Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

El futuro es de todos

DNP Departamento Nacional de Planeación

CARACTERIZACIÓN ESCOLAR EN LA CONTINUACIÓN DE ESTUDIOS EN LA TRANSICIÓN DEL COLEGIO A LA UNIVERSIDAD EN COLOMBIA

Entidad

Departamento Nacional de Planeación

- Dirección de Desarrollo Digital.
- Dirección de Desarrollo Social – Subdirección de Educación

Sector	**Lenguaje**	**Fuente de datos**
Educación	Python	MEN y DANE

Presentación

La gran variedad de planteles educativos existentes en Colombia hace que las condiciones económicas y sociales de cada uno pueda ser muy distinto del uno al otro. Las diferencias sustanciales en los planteles educativos pueden determinar o influenciar la penetración de este en la inclusión universitaria de sus estudiantes. Para el Estado es importante conocer e identificar características en la transición escolar del colegio a la universidad y su grado de penetración en la educación superior que permita el desarrollo de política pública en el marco de aumentar la cobertura en la transición de estudiantes salientes de la educación media a la educación superior. De acuerdo con lo anterior, este proyecto busca, a partir de herramientas de aprendizaje automático e inteligencia artificial, dar una caracterización escolar de los diferentes planteles de educación básica en cuanto a su probabilidad de egresar estudiantes que entren directamente a la educación superior. De igual manera, se busca caracterizar y encontrar la relación por parte de las universidades en recibir estudiantes de uno u otro tipo de colegio.

Abstract: *The variety of educational institutions of basic schools in Colombia makes to economic conditions and social conditions of each could be different one to each other. Sustancial differences between educational institutions could determinate of inflience in higher education inclusión of the students. For State is important knows and identify features in scholar transition from basic educations to higher education and their inclusion grade of higher education that allows develop public policy in the framework to increase coverge in the transición of outgoing students from basic education to higher education. This Project seeks, from tools of Machine Learning and Artificial Intelligence, give a scholar caracterization to the differents educative institutions of basic education as for the probability of graduating students to start study in a higher education institution. Similary, this Project seeks caracterize and find the relationship from higher education institutions in recibe students from one or another basic education institution.*

Objetivo general

Caracterizar y generar un sistema de Inteligencia Artificial que permita caracterizar los colegios con alta penetración de estudiantes en la educación superior en el periodo de transición y la capacidad de recepción por parte de las universidades de aceptar estudiantes de diferentes perfiles.

Objetivos específicos

1. Consolidación de la información de diferentes fuentes relacionadas a los colegios y a las universidades.
2. Caracterización de los individuos que egresan del colegio (ingresen o no a la universidad)
3. Caracterización de la recepción de nuevos estudiantes por parte de las universidades

Metodología

La gran variedad de fuentes de información relacionadas a colegios en Colombia recolectadas por el Ministerio de Educación (MEN) y el Departamento Administrativo Nacional de Estadística (DANE) han planteado un reto en cuanto a la consolidación estructurada de la información sobre educación básica, media y superior. Por tal motivo, como primera fase del proyecto se consolidaron diferentes fuentes de información a nivel colegio proveniente del MEN y el DANE estructurada de tal modo que describa la transición escolar de la educación media a la educación superior. Con base en ello, con los datos suministrados por el MEN de un cruce previo de información entre datos de estudiantes de colegio y universidades en Colombia para los años 2014, 2015 y 2016, se cruzó con los micro datos suministrados por el DANE de Educación Formal que contiene diferentes características de los colegios contando con más de 90 tablas.

Procesamiento de información

Utilizando técnicas de procesamiento de bases de datos, se procesaron las tablas de manera fragmentada bajo el siguiente protocolo:

1. Separadas las bases de micro datos de Educación Formal del DANE, se identificó la llave natural de cada base, siendo esta el código de colegio.
2. Con las bases de datos preprocesadas suministradas por el MEN con el cruce de información entre el SIMAT y el SNIES, identificando cuantos estudiantes graduados de un colegio, entraron a determinada universidad o si no entraron para la transición de los periodos 2014-2015, 2015-2016, 2016-2017, se identifica la llave natural.
3. Para cada base de transición escolar, se hace JOIN por separado con cada base de educación formal
4. Una vez realizado todos los JOIN para una sola base de transición escolar, se unen las bases resultantes por la llave natural generando una tabla por cada periodo de transición.
5. Se repite el proceso para cada base de transición.

