Pontificia Universidad Javeriana

Primera entrega de proyecto

Angel Eduardo Morales Abril
Andrés David Pérez Cely

Analitica de datos

Profesor: Ferney Maldonado Lopez

Septiembre, 2025

1. Introducción

1.1 Definición del problema

El programa Talento Tech Bogotá, desarrollado por Tecnalia en alianza con la Secretaría de Desarrollo Económico, busca cerrar brechas en competencias digitales mediante la formación de ciudadanos en áreas como análisis de datos, desarrollo web, inteligencia artificial y ciberseguridad. Durante los últimos años, la iniciativa ha logrado convocar a un número amplio y diverso de participantes, entre los cuales se incluyen personas en condición de vulnerabilidad social, con distintos niveles educativos y trayectorias formativas.

A pesar de contar con una base de datos detallada que recoge variables sociodemográficas (género, estrato socioeconómico, condición de víctima del conflicto, discapacidad, condición de campesino, entre otras), así como información educativa y de participación (temática elegida, horas de asistencia, estado de matrícula, puntajes y niveles alcanzados), no existe claridad sobre cómo estos factores se relacionan entre sí ni cuál es su influencia en los resultados del programa.

En particular, se desconoce qué variables determinan la elección de la temática de estudio y si esta responde a motivaciones individuales, condiciones sociales o limitaciones de acceso. También se desconoce qué características están vinculadas con la permanencia y asistencia de los beneficiarios, lo que dificulta anticipar riesgos de deserción o baja participación. Además, no se sabe qué factores explican el desempeño formativo y si existen patrones diferenciados entre grupos poblacionales, lo cual puede incidir en la efectividad de la formación recibida.

Esta falta de comprensión constituye un problema porque impide a los gestores del programa, diseñar convocatorias más focalizadas y pertinentes, ajustar los contenidos y metodologías de acuerdo con los perfiles reales de los participantes, e implementar estrategias preventivas frente a la deserción y el bajo rendimiento.

Aunque se dispone de información abundante, la ausencia de un análisis sistemático y profundo limita la capacidad de tomar decisiones basadas en evidencia. El problema central, por tanto, es la carencia de conocimiento sobre la relación entre las características sociodemográficas y educativas de los participantes y su comportamiento en el programa, lo que afecta directamente la efectividad e impacto social de Talento Tech Bogotá.

1.2 Objetivos

Objetivo general:

Aplicar técnicas de analítica de datos para explorar, analizar e inferir relaciones significativas entre variables sociodemográficas y formativas de los participantes del programa Talento Tech Bogotá, con el fin de generar información útil para la toma de decisiones y el diseño de futuras convocatorias.

Objetivos específicos:

- Identificar posibles correlaciones entre variables categóricas como estrato socioeconómico, nivel educativo y temática de estudio seleccionada.
- Analizar si existen patrones entre características poblacionales (como ser víctima de conflicto, tener discapacidad o ser campesino) y los niveles temáticos elegidos.

- Evaluar la relación entre el perfil de los participantes y su nivel de asistencia (medido en horas), con el fin de detectar factores que podrían influir en la permanencia o deserción.
- Proponer visualizaciones y reportes que faciliten la interpretación de los hallazgos por parte de los interesados en la gestión del programa.

1.3 Relevancia y motivación

El proyecto adquiere relevancia al situarse en el marco de una de las iniciativas de formación más importantes de la ciudad de Bogotá: Talento Tech Bogotá, un programa que busca garantizar el acceso de los ciudadanos a competencias digitales de alta. El análisis de los datos de los participantes se convierte en una oportunidad única para comprender de manera más profunda cómo las condiciones sociales, económicas y educativas de la población inciden en la elección de las rutas formativas, en la permanencia dentro del proceso y en los resultados de aprendizaje alcanzados.

La motivación central surge de la necesidad de transformar la información disponible en un verdadero insumo estratégico para la toma de decisiones. Aunque existe una base de datos amplia y representativa, esta no ha sido aprovechada de forma sistemática para generar conocimiento aplicable. De ahí que el proyecto tenga un alto valor, pues abre la posibilidad de orientar futuras acciones con base en evidencia y no únicamente en supuestos.

El uso de técnicas de analítica ofrece la posibilidad de identificar con mayor precisión los perfiles de los beneficiarios, comprender las dinámicas de participación y reconocer los factores que pueden limitar el aprovechamiento de la formación. Con ello, se busca aportar información que ayude a mejorar la pertinencia de las convocatorias, a diseñar contenidos y metodologías ajustados a las características reales de los estudiantes y a implementar estrategias que reduzcan los riesgos de deserción o baja asistencia.

La relevancia del proyecto radica en que contribuye a optimizar la gestión de un programa de formación clave para la ciudad. Al aprovechar de manera rigurosa la información disponible, se busca generar aprendizajes que orienten decisiones concretas y fortalezcan la efectividad de Talento Tech Bogotá. De esta manera, el proyecto apoya a los responsables del programa en la mejora de sus estrategias, y también aporta evidencia útil para garantizar que los recursos invertidos logren un mayor impacto en el aprendizaje y la permanencia de los participantes.

1.4 Alcance y limitaciones

El alcance del proyecto se centra en el análisis de la base de datos proporcionada por Talento Tech Bogotá, con el objetivo de identificar relaciones entre las características sociodemográficas y educativas de los participantes y su comportamiento dentro del programa. Esto incluye el estudio de variables como género, estrato socioeconómico, nivel educativo, condición de víctima del conflicto armado, temática de estudio elegida y horas de asistencia, entre otras. El análisis permitirá describir perfiles, reconocer patrones de participación y proponer interpretaciones útiles para la toma de decisiones.

