Analítica de Datos para el Programa Talento Tech Bogotá: Desde la Exploración hasta un Modelo Predictivo de Éxito Académico

Angel Eduardo Morales Abril, Andres David Perez Cely
Programa de Analítica de Datos
Talento Tech Bogotá

***Abstract*—Este trabajo presenta el desarrollo completo de un proyecto de analítica de datos aplicado al programa Talento Tech Bogotá. Se analiza una base de más de 34 mil participantes con información sociodemográfica, educativa y de desempeño formativo, con el fin de comprender los factores asociados al éxito académico y apoyar la toma de decisiones basada en datos. Se abordan de forma integrada las fases de recolección y entendimiento de datos, preprocesamiento, análisis exploratorio y modelado predictivo mediante técnicas de machine learning (Regresión Logística, Random Forest y XGBoost), así como su implementación en un prototipo web (FastAPI + Angular). Los resultados muestran que variables relacionadas con la asistencia, el nivel de desempeño inicial y ciertas características académicas son más relevantes que los factores puramente sociodemográficos para explicar la probabilidad de finalización exitosa.**

***Index Terms*—Analítica de datos, educación, éxito académico, Random Forest, XGBoost, Talento Tech Bogotá.**

I. Resumen Ejecutivo

I-A. Resumen del proyecto

El programa Talento Tech Bogotá, liderado por Tecnalia y la Secretaría de Desarrollo Económico, busca cerrar brechas en competencias digitales mediante formación en análisis de datos, desarrollo web, inteligencia artificial y ciberseguridad, entre otras rutas. Aunque se dispone de una base de datos rica en variables sociodemográficas, educativas y de participación, no existía un análisis sistemático que permitiera comprender cómo estas variables se relacionan con la elección de temáticas, la permanencia y el desempeño de los participantes.

En este proyecto se construyó un pipeline de ciencia de datos que incluye limpieza, transformación, análisis exploratorio, ingeniería de características y entrenamiento de modelos de clasificación para predecir el éxito académico (*formado* vs. *no formado*), culminando en un servicio de predicción desplegado mediante una API en FastAPI y una interfaz web en Angular.

I-B. Conclusiones importantes e impacto

Los análisis muestran que:

- La población participante está concentrada en estratos 2 y 3, con predominio masculino, lo que evidencia el alcance social del programa.
- Las variables de asistencia efectiva, puntajes iniciales en el eje temático y promedios de competencias digitales son los predictores más fuertes del éxito académico.
- Variables como género y estrato influyen más en la participación que en el desempeño final, lo que sugiere que las brechas de acceso no necesariamente se traducen en brechas de logro cuando se ofrece una formación estructurada.
- Un modelo Random Forest con selección de características (SelectKBest) alcanza una exactitud cercana al 64% y AUC alrededor de 0,65, superando tanto una Regresión Logística como la línea base de un clasificador mayoritario.

Estos hallazgos permiten orientar estrategias de retención (por ejemplo, enfocadas en asistencia temprana) y ajustar la oferta formativa según niveles iniciales y perfiles de los beneficiarios.

I-C. Resumen de los métodos y resultados

El pipeline implementado comprende:

1. Preprocesamiento avanzado: limpieza de valores faltantes, tratamiento de outliers, ingeniería de variables (promedios de líneas y áreas, ratio de asistencia), reducción de dimensionalidad con PCA y estandarización.
2. Análisis exploratorio: estadística descriptiva, visualización de distribuciones, análisis de correlaciones (Pearson) y patrones de éxito por género, nivel educativo y temática.
3. Modelado: comparación de Regresión Logística, Random Forest y XGBoost con selección de variables mediante SelectKBest, búsqueda de hiperparámetros y evaluación con métricas de clasificación.
4. Implementación: despliegue del mejor modelo en un backend FastAPI consumido desde un frontend Angular.

El modelo final (Random Forest) ofrece un compromiso razonable entre rendimiento, robustez e interpretabilidad, adecuado

como prototipo para apoyar decisiones de seguimiento académico.

