Análisis de la Oferta de Programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia mediante Técnicas de Aprendizaje No Supervisado

Integrantes

Jazmin Elena Pesca Muñoz

Yudi Elena Ruiz Hernandez

Ejecutores

Natalia Betancur Herrera

Frank Yesid Zapata Castaño

Margarita Maria Orozco

Universidad de Antioquia, Universidad de Caldas

Talento TECH

BOOTCAMP Inteligencia Artificial

marzo 2025

ÍNDICE

INTRODUCCIÓN

En el contexto de la educación para el trabajo y el desarrollo humano (ETDH) en Colombia, es fundamental comprender la diversidad, calidad y distribución geográfica de los programas ofrecidos. La creciente necesidad de mejorar la pertinencia educativa y fortalecer la toma de decisiones en este ámbito exige un enfoque basado en datos y técnicas analíticas avanzadas.

Este estudio propone el uso de modelos de aprendizaje no supervisado para identificar patrones de agrupación natural en los programas de Educación para el Trabajo y el Desarrollo Humano (ETDH). La metodología empleada incluye herramientas como el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y visualizar tendencias, así como algoritmos de clustering, como K-Means, para agrupar los programas en función de sus similitudes. A través de estas técnicas, se busca analizar la distribución geográfica de los programas, identificar sus características comunes y determinar los factores que pueden influir en su calidad y diversidad.

La identificación de estas agrupaciones permitirá a los responsables de la formulación de políticas educativas y a las instituciones de formación contar con información valiosa para la toma de decisiones estratégicas. Esto contribuirá al diseño de estrategias de mejora continua en la educación para el trabajo, garantizando que los programas sean más accesibles, relevantes y alineados con las necesidades del mercado laboral y la sociedad en general.

Se espera que los hallazgos contribuyan al entendimiento en términos de diversidad, calidad y distribución geográfica de los programas de Educación para el Trabajo y el

Desarrollo Humano (ETDH), permitiendo así la identificación de oportunidades de mejora y la implementación de estrategias para una oferta educativa más equitativa y alineada con las necesidades del mercado laboral.

PLANTEAMIENTO DEL PROBLEMA

La Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia desempeña un papel crucial en la formación de competencias técnicas y ocupacionales, facilitando la inserción laboral y contribuyendo al desarrollo económico del país. Sin embargo, a pesar de su importancia, la oferta de programas de ETDH presenta **desafíos significativos en términos de diversidad, calidad y distribución geográfica, lo que dificulta su alineación con las necesidades del mercado laboral y de la población que busca acceder a estos programas.**

Uno de los principales problemas radica en la falta de un análisis sistemático de los programas ofrecidos, lo que impide identificar patrones comunes que permitan evaluar su pertinencia y calidad (Pontificia Universidad Javeriana, 2022). A pesar de la existencia del Sistema de Información de la Educación para el Trabajo y el Desarrollo Humano (SIET), la información almacenada en esta plataforma no ha sido ampliamente utilizada para extraer conocimientos que faciliten la toma de decisiones en el ámbito educativo y laboral. La ausencia de herramientas analíticas avanzadas para procesar estos datos limita la capacidad de las instituciones educativas para planear estrategias que optimicen la oferta de ETDH (Datacenter Market, 2024).

Por otro lado, la distribución geográfica de estos programas no siempre responde a las necesidades específicas de las regiones, lo que genera inequidades en el acceso a la

formación para el trabajo. Algunas zonas cuentan con una oferta educativa abundante, mientras que otras carecen de programas adecuados, lo que impacta negativamente en la empleabilidad de sus habitantes y en el desarrollo regional. Así mismo, la falta de mecanismos para evaluar la calidad y relevancia de los programas impide garantizar que los egresados adquieran las competencias demandadas por el sector productivo (UNIR Ecuador, 2024).

En este contexto, surge la necesidad de aplicar técnicas de aprendizaje no supervisado para analizar la oferta de ETDH en Colombia. Mediante la identificación de patrones en los datos registrados en el SIET, se puede lograr una comprensión más profunda de la diversidad, calidad y distribución de estos programas. Este enfoque permitiría mejorar la toma de decisiones basada en evidencia, optimizar la asignación de recursos y diseñar estrategias que potencien el impacto de la ETDH en el país.

Por lo tanto, este estudio busca abordar la falta de un análisis estructurado de la oferta de programas de ETDH mediante la aplicación de metodologías de análisis de datos avanzadas. Con ello, se pretende contribuir a la formulación de políticas educativas más efectivas y a la mejora continua de la educación para el trabajo en Colombia, asegurando su alineación con las dinámicas del mercado laboral y las necesidades de la población.

Pregunta de investigación:

¿Existen patrones naturales de agrupación entre los programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia, basados en variables que permitan comprender su diversidad, calidad y distribución geográfica?

OBJETIVOS

Objetivo General

Analizar la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia, a partir de la identificación de patrones comunes mediante técnicas de aprendizaje no supervisado, con el fin de entender la diversidad, calidad y distribución geográfica de estos programas.

Objetivos Específicos

- Realizar una revisión exhaustiva de la literatura y bases de datos existentes sobre la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia, para identificar variables más relevantes para el estudio.
- Analizar datos sobre la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia, utilizando una metodología de aprendizaje no supervisado.
- Identificar patrones de diversidad, calidad y distribución geográfica a través de un modelo de aprendizaje no supervisado.

