distancia implica mayor diferencia de estos.

Figura N.° 14, Clustering basados en alumnos retirados y población 2020 con 4 grupos

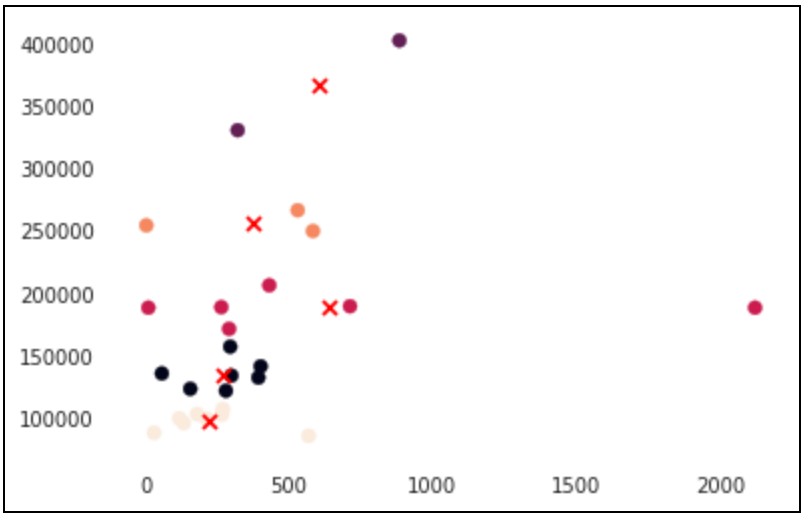



Fuente: Elaboración propia en base al código del proyecto

### 4.1.3 KNN para 5 Grupos:

A partir de la figura N°15, se visualiza el gráfico respecto al clustering con 5 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides, destacando que mayor distancia implica mayor diferencia de estos.

Figura N.° 15, Clustering basados en alumnos retirados y población 2020 con 5 grupos

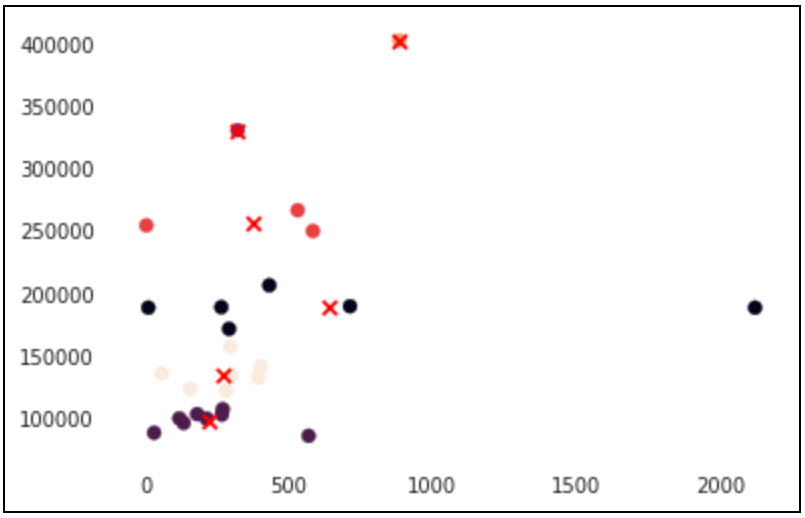



Fuente: Elaboración propia en base al código del proyecto

### 4.1.4 KNN para 6 Grupos:

A partir de la figura N°16, se visualiza el gráfico respecto al clustering con 6 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides, destacando que mayor distancia implica mayor diferencia de estos.

Figura N.° 16, Clustering basados en alumnos retirados y población 2020 con 6 grupos

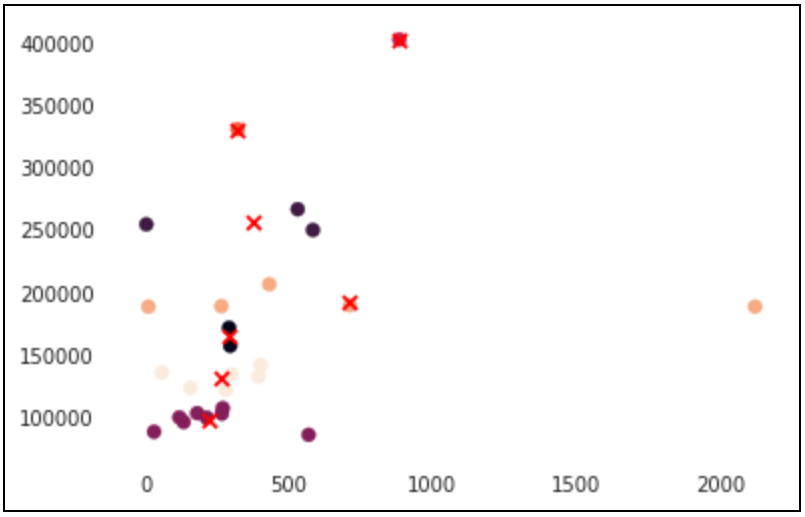



Fuente: Elaboración propia en base al código del proyecto

### 4.1.5 KNN para 7 Grupos:

A partir de la figura N°17, se visualiza el gráfico respecto al clustering con 7 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides, destacando que mayor distancia implica mayor diferencia de estos.

Figura N.° 17, Clustering basados en alumnos retirados y población 2020 con 7 grupos



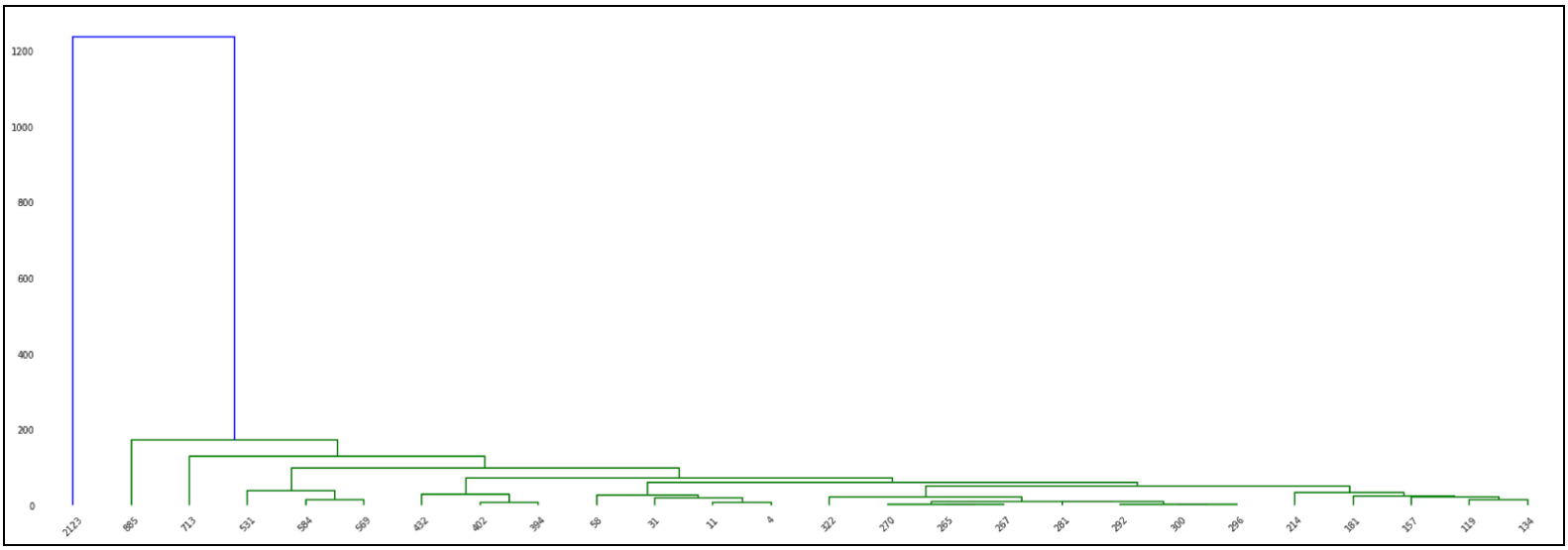

Fuente: Elaboración propia en base al código del proyecto

## 4.2 Agrupación Jerárquica

Para poder organizar los datos jerárquicamente, se utiliza la estructura de árbol o dendrograma, lo cual este tipo de estructura permite determinar el número adecuado de agrupamientos como también para la detección de Outliers.

Para el caso de agrupación respecto la cantidad de alumnos que se retiraron del establecimiento, a partir de la figura N°18 se visualiza el dendograma, lo cual al principio solo se visualiza un cluster donde contiene todos los datos (1200 alumnos retirados aproximadamente), y para cada paso se van dividiendo los clusters hasta formar uno solo como parte de la totalidad, siendo que desde el segundo paso de la división los valores de otros clusters se vuelven muy pequeños como resultado.

Figura N.° 18, Dendograma para el caso de la cantidad de alumnos retirados




Fuente: Elaboración propia en base al código del proyecto

# 5. Análisis de Componentes Principales (PCA) y Factorial

## 5.1 Análisis de PCA

El Análisis de Componentes Principales (PCA), por sus siglas en inglés *Principal Components Analysis*, este es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez, pero con la finalidad de conservar la información y si es posible perder los mínimos datos de estos. Este método permite, por lo tanto, "condensar" la información aportada por múltiples variables en solo unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas, tales como regresión, clustering, etc. Aun así no hay que olvidar que sigue siendo necesario disponer valores de las variables originales para calcular dichos componentes.

Para este caso, se ocupará un Dataset sobre modelo predictivo de rendimiento escolar, lo cual se realizará en el lenguaje de programación Python para su mayor y mejor comprensión de datos. Por lo tanto, para adquirir un buen PCA se deben realizar los siguientes pasos:

1. **Normalizar los Datos:** Para efectos de codificación, se debe realizar este paso con el fin de aplicar después una distribución normal en los datos, para esto, todos y cada uno de los datos del Dataset deben ser de tipo numérico, en este caso, existía una columna de "Comunas", lo cual representaba valores string y que luego fueron transformadas a valores enteros como se dio a conocer en la tabla anterior.

2. **Calcular la matriz de Covarianza:** Luego de normalizar los datos a utilizar, se debe calcular la matriz de covarianza para obtener los valores de ella, ver la distancias y además verificar que estos valores sean numéricos y viables para ser utilizados.

3. **Calcular el vector con los valores eigen:** Estos corresponden a números (eigenvalores) y vectores (eigenvectores) asociados a matrices cuadradas, los cuales proporcionan la información sobre cuando son transformados por el operador, dando el lugar a un

múltiplo escalar de sí mismos, con lo que no cambian su dirección. Cabe mencionar que una transformación queda completamente determinada por sus vectores propios y valores propios, por lo que, para poder finalizar esta transformación de datos y este análisis PCA, se deben obtener estos valores y vectores para finalizar este proceso.

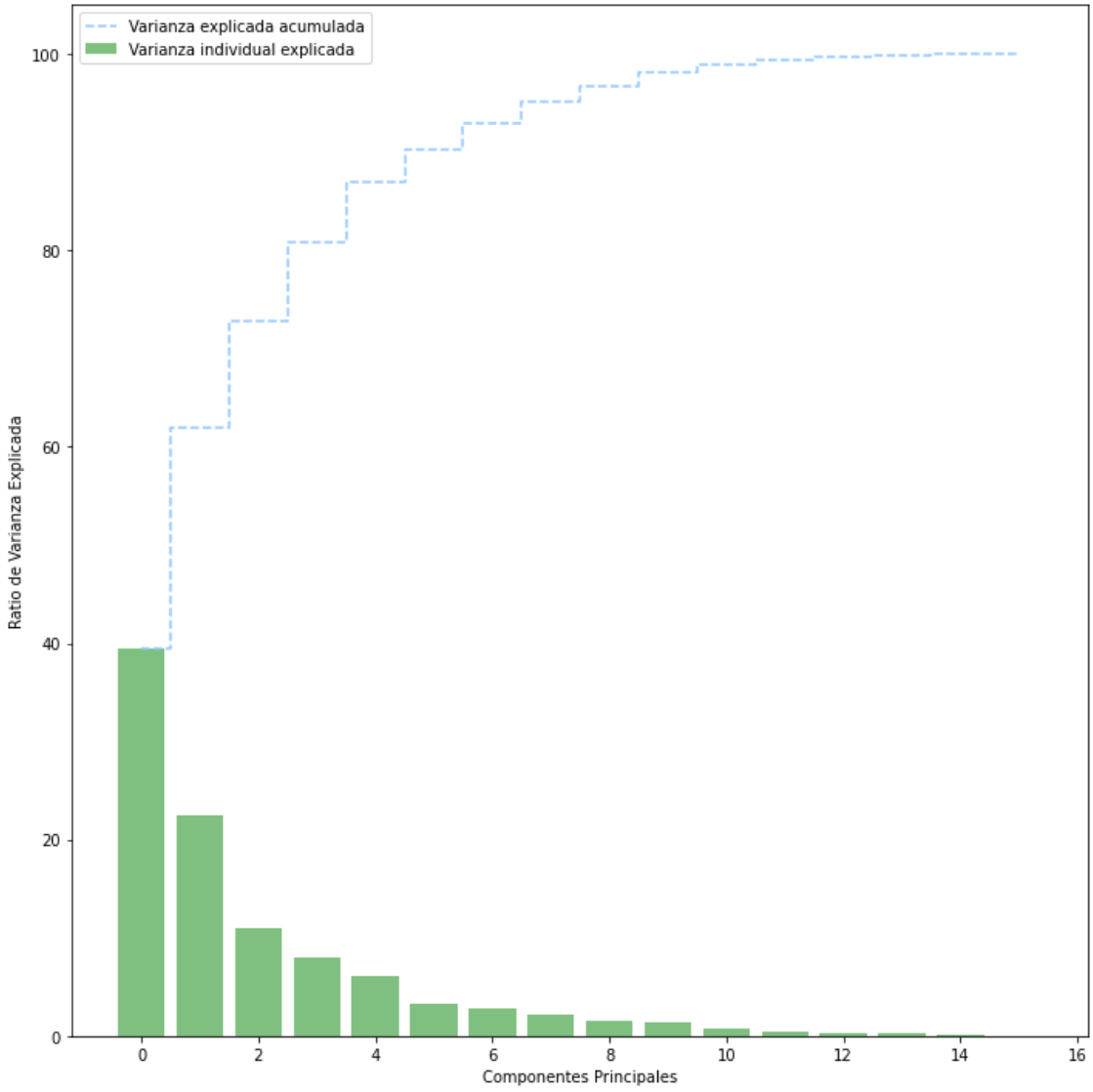
En el método PCA, cada una de las componentes se corresponde con un eigenvector, y el orden de componente se establece por orden decreciente de eigenvalue. Así pues, la primera componente es el eigenvector con el eigenvalor asociado más alto.

Tras estos pasos fundamentales para un análisis de componentes principales y a su vez esta transformación de data, se deben listar los valores y vectores propios que se obtuvieron, para que a partir de estos se calcule la varianza, lo cual representará en un diagrama de barras la varianza por cada autovalor y la acumulada de estos. Por consiguiente, se generó una matriz a partir de los pares autovalor-autovector y se representó gráficamente, obteniendo el resultado de este PCA y la transformación de los dichos datos utilizados.

## 5.2 Resultados del Análisis de PCA

A partir de la figura N.°19 se puede observar la distribución por cada varianza de los datos recaudados en el dataset, a su vez, destacando que hay 3 componentes que tienen más peso en la información. Cabe destacar que mientras un componente tenga mayor varianza individual, esto implica que este componente puede aportar mucha información en comparación con el resto, sin embargo, si este componente sobrepasa al 70% de dicha varianza significa que al datasets le faltan agregar más campos para poder buscar diferencias en otras columnas.

Figura N.° 19, Gráfico de barras de varianzas para el modelo de análisis PCA




Fuente: Elaboración propia en base al código del proyecto

## 5.3 Análisis Factorial

El análisis factorial agrupa una serie de procedimientos de análisis multivariable, que analizan la relación mutua entre variables. En este sentido, el análisis factorial permite estudiar la interdependencia entre un conjunto de variables.

