Explorando oportunidades: un enfoque de machine learning para la recomendación personalizada de carreras y programas educativos en Colombia

Erika Milena Monroy Lozano

Asesor

Eduardo Sánchez Sandoval

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencias de Datos y Analítica

2025

Dedicatoria

A mis padres, por su amor incondicional y apoyo constante, por enseñarme que la educación es el mejor camino hacia el futuro.

A mis hermanos, con la esperanza de ser fuente de inspiración en sus vidas. Que este logro les recuerde que pueden soñar en grande y que todo lo que lleguen a soñar, lo pueden cumplir. Nunca dejen de creer en ustedes mismos.

A mi pareja, por sus consejos y apoyo para aterrizar esta idea. Por creer en mí cuando dudaba, por inspirarme a ser mi mejor versión, y por enseñarme que el aprendizaje es un camino sin final. Tu apoyo incondicional ha sido el motor de este logro.

A mis gatitos, fieles compañeros de las noches largas de estudio, testigos silenciosos de cada línea de código y cada página escrita.

A los estudiantes de las escuelas rurales y urbanas de Colombia, cuyas historias y sueños me motivaron a buscar respuestas.

Que este proyecto sea un puente para que descubran sus talentos y construyan el futuro profesional que merecen. Este trabajo es para todos ustedes, que, de diferentes formas, hicieron posible lo imposible.

Agradecimientos

En primer lugar, agradezco a Dios por ser mi guía constante y por hacer posible todo lo que me he propuesto a lo largo de mi vida, incluyendo la oportunidad de seguir aprendiendo y creciendo profesionalmente. Su presencia ha sido fundamental en cada paso de este camino.

Expreso mi sincero agradecimiento a la Universidad Nacional Abierta y a Distancia (UNAD) y al programa de Especialización en Ciencia de Datos y Analítica, por brindarme las herramientas teóricas y prácticas necesarias para desarrollar este proyecto y por formarme como profesional en esta apasionante área del conocimiento.

A mi pareja, por su apoyo incondicional, paciencia infinita y motivación constante durante todo este proceso. Tu fe en mí y en este proyecto fue fundamental para convertir una idea en realidad.

Resumen

El presente proyecto desarrolla TalentAI, un sistema de recomendación inteligente basado en Machine Learning para la orientación vocacional de estudiantes de grado 10° y 11° de instituciones educativas de Bogotá D.C. en la elección de programas de educación superior. La investigación responde a la problemática de la deserción educativa en el primer año de estudios superiores en Colombia, la cual, según el Laboratorio de Economía de la Educación (LEE, 2023), alcanzó un 45,4% en programas universitarios y un 48,2% en programas tecnológicos. A esto se suma que, en Bogotá, la tasa de jóvenes que ni estudian ni trabajan (NINIs) oscila entre el 19% y el 22%, y que más del 60% de los estudiantes de grado 11 no reciben una orientación vocacional estructurada. Estas cifras evidencian la magnitud del problema y justifican la necesidad de un sistema que optimice la toma de decisiones académicas.

La metodología implementada combina un enfoque mixto, con análisis cuantitativo mediante la evaluación comparativa de cinco algoritmos de Machine Learning: K-Nearest Neighbors (KNN), Redes Neuronales, Random Forest, XGBoost y Regresión Logística. Se entrenaron los modelos con 20.000 registros sintéticos de estudiantes, caracterizados por 13 variables predictoras (5 puntajes ICFES y 8 dimensiones de competencias), para predecir 30 áreas de conocimiento. Los datos de programas educativos fueron obtenidos del Sistema Nacional de Información de la Educación Superior (SNIES) y del SENA, identificando 46.900 programas a nivel nacional y filtrando 5.138 correspondientes a Bogotá D.C., mediante scripts automatizados de extracción y limpieza de datos.

Como resultado, los modelos KNN y Redes Neuronales fueron seleccionados para su implementación en la plataforma, al evidenciar el mejor equilibrio entre precisión y

eficiencia computacional frente a las demás alternativas evaluadas. En las pruebas realizadas, el modelo KNN obtuvo un accuracy de 0.660 y un F1-Score Macro de 0.626 con un tiempo de ejecución de 9.1 segundos, destacándose por su rapidez y consistencia; mientras que la Red Neuronal alcanzó un accuracy de 0.666 y un F1-Score Macro de 0.583 en 40.9 segundos, consolidándose como el de mayor capacidad predictiva. El sistema desarrollado incluye un formulario de evaluación de 100 competencias agrupadas en 8 dimensiones, una interfaz web intuitiva para la interacción con estudiantes y un motor de recomendación que sugiere programas educativos personalizados en función del perfil multidimensional del estudiante.

Palabras clave: Machine Learning, orientación, neuronales, educación, vocacional.

Abstract

This project presents TalentAI, an intelligent recommendation system based on Machine Learning for vocational guidance of 10th and 11th grade students in Bogotá D.C., aimed at supporting the decision-making process when choosing higher education programs. The research addresses the problem of student dropout during the first year of higher education in Colombia, which, according to the Laboratorio de Economía de la Educación (LEE, 2023), reached 45.4% in university programs and 48.2% in technological programs. In Bogotá, this challenge is compounded by the fact that between 19% and 22% of young people are classified as NEETs (Not in Education, Employment, or Training), and more than 60% of high school seniors do not receive structured vocational guidance.

The methodology combines a mixed approach, with quantitative analysis through the comparative evaluation of five Machine Learning algorithms: K-Nearest Neighbors (KNN), Neural Networks, Random Forest, XGBoost, and Logistic Regression. The models were trained with 20,000 synthetic student records characterized by thirteen predictive variables (5 ICFES test scores and eight competency dimensions) to predict thirty knowledge areas. Academic program data were obtained from the National Higher Education Information System (SNIES) and SENA, identifying 46.900 programs nationwide and filtering 2,281 corresponding to Bogotá D.C., through automated data extraction and cleaning scripts.

The experimental results demonstrated that KNN achieved an accuracy of 0.666 and a weighted F1-macro of 0.626 with an execution time of 9.1 seconds, while Neural Networks reached the highest accuracy of 0.666 and a weighted F1-macro of 0.583 in 40.5 seconds, thus evidencing the best balance between predictive performance and computational efficiency. Consequently, these two models were selected for implementation

in the final platform, which integrates an assessment form of 100 competencies organized in 8 dimensions, an intuitive web interface for student interaction, and a recommendation engine that suggests personalized academic programs according to each student's multidimensional profile. This work contributes to bridging the gap in vocational orientation tools in Latin America, offering a scalable, data-driven solution for educational decision-making.

Keywords: Machine Learning, orientation, neural, education, vocational.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Introducción

La elección de un programa académico en la educación media constituye un punto de inflexión en la trayectoria personal y profesional de los estudiantes. Sin embargo, en Colombia este proceso enfrenta múltiples limitaciones, como la ausencia de orientación vocacional estructurada, la falta de información sobre la oferta educativa y las brechas entre las aspiraciones juveniles y las demandas del mercado laboral (Laboratorio de Economía de la Educación [LEE], 2023; Rodríguez, 2022). Estas dificultades se reflejan en tasas de deserción universitaria superiores al 45% durante el primer año y en el aumento del número de jóvenes que ni estudian ni trabajan (NINIs), fenómeno que afecta entre el 19% y 22% de la población juvenil en Bogotá (LEE, 2024; Secretaría de Educación de Bogotá, 2023).

En este escenario, las tecnologías de la información y, en particular, la inteligencia artificial (IA) y el aprendizaje automático (Machine Learning), representan oportunidades clave para transformar los procesos de orientación vocacional. Diversas investigaciones han mostrado que los sistemas de recomendación basados en IA permiten personalizar las sugerencias educativas de acuerdo con intereses, competencias y valores de cada estudiante, mejorando así la pertinencia de las decisiones académicas (Smith & Johnson, 2018; Singh & Dhir, 2021; Jamieson & O'Mara, 2022). Además, la literatura evidencia que la integración de tecnologías inteligentes en la orientación educativa no solo incrementa la satisfacción estudiantil, sino que también contribuye a reducir la deserción en los primeros semestres (Tinto, 2017; Bennett, 2018; Kift, 2018).

No obstante, en Colombia los avances en esta materia siguen siendo incipientes. Estudios recientes destacan que más del 60% de los estudiantes de grado 11 carecen de un acompañamiento vocacional estructurado, lo que limita la elección informada de programas

académicos (Roa, 2023; Morinson, 2020). Si bien existen iniciativas como el programa "Reto a la U" (Rodríguez Cardozo, 2022), y políticas públicas como "Jóvenes a la U" y "Matrícula Cero" promovidas por la Alcaldía de Bogotá y el Gobierno Nacional, que han ampliado el acceso a la educación superior o la oferta educativa del SENA.

Estas iniciativas aún no resuelven de fondo la falta de acompañamiento vocacional informado porque aún no cuentan con herramientas tecnológicas avanzadas que integren analítica de datos y modelos predictivos (SENA, 2023; Ministerio de Educación Nacional de Colombia, 2024). Esto evidencia una brecha significativa entre el potencial de la inteligencia artificial y las necesidades reales de los jóvenes al momento de tomar decisiones críticas sobre su futuro.

Con el fin de atender esta problemática, el presente proyecto propone el diseño e implementación de TalentAI, un sistema de recomendación vocacional basado en Machine Learning, orientado a estudiantes de grado 10° y 11° de instituciones educativas de Bogotá. Para ello, se entrenaron y evaluaron cinco modelos de clasificación Redes Neuronales, K-Nearest Neighbors (KNN), Random Forest, XGBoost y Regresión Logística utilizando un dataset sintético de 20.000 registros de estudiantes con 13 variables predictoras, además de información sobre 46.900 programas académicos registrados en Colombia y 5.138 filtrado específicamente para Bogotá a partir de datos del SNIES y el SENA. Tras el análisis comparativo, los modelos KNN y Redes Neuronales fueron seleccionados por su mejor balance entre rendimiento predictivo y eficiencia computacional, alcanzando métricas de accuracy cercanas al 66% y F1-Score ponderado superiores a 0.65, con tiempos de entrenamiento adecuados para un sistema en producción.

Descripción del Problema

La transición de la educación media a la educación superior en Colombia enfrenta múltiples desafíos relacionados con la deserción temprana, la falta de orientación vocacional y la escasa alineación entre los intereses del estudiante y la oferta académica disponible. Según el Laboratorio de Economía de la Educación (LEE, 2023), cerca del 45,4% de los estudiantes de programas universitarios y el 48,2% de los programas tecnológicos abandonan sus estudios durante el primer año, lo que representa un impacto significativo en la inversión familiar, institucional y social. A esta problemática se suma que, en Bogotá, entre el 19% y el 22% de los jóvenes son clasificados como NINIs (ni estudian ni trabajan), de los cuales el 67% son mujeres (LEE, 2024; Rodríguez, 2022). Estos datos evidencian que la orientación vocacional insuficiente constituye un factor determinante en la construcción de trayectorias educativas inestables y en la pérdida de capital humano.

Planteamiento del Problema

La elección de un programa de educación superior constituye una de las decisiones más relevantes en la trayectoria académica de los estudiantes de bachillerato en Colombia. Esta decisión, que idealmente debería fundamentarse en una evaluación objetiva de intereses, competencias y oportunidades, se ve afectada por la ausencia de mecanismos efectivos de orientación vocacional personalizada y por la limitada disponibilidad de información actualizada sobre la oferta educativa y el mercado laboral.

Estudios recientes (Roa, 2023; Laboratorio de Economía de la Educación, 2022) evidencian que más del 60 % de los estudiantes de grado 11 en Bogotá no recibe orientación vocacional estructurada, lo cual limita su capacidad de tomar decisiones informadas y eleva los niveles de deserción en la educación superior. Esta falta de acompañamiento no solo impacta la continuidad académica, sino que también contribuye al aumento de jóvenes en condición NINI (ni estudian ni trabajan), fenómeno que afecta aproximadamente al 26 % de la población entre 14 y 28 años en Colombia, es decir, 3.2 millones de personas, de las cuales el 67 % son mujeres (DANE, 2023; GEIH, 2021). En Bogotá, aunque la tasa de NINIs es menor (19 %–22 %), continúa siendo un indicador preocupante.

Si bien programas y políticas como “Reto a la U”, “Jóvenes a la U” o “Matrícula Cero” han ampliado las oportunidades de acceso a la educación superior (Rodríguez Cardozo, 2022; Secretaría de Educación de Bogotá, 2023), estos esfuerzos no resuelven la raíz del problema: la falta de criterios claros y fundamentados para elegir un programa académico pertinente. Incluso iniciativas del SENA, que ofrece programas técnicos y tecnológicos gratuitos, se ven limitadas por la ausencia de herramientas que permitan a los

jóvenes contrastar sus intereses con la pertinencia laboral de las diferentes opciones (SENA, 2023).

La orientación vocacional tradicional no logra responder con agilidad a los cambios del mercado ni a la emergencia de nuevas áreas profesionales. La dispersión y desactualización de la información coloca a los estudiantes en una situación vulnerable, en la que se enfrentan dos escenarios adversos: (i) abandonar sus estudios al no sentirse identificados con el programa elegido, o (ii) egresar con un perfil académico que no se ajusta a las necesidades del mercado laboral. Ambos casos representan una pérdida significativa de capital humano y limitan el desarrollo social y económico del país.

De acuerdo con el Laboratorio de Economía de la Educación (2023), la tasa de deserción en programas universitarios alcanza el 45,4 % durante el primer año, y en programas tecnológicos llega al 48,2 %. Estos datos, sumados a la falta de acompañamiento vocacional estructurado, demuestran la magnitud de la problemática y justifican la necesidad de diseñar soluciones tecnológicas innovadoras que faciliten la toma de decisiones académicas fundamentadas.

En este escenario, surge la necesidad de desarrollar un sistema que, a partir de datos cuantitativos sobre competencias y resultados académicos, utilice algoritmos de Machine Learning para ofrecer recomendaciones personalizadas sobre programas técnicos, tecnológicos y profesionales. El proyecto busca orientar a los estudiantes de manera precisa, oportuna y contextualizada, contribuyendo a reducir los índices de deserción y la desinformación académica, así como a fortalecer el vínculo entre educación y empleabilidad juvenil.

Pregunta de Investigación

¿Cómo puede un sistema de recomendación basado en Machine Learning, implementado con modelos de Redes Neuronales y K-Nearest Neighbors, contribuir a mejorar la orientación vocacional de los estudiantes de grado 10° y 11° en Bogotá y, con ello, reducir los niveles de deserción y desinformación académica en la educación superior?

Sistematización del Problema

Derivado de la pregunta central de investigación, se plantean los siguientes interrogantes específicos:

¿Cuáles son los principales factores que explican la deserción y la desinformación vocacional en los estudiantes de educación media en Bogotá?

¿Qué variables académicas, socioeconómicas y de competencias individuales resultan más relevantes para predecir afinidad hacia un área del conocimiento?

¿Qué técnicas de *Machine Learning* (como Redes Neuronales y *K-Nearest Neighbors*) presentan un mejor desempeño en términos de precisión, confiabilidad y escalabilidad para la recomendación vocacional?

¿Cómo puede diseñarse una interfaz web accesible e intuitiva que permita a los estudiantes interactuar con el sistema y comprender de manera clara las recomendaciones recibidas?

¿Qué impacto potencial tendría la implementación de esta herramienta en la toma de decisiones académicas, en la reducción de los índices de deserción y en la articulación con las demandas del mercado laboral?

¿Cuáles son las limitaciones técnicas, éticas y de protección de datos personales que deben considerarse en el desarrollo y aplicación de este sistema con población estudiantil menor de edad?

Justificación

La elección de una carrera profesional constituye una de las decisiones más significativas en la vida de un estudiante, pues impacta directamente en su proyecto de vida, sus oportunidades de desarrollo y su aporte a la sociedad. En Colombia, este proceso se ve afectado por limitaciones como la falta de acompañamiento especializado, la desinformación sobre la oferta académica y la ausencia de herramientas personalizadas de orientación, lo que incrementa la probabilidad de deserción en la educación superior. Según el Laboratorio de Economía de la Educación (LEE, 2023), la tasa de deserción alcanza el 45,4 % en programas universitarios y el 48,2 % en programas tecnológicos durante el primer año, mientras que más del 60 % de los estudiantes de grado 11 no recibe orientación vocacional estructurada (Roa, 2023).