II. Introducción

II-A. Definición del problema

Talento Tech Bogotá es una iniciativa de formación en competencias digitales dirigida a una población amplia y diversa, que incluye personas en condición de vulnerabilidad social y con trayectorias educativas heterogéneas. La base de datos disponible contiene variables sociodemográficas (género, estrato, condición de víctima del conflicto, discapacidad, condición de campesino, autoidentificación étnica), educativas (nivel educativo, temática elegida) y de participación (horas de asistencia, estado, puntajes y niveles alcanzados).

Sin embargo, no se conoce con claridad cómo se relacionan estas características entre sí ni qué factores explican la elección temática, la permanencia o el desempeño final en el programa. Esta falta de comprensión dificulta diseñar convocatorias focalizadas, ajustar contenidos y metodologías según los perfiles reales de los participantes e implementar estrategias preventivas frente a la deserción y el bajo rendimiento. El problema central es, por tanto, la carencia de conocimiento basado en datos sobre la relación entre las características sociodemográficas y educativas y el comportamiento de los beneficiarios, lo que afecta la efectividad e impacto social del programa.

II-B. Objetivos

El objetivo general es aplicar técnicas de analítica de datos para explorar, analizar e inferir relaciones significativas entre variables sociodemográficas y formativas de los participantes de Talento Tech Bogotá, generando información útil para la toma de decisiones y el diseño de futuras convocatorias.

Los objetivos específicos incluyen:

- Identificar correlaciones entre variables categóricas como estrato socioeconómico, nivel educativo y temática de estudio seleccionada.
- Analizar patrones entre características poblacionales (víctima del conflicto, discapacidad, condición de campesino) y los niveles temáticos elegidos.
- Evaluar la relación entre el perfil de los participantes y su nivel de asistencia (en horas) para detectar factores asociados a permanencia o deserción.
- Desarrollar visualizaciones y reportes que faciliten la interpretación de los hallazgos por parte de los responsables del programa.

II-C. Relevancia y motivación

El proyecto se sitúa en el contexto de una de las iniciativas de formación digital más importantes de Bogotá, donde el análisis de datos de los participantes constituye una oportunidad para comprender cómo las condiciones sociales, económicas y educativas inciden en la elección de rutas formativas, la permanencia y los resultados de aprendizaje.

La motivación central es transformar una base de datos abundante en un insumo estratégico, pasando de información cruda a conocimiento accionable. Mediante técnicas de analítica y machine learning, se busca identificar perfiles de beneficiarios, dinámicas de participación y factores que limitan el aprovechamiento de la formación, aportando evidencias que permitan optimizar convocatorias, contenidos y estrategias de retención.

II-D. Alcance y limitaciones

El alcance se centra en la base de datos suministrada por Talento Tech Bogotá, analizando relaciones entre características sociodemográficas, educativas y comportamentales, sin pretender medir impactos de largo plazo ni evaluar la totalidad del programa. El trabajo se limita por:

- La calidad y completitud de los datos (valores faltantes, desbalances entre categorías, variables sensibles).
- La naturaleza exploratoria del análisis, que busca identificar patrones y construir modelos prototipo, más que soluciones definitivas.

Los resultados deben interpretarse como un primer acercamiento riguroso que abre la puerta a análisis futuros más profundos.

III. Revisión de la Literatura

III-A. Trabajos relacionados y estudios similares

Se revisaron estudios sobre programas de formación digital e iniciativas de inclusión tecnológica en América Latina, donde se han analizado factores asociados a permanencia, deserción y desempeño académico. La literatura sobre *learning analytics* muestra el uso de modelos estadísticos y de machine learning para predecir riesgo de deserción y segmentar participantes según su trayectoria formativa.

Investigaciones en programas como Misión TIC 2022 en Colombia o Conecta Empleo de Fundación Telefónica resaltan el papel del nivel educativo previo, la disponibilidad de recursos tecnológicos y la motivación individual como predictores relevantes, así como la necesidad de ajustar contenidos y metodologías a perfiles específicos de la población.

