Como citar este artículo: Como citar este artículo: E. J. Brand C, G. M. Ramírez, J. Diaz and F. Moreira, "Towards Educational Sustainability: An AI System for Identifying and Preventing Student Dropout," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, doi: 10.1109/RITA.2024.3381850.

Towards Educational Sustainability: An AI System for Identifying and Preventing Student Dropout

Erika J. Brand C., Gabriel M. Ramírez V., Jaime Diaz, and Fernando Moreira

***Abstract*—The design and development of a web application to identify a high or low probability of student dropout at the National Learning Service (SENA) in Colombia, aiming to streamline the process of identifying and supporting potential candidates for assistance provided by the institution through the student welfare department. Throughout the development, socioeconomic variables with the highest impact on characterized academic dropout processes to create a dataset. This dataset was then utilized with various artificial intelligence techniques explored in Machine Learning (Decision Trees, K-means, and Regression), ultimately determining the most effective algorithm for integration into the Software. The decision tree classification technique emerged as the most effective, achieving an impressive accuracy of 91% and a minimal error rate of 9%, substantiating its state-of-the-art standing. As a result, this Software has optimized processes within the Student Welfare Department at SENA and is adaptable for use in any higher education institution.**

Index Terms*— *Artificial Intelligence; Machine Learning School Dropout; Higher Education; Colombia.

I. Introduction

Colombia presenta un sinnúmero de oportunidades y opciones para acceder a la educación superior. Un estudio de 2012 de la Organización para el Desarrollo y la Cooperación Económica (OCDE) destacó el importante avance del país en la ampliación de la cobertura de programas de formación técnica, tecnológica y profesional. Sin embargo, a la par de este crecimiento en la oferta, se han presentado alarmantes tasas de deserción. En los programas universitarios, la tasa de deserción se ha disparado a 44,9%, mientras que los estudiantes de programas técnicos y tecnológicos están desertando a tasas de 62,4% y 53,8%, respectivamente, como indica [1].

El Ministerio de Educación Superior de Colombia, a través de su sistema de prevención y análisis de la deserción, ha identificado la inadecuada educación preuniversitaria como el principal precursor de los problemas de deserción en el nivel universitario. Los retos económicos y la falta de vocación profesional agravan aún más estos problemas, impactando directamente en las trayectorias vitales de los estudiantes y sus familias.

En respuesta a estos retos, el gobierno y varias instituciones de educación superior han puesto en marcha estrategias para frenar las altas tasas de abandono. El "Acuerdo por lo Superior 2034 - Propuesta de Política Pública para la Excelencia de la Educación Superior en Colombia en el Escenario de la Paz" [2] ejemplifica estos esfuerzos, haciendo énfasis en la equidad, la eficiencia y la intervención temprana para asegurar la persistencia y graduación de los estudiantes en los diferentes programas de educación superior.

El SENA, que funciona como una entidad de formación profesional, ofrece programas complementarios técnicos y tecnológicos, matriculando aproximadamente 463.653 aprendices anualmente [3]. Si bien la institución ha implementado estrategias de bienestar para los aprendices que abarcan apoyo socioeconómico, asesoría, orientación psicológica, promoción de la salud y prevención de enfermedades, aún se requieren medidas adicionales para abordar el problema de manera efectiva.

Además, el aprovechamiento de tecnologías como la Inteligencia Artificial, en concreto el Aprendizaje Automático, abre vías para la toma de decisiones y la ejecución de tareas sin instrucciones directas. La implementación de una aplicación que analice los datos socioeconómicos y académicos de los estudiantes hace posible la predicción de los factores de deserción en etapas tempranas, asegurando la retención proactiva de los estudiantes.

La estructura de este trabajo comprende las siguientes secciones: introducción, antecedentes, método, contribuciones, evaluación y resultados, conclusiones, discusiones y trabajo futuro, con anexos del proyecto al final.