En este contexto, las tecnologías de la información, y en particular la inteligencia artificial (IA) y el aprendizaje automático (Machine Learning), representan una oportunidad estratégica para innovar en los procesos de orientación vocacional. Investigaciones recientes evidencian que los sistemas de recomendación potenciados con IA pueden personalizar la orientación educativa, mejorar la afinidad entre estudiantes y programas académicos y reducir la deserción temprana (Smith & Johnson, 2018; Singh & Dhir, 2021; Jamieson & O'Mara, 2022). Asimismo, estudios internacionales destacan que la integración de estas herramientas en la educación favorece la toma de decisiones informadas y fortalece la relación entre competencias individuales y necesidades del mercado laboral (Hooley, 2017; Jackson, 2019; Kristof-Brown, Zimmerman & Johnson, 2005).

El presente proyecto se justifica por su contribución a la innovación tecnológica y educativa en Colombia mediante el desarrollo de TalentAI, un sistema de recomendación

vocacional que articula la analítica de datos con una interfaz web accesible y una base sólida de modelos predictivos. Se evaluaron varios algoritmos de aprendizaje automático, entre ellos Redes Neuronales, K-Nearest Neighbors, Random Forest, XGBoost y Regresión Logística. De esta comparación se eligieron los dos primeros, ya que ofrecieron el mejor balance entre precisión y eficiencia, con resultados viables para su implementación en el contexto colombiano.

Para ello, se consideraron más de 46.900 programas académicos a nivel nacional y se focalizó el sistema en 5.138 correspondientes a Bogotá, lo que permite asegurar la pertinencia de las recomendaciones frente a la oferta real disponible.

De esta manera, TalentAI constituye una herramienta pertinente y necesaria para estudiantes, instituciones educativas y responsables de política pública, ya que ofrece un soporte objetivo, escalable y basado en evidencia para la toma de decisiones académicas. Su aporte no solo se orienta a reducir problemáticas sociales como la deserción universitaria y la población juvenil en condición de NINI (LEE, 2024; Rodríguez, 2022), sino también a fortalecer la investigación aplicada en Ciencias de Datos y Analítica en el campo educativo, aportando un modelo replicable en otras regiones del país.

Objetivos

Objetivo General

Desarrollar un sistema de recomendación basado en técnicas de Machine Learning que oriente a estudiantes de bachillerato en Colombia en la elección de programas de educación superior, considerando sus intereses individuales y las demandas del mercado laboral.

Objetivos Específicos

Analizar la oferta de programas educativos y los perfiles estudiantiles para construir una base de datos estructurada que sustente el diseño del sistema de recomendación.

Diseñar un modelo de Machine Learning que prediga la afinidad entre los perfiles estudiantiles y las opciones educativas, utilizando métricas como precisión y F1-score.

Validar el modelo desarrollado mediante pruebas de desempeño técnico con bases de datos de prueba sintéticas y contextualmente relevantes para el ámbito educativo colombiano.

Evaluar la usabilidad del sistema a través de una interfaz web intuitiva, accesible y orientada a la interacción efectiva con los estudiantes.

Marco de Referencia

Estado del Arte

El desarrollo de sistemas de recomendación aplicados a la orientación vocacional ha cobrado relevancia en la última década como una alternativa innovadora para enfrentar los altos índices de deserción en la educación superior. Diversas investigaciones han demostrado que la integración de la inteligencia artificial (IA) y el Machine Learning permite personalizar las recomendaciones educativas y mejorar la correspondencia entre el perfil del estudiante y las opciones académicas disponibles (Smith & Johnson, 2018; Singh & Dhir, 2021).

En el ámbito internacional, Jamieson y O'Mara (2022) resaltan que los sistemas potenciados con IA han mejorado la precisión en la orientación vocacional mediante la utilización de grandes volúmenes de datos, favoreciendo la identificación de patrones en los intereses y habilidades estudiantiles. Asimismo, Hooley y Watts (2016) destacan la evolución de la orientación hacia modelos digitalizados, los cuales permiten escalar el acompañamiento a poblaciones más amplias.

En Latinoamérica se han identificado esfuerzos significativos. El Ministerio de Educación del Perú implementó en 2021 el Sistema Nacional de Orientación Vocacional, una plataforma digital que articula información académica y ocupacional para apoyar a los estudiantes en su elección de carrera (PerúEduca, 2021). En Panamá, Morinson Negrete (2020) evaluó un sistema de información para orientación vocacional en estudiantes de secundaria, reportando un nivel alto de satisfacción por parte de los usuarios, aunque con limitaciones en la personalización de resultados.

En Colombia, si bien existen iniciativas como el Observatorio Laboral para la Educación (Ministerio de Educación Nacional, 2024) y programas institucionales como "Jóvenes a la U"

(Secretaría de Educación de Bogotá, 2023), la mayoría de las estrategias de orientación vocacional permanecen en un nivel informativo y carecen de componentes de personalización basados en analítica de datos. Estudios recientes evidencian la necesidad de fortalecer la articulación entre los perfiles estudiantiles y la oferta educativa mediante herramientas tecnológicas avanzadas (Roa, 2023; Rodríguez, 2022).

En este sentido, el presente proyecto contribuye a cerrar la brecha entre los avances internacionales en sistemas de recomendación y la realidad nacional, al proponer una plataforma de orientación vocacional basada en modelos de Machine Learning particularmente Redes Neuronales y K-Nearest Neighbors, que combina métricas de rendimiento predictivo con una interfaz accesible para estudiantes de grado 10° y 11°.

En el caso específico de Bogotá, pese a que existen 5.138 programas de educación superior registrados en el SNIES y el SENA, aún no se dispone de plataformas que integren dicha información con perfiles estudiantiles mediante modelos predictivos avanzados, lo que refuerza la pertinencia del presente proyecto

Marco Contextual

La elección de programas de educación superior en Colombia se desarrolla en un contexto complejo, caracterizado por altas tasas de deserción, dificultades de orientación vocacional estructurada y una brecha creciente entre formación y mercado laboral.

En términos de permanencia, los informes del Laboratorio de Economía de la Educación (LEE, 2023) evidencian que cerca del 45,4 % de los estudiantes de programas universitarios y el 48,2 % de los matriculados en programas tecnológicos abandonan sus estudios durante el primer año. Esta tendencia refleja limitaciones en los mecanismos de acompañamiento y orientación al

ingreso, lo cual incide en la falta de correspondencia entre las expectativas estudiantiles y las exigencias académicas.

De manera paralela, el fenómeno de los jóvenes que no estudian ni trabajan (NINIs) constituye un reto estructural. Según el mismo laboratorio, la proporción de jóvenes en esta condición en Bogotá oscila entre el 19 % y el 22 %, con una marcada sobrerrepresentación de mujeres (Noticia, s. f.; Rodríguez, 2022). Esta situación no solo limita el desarrollo profesional de los jóvenes, sino que además representa un desafío para la competitividad económica y social del país (LEE, 2024).

En cuanto a la orientación vocacional, diversos estudios señalan que más del 60 % de los estudiantes de grado 11 en Bogotá no reciben un acompañamiento estructurado para elegir programas de educación superior (Roa, 2023). Aunque existen iniciativas públicas como el programa Jóvenes a la U y la política de Matrícula Cero (Secretaría de Educación de Bogotá, 2023), estas estrategias están más orientadas a la financiación y acceso que a la personalización de la elección académica.

Por otro lado, la disponibilidad de información sobre programas educativos en Colombia es amplia. El Sistema Nacional de Información de la Educación Superior (SNIES), el Servicio Nacional de Aprendizaje (SENA) y el Observatorio Laboral para la Educación del Ministerio de Educación consolidan datos relevantes sobre la oferta académica, matrícula, empleabilidad y condiciones laborales de los egresados (MEN, 2024; SNIES, s. f.; SENA, 2023). Sin embargo, el aprovechamiento de estos repositorios suele limitarse a la consulta manual y exploratoria, sin herramientas que traduzcan dicha información en recomendaciones personalizadas para los estudiantes.

En consecuencia, el contexto educativo de Bogotá y Colombia presenta una paradoja: hay abundancia de datos y programas de apoyo, pero falta la articulación inteligente entre la información y las necesidades particulares de los estudiantes. Esta brecha justifica la pertinencia de sistemas de recomendación basados en Machine Learning, que integren resultados académicos, competencias individuales y tendencias del mercado laboral, con el fin de fortalecer la toma de decisiones vocacionales y contribuir a la reducción de la deserción educativa y el desempleo juvenil.

Así, aunque el SNIES y el SENA consolidan más de 46.900 programas académicos a nivel nacional, la focalización en Bogotá muestra 5.138 programas, lo que evidencia la necesidad de herramientas que orienten a los estudiantes en una oferta amplia pero fragmentada.

Marco Teórico

Fundamentos de la Orientación Vocacional

La orientación vocacional constituye un campo de estudio fundamental en el desarrollo educativo y profesional de los individuos. Parsons (1909) estableció los primeros lineamientos al señalar que la elección profesional debía basarse en el conocimiento de uno mismo, el conocimiento de las ocupaciones y la relación lógica entre ambos elementos. Posteriormente, la teoría tipológica de Holland (1997) planteó seis tipos de personalidad (RIASEC: Realista, Investigador, Artístico, Social, Emprendedor y Convencional), cuyo análisis permite identificar afinidades hacia campos ocupacionales específicos. A su vez, Super (1990) desarrolló la teoría del desarrollo vocacional, que concibe la elección de carrera como un proceso dinámico y evolutivo a lo largo de la vida.

En un plano más contemporáneo, estudios como el de Kristof-Brown, Zimmerman y Johnson (2005) evidencian que la correspondencia entre competencias individuales y

características del entorno académico o laboral favorece la satisfacción y el rendimiento. Asimismo, Tinto (2017) y Hooley (2017) subrayan que la ausencia de acompañamiento vocacional estructurado incrementa los riesgos de deserción universitaria y de inserción laboral inadecuada. Estas perspectivas sustentan la necesidad de implementar herramientas modernas que integren información sobre intereses, competencias y oferta educativa, propósito central del presente proyecto.

Fundamentos de Machine Learning Aplicados a la Orientación

El aprendizaje automático (Machine Learning, ML), rama de la inteligencia artificial, se centra en el desarrollo de algoritmos capaces de mejorar su rendimiento a partir de la experiencia (Russell & Norvig, 2016). Dentro de sus técnicas, los modelos de clasificación supervisada son ampliamente utilizados en sistemas de recomendación, pues permiten asociar a cada estudiante con la categoría o programa más acorde con su perfil.

Entre los algoritmos más relevantes se encuentran:

1. K-Nearest Neighbors (KNN): clasifica en función de la similitud con instancias cercanas, destacándose por su simplicidad y eficiencia en aplicaciones educativas (Smith & Johnson, 2018).
2. Redes Neuronales Artificiales (ANN): modelan relaciones complejas y no lineales entre variables, logrando altos niveles de precisión en problemas de recomendación (Singh & Dhir, 2021).
3. Random Forest y XGBoost: basados en ensambles de árboles de decisión, presentan robustez frente al sobreajuste y alta capacidad de generalización (Ozdemir, 2016).
4. Regresión Logística: sirve como modelo base en tareas de clasificación, útil para establecer comparaciones respecto a modelos más avanzados (Garriga Trillo, 2012).

Investigaciones recientes han demostrado que la integración de estos algoritmos en la orientación vocacional mejora la pertinencia de las recomendaciones y la satisfacción de los estudiantes (Jamieson & O'Mara, 2022). Sin embargo, en Latinoamérica el uso de ML en este campo aún es incipiente, lo que evidencia la pertinencia de su aplicación en el contexto colombiano.

Métricas de Evaluación de Modelos de Machine Learning

La evaluación de modelos predictivos exige criterios que valoren tanto la exactitud global como la calidad de las predicciones en clases desbalanceadas y la eficiencia computacional (Russell & Norvig, 2016).

1. Accuracy: mide el porcentaje de aciertos globales, aunque puede ser engañoso en dataset con múltiples clases desiguales (Ricci, Rokach & Shapira, 2011).

2. Precision y Recall: permiten valorar la capacidad de identificar correctamente las clases, reduciendo falsos positivos y negativos.

3. F1-Score: combina ambas métricas en una media armónica. En este proyecto se emplearon dos variantes:

4. F1-Score Macro: promedia el rendimiento de todas las clases por igual, útil en contextos con distribución heterogénea.

5. F1-Score Weighted: pondera el rendimiento por la frecuencia de cada clase, reduciendo el sesgo en categorías minoritarias (Zhang, 2018).

6. Tiempo de Entrenamiento: criterio clave en aplicaciones prácticas, pues la eficiencia computacional determina la escalabilidad del sistema y la experiencia del usuario (Shneiderman, 2010).

A partir de estos criterios, los resultados experimentales mostraron que las Redes Neuronales alcanzaron un accuracy de 0.663 y un F1-Score Weighted de 0.659 en 82.5 segundos, consolidándose como el modelo de mayor capacidad predictiva. En contraste, KNN obtuvo un accuracy de 0.660 y un F1-Score Weighted de 0.656 en apenas 9.2 segundos, destacándose por su rapidez y consistencia. Por ello, ambos modelos fueron seleccionados para la implementación en la plataforma desarrollada, ofreciendo un balance entre precisión y eficiencia que garantiza su aplicabilidad en entornos educativos.

Marco Conceptual

Orientación Vocacional: Proceso pedagógico y psicológico que busca acompañar a los estudiantes en la identificación de sus intereses, habilidades y valores, con el fin de guiar la elección de un programa académico o carrera profesional coherente con su proyecto de vida (Rodríguez, 2022).

Deserción Educativa: Fenómeno que ocurre cuando un estudiante abandona sus estudios antes de finalizar el programa académico en el que está matriculado. En el contexto colombiano, suele analizarse especialmente en el primer año de educación superior, dado su impacto en la continuidad formativa y en la inversión social y económica (LEE, 2023).

NINIs: Término que describe a los jóvenes que no estudian ni trabajan (Not in Education, Employment or Training). Se considera un indicador de vulnerabilidad social y educativa, pues refleja la falta de integración de esta población en los sistemas de formación y en el mercado laboral (Secretaría de Educación de Bogotá, 2023).

Machine Learning (ML): Subcampo de la inteligencia artificial que permite a las computadoras aprender patrones a partir de datos, sin necesidad de ser programadas explícitamente para cada tarea (Russell & Norvig, 2016).

Sistema De Recomendación: Aplicación tecnológica que utiliza algoritmos de análisis de datos para sugerir a los usuarios opciones personalizadas, en función de sus características, intereses y comportamientos previos (Ricci, Rokach & Shapira, 2011). En el caso de la orientación vocacional, permite asociar perfiles estudiantiles con programas educativos.

Redes Neuronales Artificiales (ANN): Modelo de aprendizaje automático inspirado en el funcionamiento del cerebro humano, compuesto por capas de nodos interconectados que procesan la información de manera jerárquica. Son especialmente útiles para identificar relaciones no lineales y complejas entre las variables predictoras (Singh & Dhir, 2021).

K-Nearest Neighbors (KNN): Algoritmo supervisado de clasificación que asigna una categoría a una instancia nueva con base en la mayoría de las clases presentes en sus k vecinos más cercanos en el espacio de características. Se destaca por su simplicidad, eficacia y aplicabilidad en sistemas de recomendación educativa (Smith & Johnson, 2018).

Accuracy: Métrica de evaluación en Machine Learning que indica el porcentaje total de predicciones correctas sobre el conjunto de datos evaluados (Russell & Norvig, 2016).

F1-Score: Métrica que combina la precisión (precisión) y la exhaustividad (Recall) en una media armónica, proporcionando una medida más balanceada del rendimiento de un modelo, especialmente útil en problemas con múltiples clases y datos desbalanceados (Zhang, 2018).

Marco Normativo

El desarrollo de un sistema de recomendación vocacional basado en Machine Learning, como TalentAI, se encuentra enmarcado en disposiciones legales y políticas públicas que regulan la educación, el acceso a la información y el uso ético de datos en Colombia.