III-B. Brechas entre el trabajo actual y aplicaciones previas

Mientras que muchos programas han utilizado analítica descriptiva o segmentación básica, este trabajo se enfoca específicamente en la población de Talento Tech Bogotá, empleando un conjunto de datos que combina características demográficas detalladas, información de participación y resultados formativos. La integración de variables de desempeño inicial (pruebas y puntajes por línea temática) con métricas de asistencia y estado final permite un análisis más fino y accionable que el

de estudios centrados únicamente en datos de inscripción o de uso de plataformas.

III-C. Justificación de los métodos seleccionados

La elección de métodos parte de tres criterios:

- **Naturaleza de los datos**: combinación de variables categóricas y numéricas, que requiere técnicas adecuadas para datos mixtos, como correlaciones específicas y modelos que manejen bien no linealidades.
- **Objetivo del estudio**: generar conocimiento exploratorio e inferencial más que construir un producto industrial de predicción, privilegiando interpretabilidad y robustez sobre complejidad extrema.
- **Toma de decisiones**: la administración del programa necesita resultados comprensibles, por lo que se combinan modelos interpretables (Regresión Logística) con modelos de mayor capacidad (Random Forest, XGBoost).

IV. Recolección de Datos y Entendimiento

IV-A. Fuentes de datos y métodos de recolección

La información procede de la base entregada por Tecnalia Colombia, que integra datos de: formularios de inscripción, pruebas diagnósticas, sistemas de seguimiento académico y registros de asistencia. Cada fila representa un participante y las columnas incluyen información declarada y datos registrados por el programa a lo largo del proceso formativo.

IV-B. Descripción de los conjuntos de datos crudos

El conjunto original contiene más de 34 000 registros y 48 variables, organizadas en bloques:

- **Información demográfica**: identificadores, ubicación (Bogotá D.C.), género, estrato, autoidentificación étnica, campesinidad, discapacidad.
- **Educación y compromiso**: nivel educativo, tipo de formación (virtual, híbrida, presencial), aceptación de requisitos, compromiso de dedicación horaria, disponibilidad de equipo.
- **Formación y desempeño**: presentación de prueba inicial, tiempos, eje temático, puntajes por líneas (programación, IA, datos, blockchain, nube), niveles de desempeño y competencias transversales (alfabetización de datos, colaboración, seguridad, etc.).
- **Detalles del programa**: origen del participante, estado (FORMADO, NO APROBADO, EN FORMACIÓN, INACTIVO), programa de formación, cohorte, nivel (básico, intermedio, avanzado), horas de asistencia y horas posibles.

IV-C. Problemas de calidad de los datos y observaciones iniciales

Se identificaron múltiples desafíos:

- Gran proporción de variables categóricas que requieren codificación adecuada.
- Valores nulos e incompletos en puntajes, fechas y estados, que obligan a decidir entre imputación, eliminación o recategorización.
- Desequilibrios entre categorías (por ejemplo, predominio de ciertos niveles educativos o géneros) que pueden sesgar modelos si no se tratan.
- Formatos inconsistentes (valores compuestos como "Innovador – Avanzado") que exigen normalización y descomposición.

Estos aspectos justifican una fase específica de preprocesamiento antes de cualquier modelado.

V. Preprocesamiento de Datos

El preprocesamiento se desarrolló en varios cuadernos (5.data_preparation.ipynb, 5.preprocesamiento_datos.ipynb) y permitió transformar la base cruda en un conjunto de datos limpio y listo para análisis y modelado, guardado finalmente como df_processed_final.csv con 1643 registros y 95 variables numéricas.