II. Contexto

El abandono escolar es una decisión personal de abandonar o continuar la formación académica en la misma institución. Sin embargo, esta decisión está influenciada por circunstancias internas o externas tanto en una nueva institución como en el mismo programa educativo. Diversas variables afectan esta

Erika J. Brand C. is with the Servicio Nacional de Aprendizaje (SENA), Bogotá, Colombia (e-mail: bje0284@hotmail.com).

Gabriel M. Ramírez V. is with the Facultad de Ingenierías, Universidad de Medellín, Medellín, Colombia (corresponding author e-mail: gramirez@udemellin.edu.co).

Jaime Diaz is with the Departamento de Ciencias de Computación e Informatíca, Universidad de la Frontera, Temuco, Chile (jaimeignacio.diaz@ufrontera.cl).

Fernando Moreira is with the REMIT, IJP, Universidade Portucalense, Porto & IEETA, Universidade de Aveiro, Aveiro, Portugal (fmoreira@upt.pt).) (corresponding author e-mail: fmoreira@upt.pt)

decisión; entre ellas se encuentran los intereses personales, los afectos económicos, el perfil ocupacional del programa y los modelos pedagógicos institucionales, entre otros [4].

Aunque la educación es una de las mejores herramientas que puede tener una persona para garantizar su desarrollo socioeconómico y permitirle acceder a un trabajo y a un entorno productivo, se presentan muchos casos de deserción escolar, especialmente en instituciones educativas de nivel superior. Por ello, el Ministerio de Educación Nacional de Colombia y las diferentes instituciones educativas han implementado diversas estrategias para enfrentar este gran reto y reducir anualmente los índices de deserción escolar.

Sin embargo, al identificar las causas que obligan a los estudiantes a interrumpir tempranamente sus estudios, se evidencia la diversidad de variables, lo que obliga a buscar nuevas alternativas para reducir estos índices. Teniendo en cuenta lo anterior, a continuación se presentan algunas investigaciones sociales y tecnológicas que disminuyen los índices de deserción escolar en Colombia y otros países.

A. Abandono escolar en instituciones de educación superior

En diferentes países, el problema de la deserción en la educación superior se aborda desde diferentes perspectivas. En 2002, una investigación trabajó el tema de la deserción, incluyendo estrategias de retención estudiantil desde un punto de vista conceptual. Se desarrollaron modelos teóricos basados en factores psicológicos, económicos, sociológicos y organizacionales, que predicen el abandono de la formación para implementar estrategias de retención [5].

Visto desde otro punto de vista, el abandono de un programa académico suele presentarse como la última opción en una cadena de fracaso escolar para entender esta decisión se debe hacer un análisis para entender su motivación real. En otras palabras, cuál fue la causa de que esta decisión impidiera al sistema educativo alcanzar sus objetivos pedagógicos en el momento oportuno y sin malgastar sus recursos humanos y financieros [6]. Es fundamental entender que la institución debe hacer todo lo posible para mantener a los estudiantes en sus procesos de formación y evitar que abandonen, afectando su autonomía y productividad.

En países como Estados Unidos, se muestra que la raza y la condición de migrante son algunas de las variables que sobresalen al medir las tasas de deserción, a diferencia de los países latinos, cuyo fenómeno generalmente se relaciona con motivos económicos. Por otro lado, algunos estudios han demostrado que los estudiantes en riesgo de deserción tienen un perfil relacionado con la inasistencia o bajas calificaciones, lo que se refleja en el desperdicio académico en las etapas preuniversitarias [7].

Por su parte, un estudio del Ministerio de Educación Nacional [8] posiciona el capital cultural y académico con el que ingresan los estudiantes a la educación superior como el principal factor de deserción, seguido de los componentes socioeconómico, institucional y de orientación vocacional en la Figura 1.