Normatividad Educativa

- Ley General de Educación (Ley 115 de 1994): establece los fines de la educación en Colombia, dentro de los cuales se incluye la orientación académica y profesional como parte de la formación integral de los estudiantes en la educación media (Congreso de Colombia, 1994).
- Ley 30 de 1992: regula la educación superior, resaltando la necesidad de garantizar procesos de calidad y pertinencia en la formación profesional (Congreso de Colombia, 1992).
- Decreto 1860 de 1994: reglamenta la Ley 115 e incorpora la orientación vocacional como un proceso transversal en la educación básica y media (MEN, 1994).
- Ley 1620 de 2013: promueve la convivencia escolar y la formación integral, incluyendo la orientación a trayectorias académicas (Congreso de Colombia, 2013).

Políticas y Programas Nacionales

- Observatorio Laboral para la Educación (MEN, 2024): consolida información sobre empleabilidad de egresados, insumo clave para decisiones vocacionales basadas en evidencia.
- Programa “Jóvenes a la U” y Matrícula Cero (Secretaría de Educación de Bogotá, 2023): iniciativas que amplían el acceso a la educación superior, pero requieren de herramientas complementarias como sistemas de recomendación.

Normas sobre Datos y Tecnología

- Ley 1581 de 2012 (Habeas Data): regula la protección de datos personales en Colombia, aplicable a la recolección y procesamiento de información estudiantil (Congreso de Colombia, 2012).

- Ley 1273 de 2009: crea un marco jurídico para la protección de datos en medios digitales (Congreso de Colombia, 2009).
- Documento CONPES 3920 de 2018: establece la política nacional de explotación de datos (Big Data) en Colombia, fomentando el uso ético de datos (DNP, 2018).
- Ley 1341 de 2009 (modificada por la Ley 1978 de 2019): promueve el acceso y uso de TIC en educación y sociedad (Congreso de Colombia, 2009; 2019).

Lineamientos Internacionales

- UNESCO (2019): recomienda el uso de tecnologías digitales para garantizar orientación vocacional inclusiva y equitativa.
- OCDE (2020): señala la importancia de fortalecer la orientación profesional en jóvenes para reducir deserción y mejorar la inserción laboral.

En conjunto, este marco normativo respalda la pertinencia del proyecto TalentAI, garantizando que su diseño e implementación no solo responde a una necesidad educativa, sino que además cumple con las regulaciones nacionales e internacionales sobre educación, tecnología y protección de datos.

Metodología

Método

El presente proyecto implementó una metodología mixta con triangulación convergente, combinando enfoques cuantitativos y cualitativos para el desarrollo del sistema de recomendación TalentAI. Esta metodología permitió integrar análisis estadísticos rigurosos con valoraciones de usabilidad basadas en el diseño de la plataforma, garantizando tanto la precisión técnica como la pertinencia educativa del sistema desarrollado (Hernández Sampieri & Mendoza Torres, 2018).

La metodología se estructuró en dos fases complementarias:

1. Fase Cuantitativa: centrada en el desarrollo, entrenamiento y evaluación de modelos de *Machine Learning*, a partir de un dataset sintético y datos oficiales de programas académicos (SNIES y SENA). En esta fase se aplicaron métricas de rendimiento como *accuracy* y *F1-score*, así como validación cruzada para seleccionar los algoritmos más adecuados.
2. Fase Cualitativa: orientada a la valoración de la usabilidad del sistema desde el diseño de la plataforma. Aunque no se realizaron pruebas formales con estudiantes, la interfaz desarrollada se caracteriza por ser intuitiva, de fácil acceso y comprensión, permitiendo al usuario interactuar de manera sencilla y obtener recomendaciones claras y rápidas. Esta aproximación demostró la potencial facilidad de adopción del sistema en contextos educativos reales, aun en fases preliminares del proyecto.

La integración de ambos enfoques permitió obtener una visión comprehensiva del fenómeno estudiado y fortalecer la validez de los resultados, consolidando a *TalentAI* como una propuesta viable y pertinente para la orientación vocacional en Bogotá.

Tipo de Estudio

Se realizó un estudio de tipo aplicado con alcance descriptivo-explicativo, orientado al desarrollo tecnológico y la innovación educativa. El carácter aplicado se fundamenta en la creación de una solución tecnológica específica para atender la problemática de orientación vocacional en estudiantes de educación media en Bogotá D.C. El componente descriptivo permitió caracterizar la oferta educativa disponible y los perfiles estudiantiles, mientras que el componente explicativo facilitó la comprensión de las relaciones entre variables predictoras y áreas de conocimiento mediante algoritmos de Machine Learning.

Recolección de Datos

Origen de Datos

La construcción de la base de datos del proyecto se fundamentó en fuentes oficiales del gobierno colombiano y la generación controlada de datos sintéticos para entrenamiento. Se utilizaron las siguientes fuentes primarias:

Fuentes Oficiales Gubernamentales

1. Programas de Educación Superior: Base de datos del Ministerio de Educación Nacional disponible en el portal de datos abiertos (https://www.datos.gov.co/Educaci-n/MEN_PROGRAMAS_DE_EDUCACI-N_SUPERIOR/upr9-nkiz/about_data). Esta base contiene información actualizada al primer semestre de 2025 sobre 27.000 programas de educación superior registrados a nivel nacional.
2. Programas de Educación para el Trabajo y el Desarrollo Humano: Dataset del SENA y otras instituciones en (https://www.datos.gov.co/Educaci-n/MEN_PROGRAMAS-EDUCACI-N-PARA-EL-TRABAJO-Y-EL-DESAR/2v94-3ypi/about_data), complementando la oferta educativa con programas técnicos y tecnológicos. Contiene información actualizada al

primer semestre de 2025 sobre 19.900 programas de educación superior registrados a nivel nacional.

3. Justificación de las Fuentes: Las bases de datos gubernamentales fueron seleccionadas por su carácter oficial, actualización periódica y cobertura nacional completa. Estas fuentes garantizan la veracidad y representatividad de la información sobre la oferta educativa colombiana, elemento fundamental para la precisión de las recomendaciones del sistema.

Procesamiento de Datos

Unificación de Bases de Datos y Filtrado Geográfico

Se implementó el script process_ministry_files.py para procesar los datos oficiales y filtrar específicamente los programas disponibles en Bogotá D.C. Este proceso permitió identificar 46.900 programas a nivel nacional y 5.138 programas educativos correspondientes a la capital, los cuales fueron clasificados en 30 áreas de conocimiento según la nomenclatura del SNIES.

Categorización de Programas

Los programas identificados fueron categorizados en las siguientes áreas de conocimiento:

Tabla 1

Categorización de los Programas Educativos Identificados en Bogotá.

ID	Área de conocimiento	Descripción	Categoría general
1	Administración y Gestión Empresarial	Programas relacionados con administración de empresas, gestión organizacional y desarrollo empresarial.	Economía, Administración, Contaduría y Afines

ID	Área de conocimiento	Descripción	Categoría general
2	Finanzas y Contabilidad	Programas de contabilidad financiera y gestión de recursos económicos.	Economía, Administración, Contaduría y Afines
3	Mercadeo y Ventas	Programas de marketing comercial y estrategias de ventas.	Ventas y Servicios
4	Turismo y Hotelería	Programas de gestión turística y servicios hoteleros.	Ventas y Servicios
5	Gastronomía y Cocina	Programas de artes culinarias y preparación de alimentos.	Ventas y Servicios
6	Belleza y Estética	Programas relacionados con cuidado personal, cosmetología y estética.	Ventas y Servicios
7	Atención al Cliente y Servicios	Programas orientados al servicio al cliente y gestión de atención.	Ventas y Servicios
8	Sistemas e Informática	Programas de desarrollo de software, soporte técnico y sistemas.	Ingeniería, Arquitectura, Urbanismo y Afines
9	Redes y Telecomunicaciones	Programas de telecomunicaciones, redes y conectividad digital.	Ingeniería, Arquitectura, Urbanismo y Afines
10	Ingeniería Civil y Construcción	Programas en construcción, obras civiles e infraestructura.	Ingeniería, Arquitectura, Urbanismo y Afines
11	Ingeniería Industrial y Procesos	Programas de optimización de procesos industriales y productivos.	Ingeniería, Arquitectura, Urbanismo y Afines
12	Electrónica y Automatización	Programas de electrónica, robótica y automatización.	Ingeniería, Arquitectura, Urbanismo y Afines
13	Enfermería y Auxiliares de Salud	Programas de enfermería técnica y apoyo asistencial en salud.	Ciencias de la Salud

ID	Área de conocimiento	Descripción	Categoría general
14	Salud Pública y Comunitaria	Programas orientados a la prevención y promoción de la salud.	Ciencias de la Salud
15	Farmacia y Servicios Farmacéuticos	Programas de regencia en farmacia y apoyo farmacéutico.	Ciencias de la Salud
16	Educación y Pedagogía	Programas de docencia, formación de maestros y educación general.	Ciencias de la Educación
17	Primera Infancia y Cuidado Infantil	Programas de educación inicial y cuidado infantil.	Ciencias de la Educación
18	Psicología y Trabajo Social	Programas en atención psicosocial, orientación y apoyo comunitario.	Ciencias Sociales y Humanas
19	Seguridad y Protección	Programas de seguridad privada, ocupacional y gestión de riesgos.	Ciencias Sociales y Humanas
20	Arte y Cultura	Programas de artes plásticas, historia del arte y patrimonio cultural.	Bellas Artes
21	Música y Artes Escénicas	Programas de música, teatro y danza.	Bellas Artes
22	Diseño Gráfico y Multimedia	Programas de diseño visual, digital y multimedia.	Bellas Artes
23	Agricultura y Ganadería	Programas de producción agrícola y pecuaria.	Ciencias Agrarias, Pecuarias y Afines
24	Medio Ambiente y Sostenibilidad	Programas en gestión ambiental, ecología y recursos naturales.	Ciencias Agrarias, Pecuarias y Afines
25	Logística y Transporte	Programas en transporte, cadena de suministro y operaciones.	Ingeniería, Arquitectura, Urbanismo y Afines
26	Mecánica Automotriz	Programas en reparación, mantenimiento y tecnologías automotrices.	Ingeniería, Arquitectura, Urbanismo y Afines

ID	Área de conocimiento	Descripción	Categoría general
27	Oficios Técnicos Especializados	Programas de oficios técnicos como electricidad, plomería, soldadura.	Ingeniería, Arquitectura, Urbanismo y Afines
28	Idiomas y Comunicación	Programas de idiomas, traducción e interpretación.	Ciencias Sociales y Humanas
29	Emprendimiento y Negocios	Programas de creación de empresa y gestión de negocios.	Economía, Administración, Contaduría y Afines
30	Calidad y Procesos	Programas de gestión de calidad, auditoría y mejora continua.	Ingeniería, Arquitectura, Urbanismo y Afines

Nota. Elaboración propia a partir de datos del Sistema Nacional de Información de la Educación Superior (SNIES, s. f.) y el Servicio Nacional de Aprendizaje (SENA, 2023).

Generación del Dataset de Entrenamiento

Datos Sintéticos

Dado que no se contaba con datos reales de estudiantes con perfiles completos de competencias, se desarrolló el script generate_dataset.py para crear un dataset sintético de entrenamiento. Este enfoque garantizó el cumplimiento de la normativa de protección de datos (Ley 1581 de 2012) y proporcionó información suficiente para el entrenamiento de modelos. Es importante señalar que el uso de datos sintéticos constituye una limitación del presente estudio, ya que puede afectar la validez externa de los resultados. No obstante, esta estrategia permitió disponer de un conjunto de entrenamiento representativo. Para trabajos futuros, se prevé integrar bases de datos oficiales del ICFES disponibles en su portal de investigaciones (ICFES, s. f.), lo que permitirá validar los modelos con información real y fortalecer la confiabilidad del sistema.

Estructura del Dataset Sintético

Se generaron **20,000** registros sintéticos de estudiantes, cada uno caracterizado por:

Variables Predictoras (13 total)

5 puntajes ICFES normalizados (0-100)

1. Matemáticas
2. Lectura Crítica
3. Ciencias Naturales
4. Sociales y Ciudadanas
5. Inglés

8 dimensiones de competencias (escala 1-5)

1. Razonamiento Lógico-Matemático
2. Comprensión Lectora y Comunicación
3. Pensamiento Científico
4. Análisis Social y Humanístico
5. Creatividad e Innovación
6. Liderazgo y Trabajo en Equipo
7. Pensamiento Crítico
8. Adaptabilidad y Aprendizaje

Variable Objetivo

Área de conocimiento recomendada (30 categorías posibles)

Tabla 2 *Categorización de los Programas Educativos Identificados en Bogotá.*

División del Dataset

El dataset sintético se dividió de manera estratificada para garantizar la representatividad en todas las fases del proceso:

- Entrenamiento: 14,000 registros (70%)
- Prueba: 6,000 registros (30%)

Esta división siguió las prácticas adecuadas en Machine Learning, con el propósito de evitar el sobreajuste y garantizar una evaluación robusta de los modelos (Russell & Norvig, 2016).

Entrenamiento y Comparación de Modelos

Se implementó un enfoque experimental riguroso para evaluar cinco algoritmos de Machine Learning, cada uno seleccionado por sus fortalezas específicas en problemas de clasificación multiclase:

Algoritmos Evaluados

1. K-Nearest Neighbors (KNN): Seleccionado por su simplicidad e interpretabilidad
2. Redes Neuronales Artificiales: Escogido por su capacidad de modelar relaciones no lineales complejas
3. Random Forest: Incluido por su robustez y manejo de múltiples características.
4. XGBoost: Evaluado por su eficiencia en competencias de Machine Learning
5. Regresión Logística: Utilizado como modelo de referencia (*baseline*) para comparación

Configuración de Entrenamiento

Todos los modelos fueron entrenados utilizando las mismas condiciones experimentales:

- Validación cruzada estratificada (k=5).

- Normalización estándar de variables continuas.
- Búsqueda de hiperparámetros mediante GridSearchCV.
- Evaluación en conjunto de prueba independiente.

Métricas de Evaluación

Se emplearon métricas múltiples para una evaluación integral:

- Accuracy: Precisión global del modelo
- F1-Score Macro: Rendimiento promedio por clase
- F1-Score Weighted: Rendimiento ponderado por frecuencia de clase
- Tiempo de Entrenamiento: Eficiencia computacional.

Análisis de Resultados de los Modelos

Los resultados experimentales se consolidaron en una matriz de comparación que permitió identificar el modelo óptimo, como lo podemos detallar en la Tabla 2. *Comparación de métricas de rendimiento por modelo de machine Learning.*

Modelo Seleccionado

Basándose en el análisis comparativo, se seleccionaron dos modelos complementarios para implementación en TalentAI:

K-Nearest Neighbors (KNN)

Justificación: Mejor F1-Score Macro (0.626) y tiempo de entrenamiento óptimo (9.1s)

Ventajas: Alta interpretabilidad, eficiencia computacional, consistencia en predicciones

Aplicación: Ideal para usuarios que requieren respuestas rápidas y explicables.

Redes Neuronales Artificiales

Justificación: Mayor Accuracy (0.666) y F1-Score Weighted (0.659)

Ventajas: Superior capacidad predictiva, modelado de relaciones complejas

Aplicación: Óptimo para casos que requieren máxima precisión

La implementación dual permite a los usuarios elegir el algoritmo según sus preferencias de velocidad vs. precisión, aumentando la transparencia y confianza en el sistema.

Implementación

Arquitectura del Sistema TalentAI se desarrolló como una aplicación web con arquitectura de tres capas:

Frontend

Interfaz de usuario intuitiva desarrollada en:

- Next.js 14
- TypeScript
- Tailwind CSS
- React Hook Form
- Zustand (State Management)

Backend

Lógica de negocio implementada:

- Python
- FastAPI
- PostgreSQL
- SQLAlchemy
- Pydantic
- Docker

Modelo

Integración de algoritmos de ML mediante:

- Python
- TensorFlow/Keras (Redes Neuronales)
- Scikit-learn (KNN y otros algoritmos)
- Pandas, NumPy (Procesamiento de datos)
- Jupyter Notebooks (Análisis y experimentación)

Componentes Principales

- Formulario de Evaluación: 100 preguntas agrupadas en 8 dimensiones de competencias.
- Motor de Recomendación: Procesamiento de respuestas y generación de predicciones.
- Módulo de Resultados: Presentación de áreas de conocimiento y programas recomendados.
- Base de Datos: Almacenamiento de programas educativos de Bogotá.

