ANÁLISIS DE LOS RESULTADOS DE LAS PRUEBAS SABER 11 EDICIÓN 2018 USANDO TÉCNICAS DE MINERÍA DE DATOS

Yeimy Galindo Forero

[a] Tecnología en informática, Corporación Universitaria Minuto de Dios,
ygalindofor@gmail.com

***Resumen*—** Este artículo presenta un informe estadístico que describe la incidencia de un conjunto de variables socio-culturales en los resultados del área de matemáticas de las pruebas saber 11 de un segmento de estudiantes localizado en el municipio de Soacha, Colombia. En este estudio un enfoque analitico ha sido utilizado para hacer el analisis de informacion a traves de tecnicas de mineria de datos que siguen los principios de la metodología CRISP - DM.

***Palabras clave*—** Matemáticas, Soacha, Clasificación, Análisis.

***Abstract*—** *This article presents a statistical report that describes the incidence of a set of sociocultural variables in the results of the mathematics area of the 11 knowledge tests of the student segment located in the municipality of Soacha, Colombia. In this study an analytical approach has been used to make the information analysis to data mining techniques that follow the principles of the CRISP - DM methodology.*

***Keywords*—** *Data Mining, CRISP-DM, Classification, Soacha, math score.*

I. Introducción

La educación es parte fundamental del desarrollo en Colombia, ya que las formas de producción del conocimiento son lo que define a un país, esto hace que Soacha un municipio perteneciente a Colombia pueda tener la habilidad no solo de responder a los nuevos retos, sino también de adelantarse a los mismos.
En cuanto a la caracterización de las pruebas saber 11 en artículos o proyectos, se encuentran escritos con gran valor como el trabajo realizado por Juan D. Baron, Leonardo Bonilla, Lina Cardona-Sosa, Monica Ospina [1]; el cual pretende dar a entender algunas hipótesis de la probabilidad de obtener un puntaje entre los más altos o entre los más bajos y a través de ello ver si el estudiante decide o no graduarse de carreras pedagógicas. El proyecto anterior muestra resultados como "las mujeres que estudian educación son de competencia académica más baja dentro del grupo de mujeres en comparación a los hombres dentro del grupo de los hombres".[1].

Otro importante estudio que habla acerca del desempeño académico en estudiantes de primer semestre de psicología, es el artículo de Andrés Duque Castillo y José Gregorio Ortiz Rodríguez. En este artículo se habla acerca de un estudio sobre el desempeño académico en estudiantes de primer semestre de psicología, el cual tiene su observación a través de las pruebas saber 11, en este caso se realizó el estudio en "539 estudiantes de primer semestre del programa de psicología de la Corporación Universitaria Minuto de Dios en Bogotá Colombia, con promedio de edad de 18,28 años

(desviación típica 2,36), moda 17 años; 82,6% (445) de sexo femenino y 17,4% (94) hombres"[2].

Mientras en Colombia existe un gran número de estudios importantes respecto a los resultados de competencias en la educación media, es importante acotar que la gran mayoría de ellos no consideran variables transversales socio-culturales que pueden ser características familiares sobre el rendimiento (el número de personas que se encuentran en el hogar) y a su vez variables de brecha digital (el efecto entre aquellos que tienen acceso a un computador y/o aquellos que tienen acceso a internet), con respecto a los resultados del área de matemáticas directamente relacionadas con cada estudiante haciendo vigor en el municipio de Soacha el cual tiene un gran capital humano, por el contrario, se centran en aspectos globales.

Por los supuestos anteriores el presente proyecto hace una contribución al tema de análisis de datos en materia de competencias educativas en el municipio de Soacha, a través de un estudio analítico que usa un enfoque descriptivo para analizar los niveles de incidencia que un conjunto de variables socio-culturales y de brecha digital tienen en los 549.934 resultados del área de matemáticas de las pruebas saber 11 edición 2018.

El enfoque analítico-descriptivo abordado es basado en la metodología CRISP-DM.

II. MÉTODOS

El uso de la metodología analítica es lo adecuado para este proyecto, ya que el enfoque analítico descriptivo permite distinguir cada variable establecida que en este caso son el número de personas en el hogar y si el alumno maneja y tiene acceso al desarrollo tecnológico. Analizando estas variables de manera confinada y luego vinculándolas entre sí.