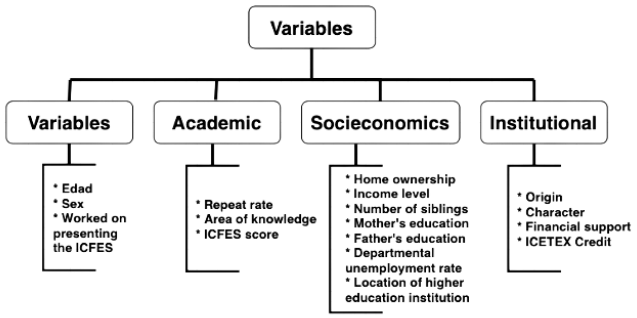

Figure 1. Variables usadas para explicar la deserción en instituciones de educación superior en Colombia

Christenson et al. [9] indican cómo prevenir el abandono escolar. Por lo tanto, es necesario iniciar una intervención para los alumnos que presentaban este riesgo y analizar el contexto social, por ejemplo, sus relaciones con profesores y compañeros, y promover el crecimiento personal en diversos aspectos, teniendo en cuenta que el objetivo principal es prevenir el abandono escolar. Finalmente, los alumnos intervenidos pudieron completar su etapa formativa.

Del análisis de los procesos de investigación realizados en el SENA en un estudio desarrollado en el centro de formación de Servicios Financieros de Bogotá, a través de una metodología descriptiva-analítica de variables sociofamiliares y económicas coincide que el factor que tiene mayor incidencia significativa en la deserción de los estudiantes de los diferentes programas de formación [10] coincidiendo con estudios anteriores.

En otro estudio, una revisión de 28 investigaciones realizadas en Colombia entre 2006 y 2016 identificó que los estudiantes con bajo nivel socioeconómico fortalecen la estructura de deserción en el país [11]. Destacaron esta problemática como una de las más difíciles de atender, a pesar de que actualmente se presentan muchos incentivos económicos a los estudiantes para garantizar la continuidad en la formación académica.

Estudios recientes sobre el abandono se centran en identificar y comparar las variables clave que influyen en el abandono temprano de la universidad en diferentes áreas de conocimiento e instituciones. Un estudio [37] profundiza en el análisis de variables relacionadas con las áreas académicas de los estudiantes, mientras que otro estudio se centra en la mejora de los procesos de análisis académico. En concreto, este último investiga las principales variables que afectan al abandono temprano entre estudiantes de primer curso de carreras técnicas, con el objetivo de mejorar los procesos de aprendizaje y prevenir el abandono en los semestres iniciales de los programas de estudio [38].

B. Aplicación de herramientas tecnológicas de IA para predecir el abandono escolar

En diferentes instituciones educativas de nivel superior en Colombia, se han implementado diversos modelos predictivos para fortalecer los procesos de retención estudiantil mediante la aplicación de técnicas y algoritmos como minería de datos, árboles de decisión, Naive-Bayes, redes neuronales, árboles de decisión y clasificación, entre otros.

La metodología CRISP-DM presenta importantes ventajas cuando se aplica a procesos asociados a la minería de datos. Su

utilización ha demostrado ser instrumental en el reconocimiento de patrones de deserción entre estudiantes de pregrado de la Universidad de Nariño y la Institución Universitaria CESMAG. Enfoca la IA en datos socioeconómicos y académicos como insumos primarios, se encontró que entre los factores críticos que contribuyen a la deserción se encuentran la pertenencia a un estrato socioeconómico bajo, ser menor de edad y mantener un promedio de calificaciones bajo [12].

Las tareas descriptivas basadas en el algoritmo K-means y la clasificación empleando árboles de decisión y Naive Bayes también han permitido predecir en el programa de Ingeniería Electrónica e Ingeniería de Sistemas de la Universidad Popular del Cesar. Los malos resultados en las calificaciones durante los primeros cuatro semestres son los factores más críticos que afectan la deserción estudiantil [13].