El trabajo no busca evaluar la totalidad de los resultados del programa ni medir impactos de largo plazo, sino más bien realizar un análisis exploratorio y estadístico a partir de la

información disponible. En este sentido, el proyecto está limitado por la calidad y completitud de los datos entregados, que presentan valores faltantes, posibles desbalances entre categorías y variables sensibles que requieren un tratamiento cuidadoso. Estas condiciones pueden restringir el nivel de generalización de los hallazgos y obligan a interpretarlos con cautela.

De igual manera, el proyecto no pretende ofrecer soluciones definitivas a los problemas de permanencia o rendimiento de los beneficiarios, sino generar evidencia inicial que pueda servir como insumo para futuras decisiones y estudios más amplios. Así, el alcance se define como un primer acercamiento riguroso a la información disponible, con limitaciones inherentes al conjunto de datos y a la naturaleza exploratoria del análisis.

2. Revisión literaria

2.1 Trabajos relacionados y estudios similares

Para solidificar los fundamentos del análisis de revisaron estudios previos sobre programas de formación digital y factores asociados a la permanencia, deserción y desempeño académico en contextos similares. Varias investigaciones en formación tecnológica sugieren que las variables de mayores relevancias para el caso son el nivel socioeconómico, la disponibilidad de recursos tecnológicos y la motivación individual.

Los estudios tratados incluyen:

- Análisis de programas de inclusión digital en ciudades de América Latina, donde se observa una relación directa entre nivel educativo previo y tasa de finalización de cursos.
- Investigaciones en analítica de aprendizaje (learning analytics) que emplean modelos estadísticos y de machine learning para predecir riesgo de deserción.
- Evaluaciones de impacto en programas de capacitación basados en competencias digitales (p. ej., iniciativas del MINTIC en Colombia) que señalan la importancia de ajustar contenidos y metodologías a perfiles sociodemográficos específicos.

Estos trabajos evidencian que la analítica de datos es crucial como herramienta para diseñar estrategias enfocadas hacia grupos específicos que tengan necesidades de recursos extra para la mejora de su aprendizaje

2.2 Brechas entre el trabajo actual o aplicación

Diversos programas de formación digital en Colombia y América Latina han implementado estregias de análisis de datos para comprender mejor la dinámica de sus participantes:

- **Programa Misión TIC 2022 (Colombia)**: utilizó análisis descriptivo y segmentación de participantes para ajustar las rutas de aprendizaje según niveles de conocimiento previos.
- **Programa Conecta Empleo (Fundación Telefónica)**: implementó analítica predictiva para estimar la probabilidad de deserción según variables de asistencia y avance en módulos.

- **Estudios de analítica educativa en Brasil y México**: reportaron que factores como género, edad y estrato socioeconómico inciden significativamente en la elección de rutas formativas en áreas de tecnología.

Este trabajo a diferencia de los mencionados se enfoca en los estudiantes de Talento Tech Bogotá, utilizando un dataset que contiene características demográficas como indicadores de desempeño formativo, lo que permitirá un análisis más fiel a la realidad y accionable a futuro.

2.3 Justificación de los métodos seleccionados

La elección de los métodos utilizados fue enfocada a los siguientes tres criterios:

- **Naturaleza de los datos**: La base cuenta principalmente con variables categóricas y numéricas, por lo que se requieren técnicas adecuadas para datos mixtos, como análisis de correlación para variables categóricas (Chi-cuadrado, Cramer's V) y continuas (coeficientes de correlación de Pearson y Spearman).
- **Objetivo del estudio**: Se busca generar conocimiento exploratorio que permita identificar relaciones y patrones, más que desarrollar modelos predictivos complejos. Por ello, se prioriza la analítica descriptiva, exploratoria e inferencial sobre el machine learning avanzado en esta primera fase.
- **Toma de decisiones basada en evidencia**: Los métodos seleccionados (EDA, visualización avanzada y análisis estadístico) ofrecen resultados interpretables y accionables para los responsables del programa, permitiendo **ajustar convocatorias, estrategias de retención y contenidos formativos** con base en hallazgos claros.

También se contempla la posibilidad dentro de fases futuras incorporar modelos de predicción de deserción o recomendación de rutas formativas si la calidad y volumen de los datos lo permiten.

3. Recolección de datos y entendimiento

3.1 Fuentes de datos y métodos de recolección

La información utilizada en este proyecto proviene de la base de datos entregada por Tecnalia Colombia, entidad encargada de ejecutar el programa Talento Tech Bogotá junto con la Secretaría de Desarrollo Económico. Esta base reúne datos recopilados durante el proceso de inscripción, matrícula y formación de los participantes, lo que permite tener un panorama amplio sobre sus características y desempeño en el programa.

Los registros se obtuvieron a partir de formularios de inscripción, pruebas diagnósticas, sistemas de seguimiento académico y controles de asistencia. El conjunto de datos incluye tanto información declarada por los participantes como datos registrados por el programa.