En la Figura 1, se muestra el diagrama de flujo del procesamiento de información donde *k* corresponde al tipo de base de datos del DANE de Educación Formal (Uso de equipos de cómputo, Docentes, Alumnos matriculados, etc) y *t* corresponde al periodo inicial de transición (si el periodo de transición son los egresados de colegio de 2014 que ingresan a la universidad en 2015, el periodo *t* corresponde al año 2014). ⋈ corresponde al JOIN entre las bases de datos entre el DANE y el MEN respectivo.

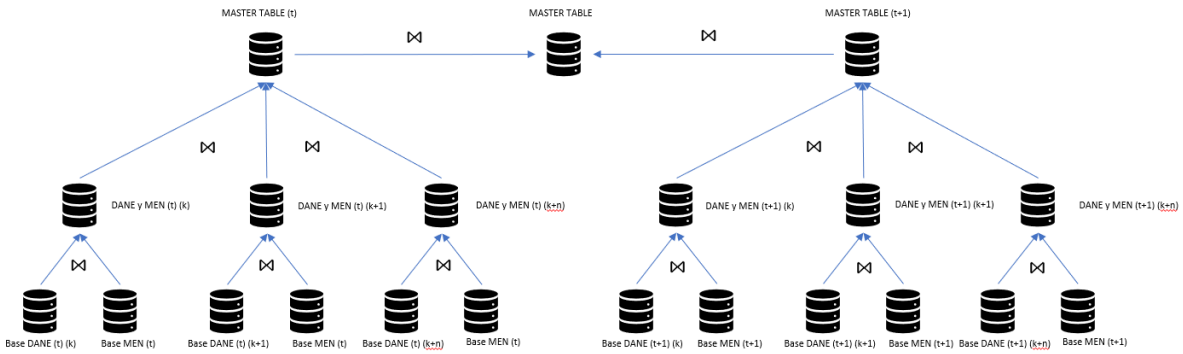

Figura 1. Diagrama de fujo del procesamiento de bases de datos de educación

Desarrollo del modelo

Con el fin de caracterizar a los colegios de acuerdo con la inserción universitaria de los estudiantes egresados del colegio a la universidad, para ello se calculó un indicador dado por:

$$IU_i = \frac{NU_i}{TB_i}$$

Donde *i* corresponde al colegio, *IU* es la inserción universitaria, *NU* es el Número de bachilleres del colegio *i, TB* es el total de bachilleres del colegio *i.* Dicha variable es la proporción de estudiantes que entraron a la universidad en determinado periodo de tiempo.

Construida la variable, se implementa un modelo no supervisado de *k-medias* adaptado a variables categorías conocido como *k-modes* el cual calcula las disimilitudes entre cada categoría definida como, la medida de disimilitud entre X y Y como el total de desacierto[1] de cada categoría de un atributo con respecto a otro, entre menos desaciertos entre las categorías, mayor la similitud entre las mismas (Huang 1998). La medida de disimilitud se define como:

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

Donde:

$$\delta(x_j, y_j) = \begin{cases} 0 \ (x_j = y_j) \\ 1 \ (x_j \neq y_j) \end{cases}$$

Una vez calculados los *clusters* considerando las categorías de los datos (ciudad, colegio, región, etc). Se calculan las estadísticas descriptivas de cada cluster estimado en el modelo anterior, esto con el fin de mostrar las diferencias estadísticas de un grupo a otro y de igual manera, encontrar aquellos colegios cuya inclusión universitaria sea parecida y las características que estos colegios tienen.