V-A. Limpieza: valores faltantes, outliers y duplicados

En la primera fase se eliminaron columnas irrelevantes (identificadores geográficos constantes, campos duplicados u homogéneos) y se analizaron valores faltantes y outliers:

- Se detectaron valores faltantes principalmente en Genero; se optó por eliminar las filas sin información de género en la versión final de modelado, tras confirmar su baja proporción relativa.
- Para la variable de puntaje temático (Puntaje_eje_tematico_selecionado) se utilizaron reglas basadas en el rango intercuartílico (IQR), identificando aproximadamente un 10% de outliers. Se decidió mantenerlos por corresponder a desempeños extremos plausibles en el contexto educativo.
- Se verificó la ausencia de valores NaN en el dataset final (df_processed_final.csv) antes del entrenamiento de modelos.

V-B. Feature engineering

Se crearon características derivadas con significado educativo y estadístico:

- promedio_lineas: promedio de puntajes en las cinco líneas temáticas técnicas, que resume el nivel de entrada en áreas como programación, IA, datos, blockchain y nube.
- promedio_areas: promedio de puntajes en áreas de competencias digitales generales (alfabetización, colaboración, contenidos, seguridad, solución de problemas).

- ratio_asistencia: cociente entre horas asistidas y horas posibles, como proxy de compromiso y permanencia.
- exito_academico: variable objetivo binaria (1 si el estudiante está en estado FORMADO, 0 en caso contrario).

V-C. *Reducción de dimensionalidad*

Se aplicó Análisis de Componentes Principales (PCA) sobre las variables numéricas:

- Se partió de 99 variables numéricas y se retuvo el 90% de la varianza acumulada, resultando en 36 componentes principales.
- Adicionalmente se generó una componente unidimensional (linea_pca) que sintetiza el desempeño en las cinco líneas temáticas, facilitando su interpretación.

V-D. *Transformación de datos y codificación*

Para garantizar compatibilidad con los algoritmos:

- Se estandarizaron variables continuas clave (puntajes, nivel educativo codificado, estrato, horas de asistencia, tiempo de prueba, etc.) usando StandardScaler.
- Se utilizaron codificaciones numéricas para variables categóricas (por ejemplo, género binarizado, recodificación de niveles educativos y tipos de formación) y variables one-hot para categorías múltiples (etnias, ejes temáticos, programas).

VI. Análisis Exploratorio de Datos

El EDA se realizó principalmente en 6.analisis_exploratorio.ipynb, permitiendo comprender la estructura del dataset y formular hipótesis.

VI-A. *Visualización y estadística descriptiva*

El dataset procesado cuenta con 1643 casos y 95 variables. Algunos hallazgos descriptivos:

- Distribución de género (filas con valores válidos 0/1): aproximadamente 65% hombres y 35% mujeres.
- La mayoría de participantes se ubica en estratos socioeconómicos bajos–medios (valores codificados correspondientes principalmente a estratos 2 y 3), reforzando el carácter inclusivo del programa.
- La variable exito_academico presenta una proporción cercana a 55% de éxito y 45% de no éxito, lo que indica un reto importante en términos de permanencia y aprobación.
- Las horas de asistencia normalizadas presentan media cercana a 0,85 y mediana 0,91, con un rango relativamente acotado y baja dispersión, evidenciando un grupo con alta asistencia y otro con participación mínima.

Fig. 1: Distribuciones por género, nivel educativo, estrato y éxito académico. (Gráficos exportados desde el EDA).

VI-B. *Formulación de hipótesis*

A partir de las visualizaciones y tablas de frecuencias se plantearon hipótesis como:

- Un mayor nivel educativo previo se asocia con mayor probabilidad de éxito académico.
- Mejores puntajes iniciales en el eje temático y competencias digitales predicen mejor desempeño final.
- La asistencia efectiva es el principal indicador de permanencia y éxito.
- Variables como género o estrato podrían influir en la participación, pero no necesariamente determinan el resultado formativo.