JUSTIFICACIÓN

En el contexto actual, donde el mercado laboral exige una actualización constante de competencias y habilidades, la Educación para el Trabajo y el Desarrollo Humano (ETDH) se consolida como un pilar fundamental para reducir la brecha entre la formación académica y las necesidades del sector productivo (Secretaría de Educación del Distrito, 2019).

Regulada por el Decreto 1075 de 2015, la ETDH forma parte del servicio público educativo en Colombia y tiene como propósito complementar, actualizar o suplir conocimientos, enfocándose en la formación académica y laboral para la obtención de certificados de aptitud ocupacional.

Su enfoque en la formación práctica y el desarrollo de competencias laborales específicas permite a los individuos mejorar su empleabilidad e integrarse de manera efectiva al mercado laboral. De acuerdo con el artículo 2.6.2.2 del Decreto 1075 de 2015, la ETDH no solo desarrolla habilidades técnicas y ocupacionales, sino que también promueve el crecimiento personal y profesional, contribuyendo así al desarrollo integral de la población.

Es fundamental diferenciar la educación formal de la no formal para comprender la relevancia de la ETDH. Mientras la educación formal es impartida en instituciones oficiales, sigue un currículo estructurado y otorga certificaciones con reconocimiento oficial, la educación no formal se orienta al desarrollo de habilidades específicas sin necesidad de una certificación oficial. La educación formal brinda mayor estructura y mejores oportunidades laborales, mientras que la educación no formal ofrece flexibilidad, accesibilidad y fomenta el desarrollo de habilidades sociales y profesionales.

Para garantizar la calidad de la oferta educativa en la ETDH, el Sistema de Información de la Educación para el Trabajo y el Desarrollo Humano (SIET) fue implementado en 2010 como una herramienta informática que permite a las Secretarías de Educación de las Entidades Territoriales Certificadas (SEC) registrar y gestionar información clave sobre licencias de funcionamiento, programas académicos, matrículas y costos. Administrado por el Ministerio de Educación Nacional, el SIET tiene como objetivos principales:

- Informar a la comunidad sobre las instituciones y programas de ETDH, garantizando la transparencia y calidad de la oferta educativa.

- Servir como insumo para la formulación de políticas educativas a nivel nacional y territorial, facilitando la planeación, monitoreo, evaluación e inspección del sector (Secretaría de Educación del Distrito, 2019).

El SIET asegura que la información sobre las Instituciones de Educación para el Trabajo y el Desarrollo Humano (IETDH) sea completa, veraz y actualizada, promoviendo una gestión más eficiente de estos programas.

Este proyecto podría ayudar a comprender la oferta de programas de ETDH en Colombia mediante técnicas avanzadas de análisis de datos. La identificación de patrones y tendencias puede contribuir a optimizar la diversidad, la calidad y distribución geográfica de estos programas, facilitando la formulación de políticas educativas alineadas con las necesidades del mercado laboral y el desarrollo humano. Al basarse en datos del SIET, se garantiza la confiabilidad y actualidad de la información, reforzando la validez y aplicabilidad de los hallazgos.

ALCANCE

El modelo de análisis propuesto tiene como objetivo explorar y comprender la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia, utilizando técnicas de aprendizaje no supervisado para identificar patrones comunes y características relevantes. A partir de los objetivos generales y específicos planteados, el alcance del modelo se define de la siguiente manera:

Cobertura Geográfica: El modelo analizará la oferta de programas ETDH en todo el territorio colombiano, permitiendo identificar diferencias y similitudes entre departamentos

y municipios, así como la distribución geográfica de la calidad y diversidad de estos programas.

Variables de Estudio: Se tomarán en cuenta variables clave como el nombre de la institución, el nombre del programa, el departamento, el municipio, la localidad, la sede, el estado del programa, el área de desempeño, el área de desempeño en salud, el tipo y subtipo de certificado, la escolaridad requerida, la jornada, el número y tipo de certificación, la calidad del certificado, el estado de la certificación y la entidad emisora de la certificación.

Resultados Esperados: El modelo generará visualizaciones claras y efectivas que permitan identificar tendencias, agrupaciones y diferencias entre programas, brindando una base sólida para entender la diversidad, calidad y distribución geográfica para el fortalecimiento y diversificación de la oferta educativa en el ámbito ETDH.

Limitaciones: El modelo estará limitado a la calidad y disponibilidad de los datos en la base de datos creada el 21 de septiembre de 2020 y actualizada el 14 de enero de 2025. Los resultados dependen de la actualización, precisión y completitud de la información registrada, lo que puede influir en la identificación de patrones y en la representatividad de las conclusiones obtenidas.

Aplicabilidad: Los resultados del modelo podrán ser utilizados por entidades u organizaciones interesadas en el desarrollo de programas ETDH, con el fin de mejorar la calidad, pertinencia y accesibilidad de la oferta educativa en el país.

METODOLOGÍA

Se emplearán técnicas de aprendizaje no supervisado, como el Análisis de Componentes Principales (PCA), que reduce la dimensionalidad de los datos conservando la mayor variabilidad posible, y algoritmos de clustering, como K-Means, para agrupar los datos en clusters basados en similitudes, o Estas herramientas facilitarán la interpretación de los patrones presentes en la oferta de programas.