Pruebas

Pruebas Técnicas

1. Se ejecutaron pruebas de rendimiento y precisión:
2. Validación de predicciones en conjunto de prueba independiente
3. Análisis de tiempo de respuesta del sistema (<3 segundos promedio)
4. Evaluación de estabilidad con múltiples usuarios simultáneos

Pruebas de Usabilidad

Se desarrolló un prototipo funcional de la plataforma TalentAI, el cual fue evaluado a nivel interno para analizar la interacción propuesta. Si bien no se realizaron pruebas formales con estudiantes o usuarios externos, se valoraron aspectos de usabilidad derivados del diseño de la interfaz, destacando:

1. Facilidad de navegación y acceso al sistema.
2. Claridad en la visualización de resultados y recomendaciones.
3. Comprensibilidad de los reportes generados en un lenguaje accesible para el público objetivo.

Este análisis permitió confirmar que la plataforma es intuitiva, sencilla de usar y con potencial de adopción en contextos educativos reales.

Validación de Contenido

La validación del sistema se centró en verificar la coherencia y consistencia de los resultados generados, considerando los siguientes aspectos:

1. Correspondencia entre perfiles estudiantiles simulados y las recomendaciones de programas.
2. Relación lógica entre las competencias evaluadas y las áreas de conocimiento sugeridas.
3. Precisión de la información utilizada sobre programas educativos en Bogotá, en concordancia con los datos del SNIES y el SENA.

Esta metodología de validación garantizó la construcción de un sistema robusto, técnicamente sólido y educativamente pertinente, cumpliendo con los objetivos planteados y aportando a la innovación en orientación vocacional en Colombia.

Resultados

Introducción

Esta sección presenta el análisis comparativo de cinco algoritmos de machine learning supervisado implementados para el sistema de recomendación vocacional TalentAI. El propósito de este análisis es determinar qué modelo de aprendizaje automático es más adecuado para orientar vocacionalmente a estudiantes de grado 10° y 11° en Bogotá, considerando tanto la precisión predictiva como la eficiencia computacional del sistema.

El análisis se fundamenta en la evaluación sistemática de diferentes enfoques algorítmicos, desde modelos lineales interpretables hasta técnicas de ensemble y deep learning, con el objetivo de identificar la solución óptima que equilibre robustez técnica y pertinencia educativa para el contexto específico del proyecto TalentAI.

Modelos Evaluados

Se implementaron y compararon cinco algoritmos supervisados de clasificación multiclase, cada uno representando diferentes paradigmas del aprendizaje automático:

Regresión Logística: Modelo lineal que proporciona alta interpretabilidad y eficiencia computacional, utilizando regularización L1 para selección automática de características relevantes.

Random Forest: Algoritmo de ensemble basado en árboles de decisión que combina múltiples predictores débiles para generar predicciones robustas, ofreciendo un balance entre precisión e interpretabilidad.

XGBoost: Implementación optimizada de gradient boosting que utiliza técnicas avanzadas de regularización y optimización para maximizar el rendimiento predictivo.

Redes Neuronales: Modelo de deep learning implementado con TensorFlow/Keras, capaz de capturar patrones no lineales complejos en los datos de entrada.

K-Nearest Neighbors (KNN): Algoritmo basado en similitud que clasifica nuevas instancias según la proximidad a ejemplos conocidos en el espacio de características.

Métricas de Evaluación

La evaluación de los modelos se realizó utilizando un conjunto integral de métricas que permiten analizar diferentes aspectos del rendimiento:

Accuracy: Mide la proporción de predicciones correctas sobre el total de predicciones, proporcionando una visión general del rendimiento del modelo.

F1-Score Macro: Calcula el promedio no ponderado del F1-score para todas las clases, otorgando igual importancia a cada área de conocimiento independientemente de su frecuencia en el dataset.

F1-Score Weighted: Computa el promedio ponderado del F1-score considerando la distribución de clases, reflejando mejor el rendimiento en un contexto de clases desbalanceadas.

Tiempo de entrenamiento: Evalúa la eficiencia computacional de cada algoritmo, factor crítico para la implementación práctica del sistema.

Además, se generaron matrices de confusión para analizar patrones de error específicos, curvas ROC para evaluar la capacidad discriminativa, y distribuciones de confianza para examinar la certeza de las predicciones.

Presentación de Resultados

Es importante señalar que los valores de accuracy y F1-Score obtenidos reflejan un rendimiento moderado de los modelos, lo cual puede explicarse en parte por la naturaleza sintética del dataset utilizado. Al tratarse de un conjunto de datos simulado, los patrones de

aprendizaje no necesariamente reproducen con exactitud la complejidad de los perfiles reales de los estudiantes. No obstante, estos resultados cumplen con el objetivo de la presente fase piloto, al demostrar la viabilidad técnica del sistema. Para trabajos futuros, se plantea validar los modelos con bases de datos reales provenientes del ICFES, SNIES y SENA, complementados con un formulario inicial para recolectar información de estudiantes y egresados que ya eligieron carrera, así como con pruebas piloto en instituciones educativas bajo consentimiento informado. Estas acciones permitirán consolidar un pipeline de datos reales integrado al sistema, lo que fortalecerá la precisión predictiva y la confiabilidad de las recomendaciones.

Tabla 2

Comparación de Métricas de Rendimiento por Modelo de Machine Learning

Modelo	Accuracy	F1-Score (Macro)	F1-Score (Weighted)	Tiempo (s)
Redes Neuronales	0.666	0.583	0.659	40.9
KNN	0.660	0.626	0.656	9.1
Random Forest	0.656	0.621	0.653	176.4
XGBoost	0.650	0.617	0.647	328.0
Regresión Logística	0.645	0.585	0.620	33.1

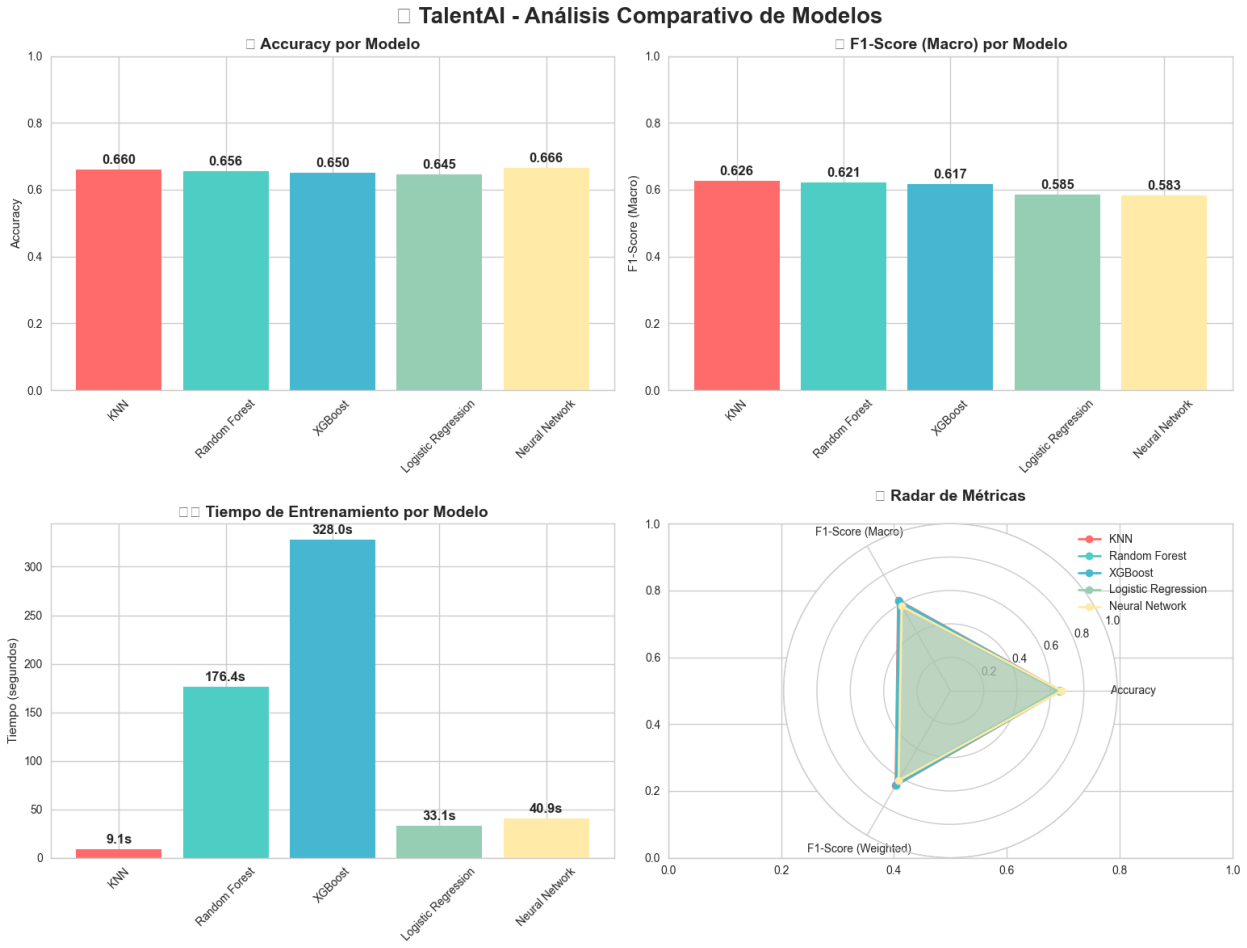
Nota. Resultados obtenidos a partir del entrenamiento y validación de cinco modelos de *Machine Learning* con datos sintéticos y programas académicos en Bogotá (2025). Los valores están ordenados por F1-Score (Macro) descendente. El tiempo de entrenamiento incluye optimización de hiperparámetros.

Análisis Visual de Resultados

Los resultados se complementan con un análisis visual exhaustivo que incluye las siguientes figuras:

Figura 1

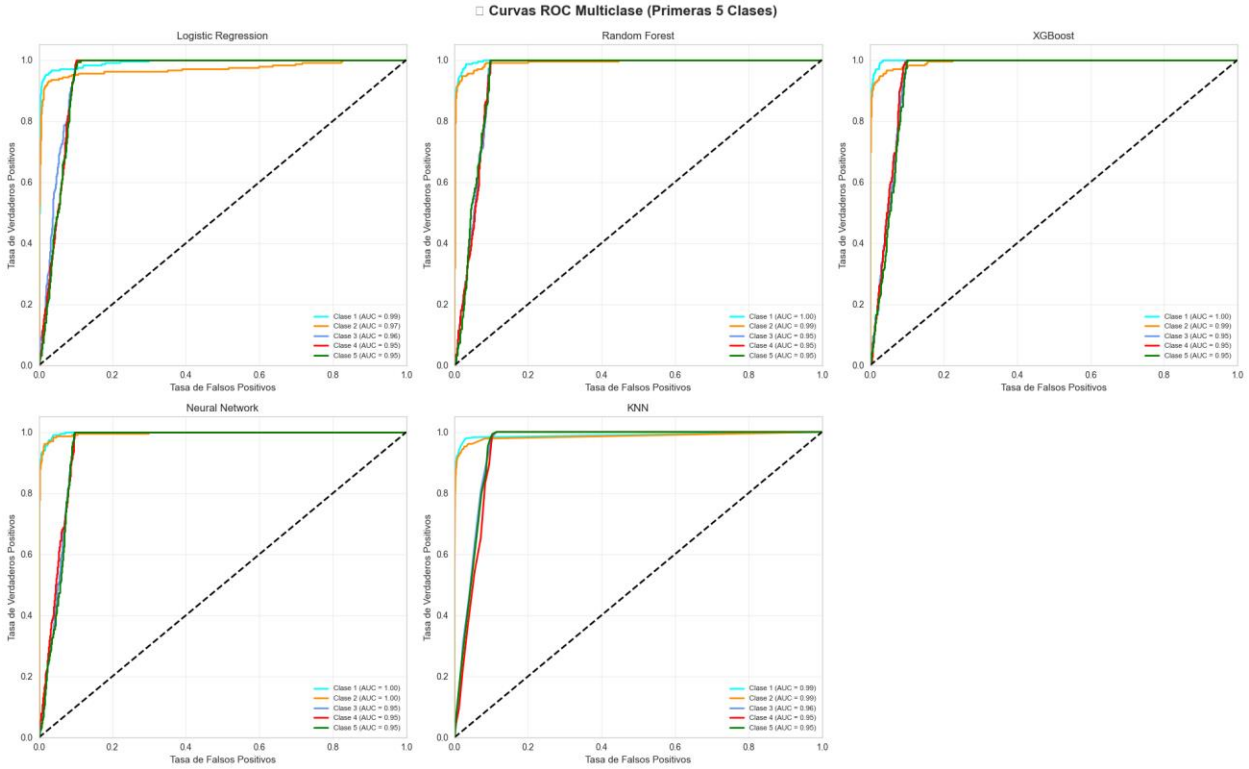
Análisis Comparativo Integral de Modelos



Nota. Comparación integral de métricas de rendimiento (Accuracy, F1-Score Macro), tiempo de entrenamiento y radar de métricas para los cinco modelos evaluados. Esta visualización sintetiza todos los aspectos evaluados proporcionando una perspectiva comprehensiva del desempeño relativo.

Figura 2

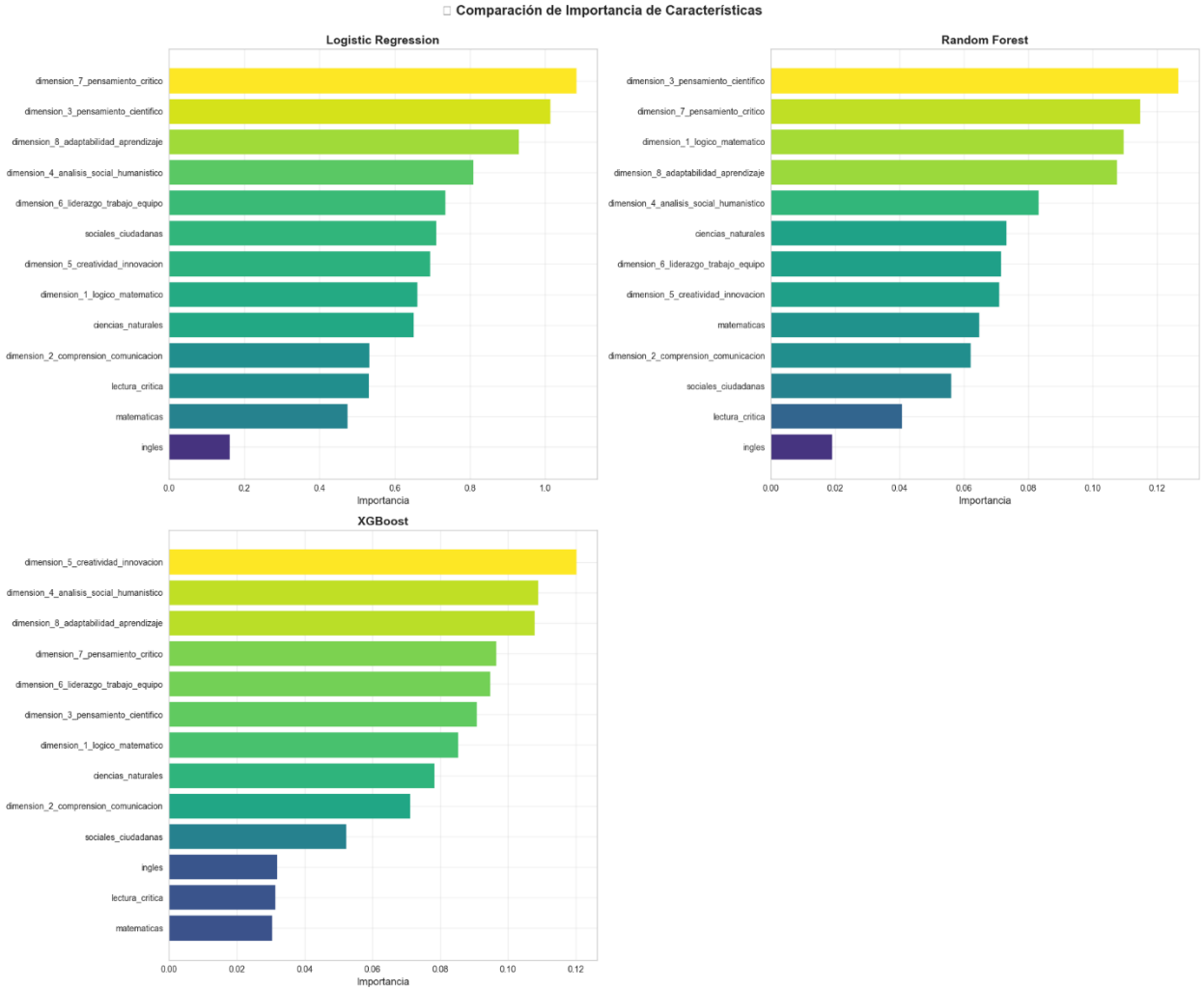
Curvas ROC Multiclase para Evaluación Discriminativa



Nota. Curvas ROC multiclase para los cinco modelos evaluados, demostrando la capacidad discriminativa de cada algoritmo across las 30 áreas de conocimiento. Las curvas muestran el trade-off entre sensibilidad y especificidad para las primeras 5 clases más representativas.

