Pontificia Universidad Javeriana

Propuesta de Proyecto

Angel Eduardo Morales Abril
Andrés David Pérez Cely

Analitica de datos

Profesor: Ferney Maldonado Lopez

Agosto 8, 2025

1. **Contexto y motivación del proyecto**

Tecnalia Colombia es una organización especializada en la prestación de servicios de formación y certificación en competencias digitales. En alianza con la Secretaría de Desarrollo Económico de Bogotá, ha liderado la ejecución del programa Talento Tech Bogotá, una iniciativa del Distrito que busca capacitar a ciudadanos en habilidades tecnológicas de alta demanda, como análisis de datos, desarrollo web, ciberseguridad, inteligencia artificial, y más.

Durante el último año, Tecnalia ejecutó procesos de convocatoria y formación en distintas temáticas del sector tech. Como parte del acompañamiento al programa, ha facilitado una base de datos que contiene información de los participantes, incluyendo variables como: condición de víctima del conflicto armado, presencia de discapacidad, condición de campesino, género, estrato socioeconómico, nivel de estudio, temática de estudio, y horas de asistencia a clase.

La motivación del proyecto radica en la posibilidad de aplicar técnicas de analítica de datos para entender mejor el perfil de los beneficiarios del programa y analizar cómo ciertos factores individuales o sociales se relacionan con la elección de temáticas de estudio y con la asistencia a las clases. Esto con el fin de generar valor para el programa al comprender los perfiles de los participantes, identificar patrones de participación, prever deserciones, evaluar la eficacia de la convocatoria, proponer mejoras basadas en datos, entre otras.

Estos análisis pueden generar información valiosa para mejorar la focalización de futuras convocatorias, adaptar contenidos formativos, y contribuir a una mejor toma de decisiones basada en evidencia dentro del programa. Además, el trabajo representa una oportunidad de aplicar los conceptos que se verán en clase a un caso real y relevante, fortaleciendo así las competencias analíticas del equipo.

2. **Objetivos del proyecto**

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

3. Interesados que se beneficiarán del proyecto

El análisis de los datos del programa Talento Tech Bogotá puede aportar valor a distintos grupos de interés, tanto dentro como fuera de la organización ejecutora. A continuación, se identifican los principales interesados y los beneficios esperados:

- **Tecnalia Colombia**: Como operador del programa, se beneficiará directamente del análisis al obtener insumos objetivos que le permitan mejorar la focalización de las convocatorias, diseñar rutas formativas más efectivas y argumentar con evidencia el impacto social de su labor ante entidades públicas.
- **Secretaría de Desarrollo Económico de Bogotá**: Entidad que financia y lidera Talento Tech. El proyecto proporciona una visión basada en datos sobre la pertinencia y cobertura del programa, facilitando el diseño de políticas públicas más inclusivas y eficientes.
- **Participantes actuales y futuros del programa**: Indirectamente beneficiados, ya que los resultados del análisis podrían traducirse en una oferta formativa mejor adaptada a sus necesidades, perfiles y condiciones sociales.
- **Equipo académico y estudiantes del curso de Analítica de Datos**: El proyecto representa una oportunidad práctica de aplicar conocimientos teóricos en un caso real con impacto social, desarrollando habilidades en exploración de datos, visualización, inferencia estadística y comunicación de hallazgos.

4. Definición de la pregunta de negocio

¿Qué factores sociodemográficos y educativos influyen en la elección de temática de formación, la permanencia de los participantes y su desempeño en el programa Talento Tech Bogotá?

5. Explicación de porque el problema no es trivial y requiere de analítica

Aunque a simple vista podría parecer un análisis descriptivo básico, el problema planteado en este proyecto no es trivial por varias razones relacionadas con la naturaleza de los datos y los objetivos del análisis.

En primer lugar, la base de datos contiene múltiples variables categóricas como el estrato socioeconómico, el nivel educativo, la condición de víctima del conflicto armado, la discapacidad, el género y la temática de estudio. Estas variables se relacionan entre sí de forma compleja, por lo que es necesario aplicar técnicas de análisis multivariado e inferencia estadística para identificar asociaciones significativas y evitar conclusiones erróneas.

Adicionalmente, muchas de estas variables representan condiciones sociales sensibles que exigen un tratamiento ético y cuidadoso. No basta con observar frecuencias o porcentajes; se requiere comprobar estadísticamente si existen diferencias reales entre grupos, lo que implica el uso de pruebas de hipótesis apropiadas para variables categóricas.

Aunque el volumen de datos no es masivo, sí es lo suficientemente estructurado y representativo como para justificar el uso de herramientas de análisis que permitan codificar, limpiar, transformar y visualizar los datos de forma rigurosa. Su distribución, escala y posibles valores atípicos obligan a realizar un tratamiento analítico cuidadoso.