Descripción base de datos

Base de Datos Utilizada: La base de datos empleada para el análisis fue creada el 21 de septiembre de 2020 y constituye la principal fuente de información para el desarrollo del modelo. Esta base de datos, proporcionada por el Ministerio de Educación Nacional y actualizada por última vez el 14 de enero de 2025, asegura una visión representativa, actualizada y estructurada de la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia.

Variables numéricas se encuentran: Código Secretaria, Código Institución, Código Programa, Cod Departamento, Cod Municipio, Registro, Costo.

Variables categóricas incluyen: Secretaria, Nombre Institución, Nombre Programa, Departamento, Municipio, Localidad, Dirección, Sede, Estado Programa, Fecha Registro, Área Desempeño, Área Desempeño Salud, Tipo Certificado, Subtipo Certificado, Escolaridad, Jornadas, Número Certificación, Tipo Certificación, Certificado Calidad, Estado Certificación y Entidad Emisora Certificación.

Esta combinación de variables proporciona una base robusta para identificar patrones y tendencias en la oferta educativa de ETDH en Colombia.

Entendimiento de los Datos

La fase de **Entendimiento de los Datos** se enfoca en la selección de los datos, exploración y validación de la información necesaria para alimentar el modelo no supervisado. Basándonos en la descripción de la base de datos ETDH, se establecen las **variables clave** y variables irrelevantes las cuales se descartan. A continuación, se describe el contenido y propósito de cada grupo de variables:

Descripción de las Variables (Tabla No 1)

VARIABLE	DESCRIPCIÓN DE LA VARIABLE	CATEGORÍA	IDENTIFICADOR	IDENTIFICADOR
Secretaria:	Nombre de la Secretaría de Educación territorial.	Categórica	Nominal	1
Nombre Institución:	Nombre oficial de la institución educativa.	Categórica	Nominal	1
Nombre Programa:	Nombre del programa de formación.	Categórica	Nominal	1
Departamento:	Nombre del departamento donde está ubicada la institución.	Categórica	Nominal	1
Municipio	Nombre del municipio donde está ubicada la institución.	Categórica	Nominal	1
Localidad:	Zona específica dentro del municipio, si aplica.	Categórica	Nominal	1
Dirección:	Ubicación física de la institución o sede.	Categórica	Nominal	1
Sede:	Sede específica de la institución donde se imparte el programa.	Categórica	Nominal	1
Estado Programa:	Indica si el programa está activo, inactivo, suspendido, etc.	Categórica	Nominal	1

Fecha Registro:	Fecha de registro del programa en el MEN. (Ministerio de Educación Nacional de Colombia)	Categórica	Nominal	1
Área Desempeño:	Sector laboral al que está orientado el programa.	Categórica	Nominal	1
Área Desempeño Salud:	Si aplica, especifica si está relacionado con el área de la salud.	Categórica	Nominal	1
Tipo Certificado:	Clase de certificación obtenida al finalizar el programa.	Categórica	Nominal	1
Subtipo Certificado:	El programa al que pertenece el tipo de certificado.	Categórica	Nominal	1
Escolaridad:	Nivel educativo requerido para ingresar al programa.	Categórica	Ordinal	1
Jornadas:	Horarios en los que se ofrece el programa (diurna, nocturna, mixta).	Categórica	Ordinal	1
Número Certificación:	(SI o NO) certifica que la institución cumple con todos los requisitos legales, administrativos y de calidad establecidos.	Categórica	Nominal	1
Tipo Certificación:	Acreditación oficial que tiene la institución para ofrecer programas de educación para el trabajo y el desarrollo humano.	Categórica	Nominal	1
Certificado Calidad:	Indica si el programa cuenta con algún tipo de certificación de calidad. (NTC)	Categórica	Nominal	1
Estado Certificación:	Vigencia o estado del certificado de calidad.	Categórica	Nominal	1
Entidad Emisora Certificación:	Organización que otorga la certificación de calidad.	Categórica	Nominal	1
Código Secretaria:	Identificador único de la Secretaría de Educación correspondiente.	Númerica	Nominal	2
Código Institución:	Código asignado a la institución educativa.	Númerica	Nominal	2
Código Programa:	Identificador único del programa educativo	Númerica	Nominal	2

	registrado.			
Cod Departamento:	Código del departamento- (Anexo)	Númerica	Nominal	2
Cod Municipio:	Código municipio	Númerica	Nominal	2
Registro :	El número de registro es el código único que se le asigna a cada programa educativo aprobado por la Secretaría de Educación.	Númerica	Nominal	2
Costo:	Valor económico del programa.	Númerica	Nominal	0
Duración Horas:	Total de horas del programa de formación.	Númerica	Nominal	2
Fecha Otorgamiento:	Se refiere al día, mes, año y hora en que la institución educativa recibió la certificación oficial por parte de la Secretaría de Educación o el ente regulador.	Numérica	Ordinal	2
Fecha Vencimiento:	Día, mes, año y hora en que expira la validez de la certificación otorgada a la institución educativa o a uno de sus programas	Númerica	Ordinal	0
Latitud / Longitud:	Coordenadas geográficas de la institución o sede.	Númerica	Nominal	0
Año Corte / Mes Corte / Fecha Corte:	Fecha de actualización de los datos.	Númerica	Ordinal	0

La categorización se hizo tomando en cuenta lo siguiente:

0	No se tiene en cuenta
1	No Ordinal o Nominal
2	Ordinal

Validación Inicial de los Datos

- **Análisis exploratorio**: La visualización de las relaciones entre variables a través de gráficos se llevó a cabo mediante un Análisis Exploratorio de Datos (EDA). Este proceso preliminar es esencial cuando se trabaja con un conjunto de datos, ya que permite examinar su estructura, detectar patrones, identificar anomalías y reconocer vínculos entre las variables. El propósito principal del EDA es obtener una comprensión profunda de la información, proporcionando una base sólida para la aplicación de modelos y el desarrollo de análisis más avanzados.

