Unificación de datos abiertos del MEC utilizando Python

Martina Irasema Martínez Bazán
FCyT UNCA
Coronel Oviedo, Paraguay
mimartinezb@fctunca.edu.py

Héctor Ramiro Estigarribia Barreto
FCyT UNCA
Coronel Oviedo, Paraguay
hestigarribia64@fctunca.edu.py

Luis Fernando Giménez Valdez
FCyT UNCA
Coronel Oviedo, Paraguay
lfgimenezv@fctunca.edu.py

***Abstract* - This study aims to unify two open datasets from the Ministry of Education and Sciences of Paraguay: the Degree Registry and the National Career Registry. By integrating these datasets, the research seeks to provide a comprehensive view of the relationship between educational offerings and awarded degrees, which is crucial for assessing educational quality and academic planning. The analysis is conducted using Python with Anaconda to leverage its advanced data management capabilities. The results highlight the need for data cleansing, standardization, and duplicate identification, revealing underlying issues in the registration processes.**

Keywords: data unification, Python, open data, educational analytics

***Resumen*— Este estudio en curso tiene como objetivo la unificación de dos conjuntos de datos abiertos del Ministerio de Educación y Ciencias de Paraguay: el Registro de Títulos y el Registro Nacional de Carreras. La integración de estos conjuntos de datos busca proporcionar una visión integral de la relación entre la oferta educativa y los títulos otorgados, lo cual es fundamental para la evaluación de la calidad educativa y la planificación académica. Para ello, se utiliza Python con Anaconda, aprovechando sus avanzadas capacidades en gestión de datos. Los resultados del análisis destacan la necesidad de realizar procesos de limpieza, estandarización y detección de duplicados, lo que pone de manifiesto problemas subyacentes en los procesos de registro.**

Palabras clave: unificación de datos, Python, datos abiertos, análisis educativo, gestión de datos.

I. Introducción

En los últimos años, Paraguay ha avanzado significativamente en la implementación de políticas de datos abiertos, promoviendo la transparencia y la participación ciudadana. Los datos abiertos permiten el libre uso, reutilización y redistribución de información pública, siempre que se mantenga la atribución original y se compartan bajo las mismas condiciones. La Ley N° 5282/14 de Libre Acceso Ciudadano a la Información Pública y Transparencia Gubernamental [1], promulgada en 2014, marcó un hito en este esfuerzo, obligando a las instituciones del Estado a publicar datos de manera proactiva.

El Ministerio de Educación y Ciencias (MEC) ha sido una de las instituciones pioneras en adoptar estas políticas, publicando desde 2012 conjuntos de datos clave sobre educación superior, como el registro de títulos y el Registro Nacional de Carreras (RNC) [2] y el Registro de títulos de la Educación Superior (RTES) [3]. Estos datos permiten un análisis detallado de la oferta educativa y los títulos otorgados, siendo fundamentales para la planificación académica y la mejora de la calidad educativa.

El portal de datos abiertos del MEC [4] facilita el acceso a información variada, como matriculaciones, nóminas de funcionarios e inventarios de bienes, promoviendo la transparencia en el sector educativo. Investigadores, periodistas y ciudadanos han utilizado estos datos para desarrollar aplicaciones tecnológicas y realizar análisis que contribuyen a la mejora del sistema educativo paraguayo.

El portal de datos abiertos mencionado está desarrollado por el Ministerio de Educación y Cultura (MEC), con el apoyo del Programa Democracia y Gobernabilidad (PDG), con fondos de USAID, y ejecutado por CEAMSO [5]. El proyecto fue desarrollado con el Framework para desarrollo ágil de aplicaciones web Ruby on Rails, la versión utilizada de ruby es la 1.9.3 con rails 3.2.13 [6]