Finalmente, este análisis tiene una aplicabilidad directa en la toma de decisiones. Las conclusiones que se obtengan podrán influir en el diseño de futuras convocatorias, estrategias de focalización y mejoras en los procesos formativos. Por todo lo anterior, se requiere una mirada analítica integral para generar conocimiento útil, confiable y accesible.

6. Descripción de los desafíos de datos identificados

El análisis de la base de datos del programa Talento Tech Bogotá presenta diversos desafíos que requieren una preparación adecuada antes de aplicar técnicas estadísticas o de machine learning.

Uno de los principales retos es el tratamiento de datos categóricos, ya que la mayoría de las variables (como género, nivel educativo, tipo de formación, condición de víctima, etc.) deben ser codificadas adecuadamente para su análisis. Se requerirá convertir estas categorías en representaciones numéricas (one-hot encoding o codificación ordinal, según el caso) para alimentar modelos de aprendizaje automático.

El conjunto de datos contiene varias variables cuantitativas, entre ellas: el tiempo en segundos de conexión, los puntajes en pruebas iniciales, las calificaciones por línea de formación, y los resultados por áreas de competencia digital. Esto incrementa la dimensionalidad del análisis y abre la posibilidad de aplicar técnicas de clustering, regresión o clasificación, pero también implica un reto de normalización de datos y detección de valores atípicos.

Otro desafío importante es la presencia de valores nulos o incompletos, especialmente en las columnas relacionadas con fechas, puntajes o estado del proceso. Será necesario realizar un análisis exploratorio para decidir si estos valores se imputan, se descartan o se tratan como categorías propias.

También se identifican posibles desequilibrios en las clases (por ejemplo, más personas con nivel educativo medio que avanzado, o más mujeres que hombres), lo cual puede afectar el rendimiento de modelos de clasificación. Se deberán considerar técnicas de balanceo como sobremuestreo o submuestreo en caso de ser necesarias.

Finalmente, el dataset contiene columnas con estructuras de texto complejas (como "Innovador - Avanzado"), que requerirán limpieza y separación para su procesamiento efectivo.

En conjunto, estos desafíos justifican el uso de herramientas de analítica avanzada, ya que no es posible extraer conclusiones significativas sólo a partir de métodos descriptivos simples. El proyecto exigirá un tratamiento cuidadoso de los datos, incluyendo procesos de limpieza, transformación, visualización y selección de modelos adecuados, para garantizar resultados válidos y útiles para todos los interesados.

7. Descripción de alto nivel de los métodos de analítica a utilizar

Para el dataset de Talento Tech Bogotá, los métodos de analítica propuestos permiten identificar patrones, relaciones y diferencias significativas entre grupos, apoyando la toma de decisiones basadas en evidencia. Los métodos propuestos en cuestión son:

1. **Análisis descriptivo (cuantitativo básico):** Resumir y visualizar la distribución de las variables, las técnicas utilizadas serán:
    a. Tablas de frecuencias y proporciones.
    b. Gráficos de barras y de pastel para variables categóricas.
    c. Medidas de tendencia central y dispersión para *horas de asistencia a clase*.
    d. Histogramas y boxplots para la variable numérica.
2. **Estadística Inferencial:** Inferir conclusiones sobre la población a partir de la muestra, las técnicas utilizadas serán:
    a. Pruebas de hipótesis para proporciones (comparar porcentajes entre grupos, por ejemplo, % de mujeres vs hombres con nivel avanzado de estudio).
    b. Pruebas Chi-cuadrado de independencia para evaluar la asociación entre variables categóricas (ej: género vs temática de estudio).
    c. Pruebas t o U de Mann–Whitney para comparar medias de horas de asistencia entre dos grupos (ej: víctimas vs no víctimas).
3. **Análisis de correlación y asociación:** Identificar qué tan relacionadas están dos variables, las técnicas utilizadas serán:
    a. Coeficiente de correlación de Spearman (para variable numérica vs categórica ordinal, ej: horas de asistencia vs nivel de estudio).
    b. Cramer’s V o Phi para medir la fuerza de asociación entre variables categóricas (ej: estrato vs temática de estudio).
4. **Modelo exploratorio:** Analizar cómo una variable puede explicar o predecir otra, las técnicas utilizadas serán:
    a. Modelos de regresión logística binaria o multinomial para predecir categorías (ej: probabilidad de nivel avanzado según horas de asistencia, género y condición socioeconómica).

    b. Análisis de varianza (ANOVA) para evaluar diferencias en horas de asistencia según categorías de estudio o condición social.
5. **Visualización avanzada:** Comunicar hallazgos de manera efectiva.
    a. Mapas de calor para matrices de correlación.
    b. Gráficos de mosaico para relaciones entre variables categóricas.
    c. Diagramas de dispersión con diferenciación por categorías.

