**REPORTE TÉCNICO**

**INTRODUCCIÓN**

El Departamento de Educación de los Estados Unidos pone a disposición del público general información de las escuelas públicas y privadas del país, datos como localización, número de campus y si es campus principal, nivel de perfil de grado de la escuela, tipo de título que ofrece la escuela, tasa de finalización y transferencias, entre otros, esto para tomar mejores decisiones al momento de elegir una institución donde se desea estudiar y aumentar la tasa de graduados para las poblaciones más desfavorecidas.

**OBJETIVO**

* Analizar los datos suministrados y realizar el debido preprocesamiento para su posterior análisis.
* Desarrollar un agrupamiento de instituciones de educación superior dependiendo de qué variables tienen mayor correlación entre sí.
* Caracterizar cada grupo y analizar las variables más significativas de cada grupo
* Entender qué hace que un grupo sea una buena opción
* Proponer cómo se podría generar un conjunto de datos que permita hacer lo mismo para Colombia a partir de la información disponible y de otra que se deba recaudar.

**BASE DE DATOS.**

Se toma la base de datos [**CollegeScorecard.csv**](https://data.world/exercises/cluster-analysis-exercise-2/file/CollegeScorecard.csv) suministrada por el Departamento de Educación de Estados Unidos () publicada en el portal educativo <https://data.world> (<https://data.world/exercises/cluster-analysis-exercise-2>), ademas se tomo de la pagina oficial del Centro Nacional de Estadística Educativa de Estados unidos (NCES, por sus siglas en ingles) la clasificación de los programas de instruccion y asi poder entender las variables columnas de la base de datos antes mencionada (<https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=56> ).

**DESCRIPCIÓN TÉCNICA**

Se realiza un preprocesamiento de la base de datos para realizar el correcto análisis, los pasos realizados son:

* Se verifica el número de columnas que tiene la base de datos y eliminan aquellas duplicadas y las que tiene más del 30% de de datos como nulos.
* Se procede a buscar las columnas con valores categóricos como o son booleanos y columnas con 3 valores únicos, esto para confirmar manualmente con el diccionario de datos si las variables son categóricas y se procede a eliminarlas.
* Por último se eliminan las variables cualitativas restantes, esto porque no se puede establecer una estadística confiable de ellas.

Al finalizar el preprocesamiento de datos quedan un total de 130 variables con las cuales se procede con la reducción de dimensión, pero antes de ello se cambian las columnas que contienen valores NaN por el promedio general de toda la variable para no afectar la reducción de dimensión.

Luego de realizar este preprocesamiento de datos, se grafican las escuelas en un mapa para verificar si la ubicación es relevante en el análisis, esto se puede observar en el apartado de VISUALIZACIÓN GEOGRÁFICA, después de realizar la reducción de dimensión empleando la factorización de la matriz no negativa obtenemos los datos sufusiones para emplear el Método del codo (Elbow Method) para determinar el número de componentes según el punto en donde la diferencia entre los valores comience a ser menos significativa (Donde la curva es más suave), Determinando así el usar 4 componentes para la función NMF.

Las 130 columnas se encuentran ahora representadas por 4 componentes principales que además están estandarizados. Se aplica de nuevo el Método del codo para determinar el número óptimo de clusters y haciendo uso de la librería Kmeans y variando nuevamente el número de componentes, Nuevamente se encuentra que el número óptimo y apropiado para el número de clusters es 6 que es dónde la gráfica se suaviza casi en su totalidad, además es pertinente elegir este valor y no el 4 debido a la gran cantidad de datos a representar.

Se aplica el Cluster y después de obtener estos resultados, se realiza de nuevo una visualización geográfica donde se puede verificar que la localización no es una variable significativa en este modelo.

Por último se realiza una caracterización de cada cluster donde se observa cuales son las variables más significativas en cada uno de ellos

**RESULTADOS**

* Se obtienen 4 clusters que separan las variables más significativas y que se correlacionen entre sí.
* Se presenta una visualización geográfica dinámica de la agrupación de los diferentes clusters en el territorio de Estados Unidos
* Significancia de ubicación de las escuelas en el territorio nacional.

**CONCLUSIÓN**

* Como primera conclusión se tiene que para este modelo no es significativo la ubicación de las escuelas para su mayor aprovechamiento, esto da a entender que faltan datos de los estudiantes como si se tuvieron que mudar para realizar sus estudios, lugar de residencia del estudiante y otras variables de localización de los estudiantes más que de las escuelas, se considera que un factor importante para tomar la decisión de donde estudiar es si se tiene que mudar o cuánto se debe desplazar el estudiante para tomar sus estudios.
* Al analizar cada cluster, el modelo nos arroja que los factores más importantes al momento de elegir la mejor opcions de qué escuela asistir se puede observar que:
  + **Cluster 0:** Las universidades incluidas en este cluster son la mejor elección para estudiantes que estén buscando títulos afines a las ciencias sociales, matemáticas o historia.
  + **Cluster 1:** Las universidades incluidas en este cluster son la mejor elección para estudiantes que estén buscando títulos afines a ciencias sociales y matemáticas, idiomas e historia, además se debe tener en cuenta la raza de estudiante específicamente el color de piel.
  + **Cluster 2:** Las universidades incluidas en este cluster tienen gran variedad de oferta de títulos, por lo que este no es un problema al momento de tomar una decisión, sin embargo, el estudiante debe tener en cuenta características como costo de matrícula, número de campus y la raza específicamente si son extranjeros no residentes o asiáticos.
  + **Cluster 3:** Las universidades incluidas en este cluster tienen gran variedad de oferta de títulos, por lo que este no es un problema al momento de tomar una decisión, sin embargo, el estudiante debe tener en cuenta características como, número de estudiantes de pregrado, número de campus y la raza específicamente si son indios americanos y nativos de alaska.
  + **Cluster 4:**
  + **Cluster 5:** Las universidades incluidas en este cluster son la mejor elección para estudiantes que estén buscando títulos afines a las ciencias exactas e historia además de tener en cuenta la raza específicamente si es hawaiano o isleño del pacifico
* Después de analizar detenidamente cada cluster, se puede observar que en 4 de los 6 clusters hay segregación racial, lo que indica que Estados Unidos debe invertir en programas educativos contra el racismo para que esto deje de ser un item a tener en cuanta al moentno de elegir la universidad donde se desea estudiar.
* Para finalizar se concluye que faltaron datos de localización y movilización de los estudiantes para que estos componentes tengan un papel más importante en el modelo.

**RECOMENDACIONES**

A la fecha no se cuenta con una única encuesta o base de datos donde se pueda analizar cuál es la mejor universidad para cada estudiante en el país.

Lo que se recomienda es realizar una unificación de diversas encuestas suministradas por el Departamento Administrativo Nacional de Estadística (DANE) como lo son, [Saber para Decidir – Sistema nacional de información de demanda laboral](https://www.dane.gov.co/index.php/estadisticas-por-tema/educacion/boletin-sinidel), [Índice de costos de la educación superior (ICES)](https://www.dane.gov.co/index.php/estadisticas-por-tema/educacion/indice-de-costos-de-la-educacion-superior-ices), [Fuerza laboral y educación](https://www.dane.gov.co/index.php/estadisticas-por-tema/educacion/fuerza-laboral-y-educacion) a estas bases de datos se le puede añadir la información que puede suministrar las universidades del país de sus graduados y demás estadísticos.