Para realizar el análisis estadístico acompañado de la metodología analítica se encuentran las técnicas de minería de datos las cuales permiten el análisis de grandes volúmenes de información donde se intenta descubrir patrones; eventualmente para hacer uso de técnicas de minería de datos existen diversas metodologías como lo son la KDD, CRISP-DM y SEMMA.

A. CRISP-DM

En este caso se utilizó la metodología CRISP-DM no solo por ser una de las más usadas Fig.1. si no ya que es una metodología de minería de datos que profundiza en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos, mientras que otras proveen sólo una guía general del trabajo a realizar en cada fase.[3].

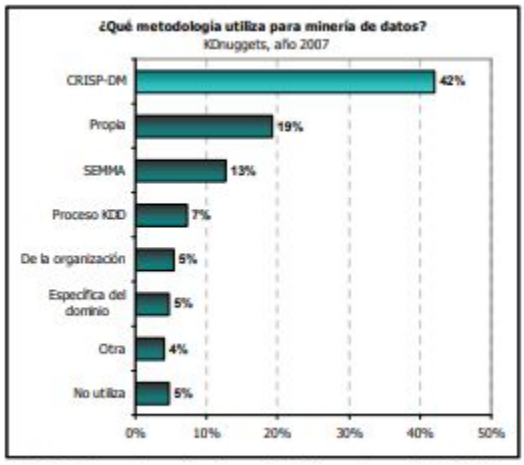

Fig 1. Encuesta realizada por la KDnuggets en el año 2007. Tomado desde https://rb.gy/neljuu.

B. FASES CRISP-DM

A continuación se mostrará las fases de la metodología CRISP - DM y su aplicación a este proyecto

1. Entendimiento del problema

   Problema: En Colombia la gran mayoría de estudios respecto a los resultados de competencias en la educación media no consideran variables transversales socio-culturales que pueden ser características familiares sobre el rendimiento (el número de personas que se encuentran en el hogar) y a su vez variables de brecha digital (el efecto entre aquellos que tienen acceso a un computador y/o aquellos que tienen acceso a internet), con

respecto a los resultados del área de matemáticas directamente relacionadas con cada estudiante haciendo vigor en el municipio de Soacha el cual tiene un gran capital humano, por el contrario, se centran en aspectos globales.

2. Comprensión de los datos

El dataset al cual se le realiza el estudio cuenta con 83 columnas como consecutivo del estudiante, departamento pertenece estudiante, municipio pertenece estudiante, tiene computador, tiene internet, personas hogar, puntaje en matemáticas, puntaje en español, entre otros. Fig 2.

En el caso de este proyecto se usa el dataset tomando como necesarias las columnas municipio pertenece estudiante, puntaje matemáticas, tiene computador y personas en el hogar. Fig 3.

Cabe aclarar que el dataset con el que se realizó este proyecto fue obtenido de la página de datos abiertos de colombia.[3].

	ESTU_TIPODOCUMENTO	ESTU_NACIONALIDAD	ESTU_GENERO	ESTU_FECHANACIMIENTO	PERIODO
0	CR	COLOMBIA	M	10/06/2002	20182
1	TI	COLOMBIA	M	22/10/2000	20182
2	TI	COLOMBIA	M	19/12/2001	20182
3	TI	COLOMBIA	M	20/10/2000	20182
4	CC	COLOMBIA	M	16/11/1998	20182

Fig 2. Lectura del Dataset. Tomado desde https://rb.gy/niosty

	ESTU_MCPIO_RESIDE	FAMI_PERSONASHOGAR	FAMI_TIENEINTERNET	FAMI_TIENECOMPUTADOR	PUNT_MATEMATICAS
0	SOLEDAD	7 a 8	Si	No	69
1	LORICA	5 a 6	Si	No	50
2	CALI	1 a 2	No	No	43
3	TUNJA	3 a 4	Si	Si	60
4	BOGOTÁ D.C.	7 a 8	Si	No	51

Fig 3. Selección de columnas o categorías. Tomado desde https://rb.gy/niosty

Para la comprensión de los datos necesitamos tener un informe conciso del marco de datos. Fig 4. Esto resulta muy útil al hacer un análisis exploratorio de los datos.

```
assert pruebasSaber11_df.notnull().all().all()
pruebasSaber11_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 549934 entries, 0 to 549933
Data columns (total 5 columns):
ESTU_MCPIO_RESIDE       549934 non-null object
FAMI_PERSONASHOGAR      549934 non-null object
FAMI_TIENEINTERNET      549934 non-null object
FAMI_TIENECOMPUTADOR    549934 non-null object
PUNT_MATEMATICAS        549934 non-null int64
dtypes: int64(1), object(4)
memory usage: 21.0+ MB
```

Fig 4. Tipos de registros y valores nulos. Tomado desde https://rb.gy/niosty

Como se puede ver en la Fig 4. en PUNT_MATEMATICAS se tienen datos de tipo numérico (int), el resto de los registros son objetos.

