Selección del conjunto de datos
PRA1

Juan María Guerrero Carrasco

Ejercicio 1 (10%)

En primer lugar, cabe especificar que los datos escogidos hacen referencia al número de alumnos que han cursado una especialidad de bachillerato u otra, especificándose el contexto de dichos alumnos; entre el que se encuentra su sexo, procedencia, residencia de estudios, así como la situación de presencialidad y si el centro es público o privado. Los motivos de escoger este conjunto de datos podrían definirse como puramente personales, ya que actualmente tengo una hermana a punto de graduarse en segundo de bachillerato, así como diversos amigos realizando prácticas de profesorado en diversos institutos a lo largo de España. Y tanto de la información que me llega de mi hermana, como de la que recibo de mis amigos, puedo ver que los motivos que llevan a un alumno a interesarse más por una especialidad o por otra pueden llegar a ser tan diversos como interesantes.

Partiendo de esta base me pareció un buen punto de partida el indagar sobre si había información de calidad al respecto en la página del gobierno de España y, aunque un poco desestructurada y con una organización caótica, obtuve el conjunto de datos que presento y que, a mi parecer, una vez tratado el dato y estructurado este de forma correcta, nos ofrece una información muy rica y que puede dar lugar a visualizaciones de calidad.

Aunque se adjuntará junto con este PDF el conjunto original, aparte del ya procesado y listo para generar las visualizaciones, cabe destacar que dicho set original puede descargarse desde el siguiente enlace: https://datos.gob.es/es/catalogo/e05073401-estudiantes-de-nuevo-ingreso-en-grado-por-tipo-y-modalidad-de-la-universidad-sexo-zona-de-nacionalidad-y-rama-de-ensenanza1 .

Ejercicio 2 (10%)

Antes de continuar, es necesario poner los datos más en contexto. En primer lugar, los datos hacen referencia al bachillerato que cursaron alumnos que finalmente han accedido a la universidad, por lo que el último año del que se podría disponer datos sería el año pasado (2022). Partiendo de esto, en el conjunto podemos encontrar datos desde el curso finalizado en 2016 (curso posterior a la fecha en la que finalicé a título personal bachillerato) hasta 2022, que sería el último dato posible por lo que podemos concluir que estarán muy actualizados, y estos últimos factores le añadirán mayor valor

personal si cabe a los resultados de las visualizaciones. Cabe destacar que la tasa de actualización es anual y la última fue el 9 de junio de 2022 (tras la EVAU).

Aparte de esto, el tema puede parecer no ser sumamente transcendental, pero nada más lejos de la realidad. Lo cierto es que la elección de la especialidad en bachillerato es la primera gran decisión a la que nos enfrentamos en nuestra vida estudiantil, siendo esta crucial a la hora de abrirnos o cerrarnos ciertos caminos para poder acceder a unas carreras universitarias u otras, y lo peor es que tenemos que tomar esta importante decisión con apenas 16 años y sin tener del todo claro a qué atenernos para tomar dicha decisión.

Partiendo de esta base, realizar un análisis visual de las razones que impulsan más a unas decisiones u otras, así como analizar tendencias de género, o generacionales; pueden ayudar tanto a los futuros estudiantes a decidir como a las personas dedicadas al ámbito de la educación para orientar o alentar hacia ciertos ámbitos a sus estudiantes.

Finalmente señalar que sí se ha tenido en cuenta la perspectiva de género, siendo este un tema donde hay que poner especial énfasis en este apartado, puesto que la educación es la base de la igualdad en prácticamente todos los aspectos.

Ejercicio 3 (25%)

El set de datos original, sin ser procesado, cuenta con una estructura algo caótica donde se impone un formato de tablas anidadas que hacen referencia a casos concretos, como suele pasar quizá más comúnmente con los datos procesados en ficheros XLM. De esta forma podemos comenzar exponiendo que el conjunto original cuenta con del orden de decenas de columnas, ya que en concreto serán 36 de estas, y de igual modo contará con del orden de miles de filas o registros, alcanzando estas las 5051.

Sin embargo, cabe destacar que entre estos datos podemos encontrar gran cantidad de información redundante, así como extremadamente mal estructurada y de poco valor a la hora de realizar una visualización con propósito de obtener conclusiones observables. Para esto se ha llevado a cabo todo el procesado de datos mediante código Python, que podemos encontrar en el fichero Guerrero_Carrasco_JuanMaria_PRA1.pynb, en el que se detalla paso a paso el proceso ETL (aunque la extracción se hace de forma independiente al fichero) que se le ha realizado al conjunto. Una vez finalizado el proceso hemos obtenido un dato de mucha mayor claridad y calidad, que consta esta vez únicamente de 8 variables, mucho más explicativas y claras, aunque habrá aumentado a cambio su número de filas / registros hasta 39888. En este último set de datos la variable a medir será el **NUMERO**, que especifica con valores enteros qué cantidad de alumnos se graduaron en esa rama y con esas condiciones de contexto concretas. Aparte de esto nuestras variables de contexto serán la **RAMA** o especialidad cursada en bachillerato y sobre la que realizamos el análisis; el **SEXO** que será ‘Hombre’

o ‘Mujer’, la **COMUNIDAD** autónoma donde se cursaron los estudios y la **NACIONALIDAD** del alumno especificando su región geográfica de origen entre 7 posibilidades, como variables de texto cualitativas. Aparte de esto podemos encontrar el **CURSO** como el año numérico en que finalizó los estudios y las variables **PRIVADA** y **PRESENCIAL** que toman valores numéricos 0 y 1 como expresiones booleanas de si se estudió en esas condiciones o no respectivamente.