  Este análisis cumple varios propósitos: comprender la estructura de los datos, examinando su tamaño, el tipo de variables (como numéricas y categóricas) y sus características generales; identificar patrones y relaciones, explorando correlaciones, tendencias o comportamientos específicos; detectar valores atípicos o errores, como registros inconsistentes o fallos en la captura de información que puedan influir en los resultados; evaluar la distribución de las variables, analizando si los datos presentan sesgos, valores extremos o se ajustan a una distribución normal; y finalmente, preparar la información para su limpieza, permitiendo reconocer valores nulos, duplicados o inconsistencias, a fin de obtener un conjunto de datos más organizado y funcional.

Herramientas para EDA

En programación, las siguientes librerías fueron utilizadas para realizar un EDA de manera rápida y visual.

Descripción de librerías y su propósito en el contexto del análisis de la base de datos ETDH:

- **pandas (pd):** Es una librería esencial en Python para la manipulación y el análisis de datos. Permite trabajar con estructuras de datos como DataFrames y Series, facilitando la limpieza, transformación y exploración de grandes volúmenes de información tabular.
- **numpy (np):** Proporciona soporte para trabajar con arreglos multidimensionales y funciones matemáticas avanzadas. Es fundamental para realizar cálculos numéricos eficientes y operaciones vectorizadas.
- **seaborn (sns):** Una librería de visualización basada en Matplotlib, especializada en la creación de gráficos estadísticos atractivos y fáciles de interpretar. Ayuda a identificar patrones, tendencias y relaciones entre variables.
- **matplotlib.pyplot (plt)**: Es una de las bibliotecas más utilizadas para crear gráficos en Python. Proporciona una amplia gama de visualizaciones, como histogramas, gráficos de dispersión y líneas, esenciales para el análisis exploratorio de datos.
- **h2o:** Una plataforma de machine learning de código abierto que permite la construcción y validación de modelos avanzados de manera eficiente y escalable. Se destaca por su capacidad de trabajar con grandes volúmenes de datos.
- **H2OKMeansEstimator:** Es un estimador de la librería H2O que implementa el algoritmo de clustering K-Means, utilizado para agrupar observaciones en clústeres basados en similitudes entre sus características.
- **sklearn.decomposition.PCA:** El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en un conjunto más pequeño de variables no correlacionadas, llamadas componentes principales.
- **sklearn.metrics.silhouette_score:** Esta métrica evalúa la calidad de un clustering midiendo qué tan cerca están los puntos dentro de un mismo clúster y qué tan

separados están de los otros clústeres. Ayuda a determinar el número óptimo de clústeres.

- **rdt.HyperTransformer:** Forma parte de la librería rdt (Real Data Transformers) y se utiliza para transformar y codificar datos, facilitando la preparación de información antes de alimentar modelos de machine learning.
- **Sweetviz:** Su principal función es generar informes visuales detallados y automatizados que permiten entender rápidamente la estructura y las características de un conjunto de datos.

- **Limpieza de datos**:

  Se realizó mediante el uso de la función **dropna** (de la librería **Pandas**); que permite eliminar filas o columnas en un DataFrame que contienen valores nulos o faltantes (NaN). Es especialmente útil en el proceso de limpieza de datos, donde a menudo es necesario deshacerse de datos incompletos para garantizar la calidad del análisis o el modelado.

Funcionalidades principales de dropna:

Eliminación de variables:

- **Costos:** Esta variable fue eliminada debido a que generaba una pérdida significativa de información en la base de datos, lo cual dificulta la correcta realización de la agrupación de los programas ETDH.
- **Otras variables eliminadas:** También se eliminaron las siguientes variables, ya que no aportan información esencial para el análisis de patrones y generaban inconsistencias en la calidad de los datos:

- Tipo de certificación
- Localidad
- Área de desempeño en salud
- Certificado de calidad
- Entidad emisora de certificación
- Estado de certificación
- Fecha de otorgamiento
- Fecha de vencimiento
- Latitud
- Longitud

Para iniciar el proceso de modelado, se eliminaron las columnas que presentaban un porcentaje de valores vacíos igual o superior al 20%, almacenando el resultado en una nueva variable llamada df_trabajo_y_desarrollo. Esta limpieza es esencial para garantizar la calidad de los datos utilizados en el análisis.

Usando la función dropna() de Pandas, que sirve para eliminar filas con valores faltantes (nulos) en el DataFrame df_trabajo_y_desarrollo. Básicamente, está filtrando y dejando solo los registros completos, es decir, aquellos donde no falta información en ninguna columna.

Después de aplicar esta limpieza y depuración, el número total de registros en la base de datos pasó de 19.867 a 16.132.