La idea fundamental en el análisis factorial es, como dice su nombre, analizar la correlación existente entre una serie de variables, con el propósito de descubrir alguna estructura latente (no

directamente observable). Se busca la reducción de la información proporcionada por "**p**" variables observadas, con la menor pérdida posible de información, en un número inferior de "**k**" variables no observadas. Además, la reducción o agrupación de variables en factores o componentes principales se caracteriza por:

- Aglutinar bajo cada factor o componente variables que estén muy correlacionadas entre ellas.
- Garantizar que las variables agrupadas en distintos factores o componentes están poco correlacionadas.

De hecho, entre factores o componentes, la correlación será igual a cero. Esta característica nos indica que cada factor o componente mide o representa una dimensión distinta en los datos.

## 5.4 Resultados del Análisis Factorial

El análisis de componentes principales (PCA) y el análisis factorial son similares porque ambos análisis se utilizan para simplificar la estructura de un conjunto de variables. Sin embargo, los análisis difieren de varias maneras importantes:

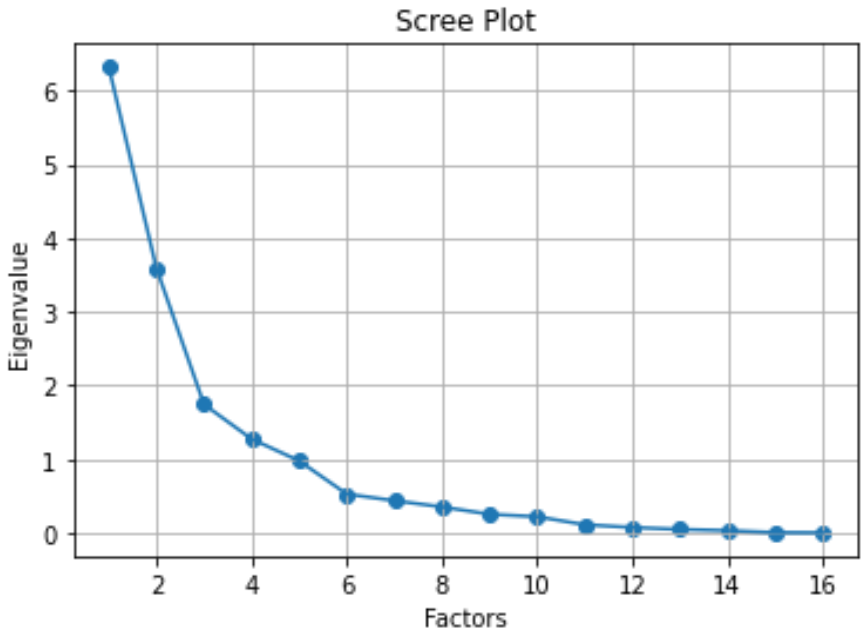
- En el análisis de componentes principales, los componentes se calculan como combinaciones lineales de las variables originales. En el análisis factorial, las variables originales se definen como combinaciones lineales de los factores.

- En el análisis de componentes principales, la meta es explicar tanta proporción de la varianza total en las variables como sea posible. La meta en el análisis factorial es explicar las covarianzas o correlaciones entre las variables.
- El análisis de componentes principales para reducir los datos a un número más pequeño de componentes. Utilice el análisis factorial para entender los constructos que subyacen a los datos.

Pero cabe mencionar que, cada uno de estos métodos nos proporcionan una estabilidad al transformar datos, es por ello que son muy útiles para trabajar en la reducción de ellos.

A partir de la figura N°20 se visualiza el gráfico de codo para obtener los números de factores, lo cual con 3 factores se produce un quiebre en la curva del gráfico.

Figura N.° 20, Gráfico de Codo para obtener número de factores



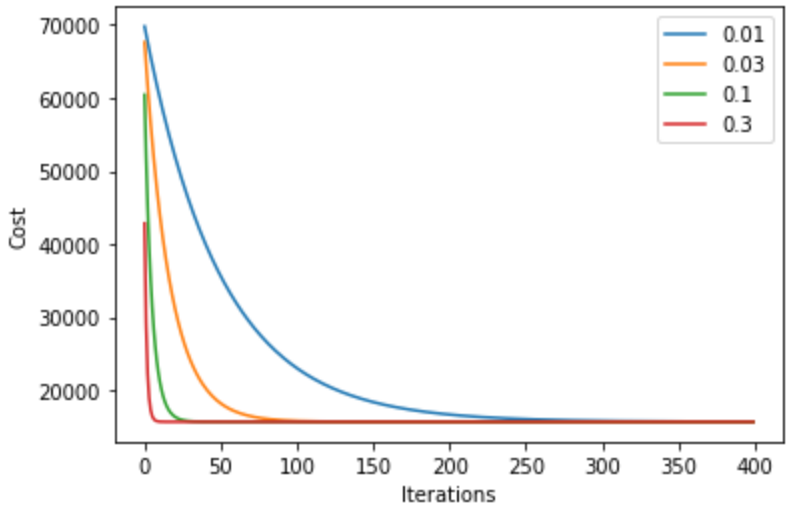

Fuente: Elaboración propia en base al código del proyecto

# 6. Regresión Lineal Multivariable

Para disponer de una mejor comprensión y predicción del modelo a estudiar, una regresión lineal multivariable nos proporciona información para entender la relación funcional entre la variable dependiente y las variables independientes y estudiar cuáles pueden ser las causas de la variación de Y.

Como el Dataset utilizado contiene múltiples variables, un modelo de regresión lineal simple no nos proporciona la información suficiente, además de arrojar error en dimensiones y número de datos, es por esto que, la regresión lineal multivariable, trabaja con un aumento de variables para determinar la precisión de estos, en este caso, obtener información sobre la cantidad total de retirados por comuna que puede haber en la región metropolitana. A partir de la figura N.° 21 se identifican las convergencias para 4 tipos distintos de casos, en costos vs. iteraciones del modelo.

Figura N.° 21, Costo vs Iteraciones para la regresión lineal multivariable




Fuente: Elaboración propia en base al código del proyecto

Tras esto, se determinó la predicción para cada una de las variables de cantidad de retirados total, observando la figura N.° 22, se determina que el modelo predice cercanamente al 100% en la mayoría, teniendo altas y bajas sin duda, pero se puede apreciar que existen casos en las que el modelo se acerca a los datos originales y esto es por el entrenamiento que se lleva a cabo, sin

embargo existen predicciones anómalas, como por ejemplo el número 7 llegando a los 1750% aproximadamente, esto quiere decir que el modelo sobrepasó al dato original mostrando más retirados de lo que debía tener, esto ocasionado por el error cuadrático medio calculado con un valor de 22.230 aproximadamente, todo esto con las mismas características y ambiente para el proyecto a modelar situándose en el año 2018.

Figura N.° 22, Porcentaje de precisión para cada variable y con error cuadrático medio

```
0 cantidad retirados 1928.0 / % 90.81488459726802
1 cantidad retirados 537.0 / % 75.31556802244039
2 cantidad retirados 384.0 / % 95.5223880597015
3 cantidad retirados 33.0 / % 56.89655172413793
4 cantidad retirados 232.0 / % 2109.090909090909
5 cantidad retirados 578.0 / % 133.7962962962963
6 cantidad retirados 319.0 / % 1029.032258064516
7 cantidad retirados 70.0 / % 1750.0
8 cantidad retirados 147.0 / % 50.342465753424655
9 cantidad retirados 150.0 / % 126.05042016806723
10 cantidad retirados 904.0 / % 102.14689265536722
11 cantidad retirados 320.0 / % 113.87900355871886
12 cantidad retirados 470.0 / % 177.35849056603774
13 cantidad retirados 327.0 / % 152.80373831775702
14 cantidad retirados 204.0 / % 63.35403726708075
15 cantidad retirados 15.0 / % 9.554140127388536
16 cantidad retirados 391.0 / % 216.02209944751382
17 cantidad retirados 300.0 / % 100.0
18 cantidad retirados 836.0 / % 143.15068493150685
19 cantidad retirados 278.0 / % 102.96296296296296
20 cantidad retirados 364.0 / % 68.54990583804143
21 cantidad retirados 222.0 / % 75.0
22 cantidad retirados 130.0 / % 48.68913857677903
23 cantidad retirados 355.0 / % 90.1015228426396
24 cantidad retirados 312.0 / % 54.83304042179262
25 cantidad retirados 30.0 / % 22.388059701492537

rmse = mean_squared_error(y,prediccioness)
print("El error (rmse) de test es: ", rmse)

El error (rmse) de test es:  22230.094606176965
```

Fuente: Elaboración propia en base al código del proyecto

# 7. Árbol de Decisión, Random Forest y AdaBoost

## 7.1 Árbol de decisión

El árbol de decisión es un método analítico que, a través de una presentación esquematizada, entrega posibles resultados de una serie de diferentes decisiones relacionadas. Este modelo permite que un individuo o una organización comparen posibles acciones entre sí, según sus costos, probabilidades y beneficios. Se pueden usar para dirigir un intercambio de ideas informal o trazar un algoritmo, que anticipe matemáticamente la mejor opción. Un árbol de decisión, por lo general, comienza con un nodo y de este se comienza a ramificar posibles resultados, de los cuales se crean nodos adicionales creando la forma similar a un árbol.

Los resultados que se obtuvieron al construir y ejecutar el programa python con Colab, con el dataset del proyecto definido, son la generación de los tres árboles, donde uno tiene 3 niveles de profundidad, otro tiene 4 niveles de profundidad y por último 5 niveles de profundidad. Todo esto se podrá ver con detalles en el anexo 13.3.

## 7.2 Random Forest

Un Random Forest (Bosque Aleatorio) es una técnica de aprendizaje automático supervisada que, a diferencia del árbol de decisión, tienen una capacidad de generalización muy alta para muchos problemas en cuanto al error. Este es un conjunto (ensemble) de árboles de decisión combinados con bagging, lo cual al usar este último puede ocurrir que distintos árboles vean distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento, por lo que esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

## 7.3 AdaBoost:

El algoritmo AdaBoost proviene del término Boosting que es un meta-algoritmo de aprendizaje automático que reduce el sesgo y varianza en un contexto de aprendizaje supervisado. Este es considerado como un clasificador débil, ya que, solamente se correlaciona con la clasificación correcta.

Ahora bien, AdaBoost es el algoritmo más popular y es quizá el más importante históricamente, ya que fue la primera formulación de un algoritmo que pudo aprender de a partir de los clasificadores débiles . Aun así, hay muchos algoritmos más recientes como LPBoost, TotalBoost, BrownBoost, xgboost, MadaBoost, LogitBoost que realizan la misma tarea. Muchos algoritmos de boosting encajan en el marco de AnyBoost, el cual muestra que los algoritmos de boosting actúan a través del descenso del gradiente en el espacio funcional, utilizando una función de coste convexa. Este algoritmo lo que hace es clasificar un conjunto de datos originales y luego los ajusta, les hace una copia adicional del clasificador en el mismo conjunto de datos, pero donde los pesos de las instancias clasificadas incorrectamente se ajustan de modo que los clasificadores posteriores se enfocan más en casos difíciles.

Como resultado de la ejecución del programa para las diferentes ramificaciones realizadas, resulta que el mejor árbol de decisión es el árbol de profundidad 3, con tratamiento de outliers y normalización scale. dando los siguientes resultados:

- Error cuadrático medio: 0,3646%
- Precisión datos prueba: 19,58%
- Precisión datos entrenamiento: 95,21%

Por lo tanto, este modelo ha logrado obtener una mayor precisión y menor error en comparación a los otros, destacando que esta puede mejorar con el uso de Adaboost.

# 8. Redes Bayesiana y Máquinas de Soporte Vectorial (SVM)

## 8.1 Redes Bayesiana

Una red bayesiana es un modelo probabilístico que relaciona un conjunto de variables aleatorias mediante un grafo dirigido, que se caracteriza por ser red gráfica sin ciclos en el que se representan variables aleatorias y las relaciones de probabilidad que existan entre ellas, lo cual permiten conseguir soluciones a problemas de decisión en casos de incertidumbre. Cabe destacar que esta red corresponde a una representación ilustrada de dependencias para razonamiento probabilístico, en la cual los nodos representan variables aleatorias y los arcos simbolizan relaciones de dependencia directa entre las variables.

## 8.2 Máquinas de Soporte Vectorial (SVM)

Las máquinas de vector de soporte, en inglés Support Vector Machines (SVM), es un clasificador discriminatorio, lo cual, dados los datos de entrenamiento en el aprendizaje supervisado, el algoritmo genera un hiperplano óptimo que clasifica en dos espacios dimensionales dentro de un conjunto de datos. Cabe destacar que, en SVM, se pretende hallar un hiperplano para la separación lineal de las clases aumentado las dimensiones del plano original, lo cual su funcionamiento es reescribir el algoritmo para simplificar y trabajar únicamente con el producto punto entre el vector de entradas y de pesos, para después reemplazar esta última ecuación por una función de Kernel.

Kernel son los encargados de obtener el hiperplano con dimensiones mayores, lo cual permite obtener una clasificación, los cuales se encuentran lineales, polinomiales, gaussianos, tangentes hiperbólicos, entre otros. Cabe destacar que el Kernel lineal corresponde a un operador de mapeo en un hiperplano con mayores dimensiones y el gaussiano, a un espacio infinito dimensional. Para este caso se utilizó tres tipos de Kernel distintos para la ejecución del programa con dataset que se ha utilizado durante el transcurso del proyecto, los cuales se explicarán a continuación.

### 8.2.1 Kernel 1, RBF:

El Kernel RBF (Radial basis function) es uno de los núcleos más utilizados debido a su cálculo similitud con la distribución de Gauss, es decir, calcula lo cerca que se encuentra entre dos puntos.

### 8.2.2 Kernel 2, Polinomial ("POLY"):

Por último, el Kernel polinomial es un núcleo que representa la similitud de vectores en un espacio de características sobre polinomios de las variables originales, lo que permite el aprendizaje de modelos no-lineales. Cabe destacar que el núcleo polinomial observa no solo las características dadas de las muestras de entrada para determinar su similitud, sino también las combinaciones de estas. En el contexto del análisis de regresión, estas combinaciones se conocen como características de interacción. El espacio de características (implícito) de un núcleo polinomial es equivalente al de la regresión polinomial, pero sin el aumento combinatorio en el número de parámetros a aprender, por lo que cuando las características de entrada tienen valores binarios, las características corresponden a conjunciones lógicas de características de entrada.

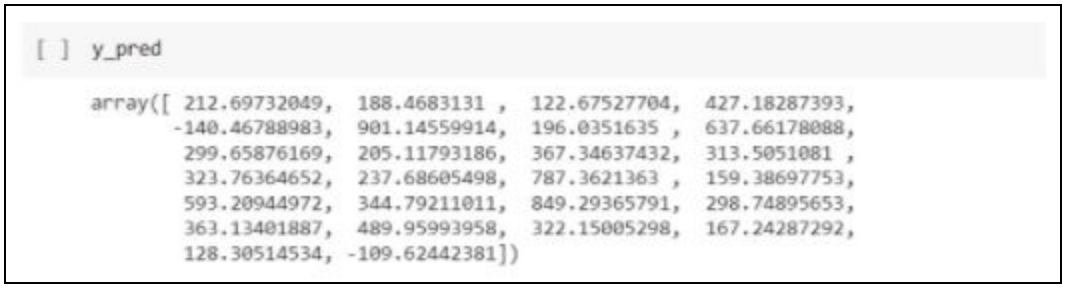
# 9. Redes Neuronales

Las redes neuronales son un conjunto de neuronas conectadas entre sí y que trabajan en conjunto, sin que haya una tarea concreta para cada una. La principal idea de estas redes es la idea de imitar el funcionamiento de las redes neuronales de los organismos vivos, ya que a medida que se va adquiriendo la experiencia, las neuronas van creando y reforzando ciertas conexiones para "aprender" algo que se queda fijo en el tejido. Cabe destacar que este es un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo, por lo que una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones.

Para este caso en particular se utilizarán redes neuronales para entrenar en 150 épocas y batch_size = 50. Con estas iteraciones se puede mejorar y predecir la cantidad de retirados que puede tener una comuna según las características mencionadas en el dataset.

La predicción arrojada con este modelo se ve en la figura N°26.