Esta comprensión de los datos se hace con el fin de determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, logrando la completitud y corrección de los datos.[5].

3. Preparación de los datos

En esta fase se procede a adaptar el dataset a las tecnicas de mineria de datos.

Ya que nos estamos centrando en el municipio de Soacha, en la preparación del dataset hay que separar los datos pertenecientes al municipio de soacha de los datos globales.

ESTU_MCPIO_RESIDE	FAMI_PERSONASHOGAR	FAMI_TIENEINTERNET	FAMI_TIENECOMPUTADOR	PUNT_MATEMATICAS
SOACHA	5 a 6	Si	No	67
SOACHA	3 a 4	No	No	46
SOACHA	9 o más	Si	Si	55
SOACHA	3 a 4	Si	Si	55
SOACHA	3 a 4	Si	Si	57

Fig 4. Tipos de registros y valores nulos. Tomado desde https://rb.gy/niosty

Para realizar la integración de estos datos hay que tener en cuenta la media a nivel global de todos los datos y la media a nivel de Soacha, correspondiente al puntaje en matemáticas.

Media a nivel Global

```
pruebasSaber11_df['PUNT_MATEMATICAS'].mean()
```

```
50.191430244356596
```

Media a nivel de Soacha

```
pruebasSaber11_df_s['PUNT_MATEMATICAS'].mean()
```

```
51.16982131039047
```

Se puede ver una diferencia de un 1.0 entre la media global y la media del municipio de Soacha

4. Modelado

El algoritmo que se usó para el modelado de los datos fue el de k-means ya que permite minimizar la suma de distancias cuadráticas de cada observación al centroide de su cluster.[6].

teniendo en cuenta que las variables FAMI_PERSONASHOGAR, FAMI_TIENEINTERNET, FAMI_TIENECOMPUTADOR, son variables de tipo objeto se reemplazan por valores numéricos para ser así más sencillo el uso de las mismas. Fig 5.
A su vez se organiza una matriz donde se podrá comparar los puntajes de matemáticas con el efecto de tener acceso a un computador o no.
Las n observaciones se pueden representar como un vector de d dimensiones, tantas como variables que representa cada observación. Fig 6.

```
[[67  2]
 [46  2]
 [55  1]
 ...
 [52  2]
 [43  2]
 [51  2]]
```

Fig 5. Reemplazar objetos por datos numéricos. Tomado desde https://rb.gy/niosty

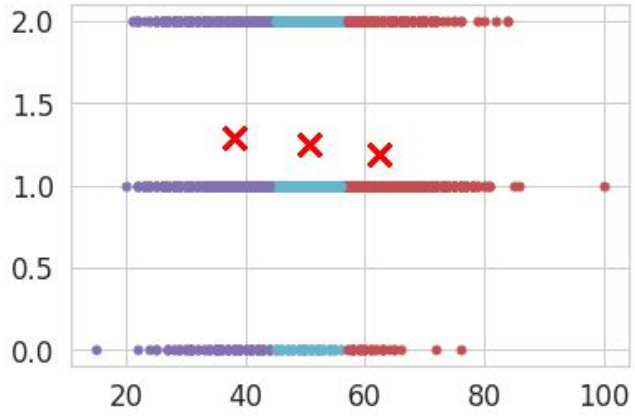

Fig 6. Visualización de los centroides. Tomado desde https://rb.gy/niosty

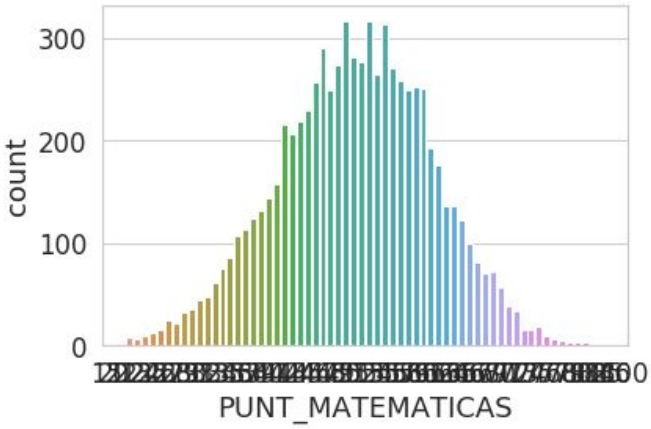

Fig 7. Recuento de los puntajes en matemáticas. Tomado desde https://rb.gy/niosty

Estos datos presentan bastante información pero

en realidad no nos muestra que es lo que realmente sucede. Por tanto debemos modelar los datos de alguna manera.