Por otro lado, para predecir la probabilidad de deserción, se utilizaron diversas técnicas de clasificación basadas en árboles de decisión y la metodología Knowledge Discovery in Database (KDD), las cuales incluyen cinco etapas específicas: selección, procesamiento, transformación, minería de datos y evaluación [14] para detectar el nivel y las calificaciones como los principales factores de deserción estudiantil.

Con estos resultados, es posible demostrar la gran aplicabilidad del Machine Learning ML y otras ramas de la IA para llevar a cabo procesos de predicción de la deserción estudiantil y permitir a las instituciones de educación superior implementar estrategias de retención de estudiantes.

En otros países también se han desarrollado diferentes proyectos para mitigar este problema. Por ejemplo, [15] utilizó los servicios de Azure Machine Learning de Microsoft para realizar el comportamiento de datos de árboles de decisión. En [16] también se utilizan árboles de decisión (CBAD). Además, utilizaron RapidMiner Software para la optimización de parámetros para la predicción, una herramienta especializada en minería de datos con resultados de precisión del 87,27% [17].

En [18], se realizó un estudio de investigación cualitativa basado en 64 artículos. Relacionan diferentes técnicas aplicadas a los intereses y riesgos en el aprendizaje de los estudiantes en diferentes niveles educativos, incluyendo la educación superior. Como resultado, en la deserción estudiantil, se encontró que los árboles de decisión son un método valioso para hacer este tipo de predicción. Este método permitió identificar los siguientes factores que ayudan a predecir la deserción: antecedentes familiares, nivel socioeconómico de las familias, grado de secundaria y resultados de los exámenes.

Al utilizar árboles de decisión con IBM SPSS, durante el proceso de predicción fue frecuente encontrar datos desequilibrados que afectaban a los resultados de la predicción. Sin embargo, fue necesario combinar las técnicas Support Vector Machine (SVM) y Synthetic Minority Oversampling Technique (SMOTE) para solucionarlo y mejorar el rendimiento.

En la actualidad, muchas instituciones educativas utilizan este sistema para identificar a los estudiantes con riesgo de abandono temprano. Los índices más elevados se dan en ciencias, tecnología, ingeniería y matemáticas, superando el 60% de riesgo de abandono en los primeros años.

A partir del análisis realizado en los diferentes proyectos y verificando las características de la población de estudiantes del SENA, se tendrán en cuenta las variables psicológicas, económicas, sociológicas y académicas para la caracterización y creación del conjunto de datos. Se aplicará minería de datos para su preparación, preprocesamiento y transformación. Se explorarán diferentes técnicas de aprendizaje automático como redes bayesianas, máquinas de vectores de soporte y árboles de decisión para identificar las más viables ya que varios métodos no funcionan de la misma manera con todos los conjuntos de datos.

C. Aprendizaje automático

El aprendizaje automático o Machine Learning, es una subárea de la Inteligencia Artificial, implica el desarrollo de algoritmos capaces de adquirir conocimientos mediante el análisis de diversos conjuntos de datos para predecir resultados predefinidos [19]. Existen varios tipos de aprendizaje automático:

1) Aprendizaje supervisado: Este enfoque permite a los algoritmos aprender a partir de datos previamente etiquetados, donde el conjunto de datos posee características y una etiqueta establecida. La etiqueta asignada a los datos existentes sirve de referencia para etiquetar los nuevos datos, que se convierten en la entrada del algoritmo.

2) Aprendizaje no supervisado: El aprendizaje no supervisado implica conjuntos de datos que carecen de etiquetas predefinidas. En su lugar, el algoritmo busca similitudes en la información y posteriormente la agrupa o clasifica basándose en relaciones inherentes.

3) Aprendizaje profundo: Representa una categoría dentro del aprendizaje automático en la que el proceso de aprendizaje continúa y la información se analiza de varias formas. El aprendizaje profundo permite diversos tipos de entrada como imágenes, audio y textos.