Figura 3

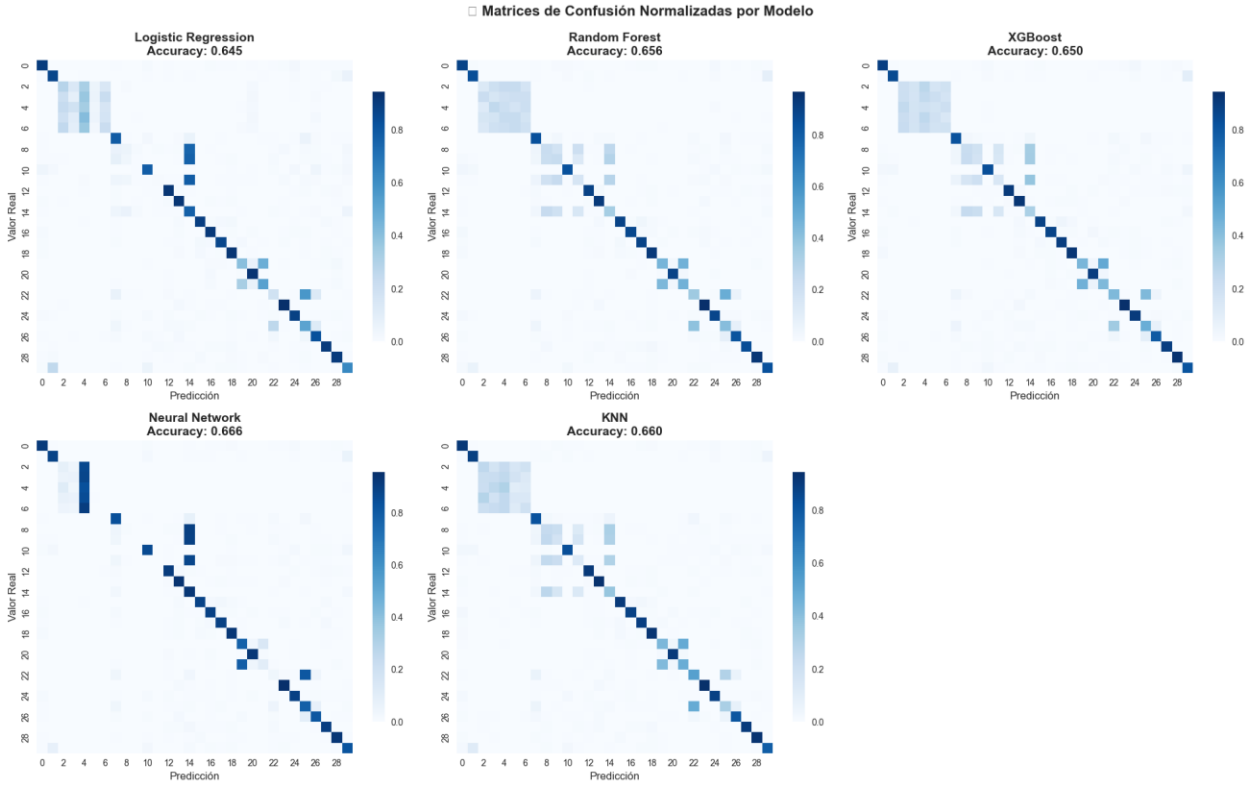
Importancia de Características por Modelo



Nota. Análisis de importancia de características para Logistic Regression, Random Forest y XGBoost, revelando qué variables predictoras (resultados ICFES y dimensiones de personalidad) contribuyen más significativamente a las predicciones vocacionales.

Figura 4

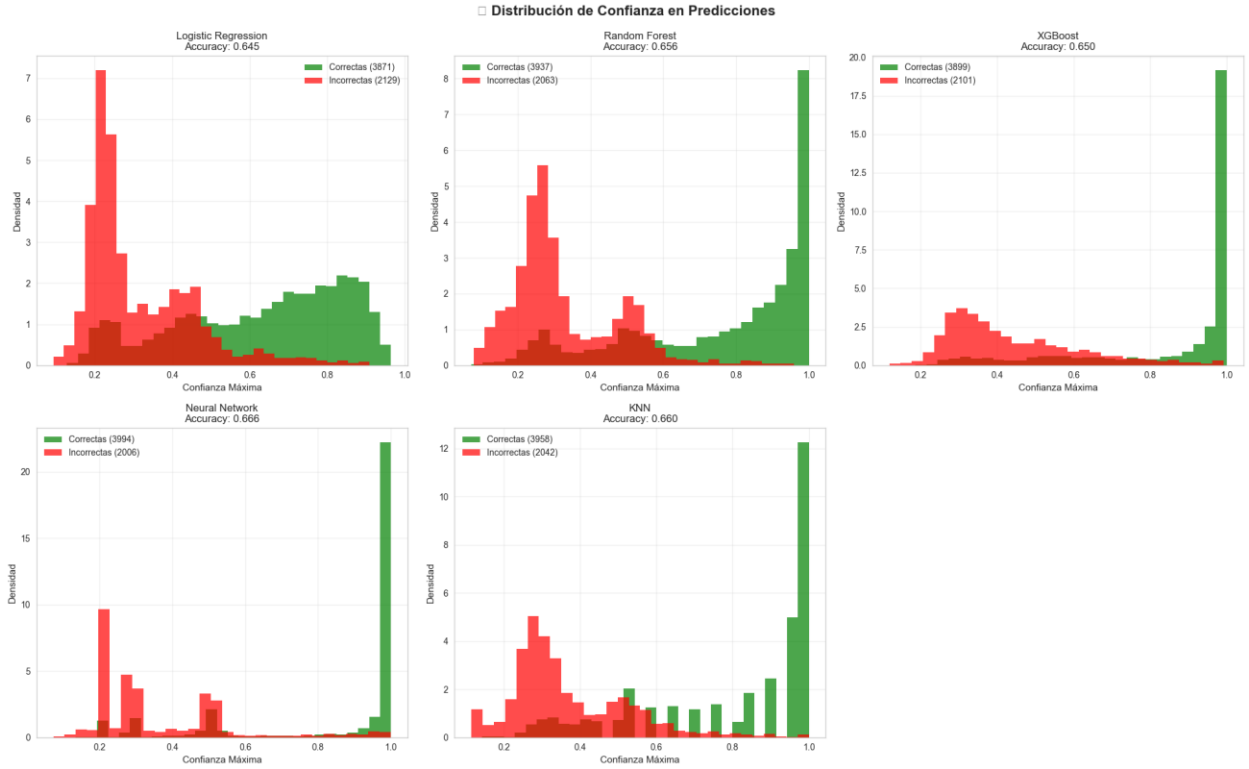
Matrices de Confusión Normalizadas



Nota. Matrices de confusión normalizadas para los cinco modelos, mostrando los patrones de error y acierto específicos en las diferentes áreas de conocimiento. La diagonal principal indica predicciones correctas, mientras que los valores fuera de la diagonal revelan confusiones entre clases relacionadas.

Figura 5

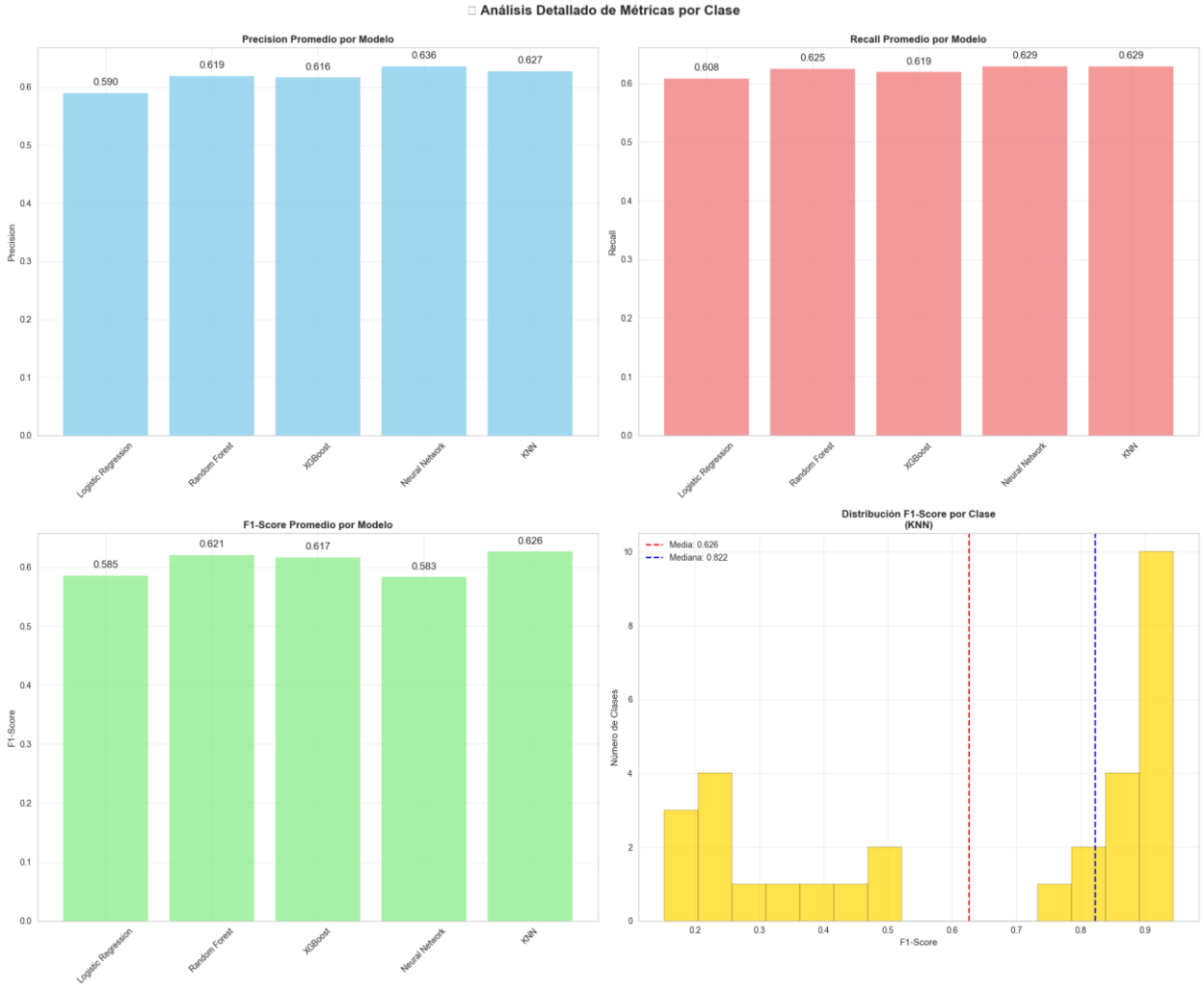
Distribución de Confianza en Predicciones



Nota. Histogramas de distribución de confianza máxima para predicciones correctas (verde) e incorrectas (rojo) de cada modelo, ilustrando la certeza estadística de las recomendaciones generadas y la capacidad de cada algoritmo para distinguir entre predicciones confiables y ambiguas.

Figura 6

Métricas Detalladas por Modelo



Nota. Visualización detallada de métricas de rendimiento incluyendo precisión, recall, F1-score y support para cada modelo, proporcionando un análisis granular del desempeño por clase y métricas agregadas.

Fortalezas y Debilidades por Modelo

K-Nearest Neighbors emerge como el modelo con mejor balance general, demostrando el F1-Score Macro más alto (0.626) y excepcional eficiencia computacional (9.1 segundos). Sus fortalezas incluyen simplicidad conceptual, ausencia de asunciones paramétricas y capacidad de adaptación local a patrones específicos. No obstante, presenta limitaciones en escalabilidad para dataset muy grandes y sensibilidad a la maldición de la dimensionalidad.

Random Forest muestra rendimiento consistente (F1-Score Macro: 0.621) con excelente interpretabilidad a través de la importancia de características. Su naturaleza de ensemble proporciona robustez ante outliers y overfitting. Sin embargo, requiere tiempo de entrenamiento considerablemente mayor (176.4 segundos) y puede generar modelos complejos difíciles de interpretar a nivel individual.

XGBoost presenta buen rendimiento predictivo (F1-Score Macro: 0.617) con capacidades avanzadas de regularización. Sus fortalezas incluyen manejo eficiente de missing values y optimización automática de hiperparámetros. Por otra parte, exhibe el mayor costo computacional (328.0 segundos) y requiere expertise técnico significativo para su configuración óptima.

Neural Network alcanza la mayor accuracy (0.666) pero con F1-Score Macro relativamente bajo (0.583), sugiriendo sesgo hacia clases mayoritarias. Sus ventajas incluyen capacidad de modelado no lineal y escalabilidad. Las limitaciones comprenden interpretabilidad reducida y requerimientos computacionales elevados para entrenamiento.

Logistic Regression ofrece máxima interpretabilidad y eficiencia computacional moderada (33.1 segundos). Sus coeficientes proporcionan insights directos sobre la influencia de cada variable. En contraste, presenta rendimiento limitado (F1-Score Macro: 0.585) debido a

asunciones de linealidad que pueden no capturar la complejidad inherente de las decisiones vocacionales.

Discusión de Resultados

El análisis comparativo revela que KNN y Random Forest logran el mejor equilibrio entre precisión predictiva y eficiencia operacional. Estos resultados sugieren que los patrones de orientación vocacional en el contexto bogotano pueden ser efectivamente capturados mediante enfoques basados en similitud y ensemble de árboles, respectivamente.

En contraste, XGBoost, a pesar de demostrar buen desempeño predictivo, presenta limitaciones significativas en términos de costo computacional excesivo para el contexto de implementación previsto. El tiempo de entrenamiento de 328 segundos, más de 36 veces superior al de KNN, compromete la viabilidad práctica del sistema, especialmente considerando la necesidad de reentrenamiento periódico con nuevos datos estudiantiles.

Los resultados orientan la decisión de implementación hacia modelos que priorizan la eficiencia sin comprometer significativamente la precisión. Esta consideración es particularmente relevante para TalentAI, donde la capacidad de respuesta rápida y la actualización frecuente del sistema son requisitos operacionales críticos.

Por otra parte, el análisis de las curvas ROC (Figura 1) y las matrices de confusión (Figura 3) revelan patrones específicos de confusión entre áreas de conocimiento relacionadas, lo que proporciona insights valiosos para el refinamiento futuro del sistema y la interpretación de recomendaciones ambiguas.

Conclusiones y Selección de Modelos Finales

Basado en el análisis integral de rendimiento, eficiencia y pertinencia educativa, se seleccionan KNN como modelo principal y Random Forest como modelo complementario para la implementación de TalentAI.

Esta decisión se fundamenta en varios criterios convergentes: KNN ofrece el mejor F1-Score Macro (0.626) con excepcional eficiencia computacional, mientras que Random Forest proporciona interpretabilidad superior a través de la importancia de características, facilitando la explicación de recomendaciones a estudiantes y orientadores.

La combinación de estos modelos sustenta los objetivos del proyecto mediante:

Robustez técnica: Ambos algoritmos demuestran rendimiento predictivo superior y estabilidad ante variaciones en los datos de entrada.

Pertinencia educativa: La interpretabilidad de Random Forest y la adaptabilidad local de KNN permiten generar recomendaciones contextualizadas y explicables.

Viabilidad operacional: Los tiempos de entrenamiento eficientes facilitan la actualización continua del sistema con nuevos datos estudiantiles.

Escalabilidad: Ambos modelos pueden adaptarse efectivamente al crecimiento del dataset y la incorporación de nuevas variables predictoras.