8. Descripción del conjunto de datos (fuente, volumen, atributos)

El conjunto de datos proporcionado por Talento Tech Bogotá contiene información sobre los participantes del programa de formación Talento Tech en Bogotá, D.C., Colombia. Este incluye 34.334 filas donde cada fila representa un participante, además contiene 48 columnas que contienen datos demográficos, educativos, etc.

Los datos proporcionados provienen de Tecnalia, una empresa que ofrece cursos técnicos en Bogotá, Colombia.

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).
    - **Codigo_departamento**: Código del departamento (todos los registros son 11, correspondiente a Bogotá, D.C.).
    - **Departamento**: Nombre del departamento (todos los registros son "BOGOTÁ, D.C.").
    - **Región**: Región (todos los registros son "Región 8").
    - **Codigo_municipio**: Código del municipio (todos los registros son 11001, correspondiente a Bogotá, D.C.).
    - **Municipio**: Nombre del municipio (todos los registros son "BOGOTÁ, D.C.").
    - **Genero**: Género del participante (por ejemplo, Hombre, Mujer, No reporta).
    - **Campesino**: Indica si el participante es campesino (SÍ o NO).
    - **Estrato**: Estrato socioeconómico (1–6, No estratificado, o No reporta).
    - **Autoidentificacion_Etnica**: Autoidentificación étnica (por ejemplo, Ningún grupo étnico, Negro, Mulato, Afrodescendiente, Afrocolombiano, Indígena, o No reporta).
    - **Discapacidad**: Estado de discapacidad (vacío en la muestra proporcionada, lo que indica que no se reportaron discapacidades).
2. **Educación y Compromiso**:

- **NIvel_educacion**: Nivel educativo (por ejemplo, Educación Media, Educación Técnica Profesional, Educación Tecnológica, Educación Universitaria Pregrado, Especialización, o No reporta).
- **Compromiso_10_horas**: Compromiso de dedicar 10 horas al programa (SÍ o vacío).
- **Tipo_formacion**: Tipo de formación (Virtual, Híbrida, o Presencial).
- **Acepta_requisitos_convocatoria**: Aceptación de los requisitos de la convocatoria (SÍ o vacío).
- **Victima_del_conflicto**: Estado de víctima del conflicto (SÍ o NO).
- **Autoriza_manejo_datos_personales**: Autorización para el manejo de datos personales (SÍ o vacío).
- **Disponibilidad_Equipo**: Disponibilidad de equipo (SÍ o vacío).

3. **Formación y Desempeño**:
   - **Presento prueba inicio**: Indica si el participante presentó la prueba inicial (SÍ o NO).
   - **fecha_ini**: Fecha de la prueba inicial (en formato de fecha serial de Excel, por ejemplo, 45276.73278935185, o Null).
   - **tiempo_segundos**: Tiempo empleado en la prueba inicial en segundos (por ejemplo, 5988, o Null).
   - **Eje_tematico**: Área temática de la prueba inicial (por ejemplo, Programación, Inteligencia artificial, Datos, Ciber seguridad y Blockchain, o Null).
   - **Eje Final**: Área temática final seleccionada (por ejemplo, Programación, Inteligencia artificial, Análisis de Datos, Blockchain, o Null).
   - **Puntaje_eje_tematico_selecionado**: Puntaje en el área temática seleccionada (por ejemplo, 84.6429, o Null).
   - **linea_1_programacion a linea_5_arquitectura_en_la_nube**: Puntajes en áreas específicas (por ejemplo, 84.6429 para programación, o Null).
   - **linea_1_des_programacion a linea_5_des_arquitectura_en_la_nube**: Niveles de desempeño en áreas específicas (por ejemplo, Innovador - Avanzado, Integrador - Intermedio, Explorador - Básico, o Null).
   - **area_1_alfabetizacion_datos a area_5_solucion_de_problemas**: Puntajes en habilidades generales (por ejemplo, 94 para alfabetización de datos, o Null).
   - **area_6_ingles**: Puntaje de competencia en inglés (por ejemplo, 31, o Null).
   - **area_1_des_alfabetizacion_datos a area_5_des_solucion_de_problemas**: Niveles de desempeño en habilidades generales (por ejemplo, Innovador, Integrador, Explorador, o Null).
4. **Detalles del Programa**:
   - **Origen**: Fuente del participante (MINTIC o CYMETRIA).
   - **Matriculado (SI o NO)**: Estado de matrícula (SÍ o NO).