Existen dos diferencias significativas entre los algoritmos del aprendizaje supervisado basados en la predicción de valores y en la predicción de categorías. En primer lugar, cuando se habla de predicción de categorías, se trata de algoritmos de clasificación; cuando se predice un valor numérico, se trata de una regresión [20]. En la predicción de categorías, el objetivo principal es clasificar imágenes (por ejemplo, gato, perro, caballo), información financiera y diagnósticos médicos. Como algoritmo de predicción de valores, incluye la regresión (éxito estudiantil, precio de las acciones y comportamiento de navegación por Internet).

En el proceso de creación de sistemas de aprendizaje es necesario analizar la relación entre las características (features) y el objetivo (target) para definir el proceso de creación de sistemas de aprendizaje [21]. Se describe cada paso: 1. Desarrollo y comprensión de la tarea u objetivo; 2. Recopilación y comprensión de los datos; 3. Preparación de los datos para el modelado; 4. Creación del modelo a partir de las relaciones entre los datos; 5. Evaluación y comparación del modelo con otros modelos similares, y 6. Transición del modelo a un sistema desplegable como software.

III. METODO

El problema, los objectivos y la metodología del trabajo realizaro se presentan a continuación.

A. Problema

El Servicio Nacional de Aprendizaje SENA es una institución pública del orden nacional que ofrece formación gratuita y gradúa anualmente a miles de colombianos; sin embargo, el porcentaje de graduados se ha visto afectado negativamente desde el año 2016, evidenciando que en promedio, solo se gradúa el 29,9% de las personas que se matriculan en la entidad, las razones más ocasionales son sociales y económicas. El SENA creó la estrategia Bienestar del Aprendiz, la cual busca contribuir a la permanencia y desempeño exitoso de los aprendices en la entidad.

Sin embargo, teniendo en cuenta el alto número de aprendices que hay en cada uno de los centros de formación en todo el país, es difícil detectar a tiempo las personas que están en riesgo de desertar de su programa de formación, así como las posibles causas que pueden llevar a esta decisión.

Con este panorama, se hace complicado implementar acciones preventivas que ayuden a incrementar las tasas de graduación y, con ello, mejorar el desarrollo social y técnico de los futuros trabajadores. Entonces surge la pregunta: ¿Una aplicación web basada en machine learning que prediga el abandono de los aprendices contribuirá a la reducción de las tasas de deserción en el Servicio Nacional de Aprendizaje SENA?

B. Objetivos

Los objetivos del proyecto son los siguientes:

Objetivo general

--Desarrollar una aplicación que permita predecir la deserción en aprendices del nivel técnico y tecnológico del Servicio Nacional de Aprendizaje SENA, a partir del análisis de variables socioeconómicas y académicas utilizando un modelo basado en Machine Learning.

Se pretende desarrollar una aplicación predictiva que anticipe riesgos de deserción estudiantil dentro del SENA en Colombia.

Objetivos específicos:

--Analizar las variables socioeconómicas y académicas de los aprendices, además de las diferentes técnicas de Machine Learning utilizadas para lograr el objetivo del proyecto.

--Caracterizar a partir de minería de datos un conjunto de variables socioeconómicas y personales de una muestra de aprendices de la institución como base para la realización del modelo de Machine Learning.

--Aplicar diferentes técnicas de Machine Learning identificadas para seleccionar aquella con mejor tasa de éxito.

--Integrar el algoritmo de Machine Learning seleccionado con la interfaz gráfica y la base de datos para un entorno web.

C. Metodología

En este proyecto de investigación se utilizó un método cuantitativo. Se recopilaron medidas numéricas y datos de análisis estadístico para poner a prueba una teoría. El planteamiento comenzó con una idea, definió un problema y unos objetivos de investigación, se revisó la bibliografía y desarrolló una perspectiva teórica. Se utilizó un método cuantitativo para esta iniciativa y se recopilaron mediciones numéricas y datos de análisis estadístico para poner a prueba la teoría de base. Este enfoque parte de una idea, define un problema y unos objetivos de investigación, revisa la bibliografía y elabora una perspectiva teórica.