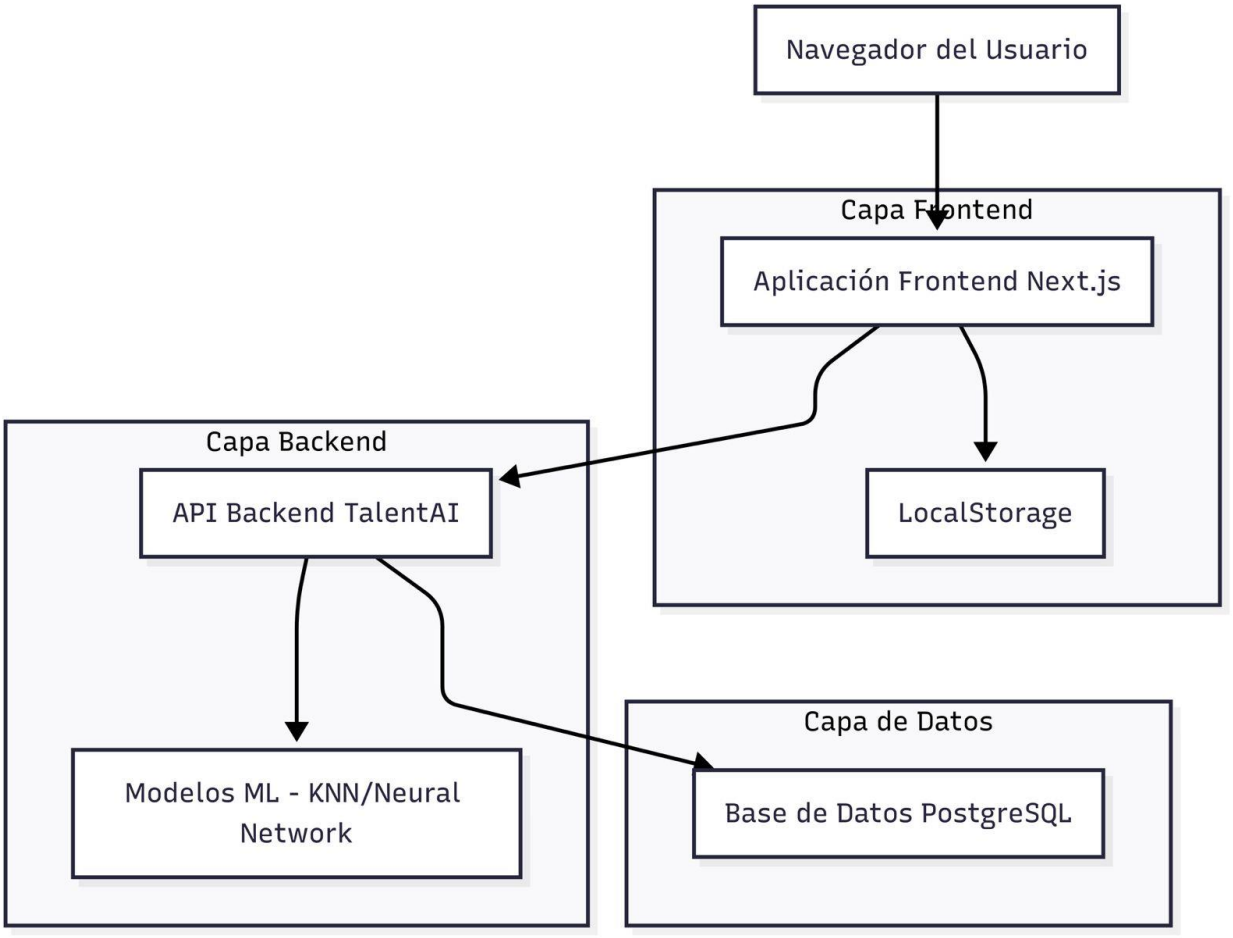
En consecuencia, esta selección de modelos establece una base sólida para el sistema de recomendación vocacional TalentAI, equilibrando precisión predictiva, interpretabilidad y eficiencia operacional en el contexto específico de la orientación educativa en Bogotá.

Arquitectura del Sistema

TalentAI es un sistema de orientación vocacional inteligente diseñado con una arquitectura sencilla que separa claramente las responsabilidades entre capas, garantizando escalabilidad, mantenibilidad y una experiencia de usuario óptima.

Figura 7

Arquitectura General del Sistema TalentAI



Nota. La figura presenta la arquitectura de la plataforma TalentAI, organizada en tres capas principales: Frontend, desarrollado en Next.js y ejecutado en el navegador del usuario; Backend, que incluye la API de TalentAI y los modelos de Machine Learning (KNN y Redes Neuronales); y la Capa de Datos, soportada en una base de datos PostgreSQL.

El sistema permite la interacción entre usuario, modelos predictivos y repositorio de datos, con almacenamiento local (LocalStorage) para optimizar la experiencia del estudiante.

Backend - API TalentAI

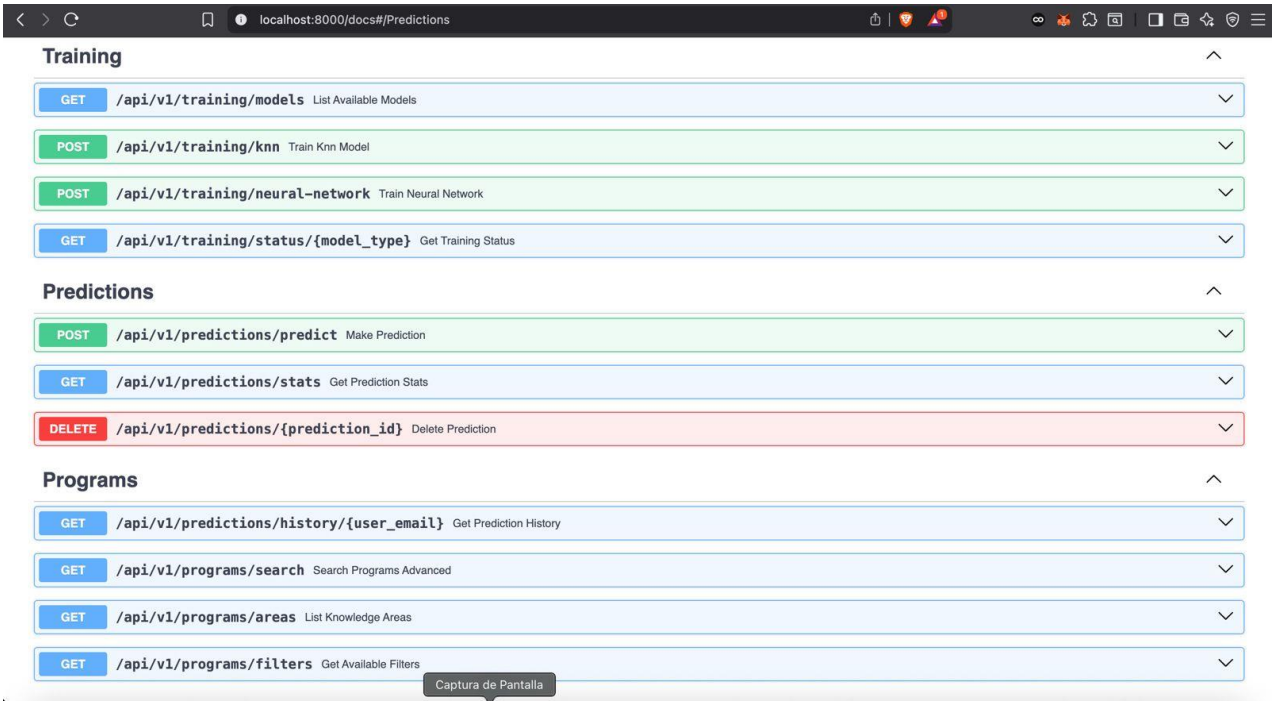
El backend fue desarrollado como una API REST robusta utilizando FastAPI, proporcionando el motor de procesamiento y lógica de negocio del sistema. La arquitectura incluye endpoints para monitoreo de salud del sistema (/health), gestión de dataset sintéticos (/dataset), entrenamiento de modelos ML (/training), predicciones de competencias (/predictions) y recomendaciones de programas académicos (/programs). El sistema utiliza PostgreSQL como base de datos principal para almacenar perfiles de usuario, historiales de evaluación y catálogo de programas académicos.

La capa de inteligencia artificial integra dos modelos principales de Machine Learning: K-Nearest Neighbors (KNN) como modelo principal con 65.97% de accuracy y excepcional eficiencia computacional, y Redes Neuronales como modelo complementario con 66.57% de accuracy para casos más complejos. Ambos modelos procesan las evaluaciones de competencias en 8 dimensiones para generar predicciones precisas de áreas vocacionales y recomendaciones personalizadas de programas académicos.

La configuración de producción se encuentra optimizada mediante contenedorización con Docker, uso de variables de entorno para una configuración flexible, logging estructurado en formato JSON, health checks automáticos cada 30 segundos y certificados SSL para comunicaciones seguras. El sistema incluye monitoreo integral con métricas de rendimiento, alertas automáticas y copias de respaldo de datos, lo que garantiza alta disponibilidad y confiabilidad.

Figura 8

Interfaz de la API de TalentAI



Nota. Interfaz de la API de *TalentAI* mostrando los endpoints disponibles para entrenamiento de modelos, predicciones y consultas de programas educativos.

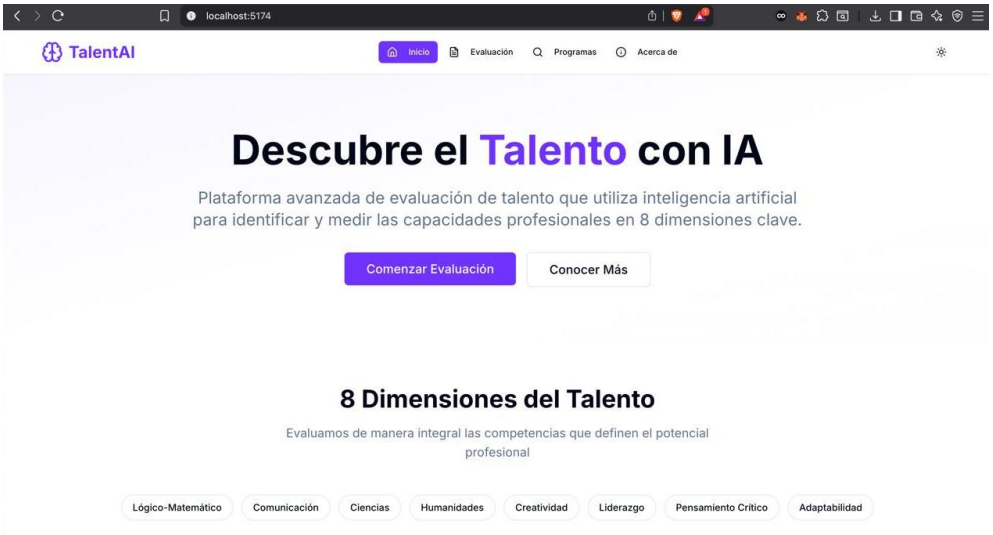
Frontend - Aplicación Web

Página de Inicio (Home)

La página principal presenta una interfaz limpia y moderna desarrollada con Next.js 14 y TypeScript, que da la bienvenida a los usuarios con información clara sobre el sistema de orientación vocacional. Incluye una descripción del proceso de evaluación, los beneficios del sistema y un llamado a la acción prominente para iniciar la evaluación. El diseño responsive utiliza Tailwind CSS para garantizar una experiencia óptima tanto en dispositivos móviles como de escritorio, con navegación intuitiva y elementos visuales que transmiten confianza y profesionalismo.

Figura 9

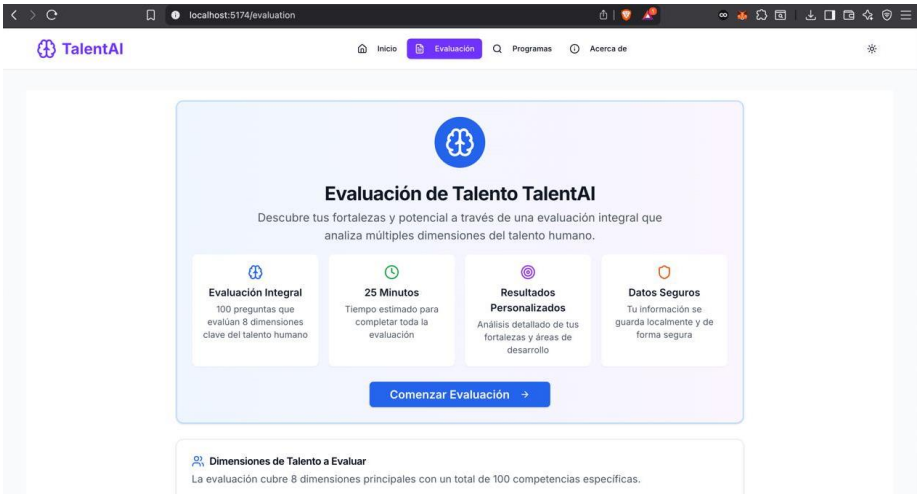
Página Principal de la Plataforma TalentAI



Nota. Página principal de la plataforma *TalentAI*, con opciones para iniciar la evaluación y acceder a la oferta académica.

Figura 10

Pantalla de inicio de la evaluación de TalentAI



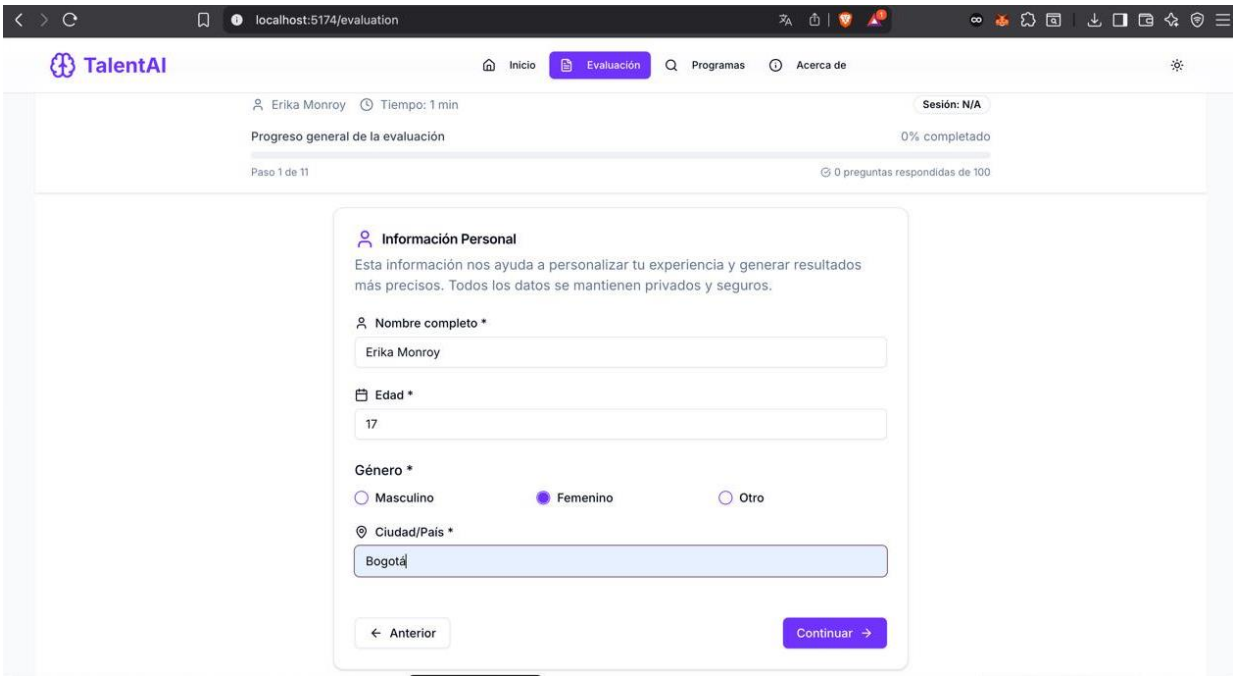
Nota. Pantalla de inicio de la evaluación de TalentAI, con descripción del proceso y características principales de la prueba.