VI-C. *Identificación de patrones y hallazgos*

Se calcularon correlaciones de Pearson entre variables numéricas clave ({Puntaje_eje_tematico_selecionado, {Nivel_educacion, {Genero, {Estrato, {exito_academico), construyendo mapas de calor. Los resultados sugieren:

- Correlaciones positivas entre puntajes iniciales y {exito_academico.
- Asociación entre {ratio_asistencia (en análisis previos) y éxito, confirmando la importancia de la permanencia.
- Correlaciones más débiles entre variables puramente sociodemográficas y el éxito, lo que refuerza la idea de que las diferencias de logro se explican más por el comportamiento dentro del programa que por el origen social.

VII. Metodología

VII-A. *Selección del modelo*

El objetivo de modelado es predecir la variable binaria {exito_academico para cada estudiante. Se planteó un problema de clasificación supervisada con etiquetas disponibles, lo que

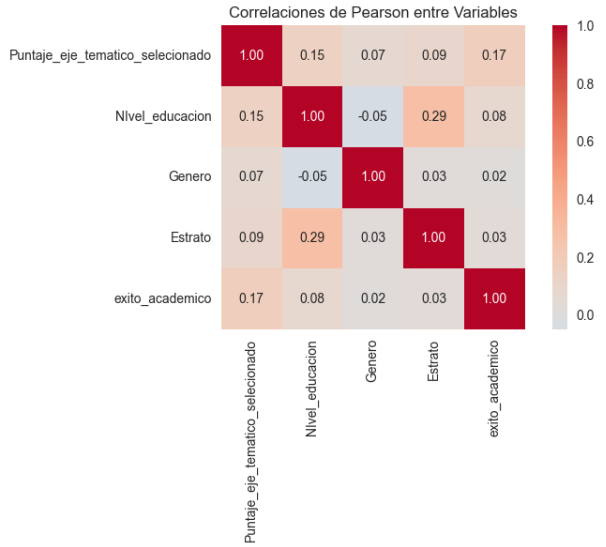

Fig. 2: Mapa de calor de correlaciones entre variables clave y éxito académico.

permite entrenar y evaluar modelos que anticipen el riesgo de no aprobación.

Se seleccionaron dos familias de modelos principales: Regresión Logística como línea base interpretable y Random Forest como modelo capaz de capturar relaciones no lineales. En una fase posterior se incorporó XGBoost como modelo de *gradient boosting* para contrastar resultados.

VII-B. Algoritmos y técnicas utilizadas

- **Regresión Logística**: modelo lineal probabilístico con penalización L1/L2 para controlar complejidad, adecuado para explicar el efecto marginal de cada variable.
- **Random Forest**: conjunto de árboles de decisión entrenados sobre subconjuntos de datos y variables, robusto ante ruido y adecuado para manejar interacciones complejas y características mixtas.
- **XGBoost**: algoritmo de *gradient boosting* que entrena árboles en secuencia, optimizando una función de pérdida y controlando sobreajuste mediante hiperparámetros como profundidad, tasa de aprendizaje y *subsampling*.
- **SelectKBest (ANOVA F)**: selección de características univariada que permite elegir las variables con mayor capacidad discriminativa respecto a la clase objetivo.

VII-C. Justificación de los hiperparámetros

En {7_metodologia_ml.ipynb y {8-9. tecnica_ml.ipynb se definieron rejillas de hiperparámetros:

- Para Regresión Logística: penalización ({l1, {l2), parámetro de regularización $C \in \{0{,}01, 0{,}1, 1, 10, 100\}$, número máximo de iteraciones y pesos de clase ({None, {balanced).
- Para Random Forest: número de árboles (100 a 1000), profundidad máxima, mínimo de muestras por división y hoja, número máximo de variables consideradas por nodo y uso o no de *bootstrap*.
- Para XGBoost: número de estimadores, profundidad de los árboles, tasa de aprendizaje, fracción de muestreo por filas y columnas y parámetro γ de regularización adicional.

Los hiperparámetros se optimizaron mediante búsqueda exhaustiva sobre una rejilla (*grid search*) combinada con selección de características SelectKBest para valores $k \in \{10, 15, 20, 25, 30\}$.