Sin embargo, no abundan los trabajos científicos realizados utilizando la información disponibilizada por el MEC, y los pocos que existen destacan la necesidad de limpieza, estandarización y unificación de la información dispersa en ambos datasets para obtener un conocimiento de utilidad sobre el mismo. Por ejemplo, Estigarribia [7] para obtener el dato del porcentaje de egresados pertenecientes a carreras TIC del Paraguay utilizó los registros de títulos con la palabra “informática” del RTES y los cruzó con los registros del RNC en el cual figuraba el dato de área del conocimiento donde se define que carreras son “TIC”. Sin embargo, no todas las carreras del RNC tienen ese campo completado, por lo que es probable que se hayan excluido datos importantes referentes al objetivo del trabajo. El mismo además menciona que para realizarlo se tuvo que descargar el archivo csv de cada dataset e importarlo en Excel, donde se obtuvo la información publicada. Otro trabajo [8] relacionado a estos datasets menciona que la herramienta utilizada fue PowerBI siguiendo la misma metodología.

En este trabajo se propone utilizar scripts de Python que al ser ejecutados permitan obtener siempre la versión más reciente de los datos, limpiarlos, estandarizarlos, y crear una nueva base de datos con la unificación los datos de los

datasets mencionados para poder extraer información valiosa de forma rápida, segura y actualizada.

Este artículo está organizado en cuatro secciones principales. La Introducción presenta el contexto y relevancia de la unificación de datos en el ámbito educativo. La sección de Metodología describe en detalle el proceso seguido para la unificación de los conjuntos de datos y los pasos realizados en cada script. En Resultados, se resumen las mejoras alcanzadas con el modelo propuesto. Finalmente, la sección de Conclusiones discute los hallazgos y su impacto en la calidad y planificación educativa en Paraguay.

II. METODOLOGÍA

La metodología para la unificación de los conjuntos de datos del "Registro de Títulos" y el "Registro Nacional de Carreras" del Ministerio de Educación y Ciencias (MEC) de Paraguay se llevará a cabo utilizando Python en un entorno de Jupyter Notebook, gestionado con Anaconda. En este trabajo se utilizan datasets en formato CSV por la simplicidad y compatibilidad con las herramientas de análisis de datos, como Python y pandas. Aunque los datos originales se encuentran estructurados en sistemas relacionales más complejos, el formato CSV permite una manipulación rápida y sencilla para los objetivos de unificación y análisis. Los resultados permitirán una visión integral del panorama educativo en Paraguay, facilitando análisis posteriores sobre la relación entre la oferta educativa y los títulos otorgados.

El script utilizado [9] se encuentra disponible para revisión, por lo que no se incluirán detalles del código en este documento. A continuación, se explican los procesos mas detallados realizados durante el procesamiento de los datos, respetando los principios de anonimato, esto es solo dentro del enfoque científico pues se trata de datos abiertos.