Una buena visualización nos puede revelar cosas que es probable que no podamos ver en una tabla de números y nos ayudará a pensar con claridad acerca de los patrones y relaciones que pueden estar escondidos en los datos.[7].

Una buena forma de analizar dos variables categóricas en forma conjunta, es agrupar los recuentos en una tabla de doble entrada; este tipo de tablas se conocen en estadística con el nombre de tabla de contingencia.

A continuación se ve los porcentuales y las comparaciones de los datos

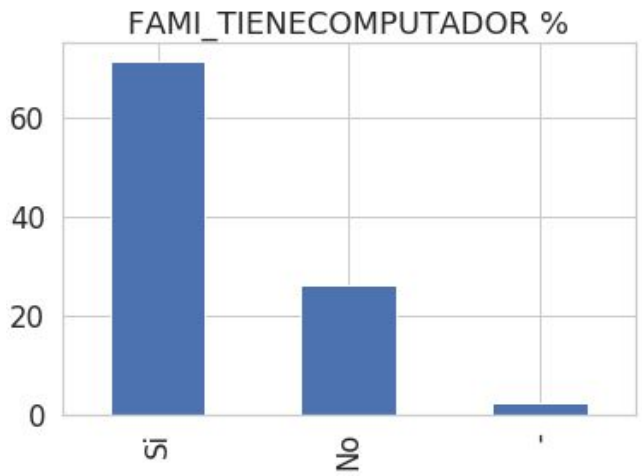

Fig 8. Totales FAMI_TIENECOMPUTADOR. Tomado desde https://rb.gy/niosty

PUNT_MATEMATICAS	15	20	21	22	23	24	25
FAMI_TIENECOMPUTADOR							
-	0.013236	0.000000	0.000000	0.013236	0.000000	0.013236	0.026473
No	0.000000	0.000000	0.026473	0.052945	0.013236	0.026473	0.066181
Si	0.000000	0.013236	0.000000	0.039709	0.066181	0.079418	0.066181
All	0.013236	0.013236	0.026473	0.105890	0.079418	0.119126	0.158835

Fig 9. Totales FAMI_TIENECOMPUTADOR comparación numérica porcentajes. Tomado desde https://rb.gy/niosty

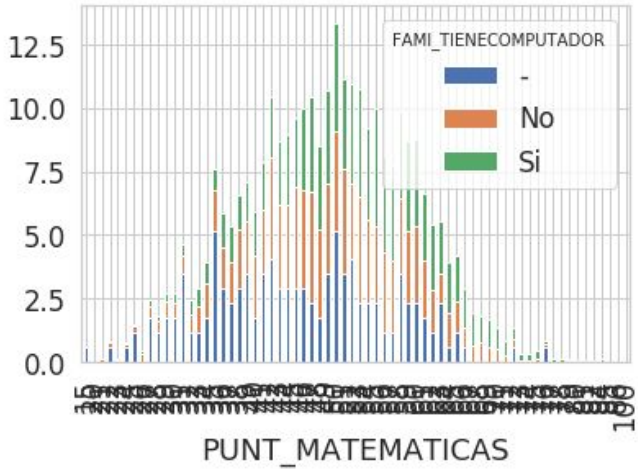

Fig 10. Comparación entre sí la familia tiene computador y el puntaje en matemáticas. Tomado desde https://rb.gy/niosty