Resultados

De acuerdo con cada *cluster*, se hizo una caracterización de cada tipo de colegio de acuerdo a sus características. Se separaron en tres *clusters* con el fin de dividir la población en colegios con menor, media y mayor índice de inclusión universitaria, una vez dividida la población con el algoritmo se procedió a buscar las características de cada población a partir de estadísticas descriptivas de las variables previamente consolidadas.

La figura 1 muestra la distribución la media de la inclusión universitaria para cada *cluster* en cada año, teniendo que el primer *cluster* corresponde a aquellos colegios con inclusión universitaria más baja en 2014, la media en

[1] Traducción del término en inglés *Mismatching* el cual se refiere al grado de similitud de una categoría frente a otra en cuanto a su composición semántica.

2015 y muy parecida al *cluster* 3 en 2015. Caso contrario para el *cluster* 3 y para el *cluster* 2, conserva los colegios cuya inclusión universitaria es mayor.

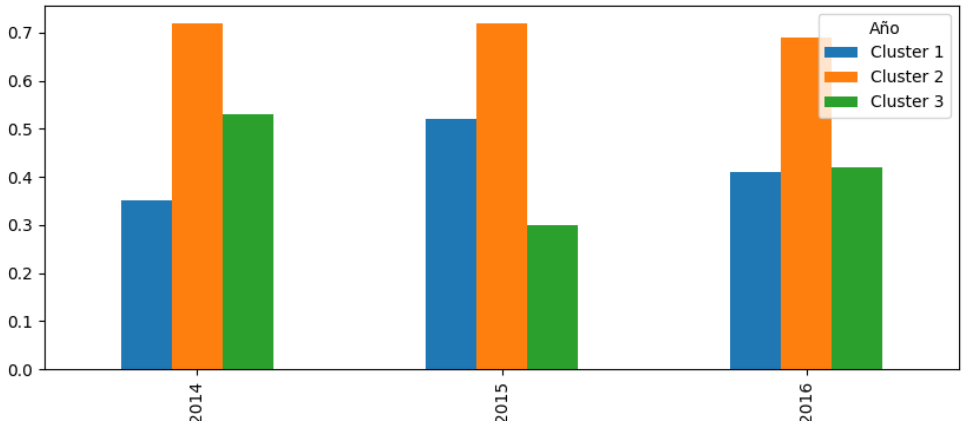

Figura 1: Media de inclusión universitaria por *cluster*

Por otro lado, las siguientes figuras muestran la distribución de diferentes características de acuerdo con su *cluster* de colegios. La relevancia del siguiente análisis se centra en la diferenciación de los cambios en las características de un *cluster* a otro, teniendo en cuenta que cada *cluster* representa un nivel distinto de colegios en cuanto a su penetración a la educación superior.