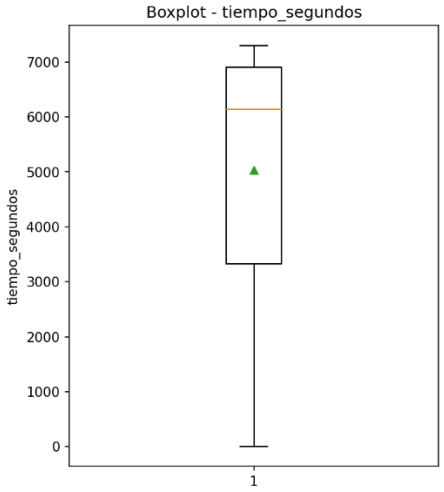
3.2 Descripción de los conjuntos de datos crudos

El conjunto de datos proporcionado por Talento Tech Bogotá incluye 34.333 filas donde cada fila representa un participante, además contiene 48 columnas que contienen datos

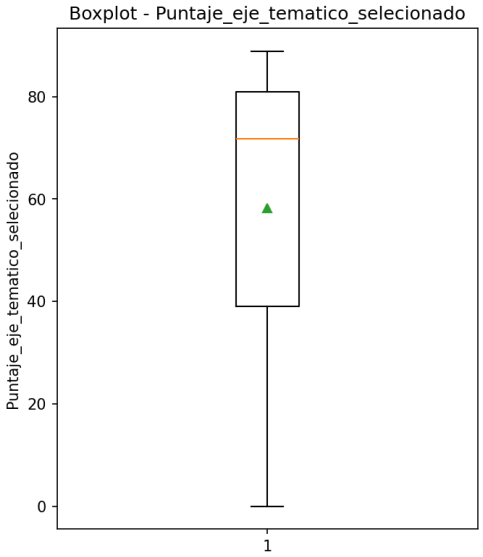
demográficos, educativos, etc. A continuación, se presenta un resumen de los sectores clave y las columnas que forman parte de estos.

1. **Información Demográfica**:
    - **id**: Identificador único para cada participante (por ejemplo, 1, 2, ..., 34333). Variable de tipo nominal.
    - **Codigo_departamento**: Código del departamento (todos los registros son 11, correspondiente a Bogotá, D.C.). Variable de tipo nominal.
    - **Departamento**: Nombre del departamento (todos los registros son "BOGOTÁ, D.C."). Variable de tipo nominal.
    - **Región**: Región (todos los registros son "Región 8"). Variable de tipo nominal.
    - **Codigo_municipio**: Código del municipio (todos los registros son 11001, correspondiente a Bogotá, D.C.). Variable de tipo nominal.
    - **Municipio**: Nombre del municipio (todos los registros son "BOGOTÁ, D.C."). Variable de tipo nominal.
    - **Género**: Género del participante (por ejemplo, Hombre, Mujer, No reporta). Variable de tipo nominal.
    - **Campesino**: Indica si el participante es campesino (SÍ o NO). Variable de tipo binario.
    - **Estrato**: Estrato socioeconómico (1–6, No estratificado, o No reporta). Variable de tipo ordinal.
    - **Autoidentificacion_Etnica**: Autoidentificación étnica (por ejemplo, Ningún grupo étnico, Negro, Mulato, Afrodescendiente, Afrocolombiano, Indígena, o No reporta). Variable de tipo nominal.
    - **Discapacidad**: Tipo de discapacidad que reportan las personas inscritas (por ejemplo, Discapacidad auditiva, Discapacidad física, Discapacidad intelectual, Discapacidad múltiple, Discapacidad psicosocial (mental), Discapacidad visual, Discapacidad Sordoceguera). Variable de tipo nominal.
2. **Educación y Compromiso**:
    - **NIvel_educacion**: Nivel educativo (por ejemplo, Educación Media, Educación Técnica Profesional, Educación Tecnológica, Educación Universitaria Pregrado, Especialización, o No reporta). Variable de tipo ordinal.
    - **Compromiso_10_horas**: Compromiso de dedicar 10 horas al programa (SÍ). Variable de tipo nominal.
    - **Tipo_formacion**: Tipo de formación (Virtual o Híbrida). Variable de tipo binario.
    - **Acepta_requisitos_convocatoria**: Aceptación de los requisitos de la convocatoria (SÍ o NO). Variable de tipo binario.
    - **Victima_del_conflicto**: Estado de víctima del conflicto (SÍ o NO). Variable de tipo binario.
    - **Autoriza_manejo_datos_personales**: Autorización para el manejo de datos personales (SÍ). Variable de tipo nominal.
    - **Disponibilidad_Equipo**: Disponibilidad de equipo (SÍ o NO). Variable de tipo binario.
3. **Formación y Desempeño**:
    - **Presento prueba inicio**: Indica si el participante presentó la prueba inicial (SÍ o NO). Variable de tipo binario.
    - **fecha_ini**: Fecha de la prueba inicial (16/12/2023, 17/12/2023, 18/12/2023 o Null). Variable de tipo ordinal.
    - **tiempo_segundos**: Tiempo empleado en la prueba inicial en segundos (por ejemplo, 5988, o Null). Variable de tipo numérico.