Figura N.° 26, Resultado de predicción a través de redes neuronales



```
[ ] y_pred

    array([ 212.69732049,  188.4683131 ,  122.67527704,  427.18287393,
           -140.46788983,  901.14559914,  196.0351635 ,  637.66178088,
            299.65876169,  205.11793186,  367.34637432,  313.5051081 ,
            323.76364652,  237.68605498,  787.3621363 ,  159.38697753,
            593.20944972,  344.79211011,  849.29365791,  298.74895653,
            363.13401887,  489.95993958,  322.15005298,  167.24287292,
            128.30514534, -109.62442381])
```

Fuente: Elaboración propia en base al código del proyecto

# 10. Evaluación del Modelo

Tras un análisis exhaustivo de cada tipo de ramificación y modelo, de cada dato relevante para tomar alguna decisión sobre el retiro de un estudiante, se llegaron a precisiones diversas, como se muestra a continuación.

- Tratamiento de outlier: C.TO (con tratamiento de outlier)/ S.TO (sin tratamiento de outliers)
- Normalización : S (scale)/ N (normalize)/ MM( MixMax) / S.N (Sin normalización)
- PCA: C.PCA (con PCA) / S.PCA( sin PCA)

- **Regresión Multivariable:**

Tabla N.° 3, Resultados en la regresión multivariable, error cuadrático medio (RMSE)

| RAMAS | ERROR CUADRÁTICO MEDIO |
|---|---|
| C.TO/S.N/C.PCA | 145457,34 |
| C.TO/S.N/S.PCA | 22230,09 |
| C.TO/N/C.PCA | 157890,67 |
| C.TO/N/S.PCA | 59948,35 |
| C.TO/MM/C.PCA | 45638,18 |
| C.TO/MM/S.PCA | 22230,09 |
| C.TO/S/C.PCA | 119823,33 |
| C.TO/S/S.PCA | 22230,09 |
| S.TO/S.N/C.PCA | 77816,74 |
| S.TO/S.N/S.PCA | 33478,98 |
| S.TO/N/C.PCA | 156457,43 |
| S.TO/N/S.PCA | 115080,32 |

| | |
|---|---|
| S.TO/MM/C.PCA | 45638,18 |
| S.TO/MM/S.PCA | 33478,98 |
| S.TO/S/C.PCA | 46279,42 |
| S.TO/S/S.PCA | 33478,98 |

Fuente: Elaboración propia

Como se puede observar en la tabla anterior, claramente no es buen modelo para predecir la cantidad de alumnos que se retiran, siendo la rama de normalización con mejor resultado es Scale, con dataset ya tratados con outliers y sin PCA.

- ÁRBOLES:
  - TIPO: P3 (profundidad 3)/ P4 (profundidad 4) / P5 (profundidad 5) / P4A (profundidad 4 adaboost) / RF( random forest), considerando la misma nomenclatura de regresión multivariable

Tabla N.° 4, Resultados respecto árbol de decisión de profundidad 3, error cuadrático medio (RMSE)

| RAMAS | Error cuadrático medio | Precision test | Precisión train |
|---|---|---|---|
| P3/C.TO/S.N | 148,77 | 19,58% | 95,21% |
| P3/C.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P3/C.TO/MM | 0,07 | 19,58% | 95,21% |
| P3/C.TO/S | 0,36 | 19,58% | 95,21% |
| P3/S.TO/S.N | 148,89 | 19,44% | 94,96% |
| P3/S.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P3/S.TO/MM | 0,07 | 19,44% | 94,96% |
| P3/S.TO/S | 0,36 | 19,44% | 94,96% |

Fuente: Elaboración propia

Tabla N.° 5, Resultados respecto árbol de decisión de profundidad 4, error cuadrático medio (RMSE)

| RAMAS | Error cuadrático medio | Precision test | Precisión train |
|---|---|---|---|
| P4/C.TO/S.N | 157,12 | 10,3% | 96,92% |
| P4/C.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P4/C.TO/MM | 0,07 | 10,3% | 96,92% |
| P4/C.TO/S | 0,38 | 10,3% | 96,92% |
| P4/S.TO/S.N | 184,54 | -23,74% | 97,82% |
| P4/S.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P4/S.TO/MM | 0,08 | -23,74% | 97,82% |
| P4/S.TO/S | 0,45 | -23,74% | 97,82% |

Fuente: Elaboración propia

Tabla N.° 6, Resultados respecto árbol de decisión de profundidad 4 con adaboost, error cuadrático medio (RMSE)

| RAMAS | Error cuadrático medio | Precision test | Precisión train |
|---|---|---|---|
| P4A/C.TO/S.N | 153,79 | 14,06% | 99,58% |
| P4A/C.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P4A/C.TO/MM | 0,07 | 10,86% | 99,68% |
| P4A/C.TO/S | 0,38 | 11,93% | 99,21% |
| P4A/S.TO/S.N | 167,55 | -2,01% | 99,88% |
| P4A/S.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P4A/S.TO/MM | 0,08 | -6,78% | 99,88% |
| P4A/S.TO/S | 0,41 | -4,38% | 99,79% |

Fuente: Elaboración propia

Tabla N.° 7, Resultados respecto árbol de decisión de profundidad 5, error cuadrático medio (RMSE)

| RAMAS | Error cuadrático medio | Precision test | Precisión train |
|---|---|---|---|
| P5/C.TO/S.N | 174,78 | -11% | 98,15% |
| P5/C.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P5/C.TO/MM | 0,07 | 19,58% | 95,21% |
| P5/C.TO/S | 0,42 | -11% | 95,15% |
| P5/S.TO/S.N | 177,25 | -14,16% | 99,35% |
| P5/S.TO/N | Sin resultado | Sin resultado | Sin resultado |
| P5/S.TO/MM | 0,08 | -14,16% | 99,35% |
| P5/S.TO/S | 0,43 | -14,16% | 99,35% |

Fuente: Elaboración propia

Tabla N.° 8, Resultados respecto random forest, error cuadrático medio (RMSE)

| RAMAS | Error cuadrático medio | Precision test | Precisión train |
|---|---|---|---|
| RF/C.TO/S.N | 182,05 | -20,43% | 83,6% |
| RF/C.TO/N | Sin resultado | Sin resultado | Sin resultado |
| RF/C.TO/MM | 0,09 | -35,58% | 83,85% |
| RF/C.TO/S | 0,46 | -29,37% | 83,84% |
| RF/S.TO/S.N | 169,90 | -4,89% | 84,86% |
| RF/S.TO/N | Sin resultado | Sin resultado | Sin resultado |
| RF/S.TO/MM | 0,07 | 3,86% | 84,47% |
| RF/S.TO/S | 0,40 | -0,9% | 87,68% |

Fuente: Elaboración propia

Como se puede observar, todas las normalizaciones con "Normalize" no entregan un resultado debido que este no ha funcionado con los algoritmos relacionados con árboles de decisión. La rama de normalización con mejor resultado fue Scale, con dataset ya tratados con outliers, en cuanto a la profundidad de árbol que es 3. Al visualizar estos resultados, es muy recomendable utilizar este modelo sobre árbol de decisión ya que su error cuadrático medio es más bajo y con mayor precisión (test y train).

- REDES BAYESIANAS

Tabla N.° 9, Resultados respecto redes bayesianas, error cuadrático medio (RMSE)

| RAMAS | ERROR CUADRÁTICO MEDIO |
|---|---|
| C.TO/S | 77511,5 |
| C.TO/N | 77511,5 |
| C.TO/MM | 77511,5 |
| C.TO/S.N | 441205,875 |
| S.TO/MM | 64955,5 |
| S.TO/S | 64955,5 |
| S.TO/S.N | 600759,25 |
| S.TO/N | 72476,12 |

Fuente: Elaboración propia

- SVM
    - TIPO: RBF (rbf) / PL (poly)

Tabla N.° 10, Resultados respecto SVM, error cuadrático medio (RMSE)

| RAMAS | ERROR CUADRÁTICO MEDIO |
|---|---|
| RBF/C.TO/S.N | 145,61 |
| RBF/C.TO/N | 305,49 |
| RBF/C.TO/MM | 305,49 |
| RBF/C.TO/S | 305,49 |

| | |
|---|---|
| RBF/S.TO/S.N | 687,79 |
| RBF/S.TO/N | 134,67 |
| RBF/S.TO/MM | 306,19 |
| RBF/S.TO/S | 306,19 |
| PL/C.TO/S.N | 145,82 |
| PL/C.TO/N | 279,13 |
| PL/C.TO/MM | 279,13 |
| PL/C.TO/S | 279,13 |
| PL/S.TO/S.N | 732,60 |
| PL/S.TO/N | 130,90 |
| PLS.TO/MM | 284,96 |
| PL/S.TO/S | 284,96 |

Fuente: Elaboración propia

- REDES NEURONALES

Tabla N.° 11, Resultados respecto redes neuronales, error cuadrático medio (RMSE)

| RAMAS | ERROR CUADRÁTICO MEDIO |
|---|---|
| C.TO/S.N | 527,70 |
| C.TO/N | 385,85 |
| C.TO/MM | 511,80 |
| C.TO/S | 524,86 |
| S.TO/S.N | 5431,50 |
| S.TO/N | 383,52 |
| S.TO/MM | 459,88 |
| S.TO/S | 2872,45 |

Fuente: Elaboración propia

Al analizar cada una de las diferentes ramificaciones, se determinó cual es la más apropiada para cada tipo de modelo, las cuales son las siguientes:

- **REGRESIÓN:** Con tratamiento de outliers (C.TO), con normalización scale (S), y sin uso de PCA (S.PCA). Para esta ramificación se tiene un error cuadrático medio de 22.230,09.
- **ARBOLES:** Profundidad 3 (P3), con tratamiento de outliers (C.TO) y normalización scale (S). Para esta ramificación se tiene un error cuadrático medio de 0,3646, una precisión para los datos test de 19,58% y una precisión de 95,21% para los datos de entrenamiento.
- **REDES BAYESIANAS:** Sin tratamiento de outliers (S.TO), normalización scale (S). Para esta ramificación se obtiene un error cuadrático medio de 64.955,5.
- **SVM:** Sin tratamiento de outliers (S.TO), normalización normalize (N), y kernel poly (PL). Para esta ramificación se determinó un error cuadrático medio de 130, 90
- **REDES NEURONALES:** Sin tratamiento de outliers (S.TO), normalización normalize (N). Para esta ramificación se obtuvo un error cuadrático medio de 383,5214.

De las ramificaciones anteriores, se puede decir que los modelos óptimos son las de árboles y SVM, si bien para la primera tiene el error cuadrático medio más pequeño, la precisión que este tiene en base a los datos de prueba en relativamente bajo siendo menor a un 20%, por lo cual es recomendable utilizar ambos modelos para una mejor estimación.

# 11. Conclusión

Un modelo predictivo para análisis del rendimiento escolar, en este caso el rendimiento de retiros escolares de establecimientos educacionales, proporciona una toma de decisiones para tiempos futuros.

Particularmente se logró el objetivo de predecir la cantidad de alumnos retirados, con un error de ±130 alumnos, con respecto al año estudiado, 2018, para proporcionar una visión de desempeño de establecimientos de comunas de la región metropolitana. A su vez, se recogieron datos confiables para determinar este número de alumnos, además se generaron varios modelos predictivos, lo cual el más acertado fue svm, haciendo mención al kernel poly.

Dicho y estudiado lo anterior, se puede generar conciencia y toma de decisiones como las siguientes propuestas:

- **Como la mayor cantidad de alumnos retirados se concentra en santiago**, tiene relación con los gastos educacionales y a su vez, con las denuncias de violencia en el sector, es por esto que se puede generar un resguardo para la comuna, en lugares cercanos a establecimientos educacionales, por otra parte, económicamente hablando, se tiene más gasto de lo normal, por lo que es preferible adquirir utensilios escolares justos y no tener de más, además lo ideal sería no adquirir prendas de vestir idealizando un establecimiento, ya que este genera más gastos aun.

- Por otro lado, **Quilicura es la comuna con menos alumnos retirados**, esto se debe a su baja violencia en el sector, su amplia armonía en áreas verdes, estratos sociales, además de un bajo porcentaje de hogares carentes a los servicios básicos, por lo que sería una comuna a seguir para el caso principal nombrado anteriormente.

Si bien existen más características que determinan el retiro de alumnos como el ámbito económico, el estrato social, la sociedad en la cual está inmersa la persona, simplemente se

requiere una mejor apreciación para conocer las situaciones académicas de cada alumno, por lo que es recomendable prepararse para años posteriores, analizar a los estudiantes, ponerse en su lugar y no solo verlos como un valor monetario, ya que la educación es primordial para la vida, y si los establecimientos educacionales no proporcionan esta ayuda necesaria, los alumnos seguirán retirándose, abran más baja, aumento de violencia y por último, es necesario dar ayuda económica para las familias que lo necesitan, no solamente mirar la clase baja, sino la media también, ya que a pesar de ser una clase más alta, no deja de tener escasez de economía, de problemas familiares, entre otros.

Para el modelo predictivo en general, es importante utilizar las herramientas y realizar estructuración de codificación correctamente, ya que este tipo de contexto corresponde a problemas de regresión, por lo que si se utiliza la codificación para clasificación puede ocurrir problemas en cuanto a las sintaxis y generación de resultados. Cabe destacar que en la predicción, dentro de este proyecto, en ningún caso puede estimar la cantidad de alumnos y alumnas que se retiran en sus respectivos establecimientos con total exactitud, pero si se deben considerar lo probable que esto puede ocurrir y así anticipar con un plan de acción para mitigar estos casos.

# 12. Referencias

[1] Mineduc. (2020). Bases de datos resumen de rendimiento por UE. Recuperado de http://datos.mineduc.cl/dashboards/19741/bases-de-datos-resumen-de-rendimiento-por-ue/

[2] BCN. (2020). Región metropolitana de Santiago. Recuperado de https://www.bcn.cl/siit/nuestropais/region13

[3] BARNETT, V. y Lewis, T. (1994). Outliers in Statical Data. John Wiley & Sons. New York.

[4] Máxima Formación. (2020, 7 septiembre). ¿Cómo Lidiar Con Los Datos Atípicos (Outliers)?. Recuperado 23 de octubre de 2020, de https://www.maximaformacion.es/blog-dat/como-lidiar-con-los-datos-atipicos-outliers/

# 13. Anexo

## 13.1 Método basado en el recorrido intercuartílico:

Tabla N.° 12, Datos para detectar Outliers para total de retirados

| CANTIDAD RETIRADOS TOTAL | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -862 |
| **Primera cota inferior** | CI-1,5DI | -358 |
| **Primera cota superior** | CS+1,5DI | -358 |
| **Segunda cota superior** | CS+3DI | 1490 |
| **MÍNIMO** | 809 | |
| **CUARTIL INFERIOR (CI)** | 146 | |
| **MEDIANA** | 287 | |
| **CUARTIL SUPERIOR (CS)** | 482 | |
| **MÁXIMO** | 2.123 | |
| **RANGO INTERCUARTIL(DI)** | 336 | |
| **OUTLIER** | - | |
| **OUTLIER SEVERO** | 2123 | |