2014

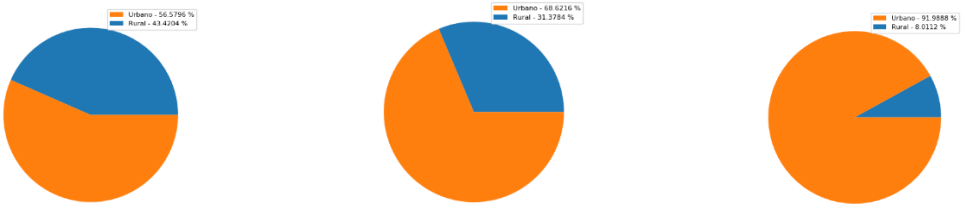

Figura 2: Proporción de colegios por ubicación de colegio urbana o rural

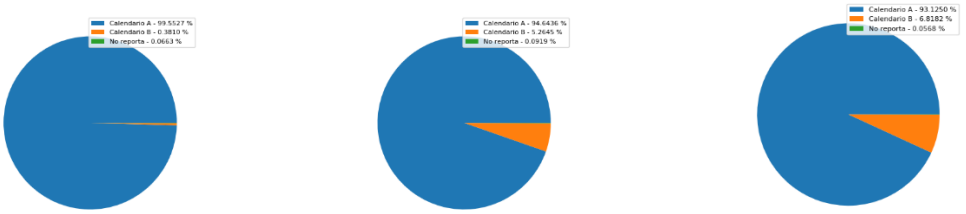

Figura 3: Proporción de colegios por tipo de calendario en el colegio A o B

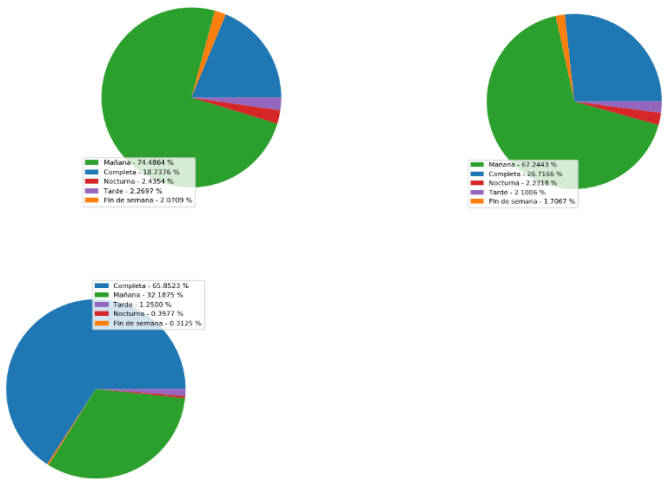

Figura 4: Proporción de colegios por tipo de jornada

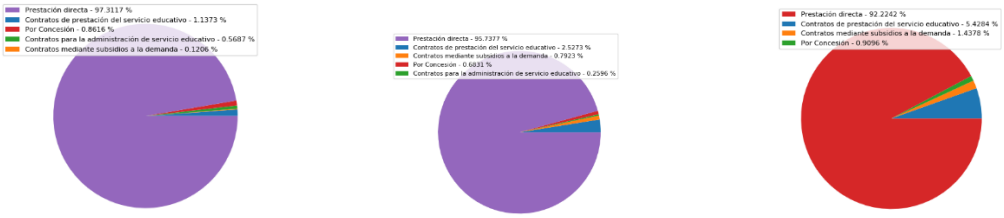

Figura 5: Proporción de colegios por modalidad de colegio

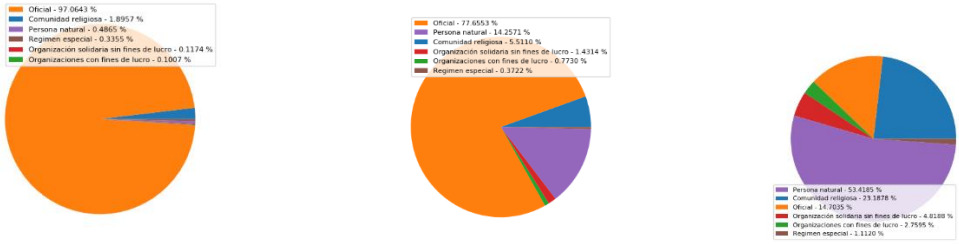

Figura 5: Proporción de colegios por naturaleza

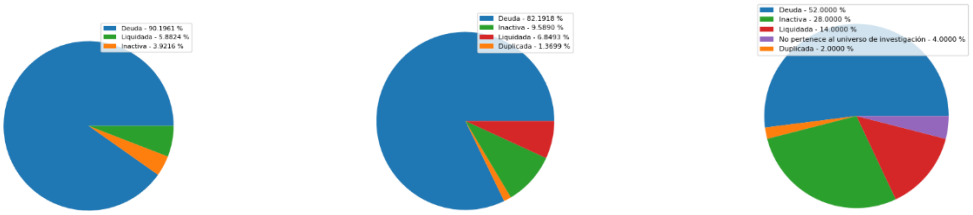

Figura 6: Proporción de colegios por tipo de novedad

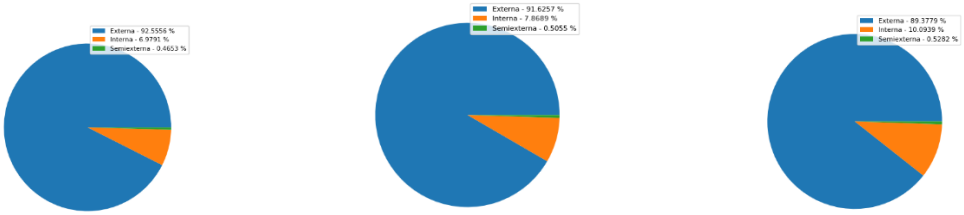