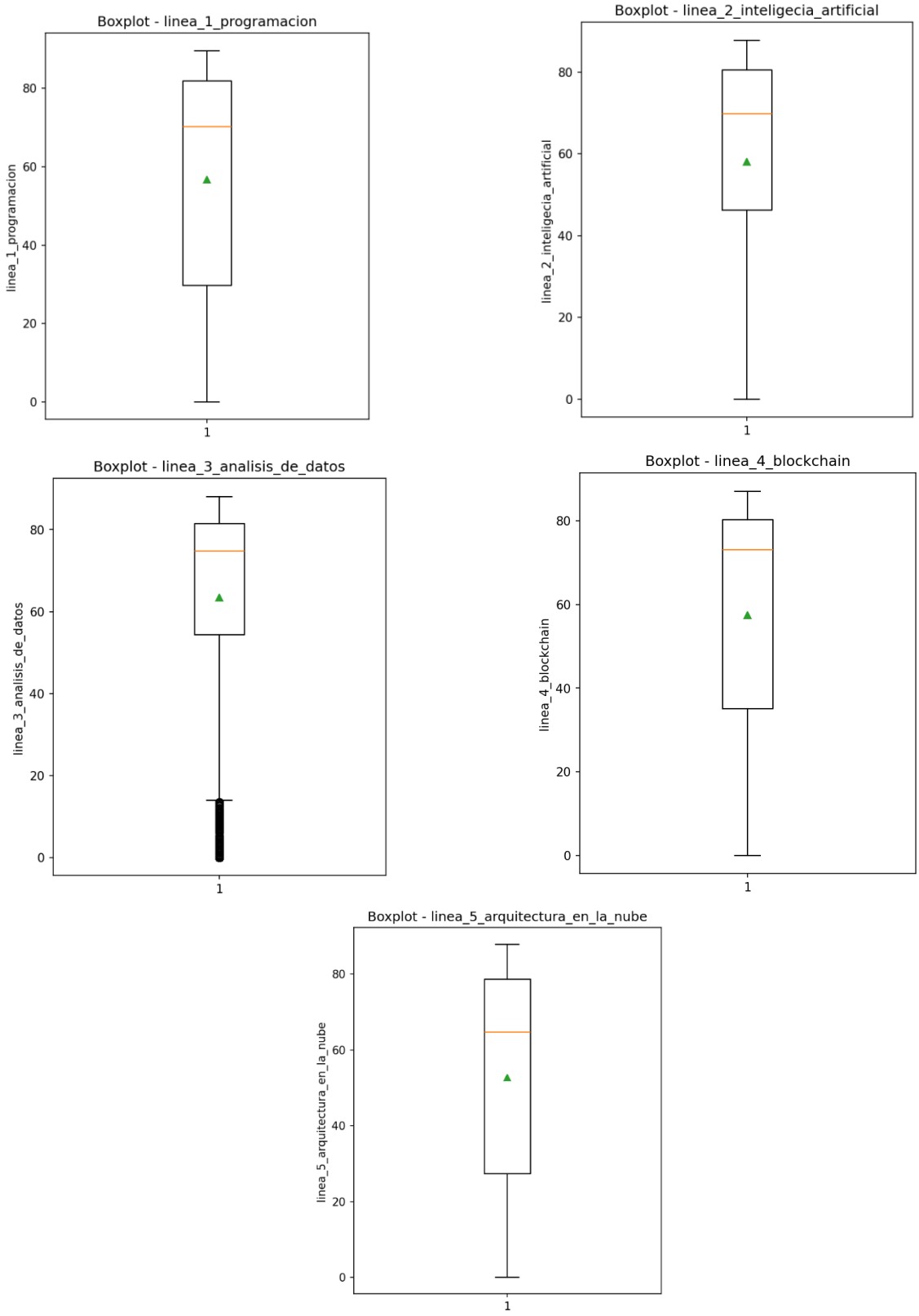
      Boxplot para visualizar la distribución de los datos:



- **Eje_tematico**: Área temática de la prueba inicial (por ejemplo, Programación, Inteligencia artificial, Datos, Ciber seguridad y Blockchain, o Null). Variable de tipo nominal.
- **Eje Final**: Área temática final seleccionada (por ejemplo, Programación, Inteligencia artificial, Análisis de Datos, Blockchain, o Null). Variable de tipo nominal.
- **Puntaje_eje_tematico_selecionado**: Puntaje en el área temática seleccionada (por ejemplo, 84.6429, o Null). Variable de tipo numérico.
  Boxplot para visualizar la distribución de los datos:

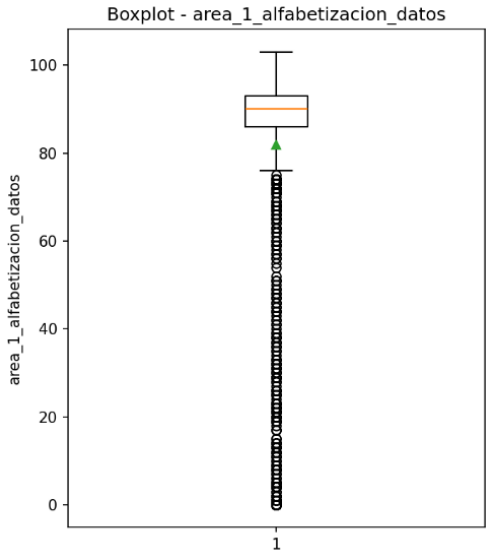


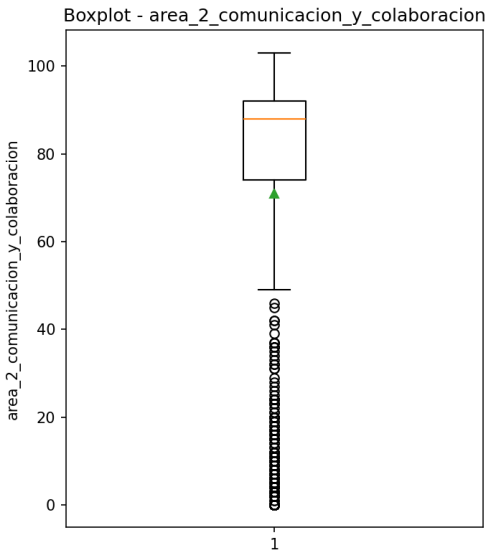
- **linea_1_programacion, linea_2_inteligecia_artificial, linea_3_analisis_de_datos, linea_4_blockchain y linea_5_arquitectura_en_la_nube**: Puntajes en áreas específicas (por ejemplo, 84.6429 para programación, o Null). Variables de tipo numérico.
  Boxplots para visualizar la distribución de los datos:

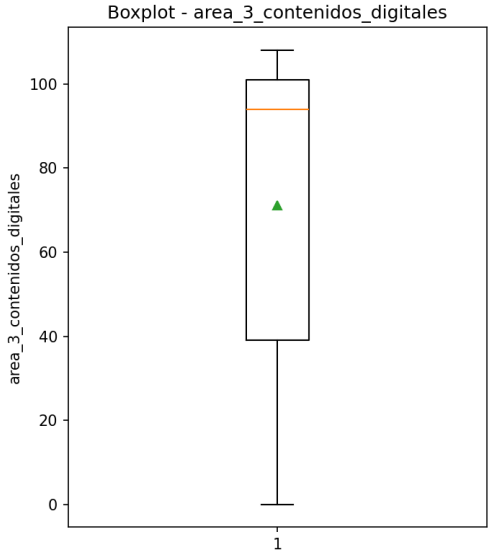


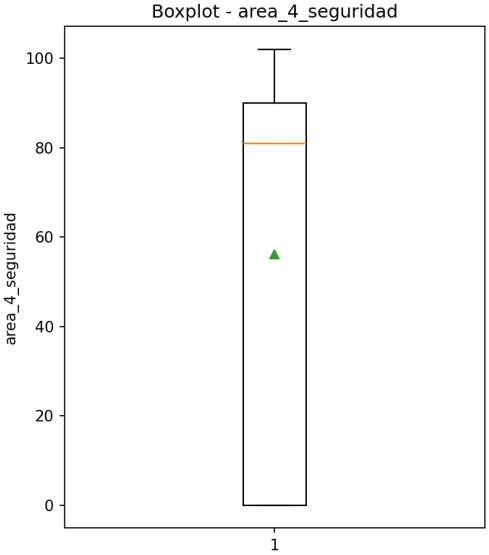
- **linea_1_des_programacion, linea_2_des_inteligecia_artificial, linea_3_des_analisis_de_datos, linea_4_des_blockchain, linea_5_des_arquitectura_en_la_nube**: Niveles de desempeño en áreas específicas (por ejemplo, Innovador - Avanzado, Integrador - Intermedio, Explorador - Básico, o Null). Variables de tipo ordinal.

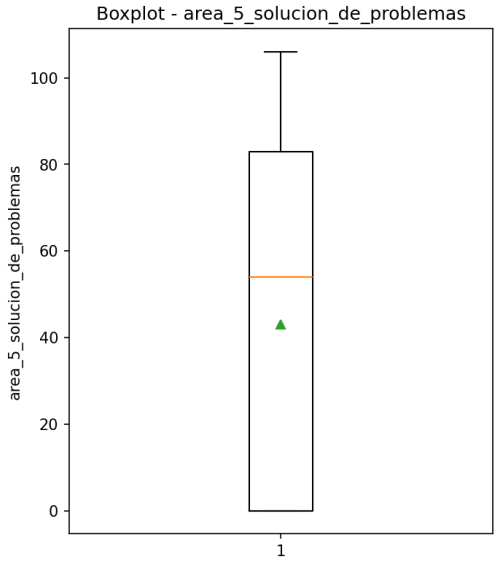
- **area_1_alfabetizacion_datos, area_2_comunicacion_y_colaboracion, area_3_contenidos_digitales, area_4_seguridad, area_5_solucion_de_problemas**: Puntajes en habilidades generales (por ejemplo, 94 para alfabetización de datos, o Null). Variables de tipo numérico. Boxplots para visualizar la distribución de los datos:



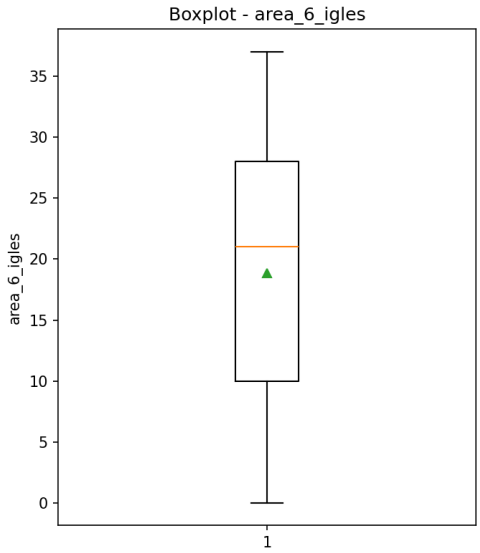








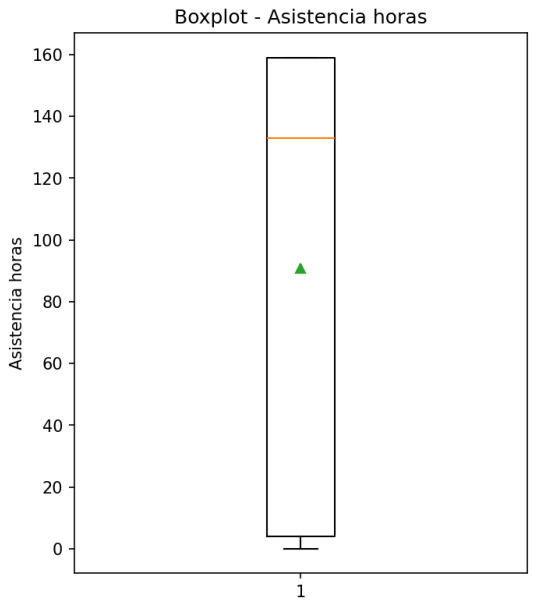
- **area_6_igles**: Puntaje de competencia en inglés (por ejemplo, 31, o Null). Variable de tipo numérico.
  Boxplot para visualizar la distribución de los datos:



- **area_1_des_alfabetizacion_datos, area_2_des_comunicacion_y_colaboracion, area_3_des_contenidos_digitales, area_4_des_seguridad y area_5_des_solucion_de_problemas**: Niveles de desempeño en habilidades generales (por ejemplo, Innovador, Integrador, Explorador, o Null). Variables de tipo ordinal.

4. **Detalles del Programa**:
   - **Origen**: Fuente del participante (MINTIC o CYMETRIA). Variable de tipo binario.
   - **Matriculado (SI o NO)**: Estado de matrícula (SÍ o NO). Variable de tipo binario.
   - **Estado**: Estado del participante (FORMADO, NO APROBADO, INACTIVO, o Vacío). Variable de tipo nominal.

- **Programa de Formación**: Nombre del programa (por ejemplo, Programación, Inteligencia artificial, Análisis de datos, Blockchain, Arquitectura en la nube). Variable de tipo nominal.
- **Cohorte**: Número de cohorte (por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, o combinado como 2_3). Variable de tipo ordinal.
- **Tipo de formación**: Modalidad de formación (VIRTUAL, HÍBRIDA, o Vacías). Nominal.
- **Nivel**: Nivel del programa (Básico, Intermedio, o Avanzado). Variable de tipo ordinal.
- **Asistencia horas**: Horas asistidas (por ejemplo, 159, 0, u otros valores). Variable de tipo numérico.
  Boxplot para visualizar la distribución de los datos:



- **Total_horas posibles**: Total de horas posibles (159 en todos los casos). Variable de tipo nominal.

3.3 Problemas de calidad de los datos y observaciones iniciales

Al revisar la base de datos se identificaron varios aspectos que pueden afectar el análisis si no se tratan adecuadamente. En primer lugar, la mayoría de las variables son categóricas y requieren procesos de codificación antes de aplicar técnicas estadísticas o de machine learning. También se encontraron valores nulos e incompletos, especialmente en campos relacionados con puntajes, fechas y estados de los participantes, lo que obliga a decidir si se imputan, se eliminan o se consideran como categorías aparte.

Otro punto a destacar es la existencia de desequilibrios entre categorías, por ejemplo, mayor número de registros en ciertos niveles educativos o predominio de un género frente al otro, lo cual puede sesgar los resultados de algunos modelos. Adicionalmente, existen registros con formatos poco consistentes, como valores compuestos en una sola celda (“Innovador – Avanzado”), que deben ser separados para su correcto procesamiento.

Igualmente, los Boxplot permiten ver gráficamente que la distribución de la mayoría de los datos numéricos tiene pocos outliers. Esto permite entender que se puede confiar en la representatividad de las medidas de tendencia central como la media y la mediana, gracias a que no son fuertemente distorsionados por valores extremos. Sin embargo, esto no aplica para todas las variables, dado que las variables area_1_alfabetizacion_datos, area_2_comunicacion_y_colaboracion y linea_3_analisis_de_datos si presentan datos atípicos. Con estas variables se puede considerar eliminar algunos datos outliers en el momento del preprocesamiento.

Estos detalles muestran que, aunque la base de datos es amplia y representativa, necesita una fase de limpieza y estandarización para asegurar que el análisis sea confiable y genere resultados útiles.

3.4 Análisis de frecuencias.

Para complementar la comprensión inicial del conjunto de datos, se realizó un análisis exhaustivo de frecuencias para cada una de las 54 variables del dataset. Este análisis permite identificar la distribución de valores en cada variable y detectar patrones iniciales que guiarán análisis posteriores más profundos.

3.4.1 Metodología de análisis de frecuencia.

Se clasificaron las variables según su naturaleza estadística y se aplicaron técnicas de visualización específicas para cada tipo:

Variables Numéricas (14 variables): Histogramas con estadísticas descriptivas

Variables Ordinales (15 variables): Gráficos de barras con ordenamiento lógico

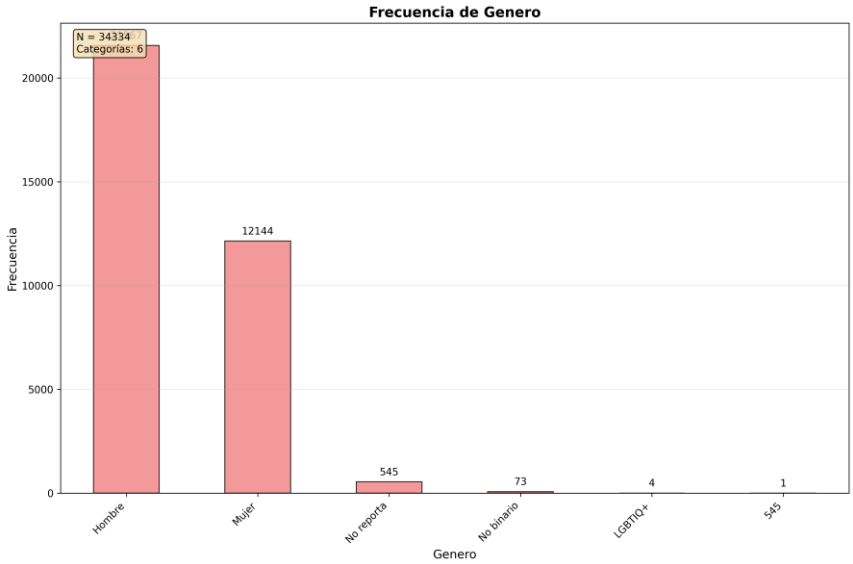
Variables Nominales (17 variables): Gráficos de barras ordenados por frecuencia

Variables Binarias (8 variables): Gráficos de barras simples

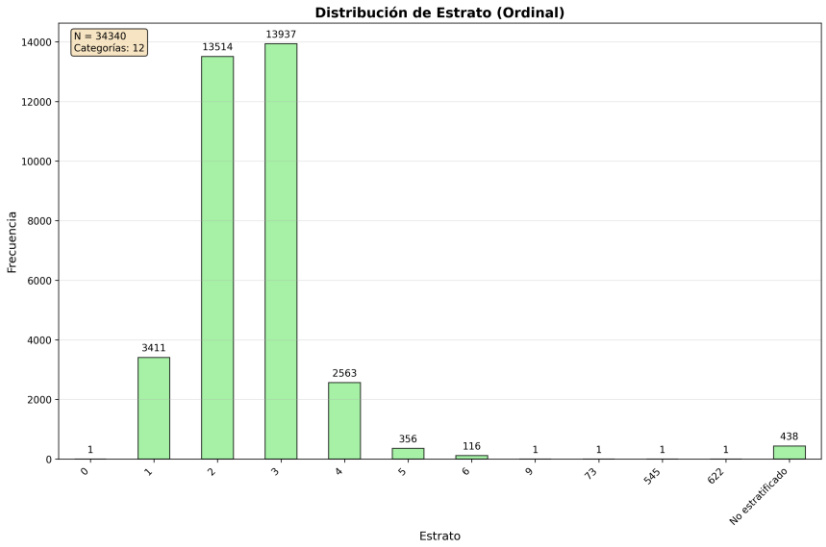
3.4.2 Hallazgos principales por tipo de variable.

Variables Demográficas Clave:

El análisis de frecuencias reveló patrones importantes en las características demográficas de los participantes.

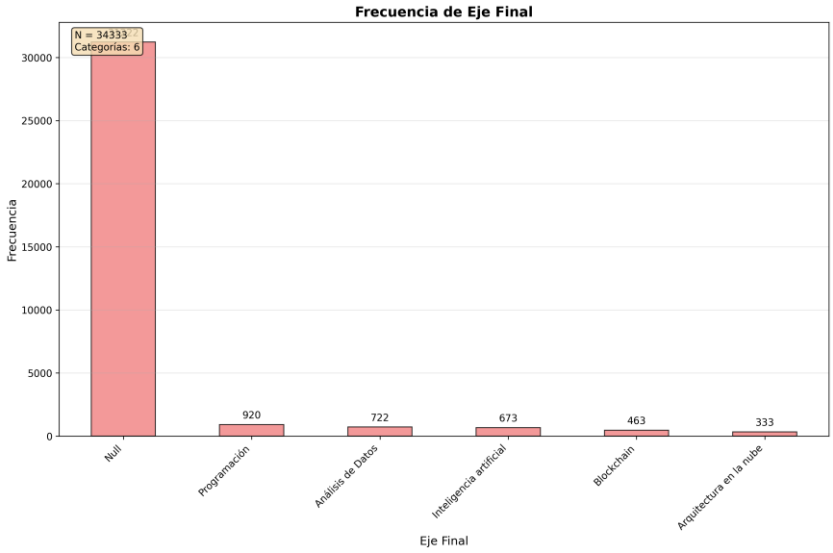


La distribución por género muestra una participación mayoritariamente femenina en el programa, lo que sugiere que las mujeres tienen mayor interés o acceso a las convocatorias de formación digital. Este patrón es relevante para diseñar estrategias de inclusión que equilibren la participación por género.

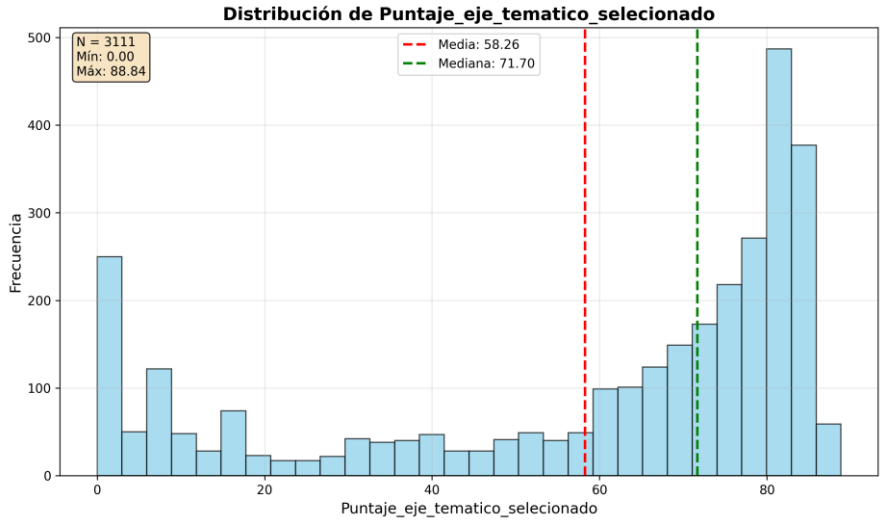


La concentración de participantes en estratos socioeconómicos bajos (1-3) confirma que el programa está cumpliendo su objetivo de llegar a población vulnerable. Sin embargo, la baja participación de estratos altos sugiere oportunidades de diversificación socioeconómica.