Fuente: Elaboración propia

Tabla N.° 13, Datos para detectar Outliers para superficie Kilómetros al cuadrado

| SUPERFICIE KM2 | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -86,9 |
| **Primera cota inferior** | CI-1,5DI | -38,45 |
| **Primera cota superior** | CS+1,5DI | 91 |
| **Segunda cota superior** | CS+3DI | 139 |
| **MÍNIMO** | 7 | |
| **CUARTIL INFERIOR (CI)** | 10 | |
| **MEDIANA** | 16 | |
| **CUARTIL SUPERIOR (CS)** | 42 | |
| **MÁXIMO** | 1.024 | |

| | | |
|---|---|---|
| **RANGO INTERCUARTIL(DI)** | 32 | |
| **OUTLIER** | 99/135 | |
| **OUTLIER SEVERO** | 1024 | |

Fuente: Elaboración propia

Tabla N.° 14, Datos para detectar Outliers para población 2020

| **Población 2020** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -349491 |
| **Primera cota inferior** | CI-1,5DI | -123766 |
| **Primera cota superior** | CS+1,5DI | 478168 |
| **Segunda cota superior** | CS+3DI | 703894 |
| **MÍNIMO** | 172 | |
| **CUARTIL INFERIOR (CI)** | 101960 | |
| **MEDIANA** | 135.502 | |
| **CUARTIL SUPERIOR (CS)** | 252443 | |
| **MÁXIMO** | 578.605 | |
| **RANGO INTERCUARTIL(DI)** | 150484 | |
| **OUTLIER** | 503147/578605 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 15, Datos para detectar Outliers para tasa de pobreza

| **Tasa de pobreza por ingresos %** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -9,12 |
| **Primera cota inferior** | CI-1,5DI | -3,21 |
| **Primera cota superior** | CS+1,5DI | 13 |
| **Segunda cota superior** | CS+3DI | 18 |
| **MÍNIMO** | 0,13 | |
| **CUARTIL INFERIOR (CI)** | 2,71 | |
| **MEDIANA** | 4,80 | |
| **CUARTIL SUPERIOR (CS)** | 6,65 | |
| **MÁXIMO** | 14,14 | |
| **RANGO INTERCUARTIL(DI)** | 3,94 | |

| | | |
|---|---|---|
| **OUTLIER** | 14,14 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 16, Datos para detectar Outliers respecto hogares en carentes de servicios básico

| **Personas en hogares carentes de servicios básico %** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -14,4 |
| **Primera cota inferior** | CI-1,5DI | -5,5 |
| **Primera cota superior** | CS+1,5DI | 18 |
| **Segunda cota superior** | CS+3DI | 27 |
| **MÍNIMO** | 0,2 | |
| **CUARTIL INFERIOR (CI)** | 3,5 | |
| **MEDIANA** | 6,1 | |
| **CUARTIL SUPERIOR (CS)** | 9,4 | |
| **MÁXIMO** | 20,1 | |
| **RANGO INTERCUARTIL(DI)** | 6,0 | |
| **OUTLIER** | 20,1 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 17, Datos para detectar Outliers respecto hogares hacinados

| **Hogares hacinados %** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -7,4 |
| **Primera cota inferior** | CI-1,5DI | 2,4 |
| **Primera cota superior** | CS+1,5DI | 29 |
| **Segunda cota superior** | CS+3DI | 38 |
| **MÍNIMO** | 3 | |
| **CUARTIL INFERIOR (CI)** | 12,3 | |
| **MEDIANA** | 17 | |
| **CUARTIL SUPERIOR (CS)** | 19 | |
| **MÁXIMO** | 25 | |
| **RANGO INTERCUARTIL(DI)** | 7 | |

| | | |
|---|---|---|
| **OUTLIER** | - | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 18, Datos para detectar Outliers respecto cantidad de establecimientos municipales

| **Cantidad de establecimientos municipales** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -45,0 |
| **Primera cota inferior** | CI-1,5DI | -22,5 |
| **Primera cota superior** | CS+1,5DI | 38 |
| **Segunda cota superior** | CS+3DI | 60 |
| **MÍNIMO** | 0 | |
| **CUARTIL INFERIOR (CI)** | 0,0 | |
| **MEDIANA** | 7 | |
| **CUARTIL SUPERIOR (CS)** | 15 | |
| **MÁXIMO** | 44 | |
| **RANGO INTERCUARTIL(DI)** | 15 | |
| **OUTLIER** | 44,0 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 19, Datos para detectar Outliers respecto cantidad de establecimientos particular subvencionado

| **Cantidad de establecimientos particular subvencionado** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -61 |
| **Primera cota inferior** | CI-1,5DI | -19 |
| **Primera cota superior** | CS+1,5DI | 94 |
| **Segunda cota superior** | CS+3DI | 136 |
| **MÍNIMO** | 1 | |
| **CUARTIL INFERIOR (CI)** | 24 | |
| **MEDIANA** | 36 | |
| **CUARTIL SUPERIOR (CS)** | 52 | |
| **MÁXIMO** | 163 | |

| | | |
|---|---|---|
| **RANGO INTERCUARTIL(DI)** | 28 | |
| **OUTLIER** | - | |
| **OUTLIER SEVERO** | 147/163 | |

Fuente: Elaboración propia

Tabla N.° 20, Datos para detectar Outliers respecto cantidad de establecimientos particular pagado

| **Cantidad de establecimientos particular pagado** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -50 |
| **Primera cota inferior** | CI-1,5DI | -25 |
| **Primera cota superior** | CS+1,5DI | 41 |
| **Segunda cota superior** | CS+3DI | 66 |
| **MÍNIMO** | 0 | |
| **CUARTIL INFERIOR (CI)** | 0 | |
| **MEDIANA** | 3 | |
| **CUARTIL SUPERIOR (CS)** | 17 | |
| **MÁXIMO** | 44 | |
| **RANGO INTERCUARTIL(DI)** | 17 | |
| **OUTLIER** | 44 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 21, Datos para detectar Outliers respecto a la cantidad de matriculados en los establecimientos

| **Cantidad de matrículas totales** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -41.026,5 |
| **Primera cota inferior** | CI-1,5DI | -12.027,0 |
| **Primera cota superior** | CS+1,5DI | 65305 |
| **Segunda cota superior** | CS+3DI | 94305 |
| MÍNIMO | 7.022 | |
| CUARTIL INFERIOR (CI) | 16.973 | |
| MEDIANA | 21.702 | |

| | | |
|---|---|---|
| CUARTIL SUPERIOR (CS) | 36.306 | |
| MÁXIMO | 81.529 | |
| RANGO INTERCUARTIL(DI) | 19.333 | |
| OUTLIER | 81.529 | |
| OUTLIER SEVERO | - | |

Fuente: Elaboración propia

Tabla N.° 22, Datos para detectar Outliers respecto a los ingresos de educación

| Ingresos Educación (M$) | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -29.016.094 |
| **Primera cota inferior** | CI-1,5DI | -9.594.624 |
| **Primera cota superior** | CS+1,5DI | 42195965 |
| **Segunda cota superior** | CS+3DI | 61617435 |
| **MÍNIMO** | 6.930.984 | |
| **CUARTIL INFERIOR (CI)** | 9.826.847 | |
| **MEDIANA** | 13.938.313 | |
| **CUARTIL SUPERIOR (CS)** | 22774494 | |
| **MÁXIMO** | 68.086.543 | |
| **RANGO INTERCUARTIL(DI)** | 12947647 | |
| **OUTLIER** | 44.285.876 | |
| **OUTLIER SEVERO** | 68.086.543 | |

Fuente: Elaboración propia

Tabla N.° 23, Datos para detectar Outliers respecto aporte municipal al sector educación

| Aporte Municipal al Sector Educación (M$) | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -8.280.250 |
| **Primera cota inferior** | CI-1,5DI | -3.710.125 |
| **Primera cota superior** | CS+1,5DI | 8476875 |
| **Segunda cota superior** | CS+3DI | 13047000 |
| **MÍNIMO** | 162.370 | |
| **CUARTIL INFERIOR (CI)** | 860.000 | |

| | | |
|---|---|---|
| **MEDIANA** | 1.733.467 | |
| **CUARTIL SUPERIOR (CS)** | 3906750 | |
| **MÁXIMO** | 10.749.110 | |
| **RANGO INTERCUARTIL(DI)** | 3046750 | |
| **OUTLIER** | 8810430/ 10749110 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

Tabla N.° 24, Datos para detectar Outliers respecto Gastos de educación

| **Gastos Educación (M$)** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -29.060.862 |
| **Primera cota inferior** | CI-1,5DI | -9.342.552 |
| **Primera cota superior** | CS+1,5DI | 43239606 |
| **Segunda cota superior** | CS+3DI | 62957915 |
| **MÍNIMO** | 6.658.703 | |
| **CUARTIL INFERIOR (CI)** | 10.375.757 | |
| **MEDIANA** | 15.056.664 | |
| **CUARTIL SUPERIOR (CS)** | 23521297 | |
| **MÁXIMO** | 68.467.876 | |
| **RANGO INTERCUARTIL(DI)** | 13145540 | |
| **OUTLIER** | 44.979.840 | |
| **OUTLIER SEVERO** | 68.467.876 | |

Fuente: Elaboración propia

Tabla N.° 25, Datos para detectar Outliers respecto a la tasa de denuncias por delito de mayor connotación social

| **Tasa de denuncias por delito de mayor connotación social** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -9.956,6 |
| **Primera cota inferior** | CI-1,5DI | -3.669,9 |
| **Primera cota superior** | CS+1,5DI | 13095 |
| **Segunda cota superior** | CS+3DI | 19381 |

| | | |
|---|---|---|
| **MÍNIMO** | 869 | |
| **CUARTIL INFERIOR (CI)** | 2.616,9 | |
| **MEDIANA** | 3.796 | |
| **CUARTIL SUPERIOR (CS)** | 6808 | |
| **MÁXIMO** | 21.170 | |
| **RANGO INTERCUARTIL(DI)** | 4191 | |
| **OUTLIER** | - | |
| **OUTLIER SEVERO** | 21.169,7 | |

Fuente: Elaboración propia

Tabla N.° 26, Datos para detectar Outliers respecto a la tasa de denuncias por violencia intrafamiliar

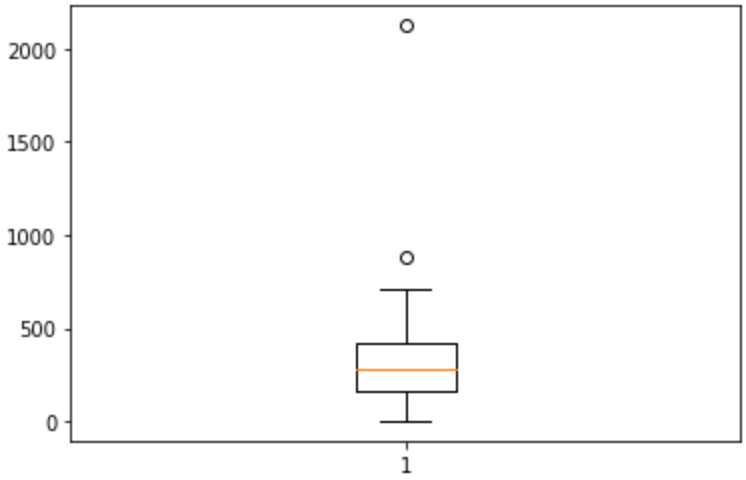
| **Tasa de denuncias por violencia intrafamiliar** | | |
|---|---|---|
| | | |
| **Segunda cota inferior** | CI-3DI | -928,7 |
| **Primera cota inferior** | CI-1,5DI | -265,2 |
| **Primera cota superior** | CS+1,5DI | 1504 |
| **Segunda cota superior** | CS+3DI | 2167 |
| **MÍNIMO** | 193 | |
| **CUARTIL INFERIOR (CI)** | 398,3 | |
| **MEDIANA** | 630 | |
| **CUARTIL SUPERIOR (CS)** | 841 | |
| **MÁXIMO** | 1.520 | |
| **RANGO INTERCUARTIL(DI)** | 442 | |
| **OUTLIER** | 1.520,3 | |
| **OUTLIER SEVERO** | - | |

Fuente: Elaboración propia

# 13.2 Diagrama Box-Plot:

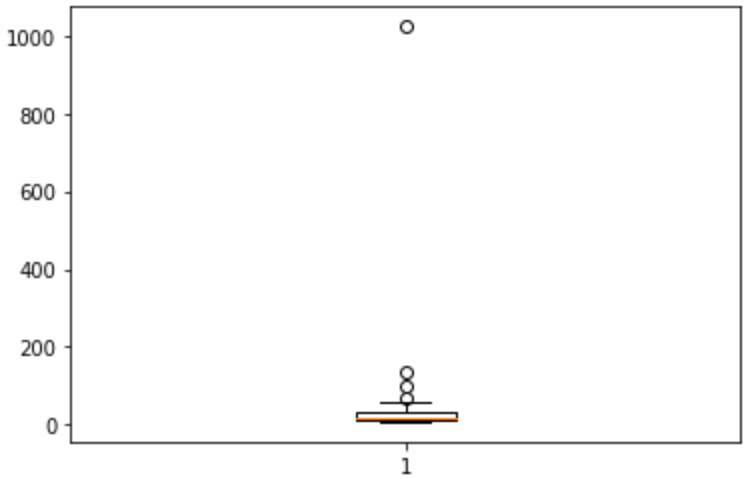
## 13.2.1 Box-Plot del dataset original:

Figura N.° 32, Boxplot respecto cantidad total de retirados



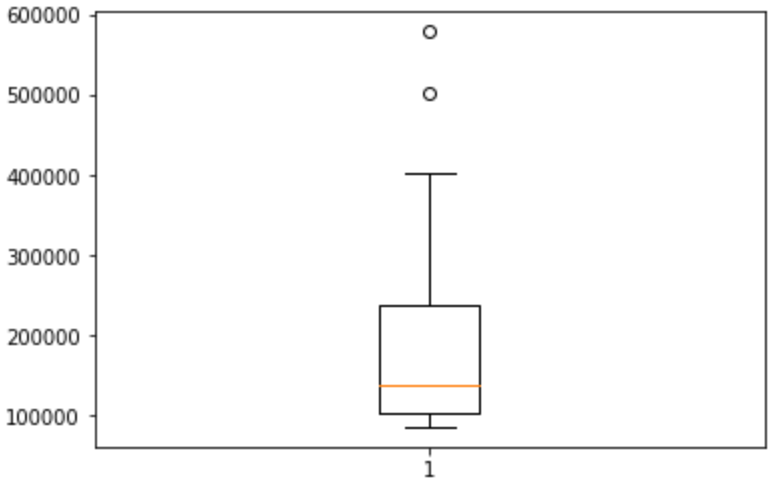

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 33, Boxplot respecto a la superficie km2



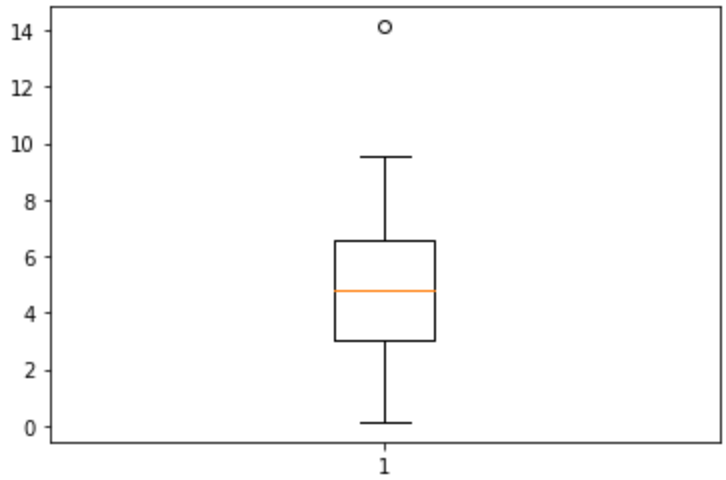

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 34, Boxplot respecto a la población 2020



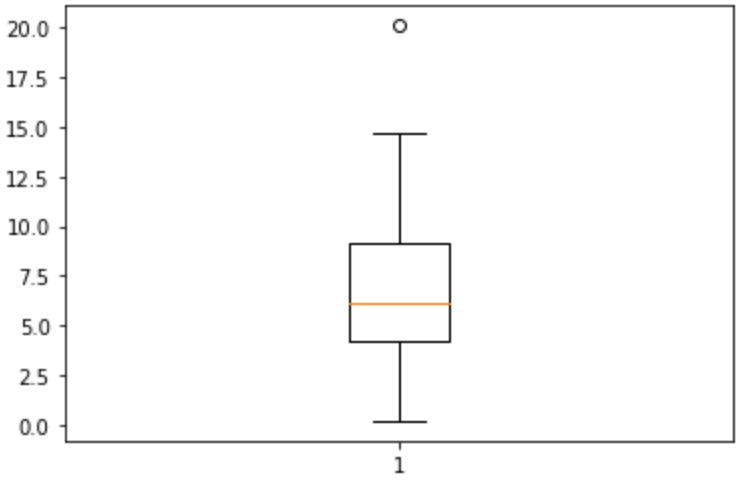

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 35, Boxplot respecto a la tasa de pobreza