- Carga y Normalización Inicial: Se importaron los datasets de carreras y títulos, Este proceso de preprocesamiento de datos utiliza dos funciones: una para limpiar y estandarizar los nombres, eliminando caracteres especiales, acentos, y espacios redundantes, y otra para expandir abreviaturas frecuentes (como "dr" por "doctor") en nombres de instituciones educativas, mejorando así la uniformidad.
  Ambas funciones se aplican a los datasets con tqdm para monitorear el progreso. Este enfoque permite una mayor consistencia en los nombres y facilita la comparación entre registros, mejorando la calidad y precisión del análisis de datos.
- Unificación de Nombres de Instituciones: Este código emplea coincidencia difusa (fuzzy matching) para identificar nombres de instituciones similares entre dos conjuntos de datos, utilizando la biblioteca fuzzywuzzy para realizar comparaciones y ajustando un umbral de similitud para determinar qué nombres se consideran equivalentes. Se realiza una búsqueda aproximada en la que se establece un umbral mínimo de similitud para aceptar coincidencias, lo que ayuda a gestionar variaciones comunes en los nombres, como abreviaciones o errores ortográficos. La función principal compara cada nombre en la primera columna de un DataFrame con los nombres en la columna correspondiente del segundo DataFrame, buscando la coincidencia más cercana en función de una puntuación de similitud calculada por token_set_ratio. Si la coincidencia supera el umbral, se considera una pareja válida, y si no, se etiqueta como "#N/D" para denotar la falta de una coincidencia.
  Además, el código incorpora un refinamiento en la coincidencia utilizando partial_ratio, que permite identificar nombres que, aunque no coincidan completamente, se acercan al umbral considerando solo partes del nombre. Si los puntajes de similitud inicial y refinado cumplen o se acercan al umbral en promedio, el nombre coincidente se clasifica como un "nombre unificado" y se agrega a los resultados. Esto facilita una visión coherente y estándar de las instituciones, reduciendo inconsistencias entre los conjuntos de datos. Los resultados se exportan a archivos Excel, donde cada nombre original se acompaña del nombre coincidente y el nombre unificado (si corresponde), lo cual facilita su revisión y ajustes posteriores. Este proceso resuelve discrepancias en los registros de instituciones, mejorando la calidad y confiabilidad de los datos y permitiendo una correspondencia precisa entre las instituciones en ambos conjuntos.
- Estandarización del Tipo de Institución: Se limpiaron y estandarizaron los campos de tipo de institución en ambos datasets, corrigiendo cualquier inconsistencia de origen en la clasificación de estas entidades.
- Creación de Identificador de Persona: En el dataset de títulos, se asignaron identificadores únicos (ID de persona) a cada registro de documento que tenía múltiples nombres asociados pero referían a distintas personas, solucionando así duplicaciones incorrectas.
- Fragmentación por Tipo de Institución: Ambos datasets fueron divididos según el tipo de institución, facilitando la aplicación de técnicas de limpieza y unificación específicas para cada sección de datos.
- Limpieza Específica por Tipo de Institución: Se aplicó un proceso de limpieza similar en cada fragmento, adaptando la metodología a las particularidades de cada tipo de institución.
- Verificación de Correspondencia de Instituciones: Se comparó la cantidad y nombres de instituciones en el dataset de títulos con las correspondientes en el dataset de carreras para cada tipo de institución. Las instituciones no coincidentes fueron evaluadas, y se unificaron los nombres donde existían variaciones mínimas o, de ser necesario, se realizaron correcciones manuales.