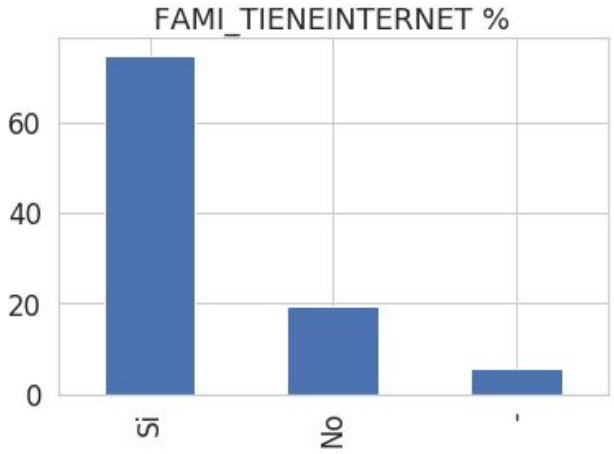

Fig 11. Totales FAMI_TIENEINTERNET. Tomado desde https://rb.gy/niosty

PUNT_MATEMATICAS	15	20	21	22	23	24	25	26
FAMI_TIENEINTERNET								
-	0.013236	0.013236	0.013236	0.013236	0.026473	0.039709	0.026473	0.026473
No	0.000000	0.000000	0.013236	0.039709	0.000000	0.039709	0.079418	0.079418
Si	0.000000	0.000000	0.000000	0.052945	0.052945	0.039709	0.052945	0.092654
All	0.013236	0.013236	0.026473	0.105890	0.079418	0.119126	0.158835	0.198544

Fig 12. Totales FAMI_TIENEINTERNET comparación numérica porcentajes. Tomado desde https://rb.gy/niosty

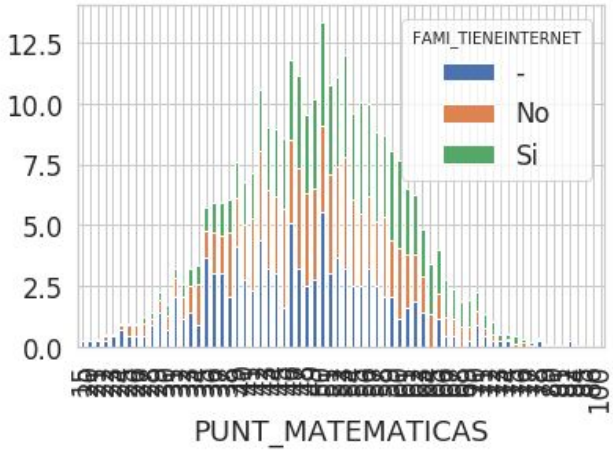

Fig 13. Comparación entre sí la familia tiene acceso a internet y el puntaje en matemáticas. Tomado desde https://rb.gy/niosty

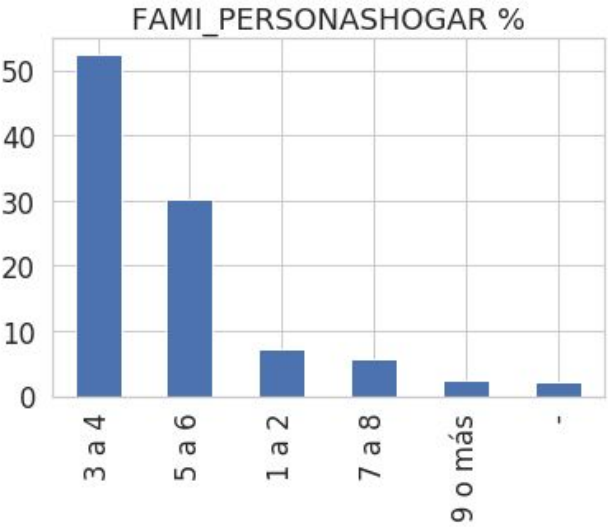

Fig 14. Totales FAMI_PERSONASHOGAR. Tomado desde https://rb.gy/niosty

PUNT_MATEMATICAS	15	20	21	22	23	24	
FAMI_PERSONASHOGAR							
-	0.013236	0.000000	0.000000	0.013236	0.000000	0.013236	0
1 a 2	0.000000	0.013236	0.000000	0.026473	0.026473	0.000000	0
3 a 4	0.000000	0.000000	0.000000	0.013236	0.026473	0.039709	0
5 a 6	0.000000	0.000000	0.013236	0.039709	0.013236	0.066181	0
7 a 8	0.000000	0.000000	0.013236	0.000000	0.000000	0.000000	0
9 o más	0.000000	0.000000	0.000000	0.013236	0.013236	0.000000	0
All	0.013236	0.013236	0.026473	0.105890	0.079418	0.119126	0

Fig 15. Totales FAMI_PERSONASHOGAR comparación numérica porcentajes. Tomado desde https://rb.gy/niosty