VII-D. Validación cruzada y técnicas de re-muestreo

Se dividieron los datos en entrenamiento (70%) y prueba (30%), estratificando por {exito_academico para preservar la proporción de clases observada (aproximadamente 55% éxito, 45% no éxito).

En la etapa metodológica se propuso el uso de validación cruzada estratificada de 5 particiones (*5-fold CV*) y técnicas de re-muestreo como SMOTE y pesos de clase; en la implementación concreta predominó el uso de:

- **class_weight='balanced'** en Random Forest, para penalizar más los errores en la clase minoritaria.
- **Validación sobre el conjunto de prueba independiente** tras seleccionar los mejores hiperparámetros según exactitud y AUC.

VIII. Implementación

VIII-A. Herramientas y bibliotecas utilizadas

El proyecto se implementó en Python utilizando:

- {pandas, {numpy para manipulación de datos.
- {scikit-learn para preprocesamiento, selección de características y modelos (Regresión Logística, Random Forest).
- {xgboost para el modelo XGBoost.
- {matplotlib, {seaborn para visualización.
- {FastAPI y {joblib para servir el modelo entrenado.
- {Angular para la interfaz web de predicción.

VIII-B. Estructura del código y pipeline

El pipeline completo se organiza en cuadernos Jupyter para las fases de datos ({5.data_preparation.ipynb, {5.preprocesamiento_datos.ipynb, {6.analisis_exploratorio.ipynb, {7_metodologia_ml.ipynb, {8-9. tecnica_ml.ipynb) y en un backend Python:

- El archivo {app.py carga el modelo final ({modelo_random_forest_final.pkl), el selector de características ({selector_kbest.pkl) y el listado ordenado de columnas originales ({columnas_originales.pkl).

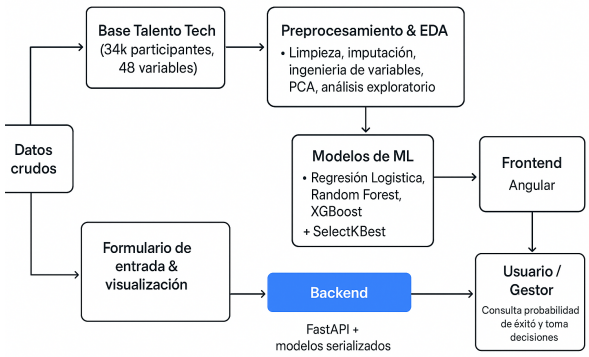

Fig. 3: Arquitectura general del pipeline: desde la base de datos hasta la API FastAPI y el frontend Angular.

- Se define un esquema de entrada ({EntradaModelo) que recoge las variables que el usuario ingresa en la interfaz (género, estrato, nivel educativo, tipo de formación, tiempo de prueba, puntaje en el eje temático, etc.).
- Una función de construcción de fila arma un {DataFrame con todas las columnas esperadas por el modelo, incluyendo codificación de categorías y uso de medias de {promedio_lineas y {promedio_areas cuando el usuario no provee dichos valores.
- El endpoint {/predict aplica el mismo {SelectKBest utilizado en entrenamiento y genera la probabilidad de éxito académico junto con la clase predicha.

VIII-C. Estrategia de experimentación

Los experimentos se estructuraron de la siguiente forma:

1. Entrenamiento de una Regresión Logística base con diferentes combinaciones de k en SelectKBest y parámetros C, penalización y pesos de clase.
2. Entrenamiento de múltiples Random Forest variando número de árboles, profundidad, mínimos de muestras y *max_features*.
3. Entrenamiento de una familia de modelos XGBoost con rejillas amplias para sus hiperparámetros principales.
4. Selección del mejor modelo de cada familia según exactitud y AUC.
5. Exportación del mejor modelo global junto con el selector y las columnas.