- Manejo de Valores Nulos: Se identificaron y trataron los valores nulos en los campos correspondientes, aplicando scripts de imputación para rellenar los campos vacíos donde fue posible.
- Unión Inicial y Detección de Inconsistencias: Este código está diseñado para identificar y extraer valores inconsistentes entre dos conjuntos de datos sobre títulos e instituciones. El proceso comienza haciendo copias de los DataFrames originales, permitiendo realizar cambios sin afectar los datos iniciales. Luego, las columnas con nombres similares en los DataFrames son renombradas para facilitar su diferenciación. Esto es importante porque al realizar un "merge" o combinación de ambos conjuntos de datos, queremos evitar confusiones entre columnas de nombre idéntico que representan categorías distintas (por ejemplo, el "título" en el registro de títulos y el "título" en el registro de carreras). El "merge" se realiza mediante una combinación externa ("outer join") que conserva todas las filas de ambos conjuntos, tanto coincidentes como no coincidentes. Un indicador especial (_merge) se incluye en el DataFrame combinado para mostrar si cada fila proviene de ambos conjuntos de datos o solo de uno de ellos.
  La estructura final permite filtrar fácilmente los valores que solo aparecen en uno de los dos DataFrames originales y que no tienen una coincidencia en el otro. Este filtrado crea un subconjunto de datos, denominado not_matching, que captura precisamente aquellas filas con valores sin par, indicando posibles inconsistencias o errores en los registros. Este subconjunto se exporta a un archivo Excel para su revisión posterior, permitiendo a los analistas inspeccionar estos datos y realizar ajustes necesarios para mejorar la integridad de los datos. Esta estrategia de filtrado y exportación facilita la depuración de los datos, resaltando aquellos valores que podrían generar problemas de precisión en los análisis si no se corrigen.
- Corrección de inconsistencias: Este script implementa un proceso de unificación de datos textuales utilizando técnicas de coincidencia difusa (fuzzy matching), aplicado a nombres de carreras y títulos académicos en distintas instituciones educativas. Para ello, hace uso de bibliotecas de Python como pandas para la manipulación de datos, fuzzywuzzy para el cálculo de similitud entre textos, y tqdm para el seguimiento del progreso en la ejecución. La unificación se basa en una lista predefinida de palabras clave relevantes (como "CIENCIAS DE LA EDUCACION" y "ADMINISTRACION"), que se utilizan para dividir los textos en partes anteriores y posteriores a dichas palabras, facilitando una comparación más precisa entre los nombres de las carreras. El script también incluye una función que evalúa si las palabras clave están presentes en uno de los textos pero ausentes en el otro, lo que puede reducir la puntuación de similitud entre las cadenas.
  La función central del proceso, fuzzy_unify, evalúa la similitud de las filas del dataset comparando los nombres de las carreras y títulos mediante coincidencia parcial difusa, asignando una puntuación basada en la cercanía de los términos, y ajustando esta puntuación si hay discrepancias en las palabras clave. Si la similitud calculada supera un umbral predefinido (85), los datos se consideran coincidentes y se unifican; en caso contrario, se conservan los valores originales. Finalmente, el script organiza y exporta los resultados en un archivo Excel para la verificación manual de los resultados antes de pasar al proceso de unificación, contiene columnas que incluyen los nombres originales de las carreras y títulos, las versiones unificadas, las coincidencias encontradas, y una puntuación de similitud que permite la evaluación de la calidad del proceso de unificación.
- Validación de Consistencias: Luego de corregir las inconsistencias, se ejecutaron nuevamente los scripts de detección de inconsistencias para verificar la reducción de discrepancias, asegurando que los datos finales se mantuvieran dentro de un rango aceptable de consistencia.
- Unificación Limpia por Tipo de Institución: Los datasets de cada tipo de institución, ahora limpios, fueron unificados en bloques según su clasificación.
- Concatenación Final: Se concatenaron todas las uniones limpias de cada tipo de institución, consolidando todos los datos en un único dataset final optimizado.

III. RESULTADOS

En este análisis se utilizaron cuatro conjuntos de datos clave:
títulos: Datos del registro nacional de títulos otorgados en diferentes áreas.
carreras: Información detallada del registro nacional de carreras, con datos sobre especialización y la institución de estudio.
merged: Resultado inicial de la combinación de titulos y carreras en un solo dataset para facilitar el análisis cruzado, aunque con duplicados y datos faltantes.
dataset_final: Versión optimizada del merged. Incluye mejoras en la limpieza y completitud de datos y un identificador único para cada persona (persona_id), proporcionando una base de datos lista para análisis.

El análisis comparativo de los datasets titulos, carreras, merged y dataset_final revela las siguientes mejoras significativas:

1. Tamaño y Estructura: El dataset_final cuenta con 553,407 filas y 13 columnas, manteniendo el volumen del conjunto merged mientras introduce la columna persona_id, lo que facilita la identificación única de registros.

2. Reducción de Duplicados: El número de valores únicos en la columna documento ha disminuido a 406,360, lo que sugiere una efectiva consolidación de registros duplicados y mejora la precisión en la identificación de individuos.

3. Completitud de Datos: Las columnas críticas como tipo_gestion, nivel_titulacion y clasificacion_campo_amplio han reducido sus valores nulos a 6.1%, un avance notable en comparación con el dataset merged, donde estas columnas presentaban un 100% de valores nulos.

4. Diversidad en Clasificación: La columna clasificacion_campo_amplio incluye ahora 25 valores únicos, lo que proporciona una categorización más exhaustiva de los campos de estudio. Esto permite identificar tendencias educativas específicas, como la predominancia de áreas como Administración de Empresas y Derecho, con 146,051 registros (26.4%).