- **Codificación**:

  Se llevó a cabo el proceso de transformación y codificación de datos, preparando la información de manera adecuada para su uso en modelos de machine learning.

Para esta tarea, se utilizó el **HyperTransformer** de la librería **rdt (Real Data Transformers)**, una herramienta especializada en convertir variables categóricas y numéricas en formatos estandarizados y optimizados. Este proceso asegura una correcta interpretación y procesamiento de los datos, facilitando la construcción de modelos más eficientes y precisos.

Esta técnica facilita la estandarización de la base de datos y asegura una representación precisa y coherente de las distintas características, optimizando así el rendimiento y la interpretación de los modelos no supervisados.

El proceso de transformación y codificación de datos ofrece varias ventajas en el contexto de machine learning. Estas incluyen la **optimización del rendimiento del modelo**, ya que al transformar y codificar los datos adecuadamente, se mejora la eficiencia y precisión de los algoritmos. Además, permite la **estandarización de la información**, uniformando el formato de las variables y facilitando su interpretación y análisis. También se logra un **manejo eficiente de variables categóricas**, convirtiendo categorías en representaciones numéricas sin perder información relevante. Esto **reduce las inconsistencias**, disminuyendo el riesgo de errores asociados a datos mal estructurados. Finalmente, los modelos construidos sobre datos transformados y codificados tienden a tener una **mejor capacidad de generalización** en diferentes conjuntos de datos**.**

MODELADO

Después, usa shape para mostrar el tamaño del DataFrame limpio (cantidad de filas y columnas) y head() para ver las primeras filas del DataFrame ya depurado. Este paso es clave para el proyecto, ya que asegura que el análisis y la agrupación se realicen sobre información completa y precisa, evitando sesgos o errores debidos a datos faltantes; quedando 26 columnas de 36 que se venían manejando.

Preparación del Modelo

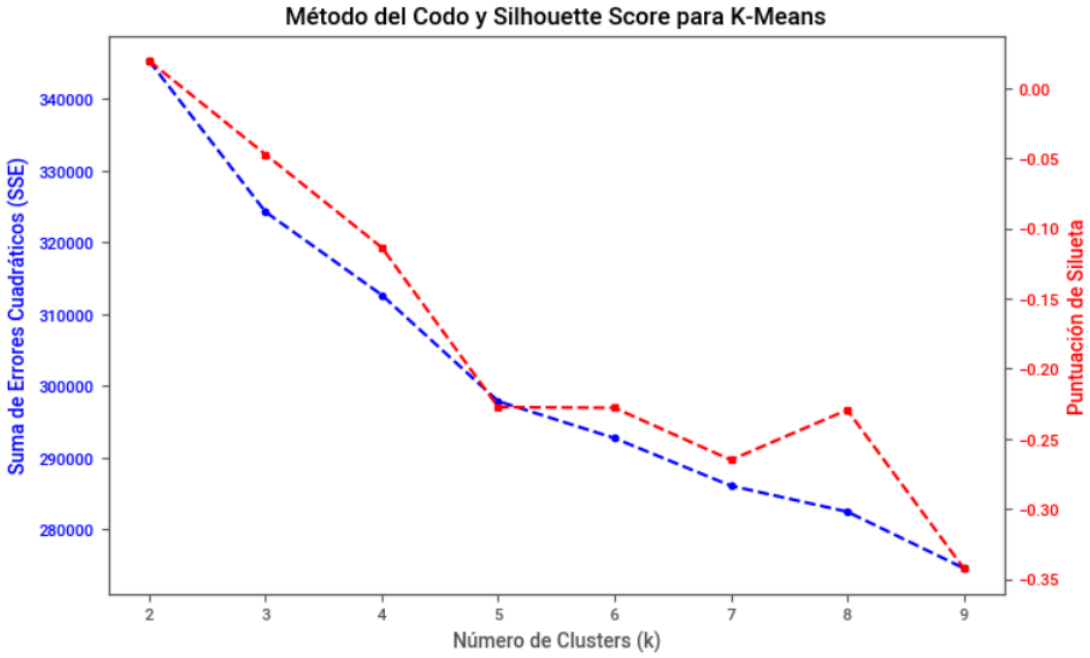
Este conjunto de instrucciones implementa un proceso completo de modelado y análisis de clustering usando técnicas de aprendizaje no supervisado, aplicadas a la base de datos df_trabajo_y_desarrollo_cleaned. A continuación, se describe cada paso y su propósito:

- **Inicialización de H2O:** Se inicia el entorno de H2O, una plataforma eficiente para el aprendizaje automático distribuido, permitiendo el manejo de grandes volúmenes de datos.
- **Carga y preparación de datos:**
    - Se crea una copia del DataFrame original df_trabajo_y_desarrollo_cleaned.
    - Se eliminan filas con valores nulos usando dropna().
    - Se descarta la columna Costo al no ser relevante para el análisis.
- **Transformación de datos con HyperTransformer:**
    - Se instancia el objeto HyperTransformer() de la librería rdt.
    - Se detecta la configuración inicial de las variables del DataFrame.