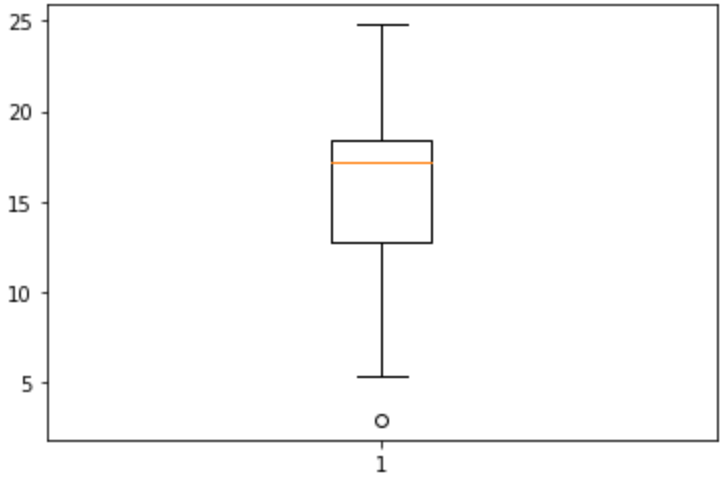

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 36, Boxplot respecto a los hogares carentes de servicios básicos




Fuente: Elaboración propia en base al código del proyecto

Figura N. °37, Boxplot respecto a los hogares hacinados

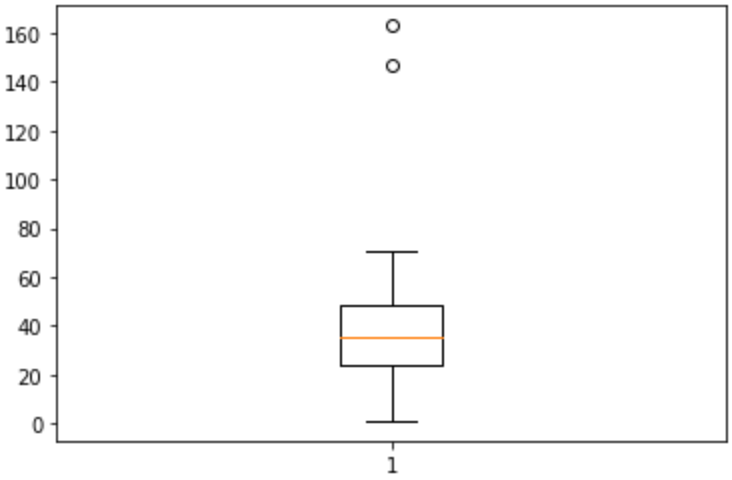



Fuente: Elaboración propia en base al código del proyecto

Figura N.° 38, Boxplot respecto a la cantidad de establecimientos municipales

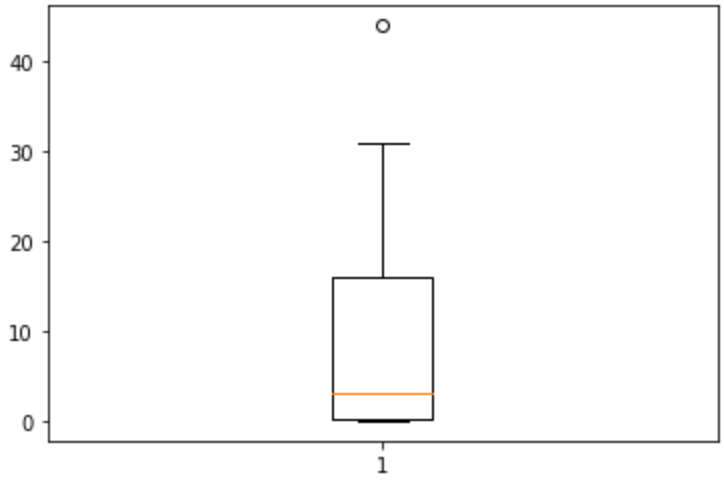

Fuente: Elaboración propia en base al código del proyecto

Figura N.°39, Boxplot respecto a la cantidad de establecimientos particular subvencionado



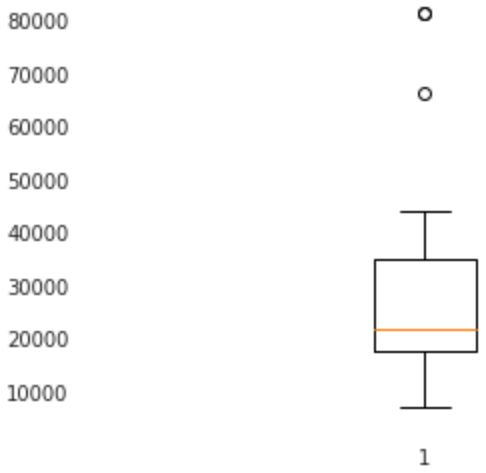

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 40, Boxplot respecto a la cantidad de establecimientos particular pagado



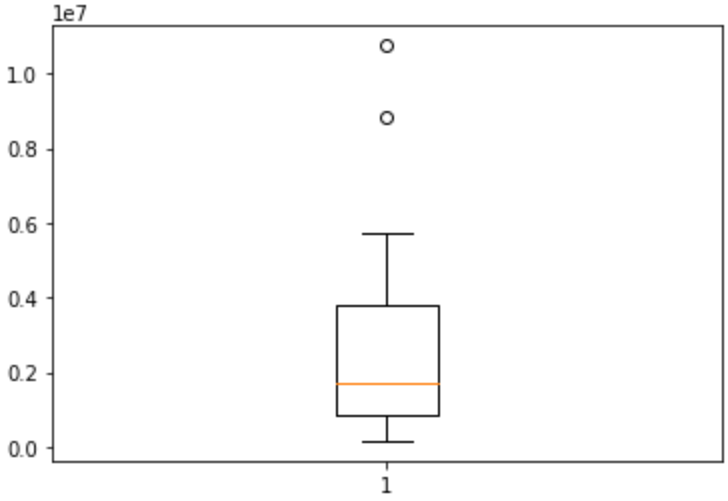

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 41, Boxplot respecto a la cantidad total de matrículas en los establecimientos



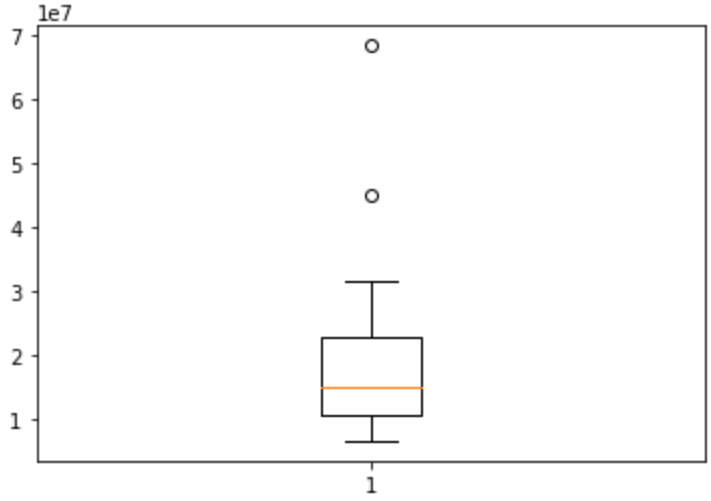

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 42, Boxplot respecto al aporte municipal al sector de educación (M$)



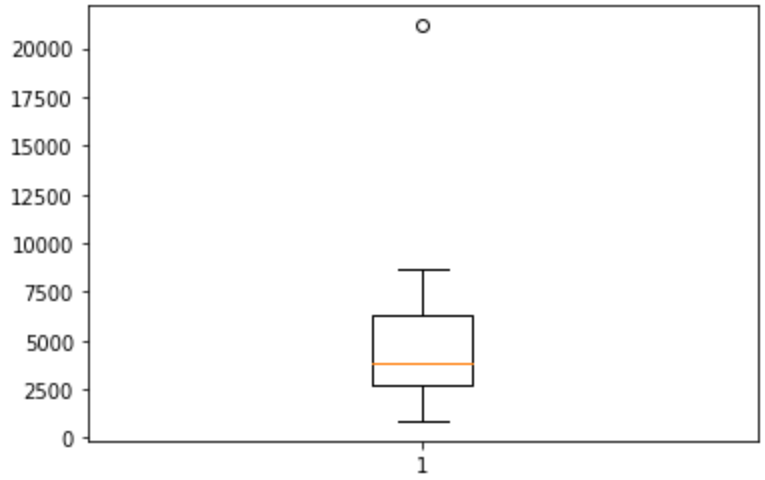

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 43 , Boxplot respecto a los gastos en la educación (M$)



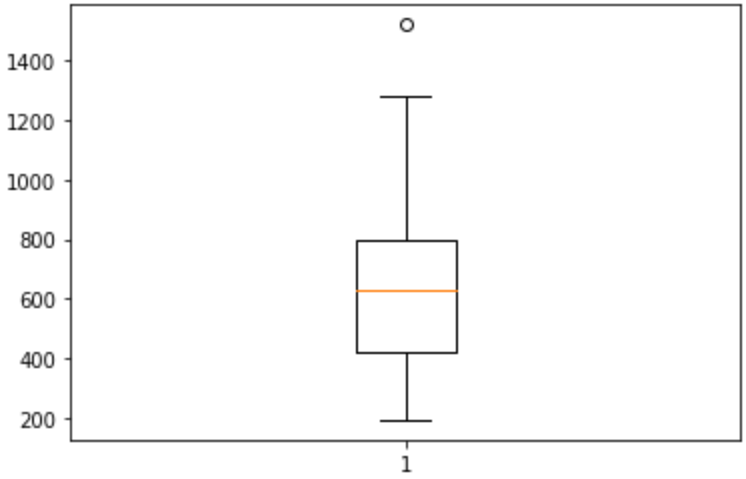

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 44, Boxplot respecto a la tasa de denuncias por delito de mayor connotación social




Fuente: Elaboración propia en base al código del proyecto

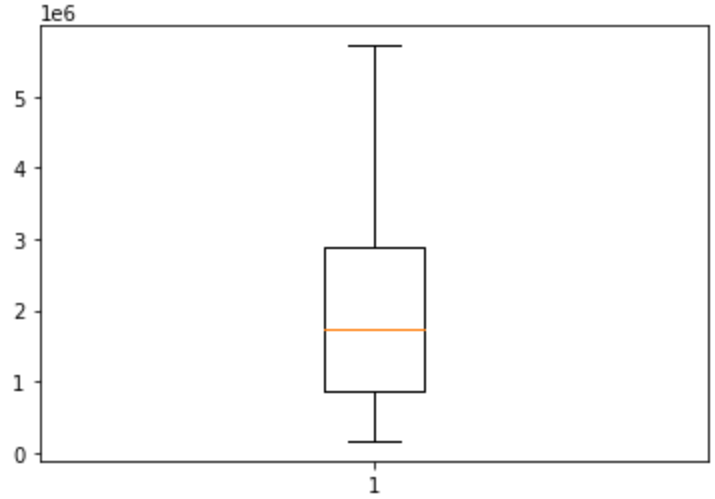
Figura N.° 45, Boxplot respecto a la tasa de denuncias por violencia intrafamiliar




Fuente: Elaboración propia en base al código del proyecto

### 13.2.2 Box-Plot del dataset con los outliers tratados:

Figura N.°46, Boxplot respecto Aporte Municipal al Sector Educación (M$), segunda versión

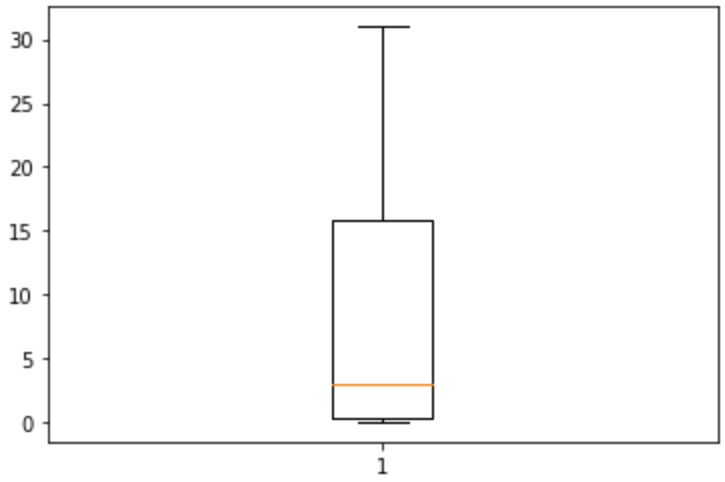



Fuente: Elaboración propia en base al código del proyecto

Figura N.° 47, Boxplot respecto cantidad de establecimientos municipales, segunda versión

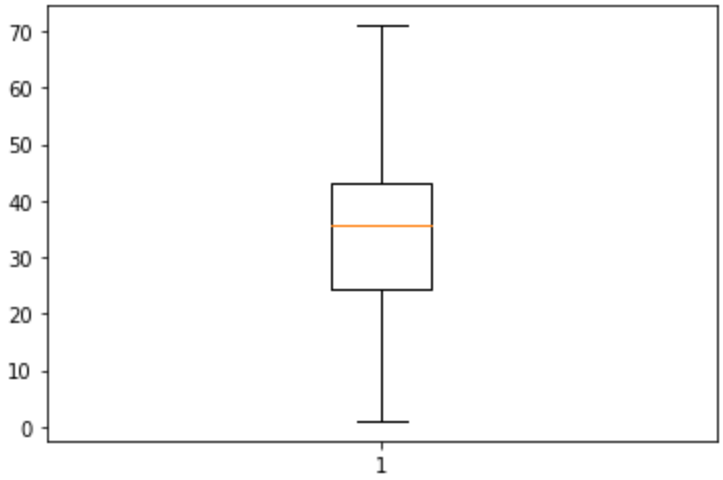

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 48, Boxplot respecto cantidad de establecimientos particular pagado, segunda versión



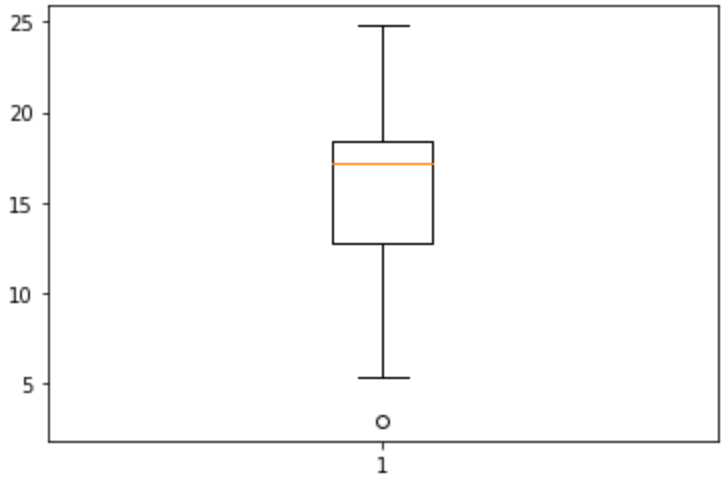

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 49, Boxplot respecto cantidad de establecimientos particular subvencionado, segunda versión



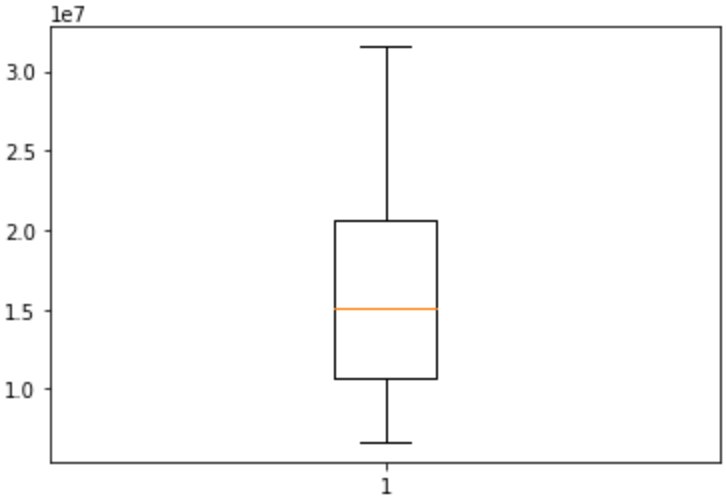

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 50, Boxplot respecto hogares hacinados, segunda versión