5. Consistencia de Datos: El nivel de consistencia en la información de género se mantiene bajo un 0.18% de valores nulos, similar al de otros datasets. Esta consistencia es fundamental para garantizar la fiabilidad de los análisis demográficos futuros.

IV. Conclusiones preliminares

El proceso de unificación y optimización de los conjuntos de datos títulos y carreras, mediante un enfoque sistemático de limpieza y consolidación en Python, ha resultado en un dataset final robusto y confiable. Este dataset final, surgido de la experiencia acumulada en el desarrollo y ajuste del código, ha demostrado ser el enfoque más efectivo para alcanzar los objetivos de análisis educativo en Paraguay. Las mejoras logradas en términos de estructura, calidad y completitud de datos hacen de este dataset una herramienta valiosa para investigadores, responsables de políticas y administradores educativos.

Con una estructura de 553,407 registros y la incorporación de un identificador único persona_id, el dataset final asegura una representación precisa y única de cada individuo. Además, la reducción de duplicados, visible en la disminución de valores en la columna de documento a 406,360, optimiza la precisión y la consolidación de los registros. La reducción de valores nulos en columnas críticas como tipo_gestion, nivel_titulacion y clasificacion_campo_amplio a un 6.1%, desde un 100% en el dataset inicial, refuerza la calidad de los datos y permite una categorización más exhaustiva de las áreas de estudio. Esto facilita la exploración de patrones específicos en la educación superior, como la predominancia en áreas de Administración de Empresas y Derecho, que representan el 26.4% de los registros.

El dataset final, con su capacidad para apoyar análisis más completos y detallados, ofrece una base sólida para que los investigadores exploren patrones, identifiquen áreas de mejora y evalúen el impacto de diversas iniciativas educativas. En conclusión, el dataset final no solo cumple con los requisitos de limpieza y organización, sino que representa una herramienta fundamental para mejorar la planificación educativa y el análisis de datos en el sector educativo de Paraguay.

V. Referencias

[1] CONGRESO DE LA NACION PARAGUAYA, «Ley Nº 5282 / LIBRE ACCESO CIUDADANO A LA INFORMACIÓN PÚBLICA Y TRANSPARENCIA GUBERNAMENTAL,» 2014. [En línea]. Available: https://bit.ly/ley528214.

[2] MEC, «Registro Nacional de carreras,» [En línea]. Available: https://datos.mec.gov.py/data/rnc.

[3] MEC, «Registro de títulos,» [En línea]. Available: https://datos.mec.gov.py/data/registros_titulos.

[4] MEC, «Lista de datos abiertos del MEC,» [En línea]. Available: https://datos.mec.gov.py/data.

[5] MEC, «Acerca de datos abiertos,» [En línea]. Available: https://datos.mec.gov.py/about.

[6] MEC, «Github mec-opendata,» [En línea]. Available: https://github.com/mecpy/mec-opendata.

[7] H. R. Estigarribia Barreto, «Análisis de la oferta académica y egreso de carreras TIC en la educación superior en Paraguay, basado en el estudio de datos abiertos,» Revista ACADEMO, Universidad Americana Vol. 11 Núm. 2 (2024): Mayo - Agosto, 2024. [En línea]. Available: https://doi.org/10.30545/academo.2024.may-ago.9.

[8] Á. F. Silvera, «Estudio descriptivo de los registros de títulos de la Educación Superior en Paraguay, corte longitudinal 2012 al 2020,» Revista de investigación científica y tecnológica, Universidad Maria Serrana, Vol. 5 Núm. 1 (2021), [En línea]. Available: https://revista.serrana.edu.py/rict/article/view/258.

[9] H. R. Estigarribia Barreto, «Unificacion dataset MEC v6.ipynb» Cuaderno de Jupyter, octubre 2024. [En línea]. Available: https://github.com/luisg1337/Unificacion-Dataset-MEC/.