    - Se ajusta el transformador con el método fit() y se transforman los datos con transform(), convirtiéndolos en un formato numérico y estructurado para facilitar el análisis.
    - Se reconstruye un nuevo DataFrame df_transformed con las columnas originales transformadas.
- **Conversión a H2OFrame:** Se convierte el DataFrame transformado en un H2OFrame, formato requerido para ejecutar modelos en H2O.
- **Reducción de dimensiones con PCA:**
    - Se aplica el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos a dos componentes principales.
    - Esto facilita la visualización y el procesamiento eficiente de los clusters.
- **Definición y entrenamiento de K-Means:**
    - Se define un rango de valores de k (número de clusters) entre 2 y 9.
    - Para cada k, se entrena un modelo K-Means usando el estimador H2OKMeansEstimator.
    - Se obtienen las etiquetas de cluster asignadas y se calcula la Suma de Errores Cuadráticos (SSE) y el índice de Silhouette para evaluar la calidad del agrupamiento.
    - Los resultados de clustering se almacenan en el diccionario cluster_results.
- **Evaluación de resultados:**
    - Se grafican el Método del Codo (SSE) y el Silhouette Score, permitiendo identificar el número óptimo de clusters.
- **Visualización de clusters:**
    - Se selecciona el valor de k con el mejor Silhouette Score.
    - Se visualizan los clusters proyectados en dos dimensiones usando los componentes principales obtenidos con PCA.

Este proceso asegura una correcta transformación, reducción y agrupación de los datos, permitiendo identificar patrones significativos en la oferta de programas de Educación para el Trabajo y el Desarrollo Humano (ETDH) en Colombia.

GRÁFICA ANEXO 2 Gráfica del Método del Codo y Silhouette Score



La gráfica combina dos métricas para evaluar el número óptimo de clusters en un modelo de **K-Means**:

1. **La curva azul (SSE - Suma de Errores Cuadráticos)**:
    - **Eje Y izquierdo (SSE)**: Mide la distancia entre cada punto y el centroide de su cluster. Un valor más bajo indica clusters más compactos.

- La gráfica muestra cómo el SSE disminuye a medida que aumenta el número de clusters (k).
- El “**método del codo**” se usa aquí: se elige el valor de k donde la disminución del SSE se vuelve menos pronunciada, formando una especie de codo en la curva. Aquí parece estar alrededor de **k = 5**.

La curva roja (Puntuación de Silueta):

- **Eje Y derecho (Puntuación de Silueta)**: Mide qué tan bien separados están los clusters y qué tan similares son los puntos dentro de un mismo cluster.
- Esta puntuación varía entre -1 y 1, donde valores cercanos a 1 indican clusters bien definidos.
- En esta gráfica, la puntuación es negativa, lo que indica que los clusters podrían no estar muy bien separados.

Interpretación:

- Con **k = 5**, parece haber un buen balance: el SSE es relativamente bajo y la caída en el error comienza a estabilizarse, formando el "codo".
- Sin embargo, la puntuación de silueta es negativa en todos los casos, lo que sugiere que el modelo podría no estar separando bien los clusters. Quizás se necesite ajustar el preprocesamiento de los datos.

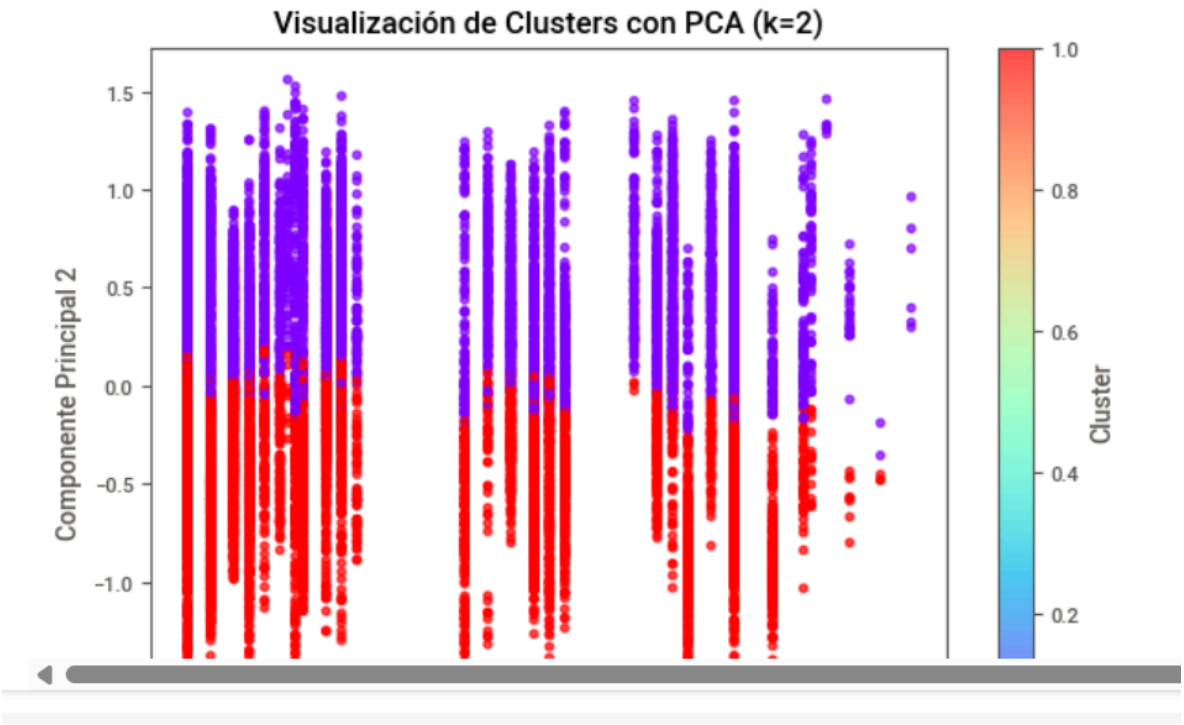
La gráfica muestra dos métricas clave para evaluar el rendimiento del modelo K-Means: la Suma de Errores Cuadráticos (SSE) en azul y el Silhouette Score en rojo, ambas en función del número de clusters (k).