A continuación, se determina algunas variables y una estrategia para ponerlas a prueba. Se midieron las variables en un contexto específico y se aplicó el método estadístico para obtener las conclusiones [22]. Los resultados obtenidos en el análisis de las diferentes variables del proyecto se utilizaron para desarrollar el software de predicción de la deserción de los estudiantes del SENA Colombia.

Se utilizó un método de desarrollo denominado modelo CRISP-DM [23], uno de los más utilizados y completos para trabajar con un enfoque de minería de datos [24]. Esta metodología consta de 6 fases:

--Comprensión del Negocio (Objetivos y requerimientos desde una perspectiva no técnica). En esta fase se debe plantear el problema, los objetivos y el instrumento para recoger la información.

--Entender los datos (Familiarizarse con los datos, teniendo en cuenta los objetivos empresariales). En esta fase se recogen, verifican y exploran los datos iniciales y se garantiza su calidad.

--Preparación de los datos (Obtener el conjunto de datos). En esta fase se seleccionan, limpian, construyen, integran y verifican los datos para obtener el conjunto de datos final para el entrenamiento del algoritmo elegido.

--Modelado (Aplicar técnicas de minería de datos a los conjuntos de datos). En esta fase se selecciona, construye y evalúa el algoritmo para obtener el modelo de Machine Learning que dará respuesta al problema planteado.

--Evaluación (Identificar si los modelos aplicados son útiles para el negocio). En esta fase se evalúan los resultados, determinando la precisión de los medios propuestos. Se determinan los siguientes pasos.

--Despliegue o implementación (Integrar los modelos en la toma de decisiones de la organización). En esta fase se desarrolla el producto final (Software), que permite aplicar el modelo en entornos naturales.

1) Entendiendo el negocio: Establecer un claro entendimiento de la deserción dentro del Servicio de Aprendizaje del SENA[25]; los escenarios específicos se definen como los siguientes: (i) no presentar el 30% de las evidencias académicas requeridas sin justificación durante un trimestre académico, (ii) acumulación de tres días continuos de ausencias o cinco días no continuos no justificados en formación presencial, y (iii) inasistencia a tres citaciones en formación virtual, incumplimiento de dos planes de mejoramiento, o no acceder al Sistema de Gestión del Aprendizaje (SGA) por más de diez días consecutivos.

A lo largo de la ejecución del proyecto, es vital la asistencia del área de Bienestar Estudiantil, que realiza un seguimiento diligente de todos los alumnos matriculados en el programa de formación, independientemente de su modalidad. Este esfuerzo de colaboración garantiza un apoyo integral a los objetivos. Además, se puede delinear las variables necesarias para el

algoritmo, integrándolas posteriormente a la perfección en el software para mejorar las capacidades predictivas y agilizar las estrategias de intervención.

Para la ejecución del proyecto, se cuenta con el apoyo del área de bienestar estudiantil, que realiza el seguimiento de todos los alumnos inscritos en un programa de formación a través de diferentes modalidades. Además, fue posible definir las variables a contabilizar para la realización del algoritmo y posterior integración en el Software.

En cuanto a la selección de variables, se realizó un estudio en el Centro de Servicios Financieros del SENA [25]: Las variables asociadas al riesgo de deserción y su relación con los programas que tienen el bienestar del estudiante se establecen en categorías como familiar, económica y laboral, tiempo y distancia. Estas variables asociadas a los aspectos familiares tienen un incremento del 18% en la probabilidad de deserción de los estudiantes, los económicos y laborales están en un 45% y el tiempo y la distancia en un 20%.

También se tuvo en cuenta la investigación de la Universidad Santiago de Cali [26]. Como resultado se identificaron diferentes factores que inciden en la deserción de los aprendices del Centro de Gestión Tecnológica de Servicios del SENA. En ese estudio se agruparon las variables más significativas en factores individuales y familiares, factores económicos, factores académicos y vocacionales y factores institucionales.