IX. Resultados y Evaluación

IX-A. Métricas de rendimiento

En el conjunto de prueba se evaluaron los modelos usando exactitud (Accuracy), área bajo la curva ROC (AUC) y matriz de confusión. Adicionalmente, a partir de la matriz de confusión pueden derivarse precisión, recobrado (Recall) y F1-Score.

TABLE I: Resumen de métricas en el conjunto de prueba.

Modelo	Accuracy	AUC	Comentario
Baseline (clase mayoritaria)	0,55	0,50	Predice siempre “éxito”
Regresión Logística	0,63	0,63	30 variables (L1)
Random Forest	**0,64**	**0,65**	30 variables, 400 árboles
XGBoost	0.64	0.64	30 variables, mejores hiperparámetros

Para la mejor Regresión Logística, se obtuvo una exactitud aproximada de 0,63 y AUC de 0,63, con una matriz de confusión que muestra un compromiso entre identificación de estudiantes exitosos y no exitosos. El mejor Random Forest alcanzó una exactitud cercana a 0,64 y AUC alrededor de 0,65, con matriz de confusión balanceada, superando ligeramente a la Regresión Logística.

IX-B. Rendimiento base vs. modelo

La línea base correspondiente a un clasificador que siempre predice la clase mayoritaria ({exito_academico=1) presenta una exactitud de aproximadamente 0,55 y AUC de 0,5. Los modelos entrenados superan significativamente esta referencia:

- Incremento de alrededor de 8–9 puntos porcentuales en exactitud con Random Forest.
- Mejora en la capacidad discriminativa (AUC > 0,6), lo que indica que el modelo es capaz de distinguir mejor entre estudiantes exitosos y no exitosos que un clasificador aleatorio.

IX-C. Visualización de resultados

Para analizar el comportamiento del modelo se propone:

- Visualizar la matriz de confusión para comprender los tipos de error (falsos positivos y falsos negativos).
- Graficar la curva ROC y comparar los modelos, evaluando el compromiso entre sensibilidad y especificidad.
- Examinar la importancia de variables en Random Forest para identificar los factores más influyentes en la predicción del éxito.

X. Interpretación de Resultados y Hallazgos

X-A. Significado de los resultados

Los resultados confirman que:

- La asistencia efectiva y los puntajes iniciales son variables clave para anticipar el éxito académico.
- El nivel educativo previo tiene efecto, pero no es determinante por sí solo; estudiantes de diferentes niveles pueden alcanzar éxito cuando participan de forma sostenida.
- Las variables demográficas puras (género, estrato, etnia) muestran menor poder predictivo directo sobre el desempeño, aunque sí pueden influir en la probabilidad de participación o en las condiciones de acceso.

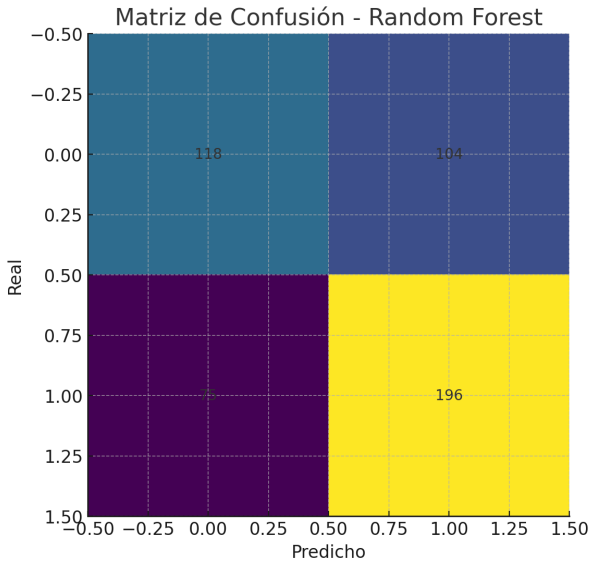

Fig. 4: Matriz de confusión del modelo Random Forest en el conjunto de prueba.

