Silvia Ampuero

IT ACADEMY – Bootcamp Data Science

Análisis exploratorio y aplicación de Clustering

Sobre datos relacionados con discapacidades neurológicas en edades comprendidas de 3 a 18 años en ámbito Nacional.

eVALUACIÓN FINAL ANTES DE PROYECTO

Propuesta de trabajo final

**Introducción de los datos**

Mi interés por explorar el autismo me lleva a investigar conjuntos de datos relevantes en este campo. Mi objetivo es contribuir, en la medida de lo posible, al entendimiento y la conciencia sobre el autismo mediante el análisis de datos.

Durante mi búsqueda, encontré que los estudios más recientes a nivel nacional se basaban en datos del año 2019. Además, al revisar algunos conjuntos de datos disponibles, me di cuenta de que las cifras relacionadas con las discapacidades no estaban actualizadas. Este retraso se debe al proceso burocrático necesario para obtener el reconocimiento oficial de la discapacidad por parte de la Generalitat.

Como resultado, los datos disponibles no reflejan completamente la realidad de las discapacidades en la población. Es evidente que existe un vacío de datos, especialmente en el ámbito infantil.

Con esta comprensión, reflexioné sobre el impacto significativo que tiene la detección temprana de cualquier discapacidad en los niños.

Es por ello que desde el ministerio de Educación, a través de la generación de un informe de discapacidad de los centros de desarrollo infantil y atención precoz (Cdiap) o (CSMIJ) Centro de Servicios Médicos Integrales para Jóvenes, se activan las necesidades especiales que pueden mejorar la calidad de vida del niño. Y así, en el ámbito escolar, que abarca desde los 3 hasta los 18 años, se asignan plazas educativas en función de estas necesidades específicas.

Este proceso me llevó a centrar mi atención en un conjunto de datos virgen y real, donde puedo aportar información valiosa sobre mi tema principal.

El conjunto de datos me muestra todas las discapacidades, como la movilidad, auditiva, motora, intelectual, autismo y trastornos graves de conducta. Sin embargo, decidí enfocarme específicamente en las discapacidades neurológicas, es decir, intelectual, autismo y trastornos graves de conducta.

Finalmente, el conjunto de datos que estoy utilizando proviene del Ministerio de Educación de España y está disponible en el siguiente enlace: <https://www.educacionyfp.gob.es/servicios-al-ciudadano/estadisticas/no-universitaria/alumnado/matriculado.html>

Este conjunto de datos proporciona un registro detallado de las plazas asignadas tanto para la educación ordinaria como para las necesidades especiales en España, abarcando edades comprendidas entre los 3 y los 18 años.

Características generales

El conjunto de datos es de naturaleza educativa y estadística, ya que proporciona información sobre la matriculación y asignación de plazas en el sistema educativo español. Pertenece al sector educativo y está relacionado con la gestión y planificación de la educación en España.

Los datos son principalmente cuantitativos, ya que incluyen cifras y estadísticas sobre la matriculación y asignación de plazas en diferentes categorías.

La fuente principal de los datos es el Ministerio de Educación de España, que recopila y publica información sobre la educación no universitaria en el país.

Los datos proporcionan una visión general de la distribución de plazas educativas en España, tanto en el ámbito ordinario como en el de necesidades especiales, y abarcan un rango de edades desde los 3 hasta los 18 años.

Es posible que los datos pueden ser complejos debido a la variedad de variables y categorías involucradas, su estructura general es bastante organizada y comprensible, lo que permite análisis estadísticos y comparaciones significativas.

El conjunto de datos abarca toda España y proporciona información a nivel nacional sobre la distribución de plazas educativas en los 5 anteriores cursos.

**Definición de las variables**

* **Comunidad Autónoma/Provincia:** Esta variable indica la región geográfica dentro de España donde se encuentra la escuela o centro educativo. Permite analizar la distribución de plazas y comparar entre diferentes regiones.
* **Año:** Representa el año académico al que pertenecen los datos. Esta variable permite analizar las tendencias a lo largo del tiempo y detectar cambios en la matriculación y asignación de plazas en diferentes años.
* **Plaza:** Indica el tipo de plaza educativa asignada, como ordinaria o de necesidades especiales. Esta variable es fundamental para comprender la distribución de recursos y la atención a la diversidad en el sistema educativo.
* **Sexo**: Esta variable indica el género del estudiante matriculado, ya sea masculino o femenino. Permite analizar las diferencias de género en la matriculación y la asignación de plazas.
* **Total**: Representa el número total de estudiantes matriculados o el total de plazas asignadas en una determinada categoría (por ejemplo, total de plazas ordinarias o de necesidades especiales).
* **Discapacidad**: Indica si el estudiante tiene alguna discapacidad y, en caso afirmativo, el tipo de discapacidad que presenta. Esta variable es crucial para comprender las necesidades específicas de los estudiantes con discapacidad y para garantizar la inclusión y la igualdad de oportunidades en el sistema educativo.

Presentación de los objetivos

Me parece interesante el análisis exploratorio de datos (EDA), dónde voy a poder explorar las relaciones y los patrones de los datos. Calcular estadísticas descriptivas y crear visualizaciones.

Examinar las relaciones entre las variables utilizando correlaciones y con técnicas de visualización.

También puedo identificar posibles tendencias a lo largo del tiempo y en diferentes regiones geográficas.

Mientras que en Cataluña aplicaré técnicas de aprendizaje no supervisado, como el clustering, para identificar segmentos de estudiantes con características similares. Por ejemplo, utilizando el algoritmo de K-means, agruparé los datos por características como la discapacidad, el sexo y el número de matrículas especiales por provincia.

Además, exploraré un modelo predictivo para predecir la asignación futura de plazas especiales en Cataluña durante periodos de 5 o 10 años.

Consideré inicialmente el modelo SARIMA, pero al no encontrar estacionalidad en los datos, optaré por un modelo de Autoregresión, adecuado para datos secuenciales con correlaciones temporales, permitiéndome predecir valores futuros basados en observaciones pasadas.

Mi objetivo final es contribuir al conocimiento sobre las necesidades educativas y la asignación de recursos para personas con discapacidad, especialmente en Cataluña, proporcionando información valiosa para la planificación.