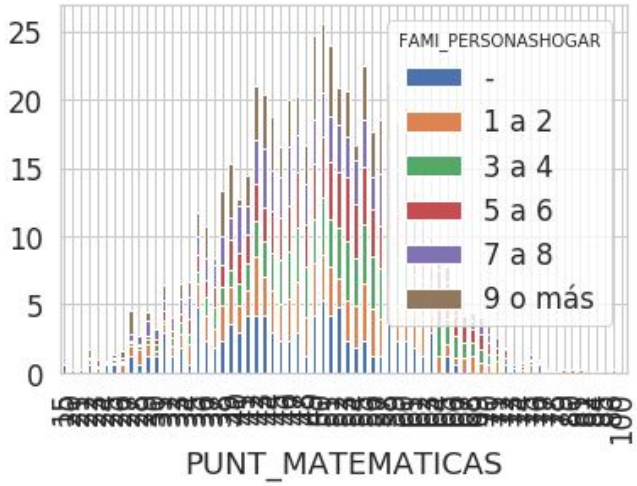

Fig 16. Comparación entre el número de personas en el hogar y el puntaje en matemáticas. Tomado desde https://rb.gy/niosty

5. Evaluación

Evaluación de la situación: Los factores relevantes en el estudio para saber qué es lo que hace que un alumno tenga un puntaje alto en el ámbito de las matemáticas son si el estudiante cuenta con un computador en su hogar o no, si tiene acceso a internet y cuantas personas son en el hogar.

Lo anterior tiene importancia al pensar en que, si un estudiante tiene un computador, cómo puede influir esto en sus resultados, será que al tener computador lo usa como una herramienta lúdica o solo como una tecnología absorbente; las respuestas a esto y más interrogantes se revelan en la parte de resultados del presente artículo.

Con lo anterior visto en la Fig 9. y la Fig. 10. se podría suponer que el 71.39% de los alumnos tienen acceso a un computador y que esté 71% se compone de la siguiente forma: del total de alumnos un 25,5% aprox tuvo un puntaje menor, mientras que un 46,3% aprox tuvo un puntaje mayor.

De igual importancia se pudo notar que el 74.69% de los alumnos tienen acceso a internet

y que esté 74% se compone de la siguiente forma: del total de alumnos un 30,5% aprox tuvo un puntaje menor, mientras que un 43,5% aprox tuvo un puntaje mayor. Fig 13.

Así mismo que el 52.29% de los alumnos viven en su hogar con 3 a 4 personas y que esté 52% se compone de la siguiente forma: del total de alumnos un 22% aprox tuvo un puntaje menor, mientras que un 30% aprox tuvo un puntaje mayor. Fig 16.

III. Conclusiones

En el municipio de Soacha la gran mayoría de los estudiantes tienen acceso a tecnologías de la información y comunicación, indagando en los hallazgos que se presentaron a través de las técnicas de minería de datos se pueden dar ciertas hipótesis como que el acceso a la tecnología no es algo en materia absorbente por el contrario viendo las gráficas obtenidas un computador en el hogar puede brindar grandes oportunidades, y se puede considerar una buena forma de producción del conocimiento.

Referencias

1. "Quiénes eligen la disciplina de la educación en Colombia." https://www.calidadeducativasm.com/wp-content/uploads/2015/10/BANREPUBLICA-quienes-eligen-disciplina-educacion-colombia.pdf. Fecha de acceso 4 nov.. 2019.
2. "Pruebas ICFES Saber 11 y su relación con el desempeño académico en estudiantes de primer semestre de psicología. Dialnet." https://dialnet.unirioja.es/servlet/articulo?codigo=5493081. Fecha de acceso 4 nov.. 2019.
3. "Estudio comparativo de metodologías para ... - CIC Digital." https://digital.cic.gba.gob.ar/bitstream/handle/11746/3525/11746_3525.pdf-PDFA.pdf?sequence=1&isAllowed=y. Fecha de acceso 19 nov.. 2019.
4. "Datos Abiertos Colombia." https://www.datos.gov.co/. Fecha de acceso 4 nov.. 2019.
5. "Metodología para el Desarrollo de Proyectos en Minería de ...." http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf. Fecha de acceso 22 nov.. 2019.
6. "Segmentación utilizando K-means en Python." http://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/. Fecha de acceso 23 nov.. 2019.
7. "Análisis de datos categóricos con Python - Raul E. Lopez Briega." 29 feb.. 2016, https://relopezbriega.github.io/blog/2016/02/29/analisis-de-datos-categoricos-con-python/. Fecha de acceso 23 nov.. 2019.