- **Estado**: Estado del participante (FORMADO, NO APROBADO, INACTIVO, o EN FORMACIÓN).
- **Programa de Formación**: Nombre del programa (por ejemplo, Programación, Inteligencia artificial, Análisis de datos, Blockchain, Arquitectura en la nube).
- **Cohorte**: Número de cohorte (por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, o combinado como 2_3).
- **Tipo de formación**: Modalidad de formación (PRESENCIAL, VIRTUAL, o HÍBRIDA).
- **Nivel**: Nivel del programa (Básico, Intermedio, o Avanzado).
- **Asistencia horas**: Horas asistidas (por ejemplo, 159, 0, u otros valores).
- **Total_horas posibles**: Total de horas posibles (159 en todos los casos).

9. Plan de ejecución.

El desarrollo del proyecto se llevará a cabo a lo largo de diez semanas, iniciando con la limpieza y preprocesamiento de los datos para garantizar su calidad y consistencia. En esta etapa se abordará el tratamiento de valores nulos, la codificación de variables categóricas y la normalización de datos, de forma que queden listos para el análisis posterior. Una vez depurada la base de datos, se procederá a la exploración inicial mediante análisis descriptivo y visualizaciones básicas, lo que permitirá identificar patrones preliminares y posibles relaciones entre variables.

En las siguientes fases se aplicarán pruebas de hipótesis y análisis estadísticos para determinar asociaciones significativas entre factores sociodemográficos y temáticas de estudio. De ser necesario, se utilizarán modelos exploratorios como regresión logística o análisis de varianza para explicar diferencias y comportamientos observados. Si el método principal no produce resultados satisfactorios o estadísticamente relevantes, se recurrirá a un plan alterno que incluye la aplicación de técnicas no paramétricas, modelos basados en árboles de decisión o random forest, y reducción de dimensionalidad para simplificar el análisis sin perder información relevante.

El avance del proyecto se organizará de la siguiente manera:

- **Semana 1:** Revisión inicial de la base de datos y comprensión de su estructura.
- **Semana 2:** Limpieza y tratamiento de valores nulos o inconsistentes.
- **Semana 3:** Codificación de variables categóricas y normalización de variables numéricas.
- **Semana 4:** Análisis exploratorio inicial (EDA) y visualizaciones básicas.
- **Semana 5:** Pruebas de hipótesis y análisis estadístico descriptivo.

- **Semana 6:** Aplicación de técnicas de asociación y correlación (Chi-cuadrado, Spearman, Cramer's V).
- **Semana 7:** Desarrollo de modelos exploratorios y regresión logística.
- **Semana 8:** Visualización avanzada de resultados (mapas de calor, gráficos de mosaico, diagramas de dispersión).
- **Semana 9:** Integración de hallazgos y redacción del reporte final.
- **Semana 10:** Preparación y entrega de la presentación final.

Al finalizar, se entregará una base de datos limpia y documentada, un informe de análisis estadístico y exploratorio, visualizaciones interpretativas y, si resulta pertinente, modelos predictivos exploratorios. La presentación final resumirá los hallazgos clave y recomendaciones para la toma de decisiones en futuras convocatorias del programa.

En cuanto a la distribución de responsabilidades, Angel Eduardo Morales Abril se enfocará en el preprocesamiento de datos, el modelado exploratorio y la redacción de resultados, mientras que Andrés David Pérez Cely se encargará del análisis estadístico, la generación de visualizaciones y la elaboración de la presentación final.

10. Herramientas requeridas

El proyecto se desarrollará en su mayoría en Python, utilizando bibliotecas como Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn y SciPy para el manejo, análisis y visualización de datos. Se trabajará en computadores personales para los cuales no debería de haber problemas en relación al rendimiento adecuado durante las etapas de procesamiento y modelado.

11. Resultados de aprendizaje esperados

Se espera que el equipo adquiera la capacidad de aplicar correctamente técnicas de limpieza y transformación de datos, ejecutar análisis estadísticos e inferenciales, desarrollar modelos exploratorios con datos mixtos y comunicar de forma clara los hallazgos obtenidos. Como valor agregado se espera fortalecer los conocimientos dentro del área de visualización de datos y presentación de resultados orientados a la toma de decisiones.

12. Resumen de la propuesta

Este proyecto tiene como objetivo analizar la base de datos del programa Talento Tech Bogotá para identificar patrones y relaciones entre variables sociodemográficas, temáticas de estudio y niveles de asistencia, utilizando técnicas de análisis descriptivo, inferencial y exploratorio. La información resultante permitirá proponer mejoras en la focalización de las convocatorias y en la adaptación de los contenidos formativos, garantizando un impacto positivo en los participantes y una mayor eficiencia en el uso de recursos. El análisis se realizará con un enfoque ético y

riguroso, priorizando la claridad y utilidad de los resultados para los interesados directos e indirectos del programa.