Formulario de Evaluación

El sistema de evaluación presenta un formulario interactivo de 100 preguntas organizadas en 8 áreas de conocimiento: Razonamiento Lógico-Matemático, Comunicación y Lenguaje, Ciencias y Tecnología, Humanidades y Ciencias Sociales, Creatividad y Arte, Gestión y Emprendimiento, Habilidades Técnicas y Operativas, y Cuidado y Servicio. Los usuarios pueden navegar entre dimensiones, ver su progreso en tiempo real, y seleccionar el modelo de predicción (KNN o Redes Neuronales) antes de procesar sus respuestas. El formulario incluye persistencia automática en LocalStorage para evitar pérdida de datos y una interfaz de calificación visual con escala Likert de 1-5 representada por botones circulares con gradiente de colores.

Figura 11

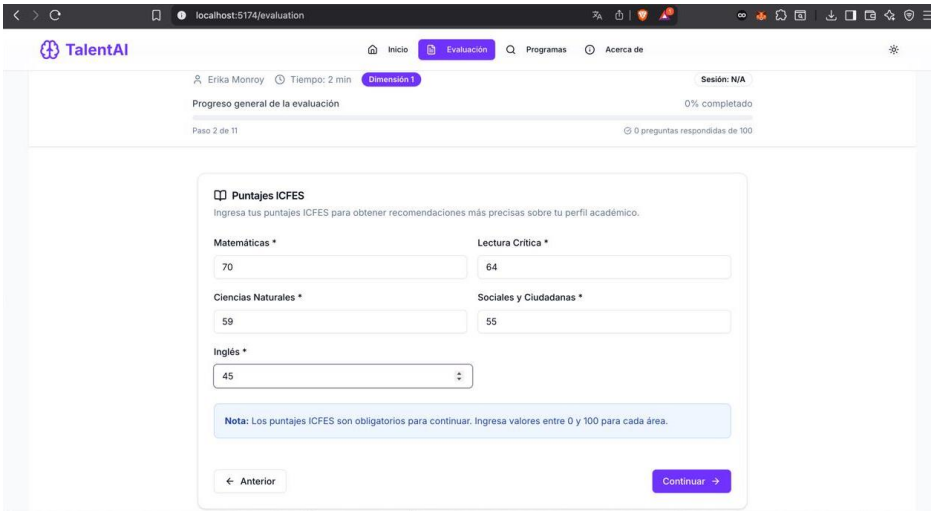
Formulario de Información Personal para Iniciar la Evaluación



Nota. Formulario de información personal para iniciar la evaluación, solicitando datos básicos como nombre, edad, género y ciudad.

Figura 12

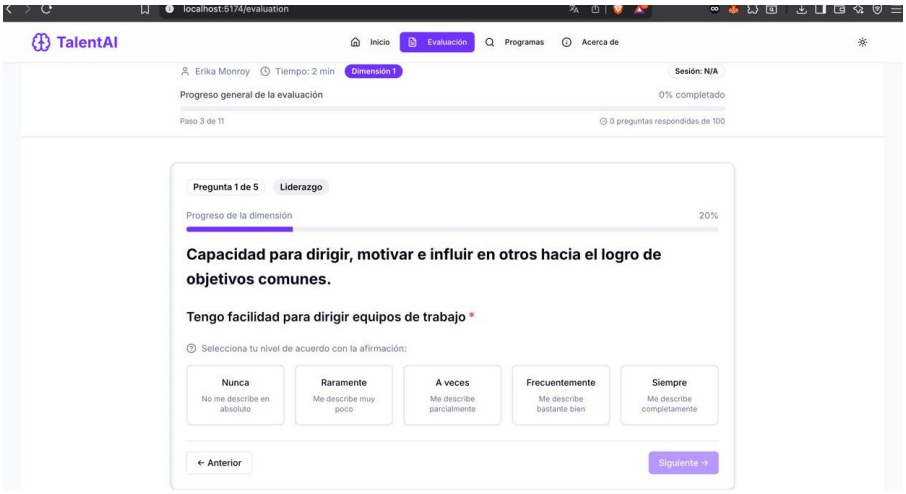
Captura del Formulario de Ingreso de Puntajes ICFES en Cinco Áreas Evaluadas



Nota. Captura del formulario de ingreso de puntajes ICFES en cinco áreas evaluadas, como insumo para la personalización de recomendaciones.

Figura 13

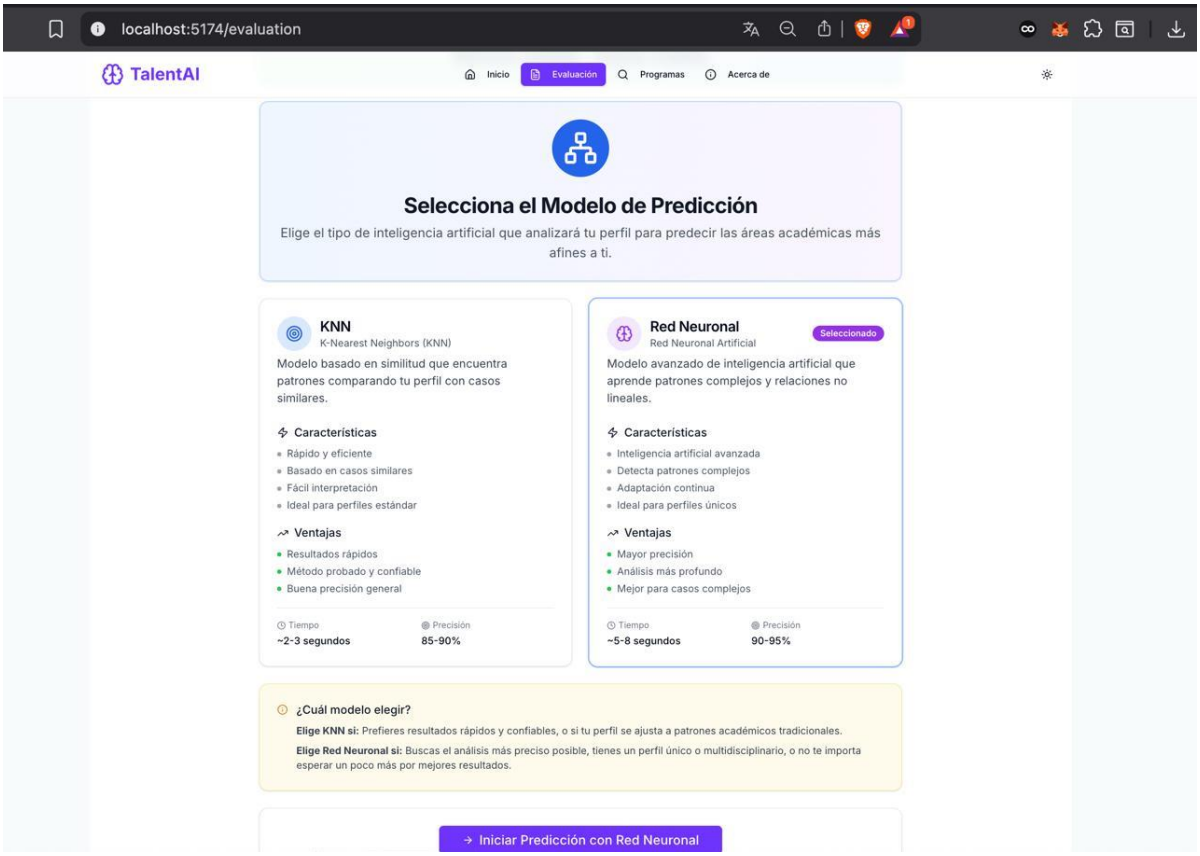
Ejemplo de Pregunta de la Dimensión



Nota. Ejemplo de pregunta de la dimensión Liderazgo en la evaluación de competencias, con escala de respuesta tipo Likert.

Figura 14

Pantalla de Selección del Modelo de Predicción



Nota. Pantalla de selección del modelo de predicción, comparando características y ventajas entre KNN y Red Neuronal.

Visualización de Resultados

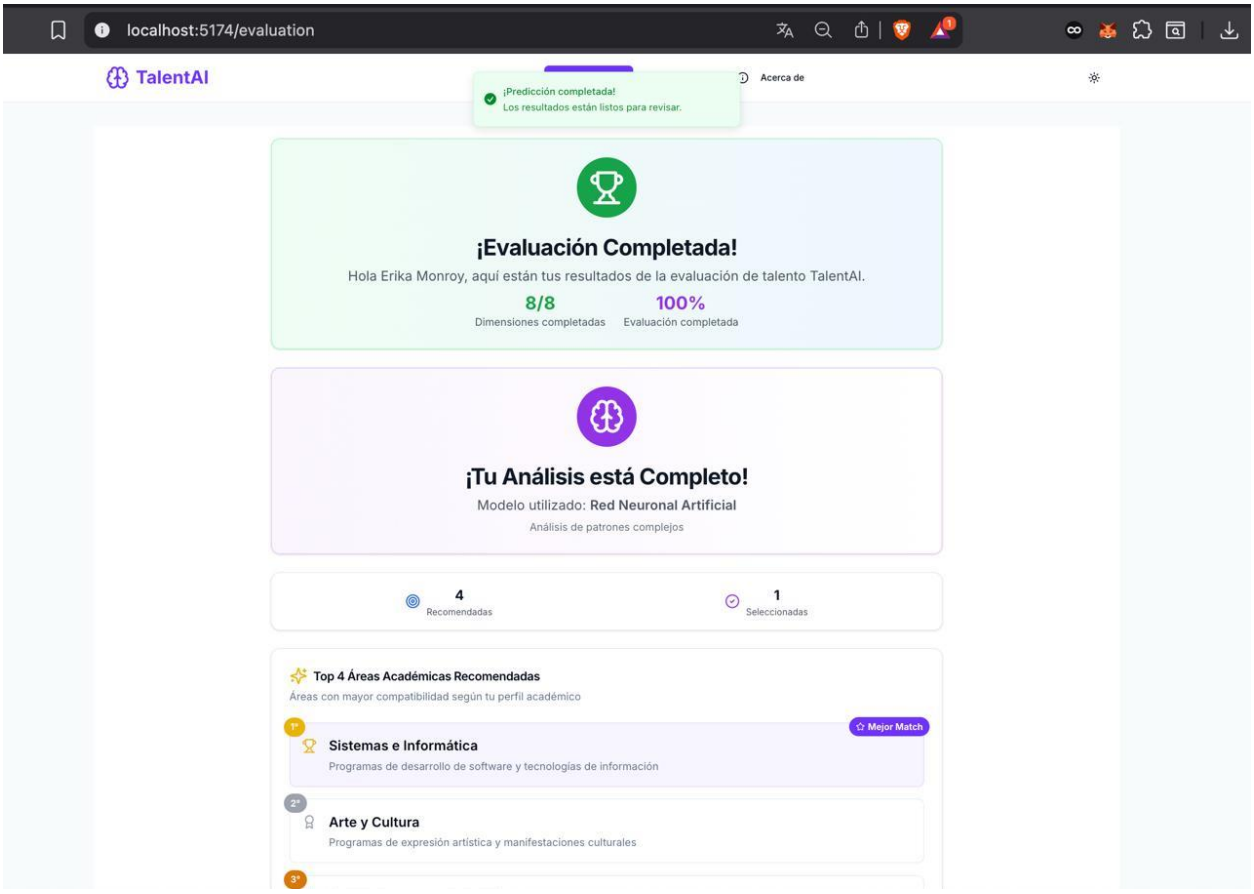
La página de resultados muestra las áreas vocacionales recomendadas a través de gráficos interactivos desarrollados con Recharts, incluyendo un radar chart de competencias por dimensión y barras de compatibilidad porcentual. Los usuarios pueden ver el top 5 de áreas recomendadas con explicaciones detalladas, comparar resultados entre ambos modelos de ML, y explorar programas académicos relacionados con filtros por institución, modalidad y duración.

La visualización incluye elementos de micro-interacción, tooltips informativos, y opciones de exportación en PDF para facilitar la toma de decisiones académicas y profesionales.

Cada sección del frontend incluirá capturas de pantalla que muestran la implementación final en la aplicación web, demostrando la interfaz de usuario, la experiencia de navegación y los resultados visuales del sistema TalentAI.

Figura 15

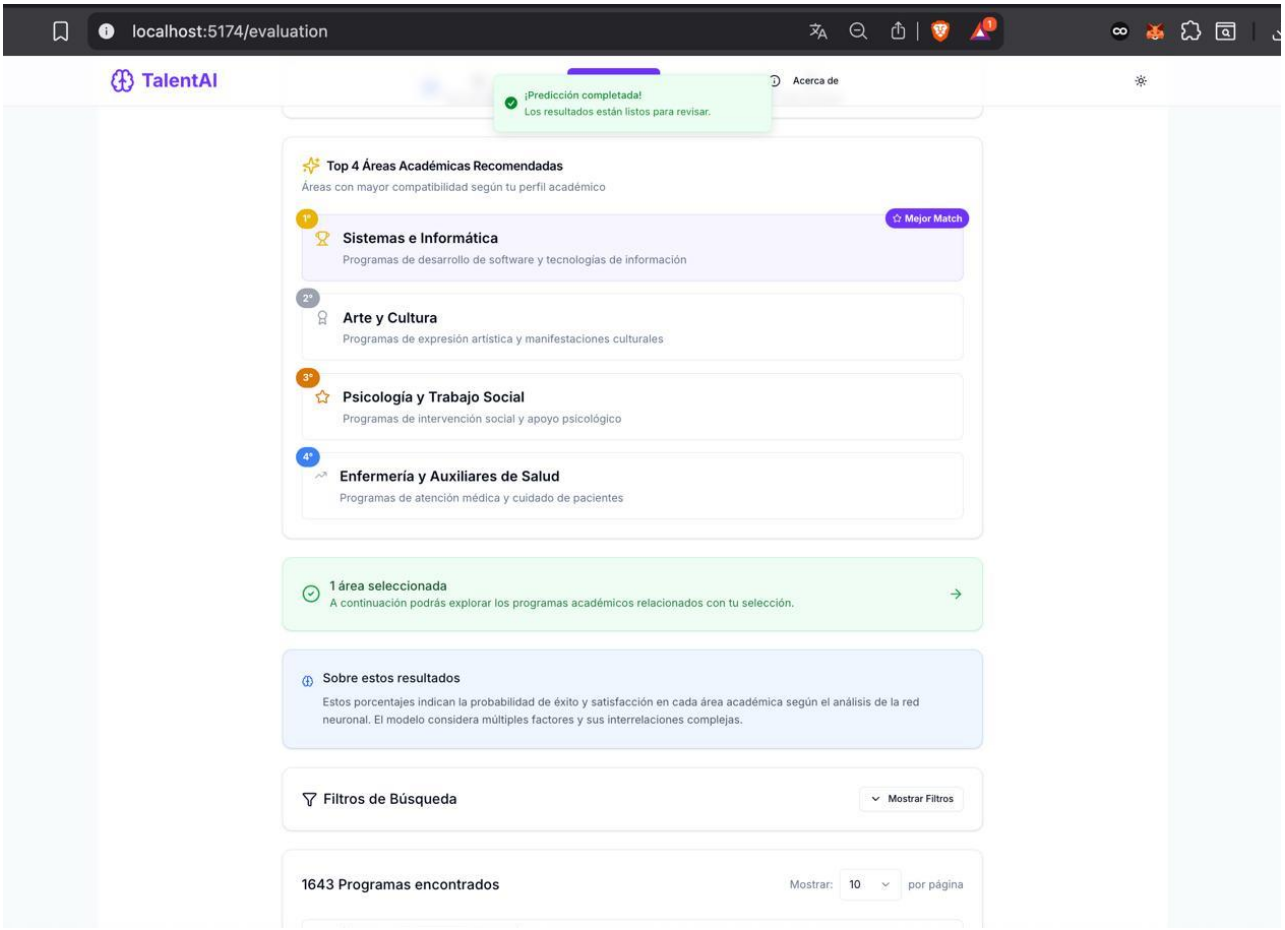
Resultados Generales De La Evaluación Completada



Nota. Resultados generales de la evaluación completada, con confirmación de dimensiones cubiertas y modelo de predicción utilizado.