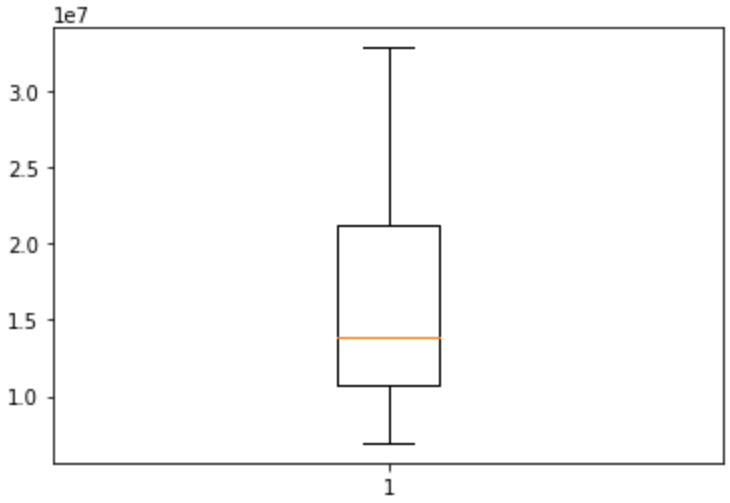

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 51, Boxplot respecto gastos educación (M$), segunda versión



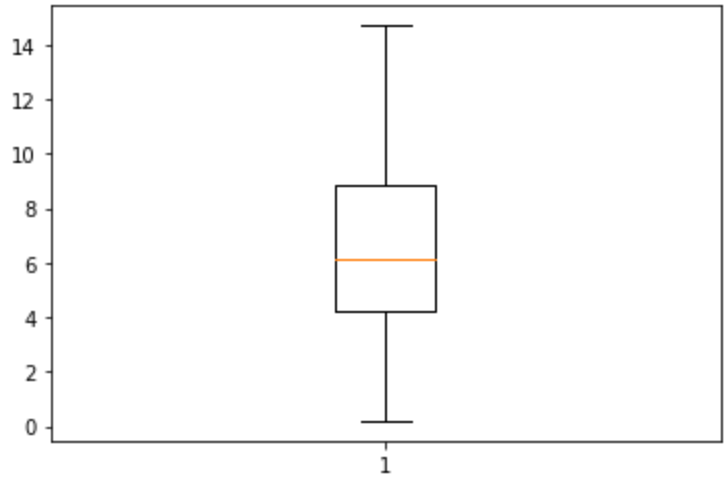

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 52, Boxplot respecto ingresos Educación (M$), segunda versión



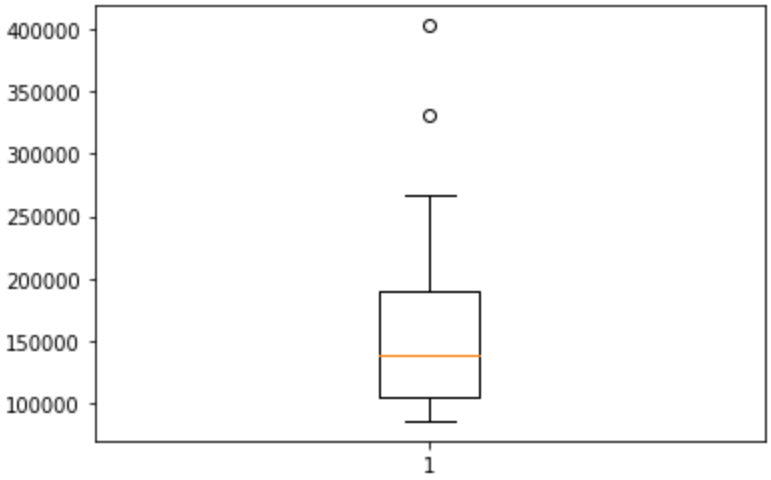

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 53, Boxplot respecto hogares carentes de servicios básico, segunda versión



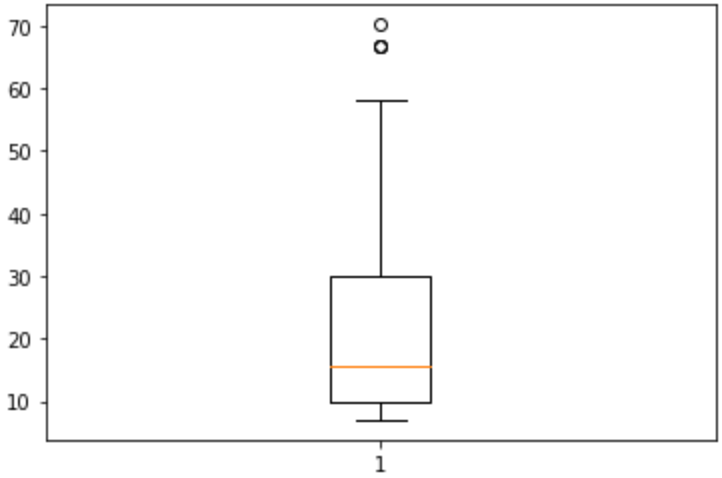

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 54, Boxplot respecto población 2020, segunda versión



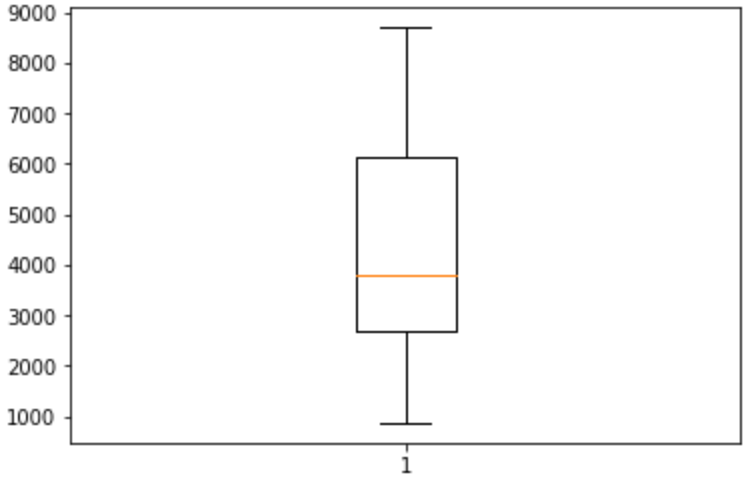

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 55, Boxplot respecto superficie km2, segunda versión



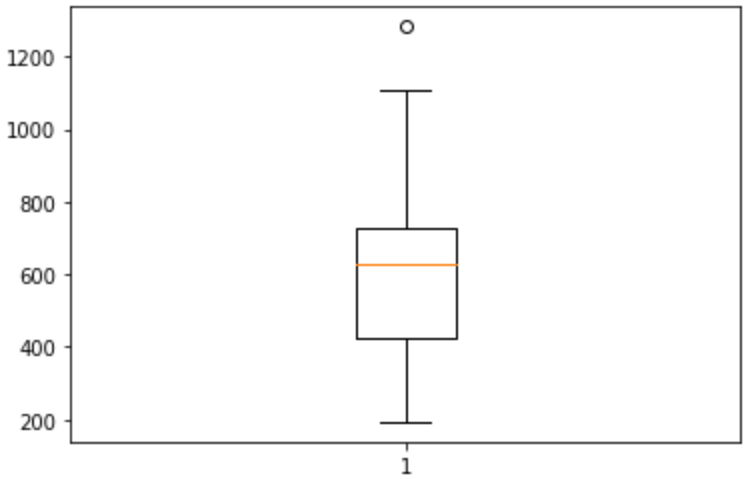

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 56, Boxplot respecto tasa de denuncias por delito de mayor connotación social, segunda versión



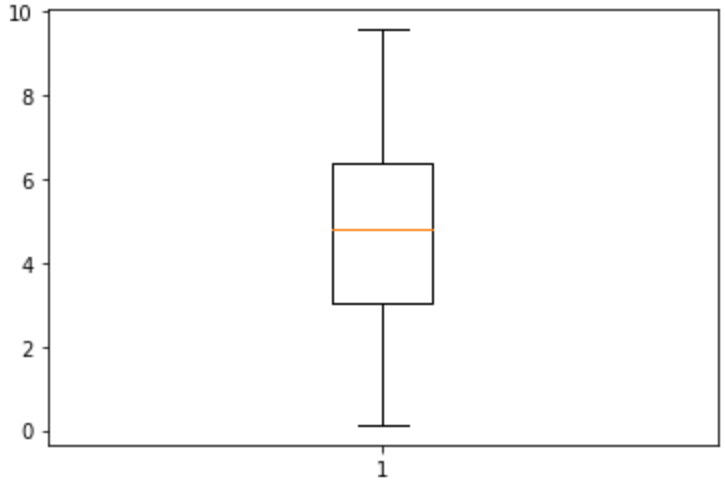

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 57, Boxplot respecto tasa de denuncias por violencia intrafamiliar, segunda versión




Fuente: Elaboración propia en base al código del proyecto

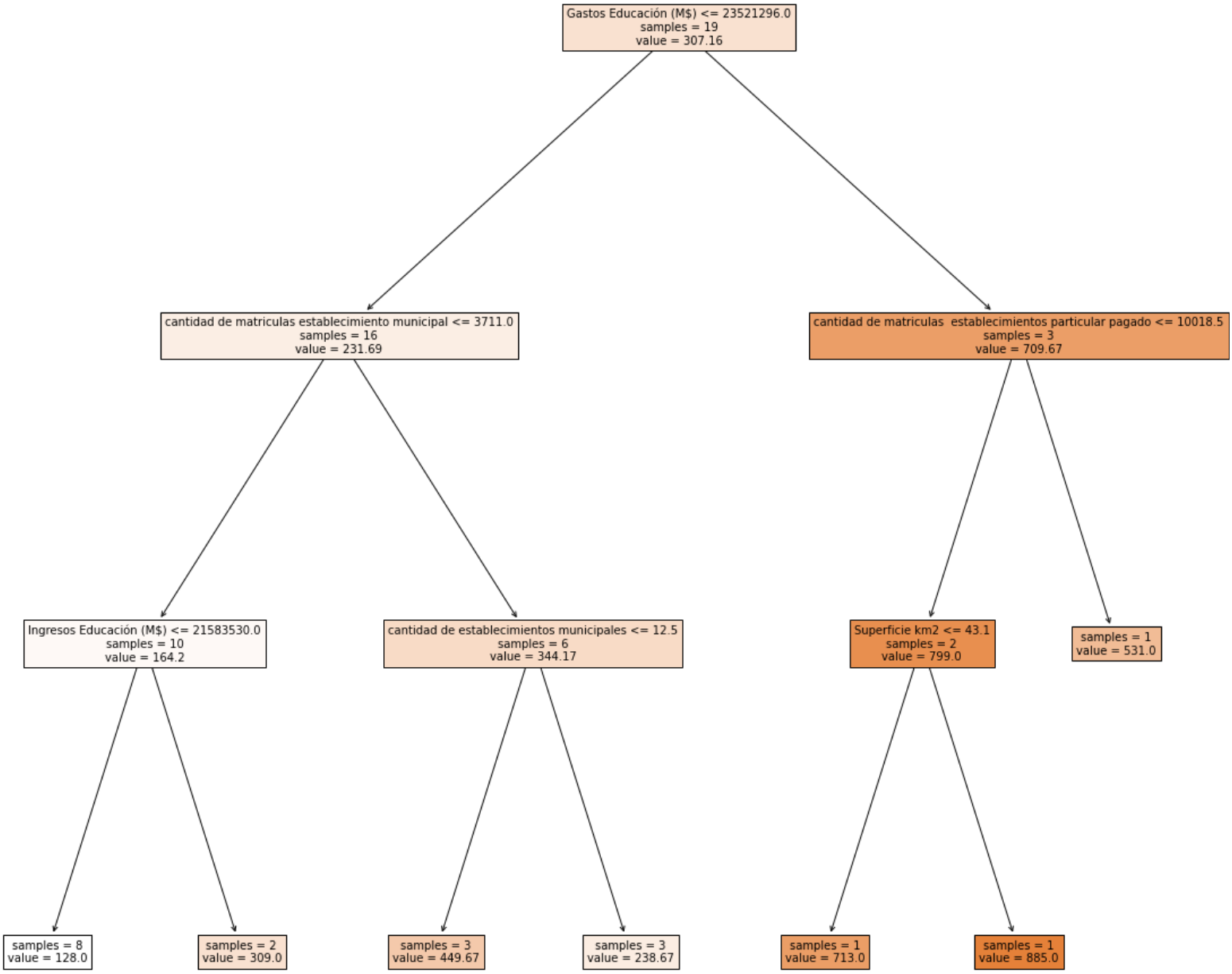
Figura N.° 58, Boxplot respecto tasa de pobreza por ingreso, segunda versión




Fuente: Elaboración propia en base al código del proyecto

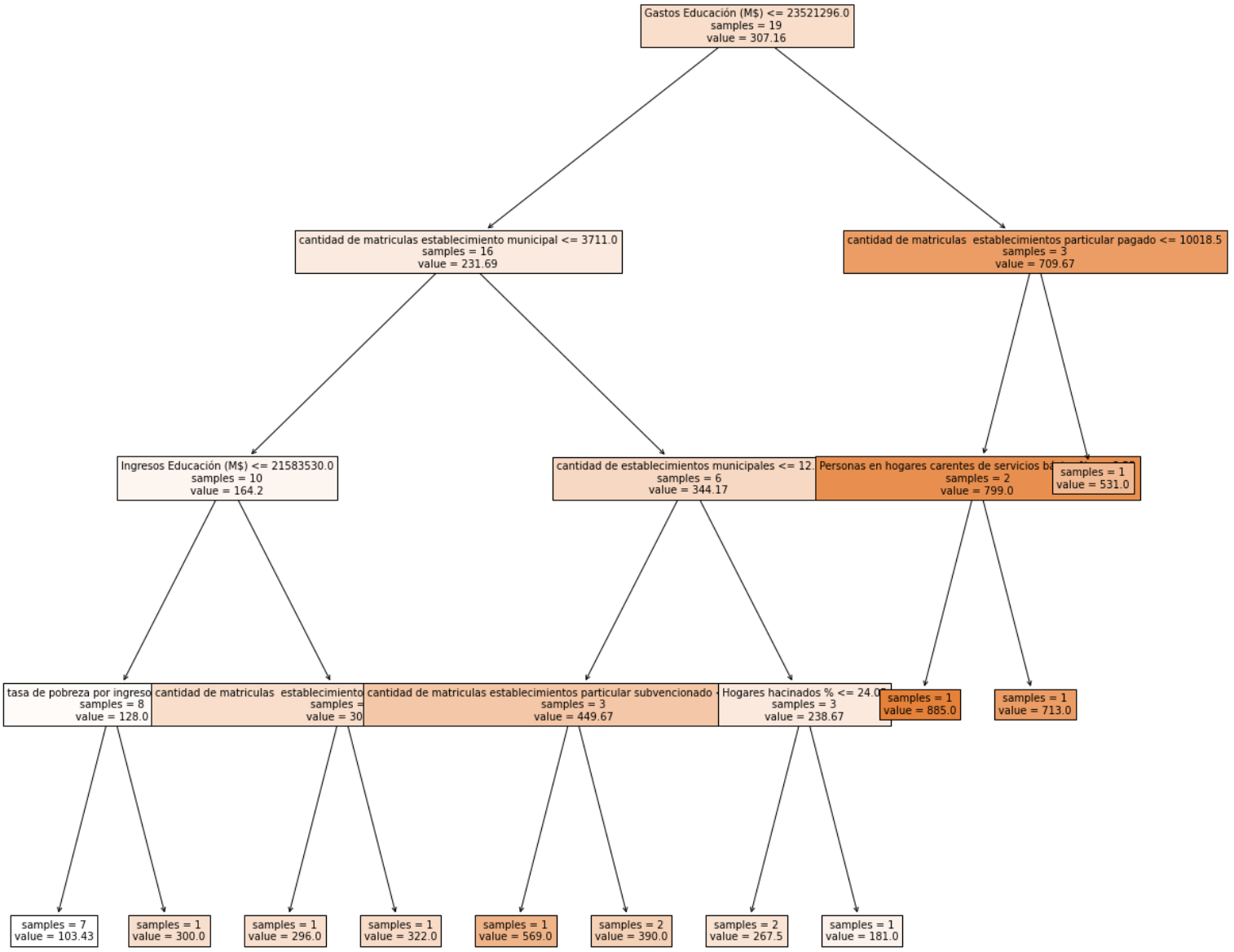
## 13.3 Diagrama de árboles:

Figura N.° 59, Árbol de decisión con 3 niveles de profundidad



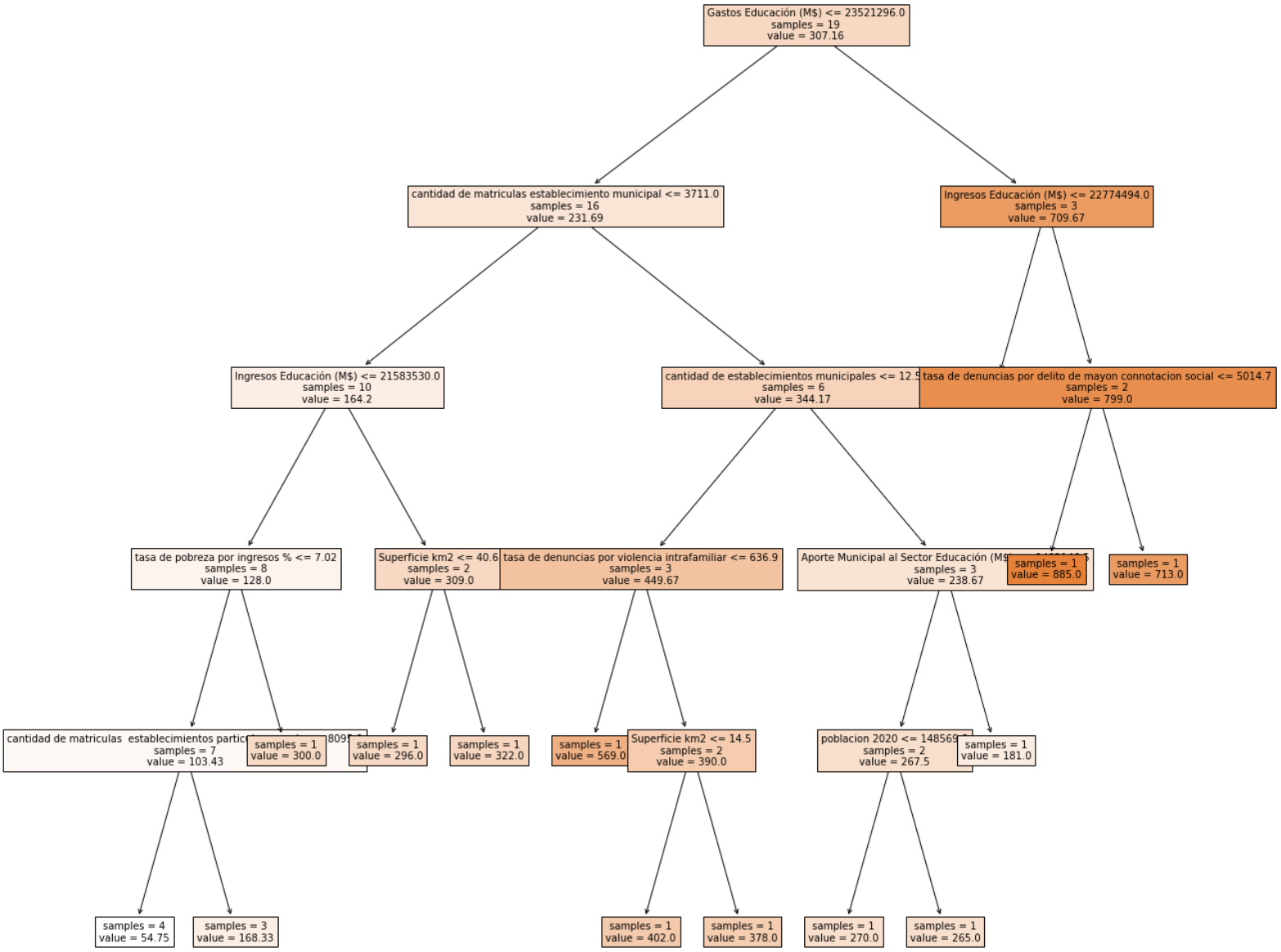

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 60, Árbol de decisión con 4 niveles de profundidad




Fuente: Elaboración propia en base al código del proyecto

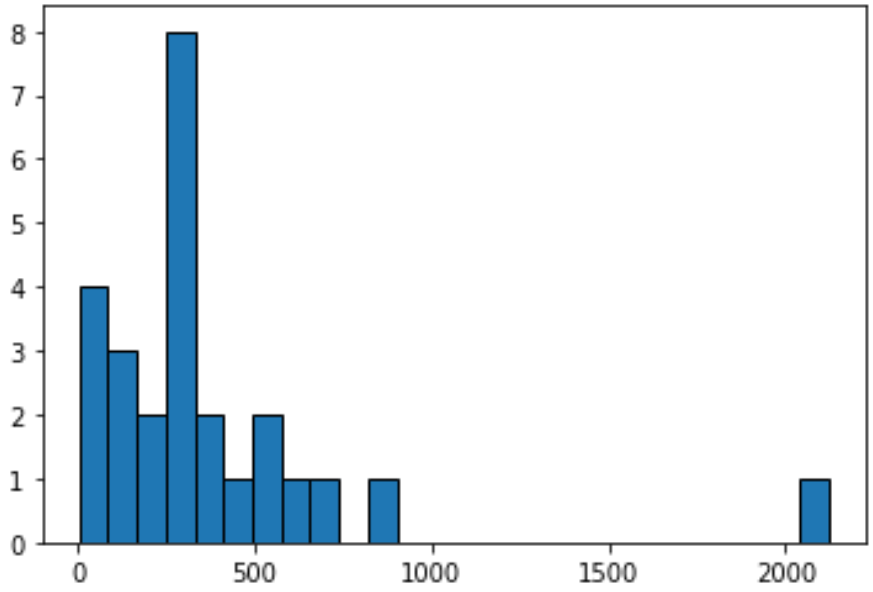
Figura N.° 61, Árbol de decisión con 5 niveles de profundidad




Fuente: Elaboración propia en base al código del proyecto

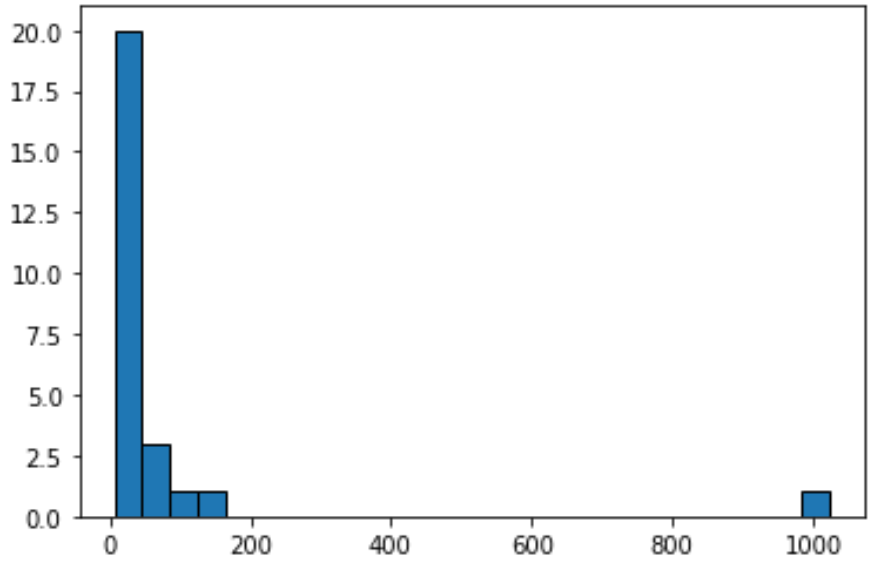
## 13.4 Histograma de los campos analizados

Figura N.° 62, Histograma sobre cantidad de alumnos retirados



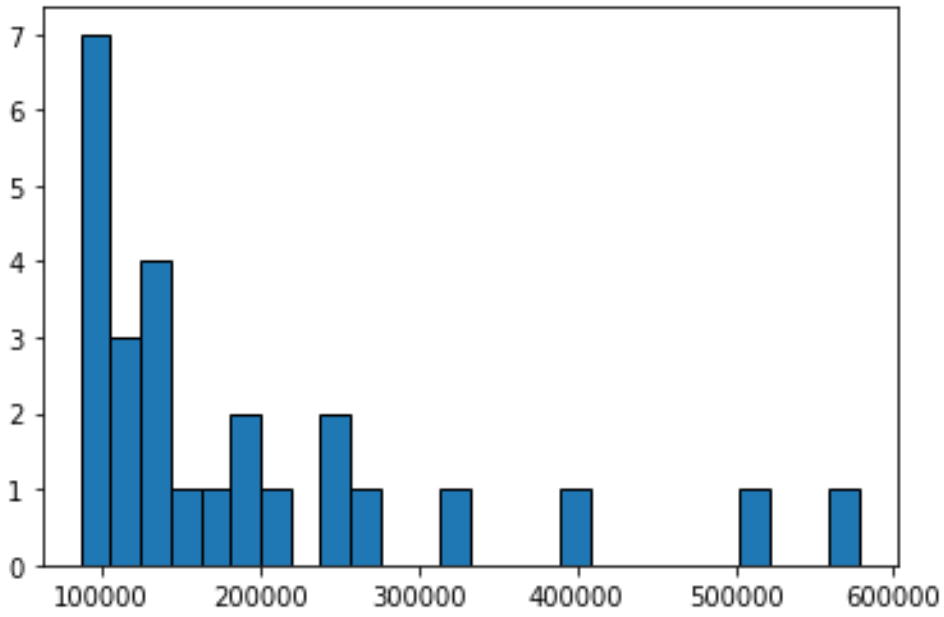

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 63, Histograma sobre superficie en km^2



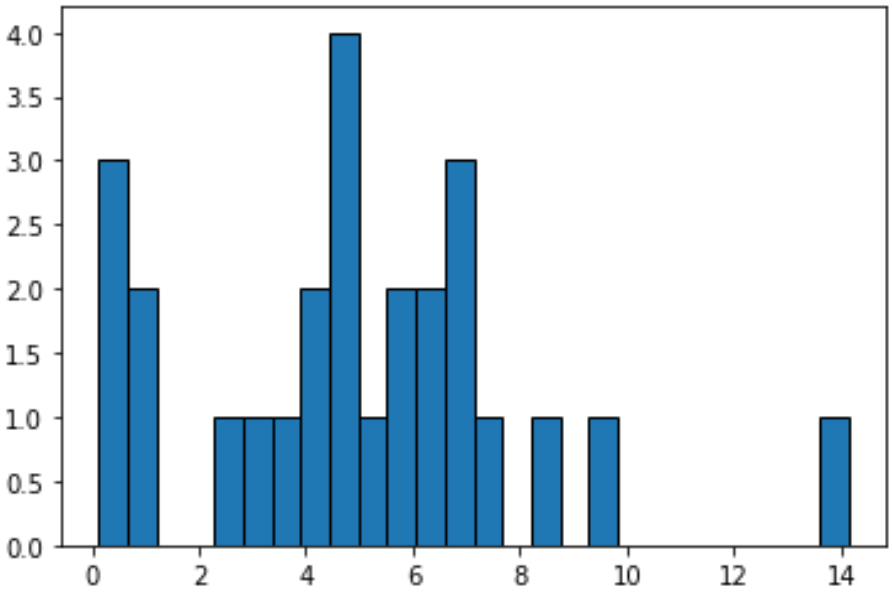

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 64, Histograma sobre la población registrada en 2020



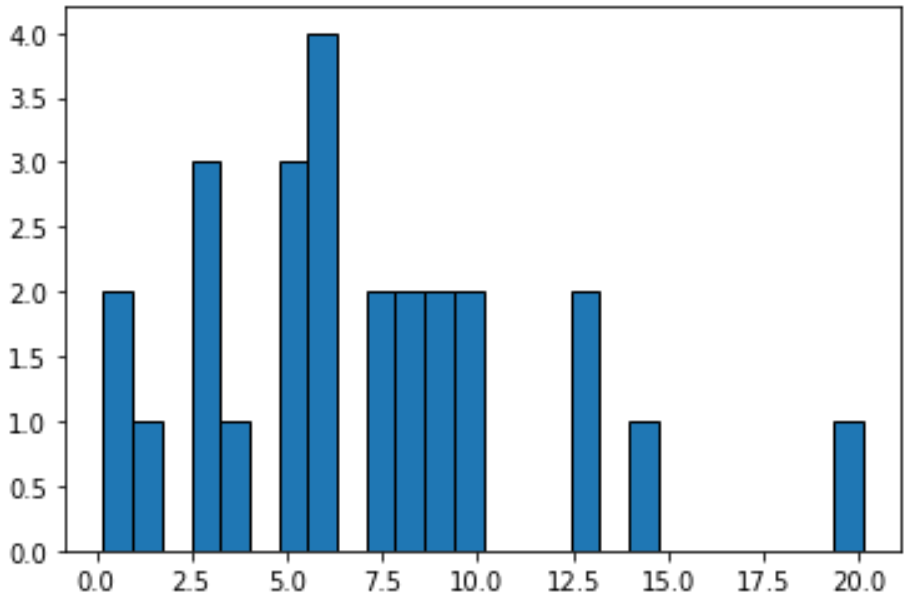

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 65, Histograma sobre la tasa de pobreza



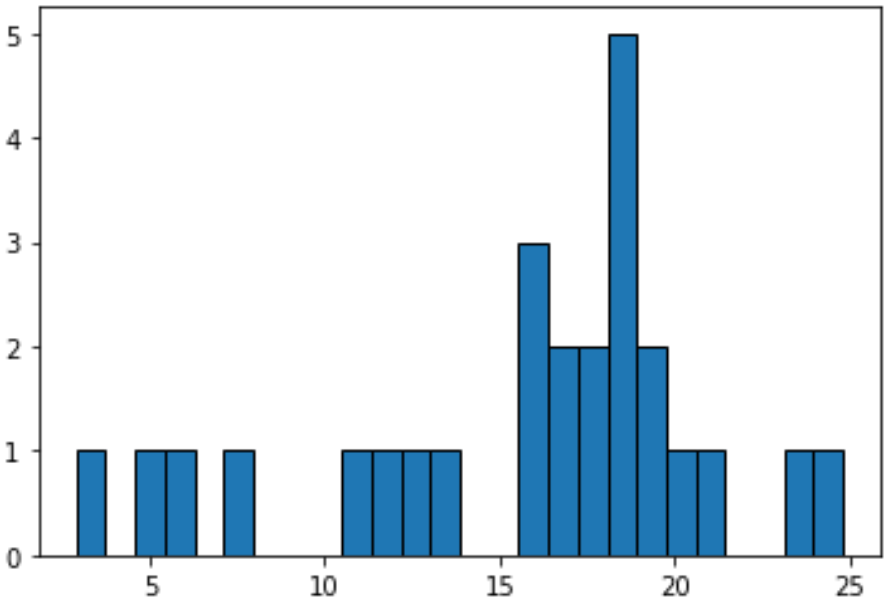

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 66, Histograma sobre hogares con carencia de servicios básicos



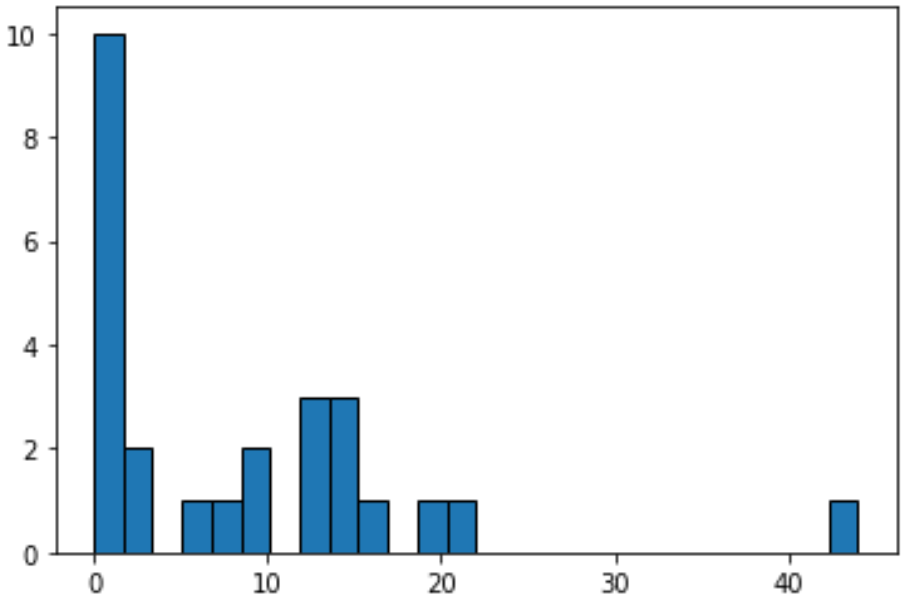

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 67, Histograma sobre hogares hacinados