- **SSE (Método del Codo):** La línea azul indica cómo disminuye el SSE a medida que aumenta el número de clusters. El "codo" de esta curva señala el punto donde añadir más clusters deja de reducir significativamente el error, sugiriendo el número óptimo de agrupaciones. En este caso, parece estar cerca de k=5, donde la curva empieza a estabilizarse.

- **Silhouette Score**: La línea roja evalúa la cohesión y separación de los clusters. Un valor más cercano a 1 indica clusters bien definidos, mientras que valores negativos sugieren mala separación entre ellos. Aquí, los valores de Silhouette Score son bajos o negativos, lo que podría indicar dificultades para encontrar agrupaciones claras en los datos.

Este modelo ayuda a descubrir similitudes entre programas, como tendencias en la oferta educativa, calidad certificada o áreas de mayor cobertura; los resultados de K-Means son el punto de partida para aplicar técnicas avanzadas como PCA, análisis de silueta o modelos jerárquicos.

Gráfica de Visualización de Clusters con PCA (k=2)



Esta gráfica muestra la visualización de clusters obtenidos con el modelo K-Means, reducido a dos dimensiones usando PCA, con **k=2** (dos clusters). Cada punto representa una observación proyectada en el espacio de las dos componentes principales.

Análisis:

- **Separación de clusters:** Los dos clusters (rojo y morado) se diferencian, pero la superposición es notable. Esto indica que la división entre grupos no es completamente clara, lo cual podría explicar los bajos valores del Silhouette Score observados antes.
- **Distribución:** Los puntos parecen organizarse en columnas, lo que sugiere una estructura en las variables originales que probablemente proviene de la transformación de datos.

- **Posible falta de diferenciación:** La mezcla de colores en la parte central sugiere que los clusters comparten características similares, lo que podría ser señal de que las diferencias entre ellos no están bien captadas en este espacio reducido.

🔹 **Ejes**: Representan las **dos primeras componentes principales (PCA)**, que son combinaciones de las variables originales y permiten visualizar los datos en 2D.

🔹 **Colores**: Representan los clusters encontrados. Rojo y morado indican los dos grupos identificados por el algoritmo.

INTERPRETACIÓN DE RESULTADOS

Diversidad:

Nombre Programa: ofrece un análisis sobre la distribución de programas técnicos laborales y su diversidad en dos conjuntos de datos.

- **Programas más comunes:**

  Dentro de la base de datos, los programas más frecuentes muestran una amplia diversidad de oferta, sin una tendencia dominante. Entre los más comunes se encuentran:

  - Técnico Laboral en Auxiliar Administrativo
  - Técnico Laboral en Auxiliar en Enfermería
  - Técnico Laboral en Auxiliar Contable y Financiero
  - Técnico Laboral en Seguridad Ocupacional

- Cada uno de estos programas tiene una participación entre el 1% y el 2% del total, lo que muestra que, aunque hay una amplia variedad de programas, algunos se

repiten con mayor frecuencia.

- **Categoría "Other":**

  La mayoría de los registros (90%) están agrupados en la categoría "Other", lo que sugiere que la mayoría de los programas tienen una frecuencia muy baja y no aparecen de forma predominante en el análisis.

Áreas de Desempeño

  - Ventas y Servicios: 35%
  - Finanzas y Administración: 25%

El análisis muestra una concentración en los sectores laborales a los que están orientados los programas, especialmente en Ventas y Servicios y Finanzas y Administración. Esto podría indicar una mayor oferta de programas en estos campos o una alta demanda laboral. En total, estos sectores representan cerca del 55% de la información analizada.

Escolaridad

El nivel educativo requerido para ingresar al programa muestra la siguiente distribución:

- Secundaria: Es el nivel predominante, con cerca del 60% de los registros.
- Media: Representa aproximadamente el 25%.
- No Aplica y Primaria: Tienen una participación mínima en comparación con los niveles superiores.

Estos datos sugieren que la mayoría de los programas están dirigidos a personas con formación secundaria, lo que puede indicar una orientación específica de la oferta educativa.

Jornada

El análisis de escolaridad indica que la mayoría de los participantes en los programas de ETDH cuentan con formación en **Secundaria**, representando cerca del 60% del total. Le siguen aquellos con formación en **Media**, con un porcentaje significativo. En menor medida, se encuentran personas con niveles educativos como **Primaria o** casos donde la escolaridad no aplica.

Este hallazgo sugiere que la mayoría de los programas están dirigidos a personas que han completado la educación secundaria, lo que podría indicar una brecha de acceso o otro tipo de población.