Como factor adicional a los anteriores, se agregaron los factores tecnológicos, teniendo en cuenta el alto impacto que este elemento ha tenido en el buen desarrollo y avance de los procesos académicos en las diferentes entidades educativas del país debido a la actual pandemia de COVID-19. Se identificaron las variables y se realizó el instrumento de recolección de información, considerando el análisis.

2) Contextualización de los datos: SENA Colombia tiene una amplia presencia, con 117 centros de formación en 33 regiones. Para el año 2020, según lo reportado por el SNIES [27], la entidad matriculó un total de 2.743 estudiantes, logrando la graduación de 137 personas en áreas críticas como "Desarrollo y Análisis de Software y Aplicaciones", "Gestión y Administración" y "Electrónica y Automatización". El conjunto de datos utilizados y recolectados para este proyecto es de 288 estudiantes a través de la herramienta Google Forms siguiendo la Ley de Protección de Datos Personales o Ley 1581 de 2012 de Colombia [36]. Es crucial resaltar que todo el proyecto se realizó en español, reflejando el idioma en el que se presentan las variables analizadas: edad, grupo étnico, discapacidad, sexo, trabajo, fuente de ingresos, ingresos totales, situación económica, estrato, número de personas en el hogar, área de residencia, selección del programa, nivel educativo del padre, nivel educativo de la madre, cadena de formación, conexión a Internet, televisión por cable, plataformas de vídeo, ordenador, libros físicos, tiempo de lectura, tiempo de navegación por Internet y probabilidad de abandono, no se tuvieron en cuenta otras variables como las psicológicas, depresión, ansiedad, estado de salud, entre otras.

3) Resultados de la aplicación del instrumento: Se realizó un análisis exploratorio, seleccionando preguntas al azar de cada categoría dentro del instrumento de recogida de información. La investigación se centró en las variables derivadas de estas preguntas y su influencia en la variable que predice la probabilidad de abandono: el análisis pretendía descubrir patrones y relaciones que pudieran ofrecer valiosas perspectivas sobre los factores que influyen en las tasas de abandono.

--95,1% de los aprendices no pertenecen a un grupo étnico minoritario. Esta variable se encuentra entre las principales causas de abandono en la enseñanza superior, ya que los contextos socioculturales afectan al rendimiento académico [14].

--Del total de estudiantes, el 85,6% no tiene trabajo, mientras que el 14,4% sí lo tiene. Esta encuesta recoge información sobre los ingresos mensuales de los estudiantes, la fuente de ingresos, la situación económica y el entorno familiar porque los factores económicos influyen significativamente en la probabilidad de que los estudiantes abandonen los estudios.

--La variable referida a la zona de residencia del estudiante es la segunda causa de deserción identificada en los centros de formación del SENA y, en general, en las demás instituciones de educación superior [16]. Para la encuesta, el 86,5% de los estudiantes vive en zona urbana y el 13,5% en zona rural.

--La continuidad de la formación es una de las variables consideradas para los procesos de deserción ya que la tasa de personas que desertan es menor cuando existe interés en continuar su proceso de formación a nivel de pregrado y posgrado. En este apartado, el 89,8% de los estudiantes piensa continuar su formación y el 10,2% no tiene interés.

--Las TIC han tenido un impacto significativo, especialmente durante la pandemia, ya que todos los procesos formativos se han trasladado de entornos presenciales a entornos online. Desgraciadamente, los estudiantes que carecen de acceso a un computador o a un teléfono con conexión a Internet (otras variables examinadas en la encuesta) han tenido que abandonar sus estudios por no poder participar en las actividades académicas. De los encuestados, el 81,9% tiene acceso a Internet, mientras que el 18,1% no lo tiene.