Podemos concluir así que, aunque una vez tratado el dato nos hemos quedado con un número relativamente reducido de variables, lo cierto es que tenemos datos categóricos, cuantitativos y booleanos, así como un muy elevado número de registros y variables de calidad, lo que nos puede ofrecer un gran abanico de posibilidades.

Ejercicio 4 (25%)

Con respecto a la originalidad de la visualización llevada a cabo, es cierto que en el ámbito de la educación existen multitud de análisis orientados a ver cómo la población se segrega en las diferentes ramas de conocimiento en base a su contexto demográfico, social y personal pero, sin embargo, el grueso de estos estudios se enfoca en etapas posteriores del ciclo formativo, es decir, la mayoría se enfoca en la etapa universitaria o de máster cuando la realidad es que es mucho antes cuando tomamos decisiones tan importantes como si orientaremos nuestro futuro a las ciencias o a las letras.

Un ejemplo de análisis que se centra en lo citado anteriormente y obtenido de la misma fuente (el gobierno de España) podría ser el siguiente: https://datos.gob.es/es/documentacion/caracterizacion-del-alumnado-de-la-universidad-espanola-y-titulaciones-mas-demandadas .

Aparta de esto sí que podemos encontrar algunas visualizaciones sin demasiada profundidad donde se trata el tema como puede ser la siguiente de statista a modo de ejemplo: https://es.statista.com/estadisticas/502118/distribucion-de-los-egresados-por-rama-de-ensenanza-espana/ , que realmente no da una evolución temporal ni analiza sesgos por género, procedencia o localización geográfica.

Dado que hemos visto que estamos ante un tema importante en el que no se suele poner el foco de análisis, merece la pena señalar las posibilidades de estudio que podemos plantear. Respecto al conjunto original se ha llevado una gran evolución, lo que nos permite llevar a cabo visualizaciones precisas que muestren la evolución temporal de cómo el alumnado se inclina más hacia unas ramas u otras en favor de un contexto social (sería interesante ver si la pandemia ha hecho a los jóvenes interesarse más o menos por el ámbito de la salud, por ejemplo). Otro enfoque interesante, y más aún en fechas cercanas a elecciones políticas, es plantearse si las políticas de igualdad de género están surtiendo efecto, puesto que la igualdad estamental y salarial en las

empresas no puede conseguirse sin la necesaria equidad en estudios científicos de calidad entre hombres y mujeres, por lo que aquí se plantea otro muy interesante caso de análisis que, de nuevo, está poco explotado en etapas formativas previas a la universidad, a pesar de que suele ser donde se decide todo.

Como conclusión, gracias al potente procesado de datos llevado a cabo, los datos se han enriquecido lo suficiente como para plantearnos explorar nuevas visualizaciones que no se habían llevado a cabo hasta ahora, pudiendo presentarse conclusiones muy valiosas.

Ejercicio 5 (30%)

Con respecto a las preguntas que pretendemos responder a partir de este conjunto de datos, como ya se adelantaban ligeramente en el ejercicio anterior, las principales serían las siguientes:

- ¿Ha variado el interés de los estudiantes en su etapa final del instituto en los últimos 6 años?

- ¿Han podido influir factores como la pandemia en un posible aumento o descenso del interés por las Ciencias de la Salud?

- ¿Es importante el factor social y cultural de la comunidad autónoma donde estudiamos para decidirnos por una especialidad? Y de ser así, ¿debería alguna comunidad incentivar de alguna forma el estudio de algunas ramas quizá más desatendidas?

- ¿Es importante el factor de origen o nacionalidad, a la hora de determinar que especialidad estudiaremos? Esta pregunta puede ser importante para determinar políticas de inclusión en caso de que ciertas nacionalidades tengan dificultades en algunos aspectos.

- ¿Funcionan las políticas de igual de género en cuanto a la inclusión de la mujer en las ciencias puras? ¿Se ha conseguido alcanzar una cifra igualitaria? ¿Se puede observar una evolución en los últimos años?

En conclusión, y basándonos en las respuestas de los apartados anteriores, parece bastante factible el poder dar una solución convincente y razonada a estas preguntas mediante técnicas de visualización y análisis del conjunto en cuestión. Cabe destacar de nuevo lo importante que ha sido el trabajo de procesamiento del dato, que en definitiva será lo que nos permita obtener un valor del conjunto de datos que no se había obtenido hasta la fecha. Las preguntas se adecúan a nuestros datos y estos han sido debidamente preparados para poder darles respuesta aportando, como se ha dicho anteriormente, un valor incalculable en el ámbito educativo.