Variables de Formación y Desempeño.

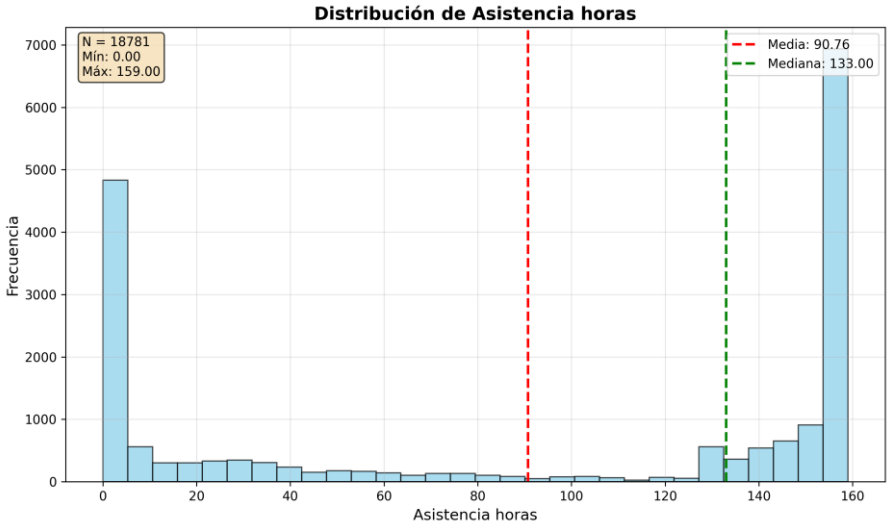


La distribución de elección de ejes temáticos revela las preferencias formativas de los participantes. Si hay concentración en ciertas áreas (como Programación o Análisis de Datos), esto indica tanto demanda del mercado como posibles brechas en otras áreas tecnológicas que requieren mayor promoción.

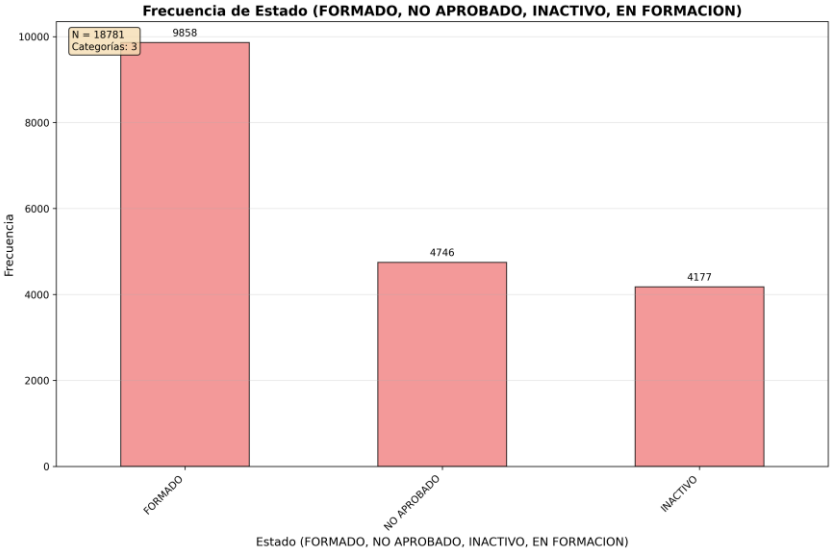


La distribución de puntajes permite identificar el nivel de conocimientos previos de los participantes. Una distribución normal sugiere heterogeneidad en las competencias iniciales, mientras que sesgos hacia puntajes bajos o altos indicarían necesidades específicas de nivelación o retos adicionales.

Variables de Participación:



La distribución de horas de asistencia es crucial para entender patrones de permanencia y compromiso. Concentraciones en valores bajos pueden indicar problemas de deserción temprana, mientras que distribuciones bimodales podrían revelar dos grupos distintos: participantes altamente comprometidos y otros con dificultades de permanencia.



La proporción de participantes en cada estado final (FORMADO, NO APROBADO, INACTIVO) proporciona una medida directa de la efectividad del programa. Altas tasas de "FORMADO" indican éxito, mientras que proporciones significativas de "INACTIVO" señalan áreas de mejora en retención.

3.4.3 Identificación de patrones críticos.

El análisis de frecuencias permitió identificar varios patrones que requieren atención especial:

Concentración en categorías específicas: Algunas variables muestran alta concentración en pocas categorías, lo que puede limitar la variabilidad para análisis posteriores.

Distribuciones sesgadas: Variables numéricas con distribuciones asimétricas que requerirán transformaciones para análisis estadísticos paramétricos.

Categorías con baja representación: Grupos poblacionales minoritarios que necesitarán técnicas especiales de análisis para evitar conclusiones sesgadas.