Figura 7: Proporción de colegios por tipo de población

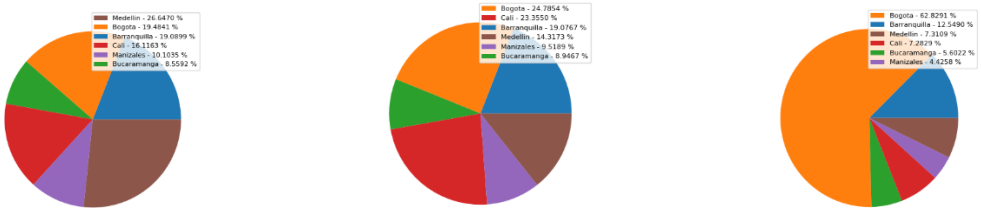

Figura 8: Proporción de colegios por territorial

Figura 9: Proporción de colegios por tipo de matricula

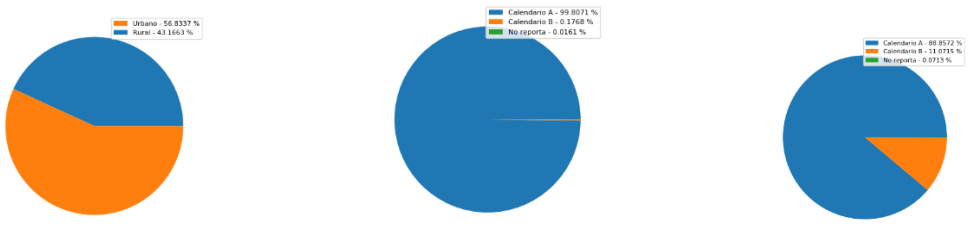

Figura 10: Proporción de colegios por tipo de área

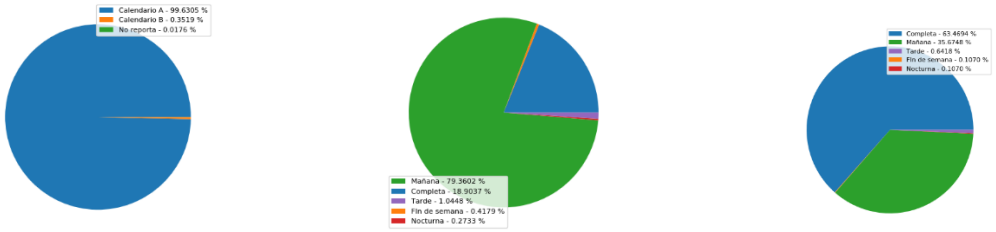

Figura 11: Proporción de colegios por tipo de calendario

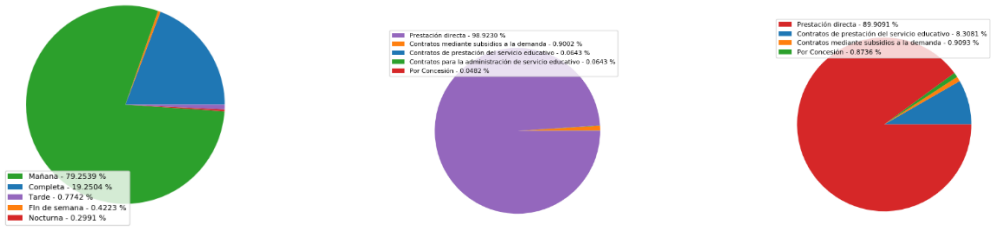

Figura 12: Proporción de colegios por tipo de jornada

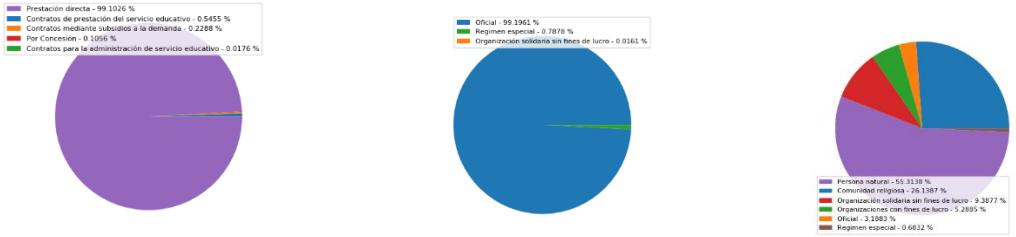

Figura 13: Proporción de colegios por modalidad de colegio

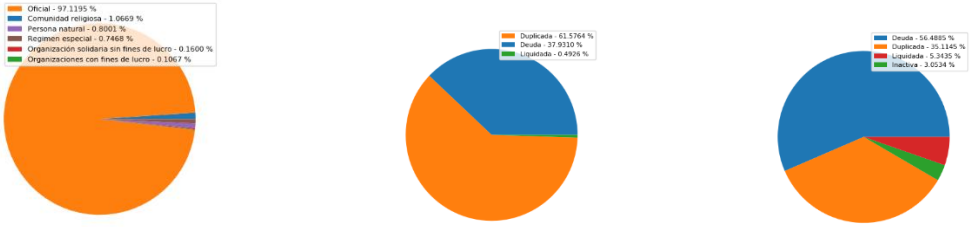