Calidad

Estado de Certificación: refleja tanto la cantidad de valores presentes como los faltantes:

- Valores presentes:
    - **Grupo 1**: 706 valores (9%)
    - **Grupo 2**: 1,382 valores (18%)
- Valores faltantes (missing):
    - **Grupo 1**: 7,585 valores faltantes (91%)
    - **Grupo 2**: 6,460 valores faltantes (82%)

Esto indica que en ambos grupos hay una cantidad muy alta de datos faltantes, especialmente en el primer grupo. Esto puede ser un problema para el análisis, ya que más del 80% de la información no está disponible.

- Valores distintos: Ambos grupos tienen 4 valores únicos, que se muestran en el gráfico de barras:
    - Renovado: Es el estado más frecuente.
    - Vencido: También tiene una presencia considerable.
    - Primera vez: Menos frecuente, pero aún relevante.
    - Cancelada: Es el estado menos frecuente en ambos grupos.

El gráfico de barras muestra que la distribución de estados de certificación es bastante similar entre los dos grupos, aunque en algunos estados como "Renovado" y "Vencido" el segundo grupo tiene una mayor proporción.

Distribución geográfica

Departamento

El análisis de la distribución geográfica muestra una mayor concentración de programas en ciertos departamentos:

- **Bogotá D.C., Antioquia, Valle del Cauca y Cundinamarca**: Presentan la mayor cantidad de registros, consolidándose como los principales centros de oferta educativa en programas de ETDH.
- **Atlántico**: Se observa un incremento en el pico en el grupo naranja (grupo 2) la participación en comparación con otros departamentos intermedios, lo que sugiere un crecimiento en la oferta educativa en esta región.

Municipio

El análisis a nivel municipal destaca la concentración de programas en algunas ciudades principales:

- **Bogotá**: Representa el 8% de los programas en un conjunto y el 10% en otro.
- **Medellín**: Participa con el 4% y el 5% respectivamente.
- **Cali**: Aporta el 4% y 5%.
- **Buenaventura**, **Barrancabermeja**, **Villavicencio** y **Cúcuta**: Tienen participaciones menores entre el 2% y 3%.
- **Otros municipios**: Concentran el 75% y el 71% del total, lo que refleja una amplia dispersión de programas en localidades menos representadas.

Este patrón subraya la importancia de considerar una descentralización más enfocada y un análisis de necesidades locales para balancear la oferta educativa.

La visualización de clusters muestra una separación moderada entre los grupos, aunque se observa cierta superposición. Esto indica que la diferenciación entre los clusters no es completamente clara, posiblemente debido a la similitud entre las características de los programas analizados.

CONCLUSIÓN

- **Uso de Clustering en el Análisis**: Se decide trabajar sobre dos clusters en lugar de cinco, como sugería la gráfica de codo, ya que con cinco clusters los datos no son representativos ni se segmentan adecuadamente.
- **Posible falta de diferenciación:** La mezcla de colores en la parte central sugiere que los clusters comparten características similares, lo que podría ser señal de que las diferencias entre ellos no están bien captadas en este espacio reducido.
- **Oportunidades Empresariales:** A partir de este análisis, se identifican oportunidades para la creación de empresas enfocadas en la oferta de programas educativos alineados con las necesidades del mercado laboral y las brechas detectadas.
- **Estrategia para la Actualización Profesional en la Oferta Educativa:** Se sugiere diseñar programas de actualización y especialización dirigidos a profesionales, con enfoques flexibles como cursos cortos, certificaciones y modalidades **virtuales o nocturnas**. Esto permitiría ampliar el acceso a personas que buscan actualizar sus conocimientos sin comprometer su jornada laboral.

ANEXOS

Análisis de Eda

Códigos

BBDD

Video elevator pitch

BIBLIOGRAFÍA

Ministerio de Educación Nacional. (2025). MEN_PROGRAMAS_EDUCACIÓN PARA EL TRABAJO Y EL DESARROLLO HUMANO. Datos Abiertos Colombia. Recuperado el 8 de marzo de 2025, de https://www.datos.gov.co/Educaci-n/MEN_PROGRAMAS-EDUCACI-N-PARA-EL-TRABAJO-Y-EL-DESAR/2v94-3ypi/about_data

Secretaría de Educación del Distrito. (2019, febrero 11). Educación para el Trabajo y el Desarrollo Humano (ETDH). Recuperado el 15 de marzo de 2025. de

https://www.educacionbogota.edu.co/portal_institucional/gestion-educativa/educacion-trabajo-desarollo-humano

Presidencia de la República de Colombia. (2015, mayo 26). Decreto 1075 de 2015: Por medio del cual se expide el Decreto Único Reglamentario del Sector Educación. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=77913

Pontificia Universidad Javeriana. (2022, 5 de septiembre). Información sobre Educación para el Trabajo y el Desarrollo Humano (ETDH) [PDF]. Portal Pontificia Universidad Javeriana. Recuperado el 8 marzo de 2025 de https://www.javeriana.edu.co/recursosdb/Info…

Datacenter Market. (2024, 4 de noviembre). Las 7 mejores herramientas de análisis de datos y cómo elegir. Recuperado el 10 marzo de 2025 de https://www.datacentermarket.es/tendencias-ti/como-elegir-la-mejor-plataforma-de-analisis-de-datos-corporativa/

UNIR Ecuador. (2024). Conoce la importancia de la evaluación de programas educativos para mejorar la efectividad de la formación académica y las habilidades necesarias para lograrlo. Recuperado el 12 de marzo de 2025 de https://ecuador.unir.net/actualidad-unir/evaluacion-programas-educativos/