Universitat Oberta de Catalunya

Luis Francisco Leandro Jiménez

Edna Espejo Osma

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.**

El contexto de la recolección es educativo. El principal objetivo es recolectar información sobre oferta académica disponible para que pueda ser utilizado en análisis como tendencias en cursos, tecnologías exploradas, categorías y subcategorías solicitadas en este momento. Por ejemplo, ver cuál es el área más explorada en tecnologías de la información, entre otros. El sitio seleccionado para este caso es https://wwww.coursera.org/, ya que es una plataforma popular en educación virtual, posee convenios con gobiernos y brinda cursos de universidad con prestigio a nivel internacional.

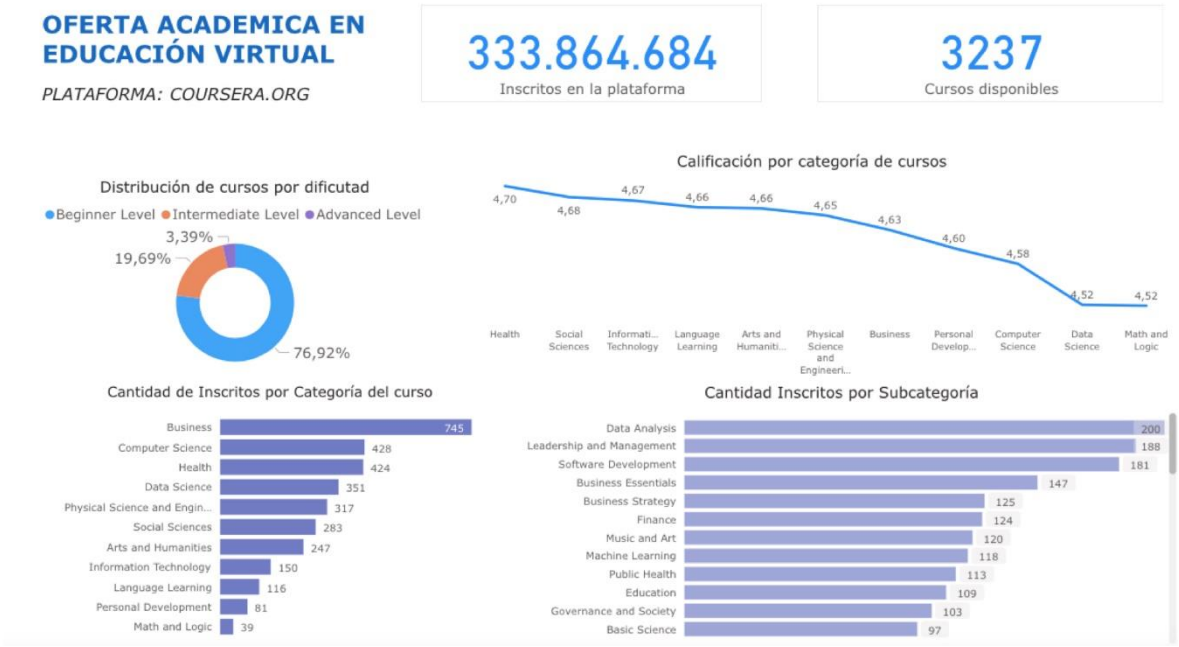
2. **Definir un título para el dataset. Elegir un título que sea descriptivo.**

Oferta académica en educación virtual

3. **Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

El dataset contiene información tomada de la página web https://www.coursera.org/, de la que se extrajo información acerca de los cursos disponibles, caracterizandola por categoria y subcategoria, cantidad de estudiantes inscritos, nivel de dificultad y calificación de estos.

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente**



5. **Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

Los campos capturados en el proceso de web scraping son:

course_name: es el nombre del curso

course_category: nombre de la categoría general, por ejemplo: business

course_subcategory: nombre de la subcategoría, en este caso se delimite el área de la categoría, por ejemplo: Entrepreneurship

students_enrolled: cantidad de estudiantes matriculados

course_difficulty: dificultad del curso, por ejemplo: Beginner Level

course_rating: ranking del curso de acuerdo con el feedback recibido

course_rating_count: cantidad de veces que recibió calificación

El web scrapping se realizó el 20 de octubre, y la validez depende de la frecuencia de cambios en los cursos, puede ser que se abran o cierren en poco tiempo.

Para la recolección se utiliza un script en python con lo siguientes paquetes:

Pandas BeautifulSoup

Se utiliza el sitemap principal de coursera, se toma el url de cursos y posteriormente se procesa cada html que contiene la información de los cursos para crear la estructura requerida.

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).**

Los datos son obtenidos por la plataforma coursera. Según Coursera en el reporte del año 2020, fue fundada en el año 2012 con el objetivo de permitir el acceso universal a la educacion. Entre los beneficios que se reportan en el informe están:

87% de los usuarios reportan mejoras en sus trayecto profesional 88% de los usuarios sin bachillerato reporta beneficios en la parte profesional. Entre otros

Existen diferentes estudios sobre el impacto de la plataforma en diferentes regiones, una de ellas es latinoamérica y el papel que ha tenido Coursera en mejorar la educación en diferentes áreas. La investigación se puede encontrar en el siguiente url: https://www.researchgate.net/publication/306028656_Los_Cursos_Masivos_en_Linea_en_Coursera_y_su_Empleo_Potencial_en_los_Programas_de_Ingenieria_en_America_Latina

Por lo tanto, se observa que la plaforma representa una oportunidad para muchas personas de mejorar su carrera profesional.

7. **Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.**

La información de los cursos ofertados por esta plataforma es interesante por su popularidad en las personas que desean aprender y es bueno estar informado cuales son las áreas en las que más se capacitan ya que permite ver cuales son las necesidades de conocimiento que se tiene actualmente para entrar al mundo laboral. Por tal razón el dataset contiene información de las calificaciones de los cursos y sus incritos para responder preguntas cómo: ¿Cuáles son los cursos con mayor demanda? ¿Cuáles son las categorias y subcategorias de estudio con mayor cantidad de inscritos? ¿Cuáles es el ranking de los cursos mejor puntuados? ¿En qué nivel de dificultad se encuentran la mayoria de los cursos?

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: ○ Released Under CC0: Public Domain License ○ Released Under CC BY-NC-SA 4.0 License ○ Released Under CC BY-SA 4.0 License ○ Database released under Open Database License, individual contents under Database Contents License ○ Other (specified above) ○ Unknown License**

Seleccionamos la licencia Public Domain License para el dataset dado que la información recopilada es información publica de libre consulta sin necesidad de estar registrado en la pagina fuente https://www.coursera.org/.

9. **Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.**

10. **Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.**

DOI: 10.5281/zenodo.4130865
URL: https://zenodo.org/record/4130865#.X5XmUZMzbBI

Coursera. 2020. 2020 Impact Report. Recuperado de: https://about.coursera.org/press/wp-content/uploads/2020/09/Coursera-Impact-Report-2020.pdf

Contribuciones	Firma
Investigación previa	LL,EE
Redacción de las respuestas	LL,EE
Desarrollo código	LL,EE