X-B. Implicaciones en el dominio del negocio

Para la gestión de Talento Tech Bogotá, estos hallazgos sugieren:

- Implementar sistemas de alerta temprana basados en asistencia y progreso inicial, priorizando acompañamientos a quienes muestran señales tempranas de riesgo.
- Diseñar estrategias de nivelación para participantes con puntajes iniciales bajos, en lugar de restringir el acceso por nivel educativo formal.
- Utilizar los modelos como herramientas de apoyo para asignar tutores o recursos adicionales a grupos con mayor probabilidad de deserción.

X-C. Consideraciones éticas, justicia y sesgos

Es fundamental que los modelos no reproduzcan ni amplifiquen sesgos:

- Las predicciones deben utilizarse para ofrecer apoyos adicionales, no para excluir o limitar oportunidades.
- Es necesario monitorear el desempeño del modelo por subgrupos (género, estrato, etnia) para detectar posibles diferencias injustificadas.
- Se recomienda actualizar y recalibrar periódicamente los modelos con datos recientes y decisiones humanas informadas.

XI. Conclusiones y Trabajos Futuros

XI-A. Resumen de los logros

Este proyecto integró de forma completa el ciclo de ciencia de datos sobre el programa Talento Tech Bogotá: desde el entendimiento y preprocesamiento de una base compleja, pasando por el análisis exploratorio y la construcción de modelos predictivos, hasta su despliegue en una aplicación web interactiva. Se comprobó la relevancia de variables de comportamiento (asistencia y puntajes iniciales) frente a factores puramente sociodemográficos.

XI-B. Desafíos presentados

Entre los principales retos se encuentran:

- Manejar adecuadamente valores faltantes y formatos heterogéneos en variables categóricas.
- Equilibrar complejidad del modelo y interpretabilidad para que los resultados sean útiles a los responsables del programa.
- Limitar el sobreajuste en rejillas de hiperparámetros amplias con tamaño de muestra moderado.

XI-C. Recomendaciones de mejora

Para etapas futuras se recomienda:

- Incorporar métricas adicionales (F1-Score, Recall por clase, métricas específicas de subgrupos) en la optimización de modelos.
- Explorar técnicas de explicabilidad de modelos (por ejemplo, SHAP) para entender con mayor detalle el efecto de cada variable.
- Integrar variables contextuales adicionales (por ejemplo, información sobre oferta de cursos, horarios, modalidad de acompañamiento).

XI-D. Ideas para trabajos posteriores o despliegue real

Algunas líneas de trabajo futuro incluyen:

- Construir un sistema de recomendación de rutas formativas personalizadas, basado en el perfil y desempeño inicial.
- Integrar el modelo en el sistema de gestión académico real del programa, con retroalimentación continua y dashboards para coordinadores.
- Extender el análisis a otras cohortes o programas similares en Colombia para comparar resultados y transferir aprendizajes.

Apéndices

Apéndice A. Diseños de módulos

En este apéndice se pueden incluir diagramas de módulos del backend FastAPI y del frontend Angular, mostrando componentes, servicios y flujos de datos.

Apéndice B. Tablas y gráficos detallados

Se recomienda incorporar:

- Tablas de distribución detallada por cohorte, temática y nivel.
- Gráficos de PCA a dos dimensiones distinguiendo estudiantes exitosos y no exitosos.
- Tablas comparativas de importancia de variables en Random Forest.

Apéndice C. Diccionario de datos

Listado detallado de variables originales y derivadas, con descripción, tipo de dato, dominio de valores y transformaciones aplicadas.

REFERENCES

[1] S. Joksimović, et al., "Learning Analytics to Improve Student Outcomes in Higher Education," in *Handbook of Learning Analytics*, 2nd ed., 2022.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016.

[4] Ministerio de Tecnologías de la Información y las Comunicaciones (MinTIC), "Programa Misión TIC 2022," 2022.

[5] Tecnalia Colombia y Secretaría de Desarrollo Económico de Bogotá, "Programa Talento Tech Bogotá: Documento de Diseño y Lineamientos de Formación," 2023.