Figura 14: Proporción de colegios por naturaleza de colegio

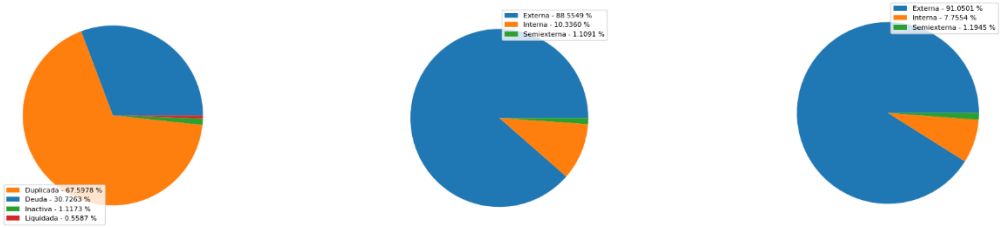

Figura 15: Proporción de colegios por novedad en colegio

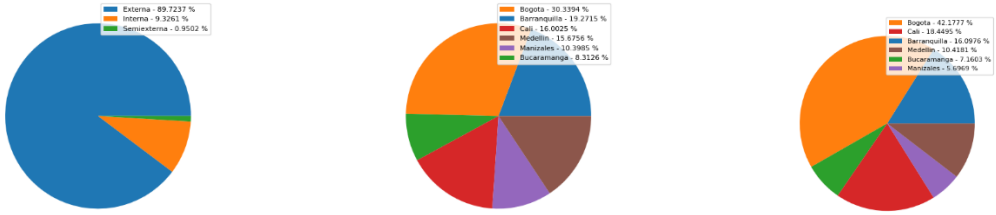

Figura 16: Proporción de colegios por tipo de población

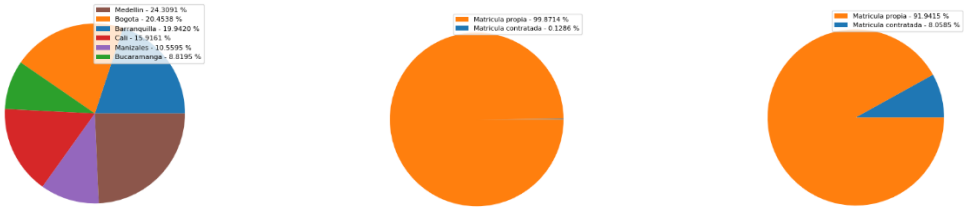

Figura 17: Proporción de colegios por territorial

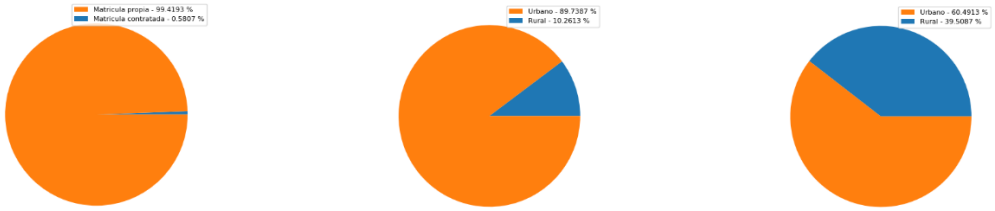

Figura 18: Proporción de colegios por tipo de matricula

2016

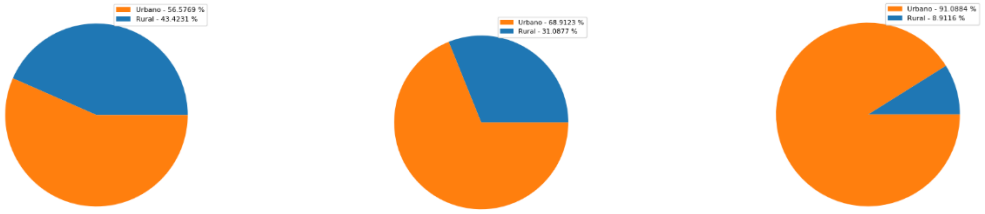

Figura 19: Proporción de colegios por tipo de área

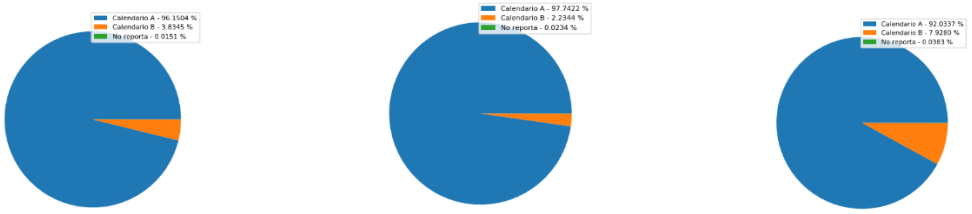

Figura 20: Proporción de colegios por tipo de calendario

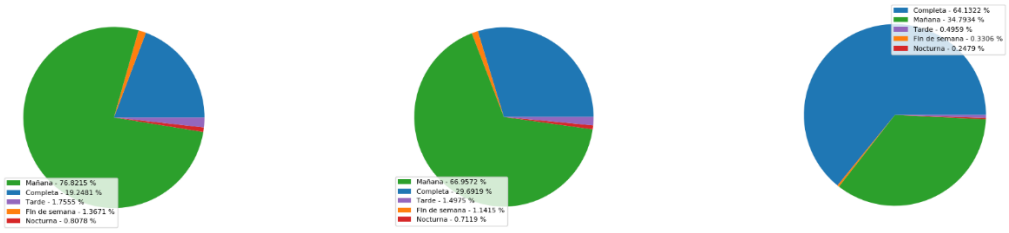