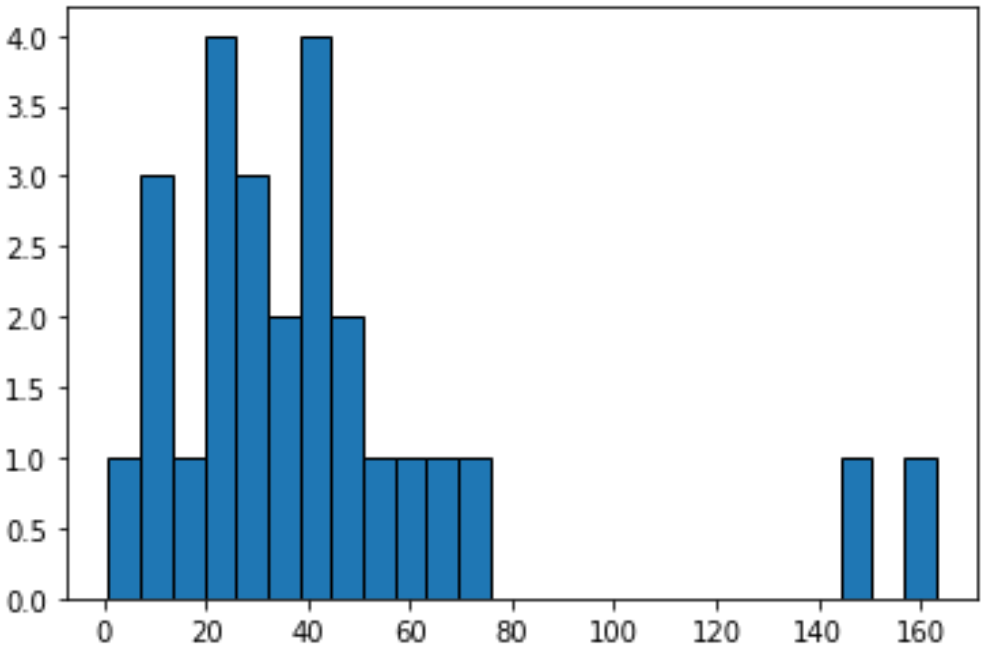

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 68, Histograma sobre cantidad de establecimientos municipales



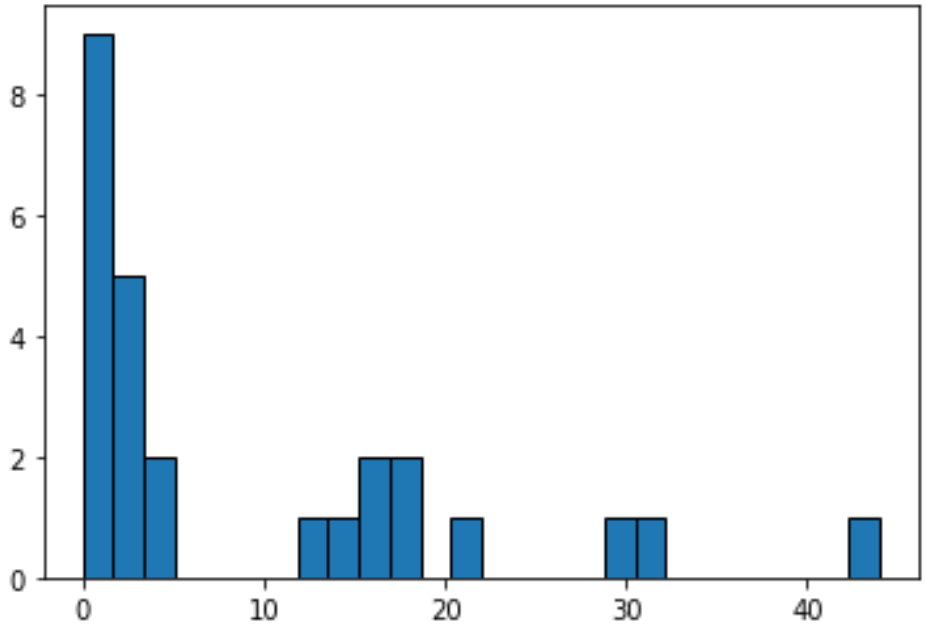

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 69, Histograma sobre cantidad de establecimientos particular subvencionado



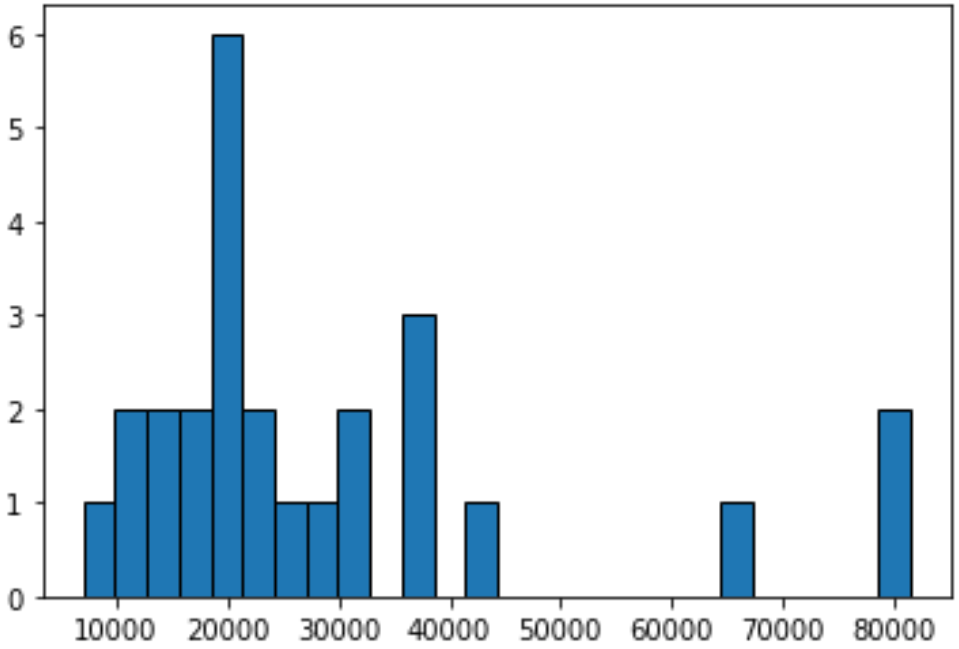

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 70, Histograma sobre cantidad de establecimientos particular pagado



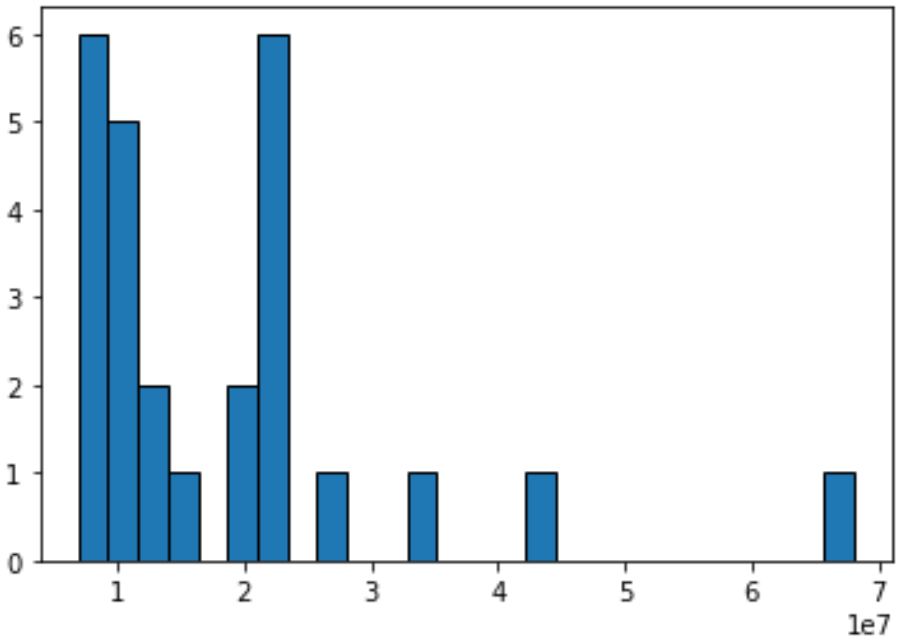

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 71, Histograma sobre cantidad de matrículas totales en los establecimientos



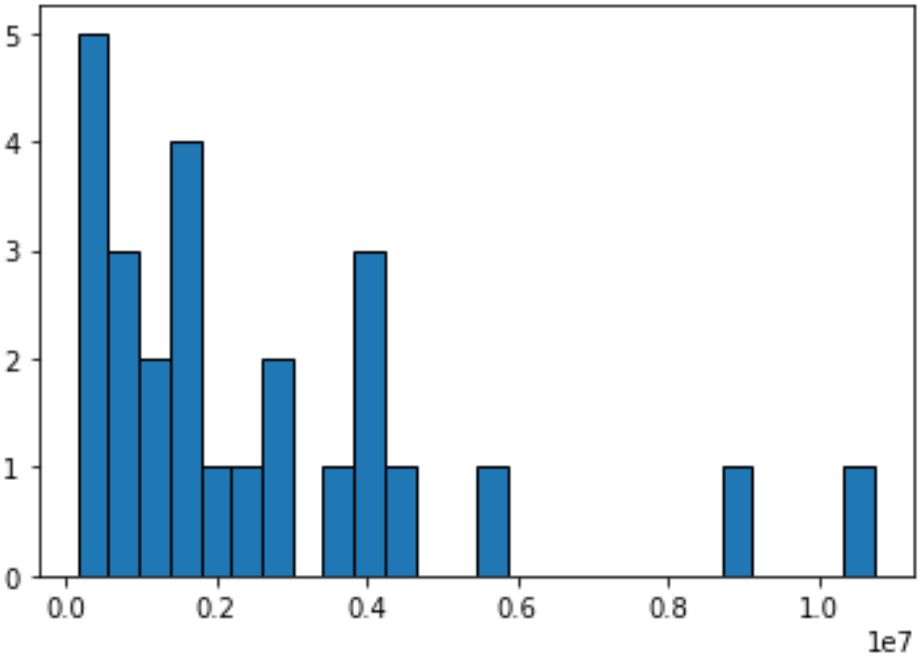

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 72, Histograma sobre ingresos de educación (M$)



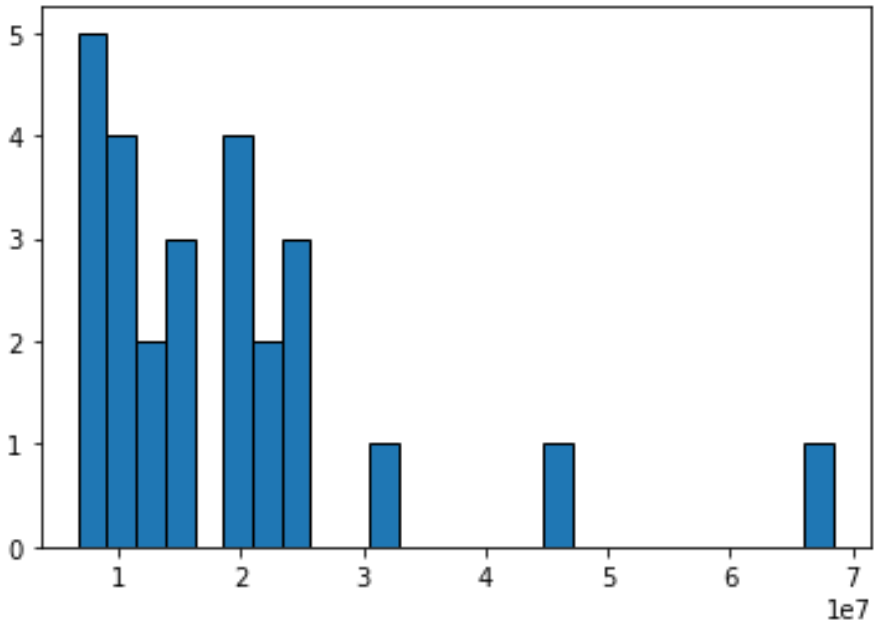

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 73, Histograma sobre aporte municipal al sector de educación (M$)



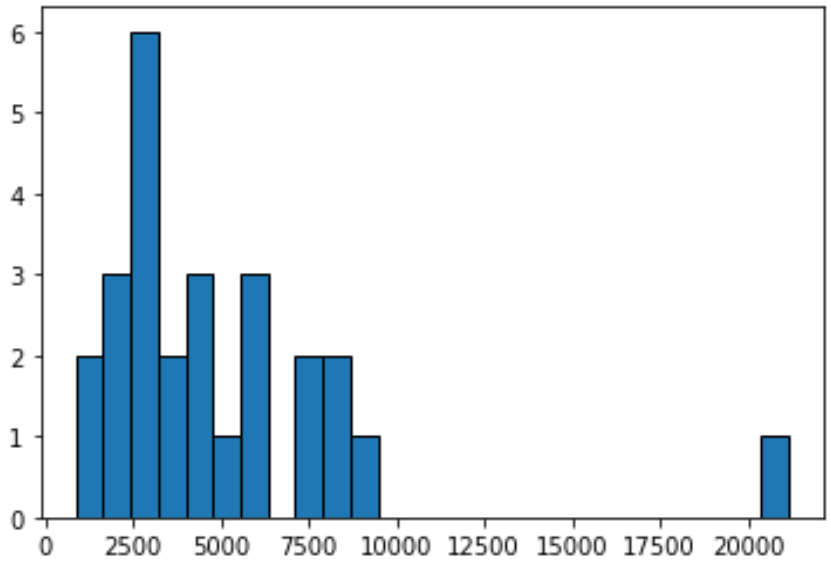

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 74, Histograma sobre los gastos de la educación (M$)



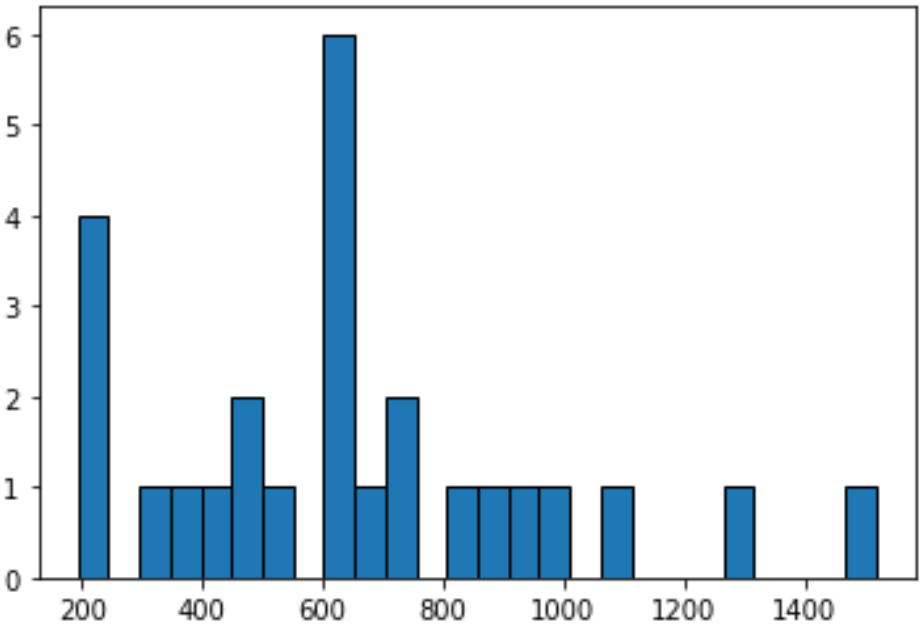

Fuente: Elaboración propia en base al código del proyecto

Figura N.° 75, Histograma sobre la tasa de denuncias por delito de mayor connotación social




Fuente: Elaboración propia en base al código del proyecto

Figura N.° 76, Histograma sobre la tasa de denuncias por violencia intrafamiliar




Fuente: Elaboración propia en base al código del proyecto