Figura 16

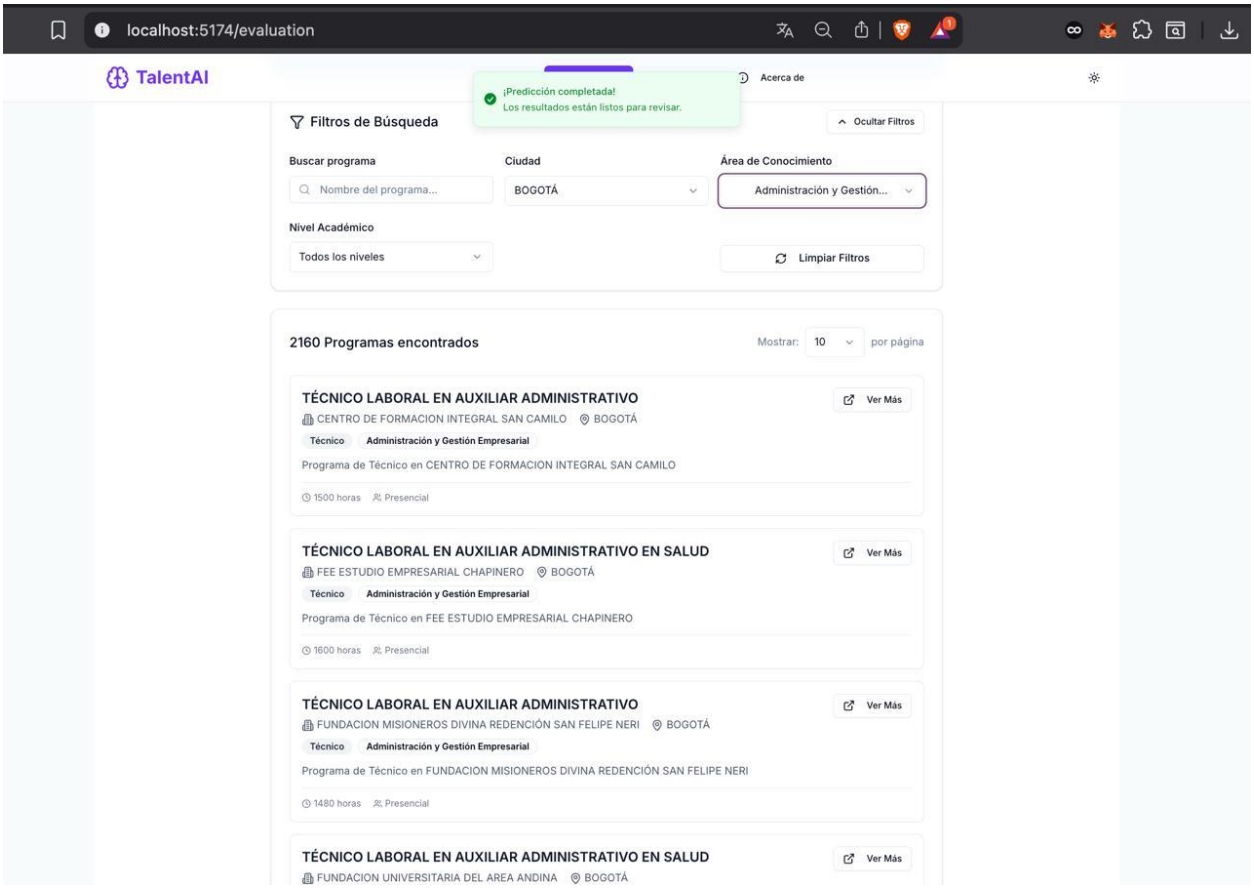
Listado de las Principales Áreas Académicas Recomendadas



Nota. Listado de las principales áreas académicas recomendadas, priorizadas según la afinidad del perfil del estudiante.

Figura 17

Exploración Detallada de Programas Académicos Sugeridos



Nota. Exploración detallada de programas académicos sugeridos según la selección del área de interés y filtros aplicados.

Conclusiones

Complementariedad de Modelos: KNN vs Redes Neuronales

Hallazgo Crítico: Los resultados de laboratorio mostraron que KNN obtuvo mejor F1-Score Macro (0.6262 vs 0.5832), mientras que las Redes Neuronales mantuvieron mejor accuracy (66.57%). Esta diferencia sugiere que las Redes Neuronales pueden ofrecer mayor estabilidad y potencial de escalabilidad para entornos de producción con grandes volúmenes de usuarios, mientras que KNN proporciona mejor balance en métricas multiclase. La evaluación en condiciones reales de producción será crucial para determinar el modelo óptimo.

Arquitectura de 4 Capas: Diseño Sólido y Escalable

La arquitectura implementada (Frontend Next.js + Backend FastAPI + ML Models + PostgreSQL) presenta un diseño sólido y bien estructurado que facilita el mantenimiento, actualizaciones y escalamiento horizontal del sistema. Esta separación clara de responsabilidades entre capas proporciona una base técnica robusta para el crecimiento futuro de la aplicación, aunque aún requiere validación en entornos de producción con usuarios reales.

Formulario de 100 Competencias: Base Sólida con Potencial de Mejora

El diseño de evaluación en 8 dimensiones con 100 preguntas representa una base sólida y funcional para el sistema, pero requiere investigación psicológica más profunda para optimizar la precisión diagnóstica. La ausencia de pruebas con usuarios reales limita la validación empírica del instrumento. Es fundamental establecer una colaboración interdisciplinaria entre psicólogos educativos, expertos en orientación vocacional para validar el instrumento psicométricamente y mejorar la calidad del dataset, lo que resultaría en predicciones más precisas, científicamente fundamentadas y contextualmente relevantes. En este sentido, el formulario de 100 competencias constituye un insumo valioso ya diseñado en TalentIA, cuya validación empírica y difusión

detallada deberá abordarse en investigaciones posteriores para fortalecer su rigor psicométrico y su utilidad práctica.

Balance Estratégico: Interpretabilidad vs Precisión y Escalabilidad en Producción

La implementación dual (KNN + Redes Neuronales) permite optimizar según el contexto de uso: KNN para casos que requieren explicabilidad inmediata y Redes Neuronales para máxima precisión en producción. Esta flexibilidad arquitectónica es clave para diferentes escenarios de orientación vocacional. Además, la comparación del rendimiento de ambos modelos en producción es fundamental para identificar cuál escala mejor bajo diferentes cargas de trabajo y volúmenes de usuarios, proporcionando datos empíricos para decisiones arquitectónicas futuras.

Recomendaciones

Evaluación y Optimización de Modelos en Producción

Aunque KNN mostró mejor F1-Score en laboratorio, las Redes Neuronales presentan mayor accuracy y potencial de escalabilidad. Se requiere evaluación en producción para determinar el modelo óptimo según diferentes escenarios de uso y cargas de trabajo.

Implementación:

1. Implementar A/B testing riguroso entre KNN y Redes Neuronales en producción
2. Optimizar hiperparámetros de ambos modelos para condiciones reales
3. Desarrollar métricas de monitoreo comparativas en tiempo real
4. Establecer criterios de selección automática de modelo según contexto de uso

Investigación Psicológica y Refinamiento del Instrumento de Evaluación

Realizar investigación psicológica profunda en colaboración con expertos en psicología educativa para rediseñar y validar científicamente el instrumento de evaluación de competencias.

Implementación:

1. Colaborar con psicólogos educativos y expertos en orientación vocacional
2. Realizar análisis factorial y validación psicométrica del instrumento actual
3. Rediseñar preguntas basándose en teorías psicológicas consolidadas
4. Implementar pruebas piloto con estudiantes reales para validación empírica
5. Establecer normas y baremos específicos para el contexto colombiano

Recolección de Datos Reales y Entrenamiento Continuo

Implementar un sistema de retroalimentación continua que capture datos reales de estudiantes y egresados (universitarios, tecnólogos y técnicos) que ya eligieron carrera, de manera que el modelo pueda entrenarse con trayectorias verificadas. Este proceso debe

complementarse con pruebas piloto en instituciones educativas bajo consentimiento informado, y con la integración de datos oficiales del ICFES, SNIES y SENA. De esta forma, se consolidará un pipeline robusto de información real que permitirá mejorar la precisión predictiva, fortalecer la confiabilidad del sistema y garantizar su pertinencia en el contexto colombiano.

Implementación:

1. Diseñar un formulario inicial para recolectar datos de estudiantes y egresados que ya eligieron carrera.
2. Ejecutar pruebas piloto en instituciones educativas, con consentimiento informado.
3. Desarrollar un pipeline de integración con bases oficiales (ICFES, SNIES y SENA) combinado con los datos piloto.
4. Crear un sistema de seguimiento longitudinal de estudiantes para evaluar trayectorias académicas y laborales reales.
5. Implementar reentrenamiento automático mensual con datos actualizados.
6. Establecer métricas de validación basadas en resultados académicos y de inserción laboral.

Combinación Optimizada de Modelos

Desarrollar un metamodelo ensemble que combine las fortalezas de Redes Neuronales (precisión) y KNN (interpretabilidad local) mediante algoritmos de voting ponderado y stacking avanzado.

Implementación:

1. Crear algoritmo de combinación adaptativa según confianza de predicciones
2. Implementar pesos dinámicos basados en características del usuario

3. Desarrollar sistema de consenso para casos de alta incertidumbre
4. Validar mejoras mediante cross-validation estratificada

Métricas de Evaluación Avanzadas y Monitoreo Integral

Implementar métricas de evaluación más robustas y específicas que vayan más allá de Accuracy y F1-Score, proporcionando una evaluación integral del rendimiento del sistema en diferentes contextos y poblaciones.

Implementación:

1. Métricas de Clasificación Avanzadas: Precision, Recall, AUC-ROC, Matthews Correlation Coefficient (MCC)
2. Métricas de Calibración: Brier Score, Expected Calibration Error (ECE) para medir confianza de predicciones
3. Métricas de Equidad: Demographic Parity, Equal Opportunity para evaluar sesgos por género, estrato socioeconómico
4. Métricas de Robustez: Stability Index, Adversarial Accuracy para medir consistencia ante variaciones
5. Métricas de Negocio: Satisfaction Score, Recommendation Acceptance Rate, Long-term Academic Success Correlation
6. Dashboard en tiempo real con alertas automáticas para degradación de rendimiento.

Referencias Bibliográficas

Bennett, S. (2018). *Technology and student retention: An overview of research and practice.* Educational Technology Journal, 58(2), 15–22.

Estupiñán, A. M. B., & Mesa, L. G. (2023). *Inteligencia Artificial: el futuro del empleo.* Revista Lecciones Vitales, lv0103. https://doi.org/10.18046/rlv.2023.6118

Garriga Trillo, A. J. (2012). *Introducción al análisis de datos: formulario y tablas.* UNED.

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta.* McGraw-Hill.

Hooley, T. (2017). *The evidence based on lifelong guidance: A guide to key studies and research.* ELGPN Publications.

Hooley, T., & Watts, A. G. (2016). *Career readiness: A critical review.* British Journal of Guidance & Counselling, 44(3), 211–223.

Hooley, T., & Watts, A. G. (2016). *The evolution of vocational guidance: Towards the era of digitalization.* Journal of Career Assessment.

ICFES. (s. f.). *Data ICFES. Instituto Colombiano para la Evaluación de la Educación.* de https://www.icfes.gov.co/investigaciones/data-icfes/

Jackson, S. (2019*). Labour market analysis for career development: A practitioner's guide. Cambridge University Press.*

Jamieson, L., & O'Mara, J. (2022). *Leveraging Artificial Intelligence for enhanced career guidance systems. Journal of Educational Technology & Society.*

Kift, S. (2018). *Student success: Why orientation and early engagement matter. Journal of University Teaching and Learning Practice,* 15(5), 1–15.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). *Consequences of individuals' fit at work: A meta-analysis. Personnel Psychology,* 58(2), 281–342. https://doi.org/10.1111/j.1744-6570.2005.00672.x

Laboratorio de Economía de la Educación (LEE). (2022). *Los "ninis" en Colombia: Una realidad preocupante y un desafío para el futuro.* Pontificia Universidad Javeriana. https://lee.javeriana.edu.co/-/lee-informe-60

Laboratorio de Economía de la Educación (LEE). (2023). *Informe No. 74: Deserción en la educación superior en Colombia.* https://lee.javeriana.edu.co/-/lee-informe-74

Laboratorio de Economía de la Educación (LEE). (2024). *Informe No. 99: Sin trabajo ni educación.* https://lee.javeriana.edu.co/publicaciones-y-documentos

Ministerio de Educación Nacional de Colombia. (2024*). Observatorio Laboral para la Educación.* https://snies.mineducacion.gov.co/portal/OBSERVATORIO/

Morinson Negrete, D. (2020). *Grado de Satisfacción de un Sistema de Información para la Orientación Vocacional en Estudiantes de la Institución Educativa Los Morales.* UMECIT, Panamá.

Noticia: En Colombia, el 67% de los jóvenes Ninis son mujeres. (s. f.). Portal Universitario Javeriana. https://www.javeriana.edu.co/unoticias/historico/-/asset_publisher/up4MWBzHuAZt/content/id/6560122

Ozdemir, S. (2016). *Principles of Data Science*. O'Reilly Media.

PerúEduca. (2021). *Sistema nacional de orientación vocacional del Ministerio de Educación del Perú.* https://www.perueduca.pe

Peset, F., & Millán González, L. (2017). *Ciencia abierta y gestión de datos de investigación (RDM).* Ediciones Trea.

Phillips, L. D., & Bana e Costa, C. A. (2007). *Transparent prioritisation with multi-criteria decision analysis. Annals of Operations Research,* 154, 51–68.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook. In Recommender Systems Handbook* (pp. 1–35). Springer.

Roa, C. A. D. (2023). *Orientación vocacional en la educación media técnica profesional. Revista de Teoría y Didáctica.* https://espacio.digital.upel.edu.ve/index.php/TD/article/view/581

Rodríguez, A. (2022). *Jóvenes NINI en Bogotá: Análisis de barreras. Laboratorio de Economía de la Educación.* https://lee.javeriana.edu.co/

Rodríguez Cardozo, L. (2022). *Evaluación del programa Reto a la U en jóvenes de Bogotá. Universidad de los Andes.*

Rodríguez, S. (2022). *Factores de incidencia para que los jóvenes bogotanos se conviertan en NINIs. Espacio Sociológico,* (2), 47–81. https://hemeroteca.unad.edu.co/index.php/sociologico/article/view/5520

Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach (3rd ed.). Pearson Education.*

Secretaría de Educación de Bogotá. (2023). *Jóvenes a la U y Matrícula Cero.* https://educacionbogota.edu.co/

SENA. (2023). Oferta académica gratuita. https://oferta.senasofiaplus.edu.co/sofia-oferta/

Shneiderman, B. (2010). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th ed.). Pearson.*

Singh, A., & Dhir, S. (2021). *AI-powered career guidance. Computers & Education,* 164, 104122. https://doi.org/10.1016/j.compedu.2021.104122

Sistema Nacional de Información de la Educación Superior - *SNIES*. (s. f.). Portal SNIES. Ministerio de Educación Nacional de Colombia. https://snies.mineducacion.gov.co/portal/

Smith, A., & Johnson, K. (2018). *Integrating machine learning in vocational guidance. Journal of Vocational Behavior.*

Smith, K., & Johnson, M. (2018). *AI-based recommendation systems in vocational guidance. Journal of Educational Computing Research,* 56(4), 685–705.

Smith, P. (2019). *Bridging the skills gap with big data. Journal of Education and Work.*

The Future of Jobs Report 2023. (2024, March 28). World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2023/

Tinto, V. (2017*). Through the eyes of students. Journal of College Student Retention,* 19(3), 254–269. https://doi.org/10.1177/1521025115621917

Zhang, Y. (2018). *User satisfaction in recommender systems. ACM Transactions on Interactive Intelligent Systems,* 8(3), 1–25. https://doi.org/10.1145/3158665

Apéndices

Apéndice A

Repositorio del Proyecto TalentAI

https://github.com/ErikaMonroy/talent_ai

Apéndice B

Notebook de Análisis Comparativo de Modelos

https://github.com/ErikaMonroy/talent_ai/blob/main/machine_learning_model/models/TalentAI_Model_Comparison.ipynb

Apéndice C

Estructura del Proyecto TalentAI

La organización de carpetas y archivos del proyecto TalentAI se estructuró de la siguiente manera:

- backend/: contiene la API y los servicios de backend.
- frontend/: incluye la aplicación web desarrollada en Next.js.
- modelo/: agrupa los modelos de Machine Learning y los análisis asociados.
- data/: almacena los datasets y archivos de datos utilizados.
- Models/: contiene la implementación de los algoritmos de Machine Learning.
- README.md: archivo principal de documentación del proyecto.
- Nota. Elaboración propia a partir del repositorio TalentAI en GitHub