Figura 21: Proporción de colegios por tipo de jornada

Figura 22: Proporción de colegios por modalidad

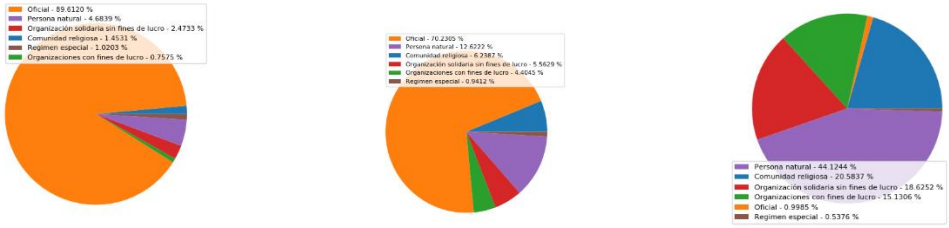

Figura 23: Proporción de colegios por tipo de naturaleza

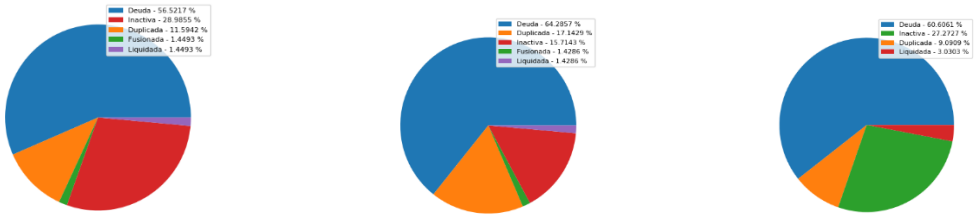

Figura 24: Proporción de colegios por tipo de novedad

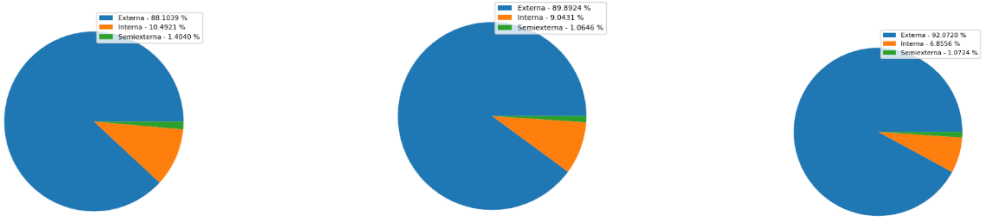

Figura 25: Proporción de colegios por tipo de población

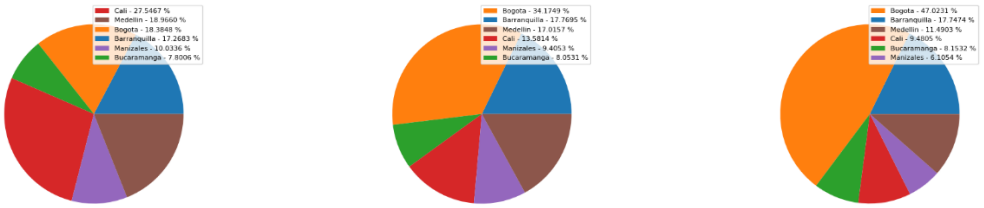

Figura 26: Proporción de colegios por territorial

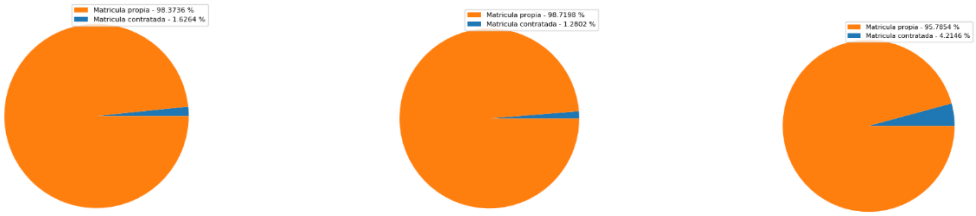

Figura 27: Proporción de colegios por tipo de matricula

Conclusiones

1. Dada la gran variedad de información relacionada con educación media y superior, se creó un algoritmo que enlaza la información proveniente del MEN en cuanto al cruce de información de colegios con respecto a universidades e información proveniente del DANE de Educación Formal
2. Se caracterizaron los colegios con metodologías de aprendizaje no supervisado en aquellos colegios cuya penetración de estudiantes a la educación superior sea baja, media y alta. Así mismo, se generó un algoritmo que crea las estadísticas de diferentes variables categóricas por cada uno de los clusters.

Socialización

Indique brevemente las entidades con las cuales se ha realizado la socialización del proyecto.