4) Análisis del dataset: Se avanzó a la siguiente fase en el análisis inicial del conjunto de datos realizando un análisis exploratorio de datos (AED) en el conjunto de datos seleccionado. En esta fase, se examinó meticulosamente diversos datos del archivo Excel (.xlsx) utilizando la herramienta Google Colab y programación Python. Este análisis exhaustivo permitió identificar patrones y discernir distribuciones estadísticas, proporcionando una base para futuros análisis y perspectivas.

--Identificación de la cantidad de datos: El conjunto de datos consta de 288 entradas distribuidas en 26 columnas, que abarcan tipos de datos como int64 y object. Un aspecto ventajoso de este conjunto de datos es la ausencia de valores nulos. Esta exhaustividad aumenta la fiabilidad del conjunto de datos para los análisis posteriores de la Figura 2.

```
#Tipo de datos
desercion_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 288 entries, 0 to 287
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Nombres y Apellidos             288 non-null    object
 1   Tipo de documento del aprendiz  288 non-null    object
 2   Número de documento del aprendiz 288 non-null   int64
 3   Edad                            288 non-null    int64
 4   Grupo_Etnico                    288 non-null    object
 5   Discapacidad                    288 non-null    object
 6   Genero                          288 non-null    object
 7   Trabajo                         288 non-null    object
 8   Fuente_ingresos                 288 non-null    object
 9   Total_ingresos                  288 non-null    object
 10  Situacion_economica             288 non-null    object
 11  Estrato                         288 non-null    int64
 12  Cantidad_personas_hogar         288 non-null    int64
 13  Area_Residencia                 288 non-null    object
 14  Selección_programa              288 non-null    object
```

Figura 2. Tipos de datos.

--Descripción de las Variables Numéricas: Se generaron resúmenes estadísticos para las variables numéricas del conjunto de datos, que proporcionan una comprensión detallada de sus distribuciones y valores estadísticos fundamentales en la Figura 3.

```
# Description Number
desercion_data.describe()
```

	Número de documento del aprendiz	Edad	Estrato	Cantidad_personas_hogar
count	2.880000e+02	288.000000	288.000000	288.000000
mean	1.054657e+09	20.732639	1.694444	4.170139
std	6.097412e+08	5.419803	0.711065	1.608626
min	7.719157e+06	16.000000	1.000000	1.000000
25%	1.005832e+09	18.000000	1.000000	3.000000
50%	1.075224e+09	20.000000	2.000000	4.000000
75%	1.080312e+09	21.000000	2.000000	5.000000
max	1.077530e+10	50.000000	4.000000	11.000000

Figura 3. Datos estadísticos de las variables numéricas del dataset.

-- Descripción de las Variables Categóricas: Se realizó un análisis de las variables categóricas, examinando las cantidades de datos de cada variable, las opciones potenciales que representan y las frecuencias dentro del conjunto de datos. Se tomó la decisión de eliminar esta columna, racionalizando el conjunto de datos para mejorar la eficiencia del modelo debido a la insignificancia de los nombres y apellidos de los alumnos para el entrenamiento del modelo.

5) *Preparación de los Datos:* Para preparar los datos, un primer paso consistió en realizar un análisis EDA, que incluía un examen de la correlación y la varianza entre las columnas del conjunto de datos. Este paso es crucial para mitigar cualquier sesgo, como una correlación del 100%, en las variables utilizadas para el entrenamiento a la hora de predecir la variable de resultado.

Al evaluar la relación entre variables, un valor más cercano a -1 o 1 indica una tendencia más fuerte. Para predecir la variable "Probabilidad_deserción", se examinó las tendencias de las otras variables del conjunto de datos. Por ejemplo, la variable "Ordenador" mostraba una sólida relación positiva con la variable que se iba a predecir, mientras que "Ingresos_totales" mostraba una relación inversa negativa. Por el contrario, la variable "Cadena_formativa" mostraba una correlación débil, cercana a 0, lo que indicaba la necesidad de establecer una correlación más fuerte con la variable a predecir en la Figura 4.

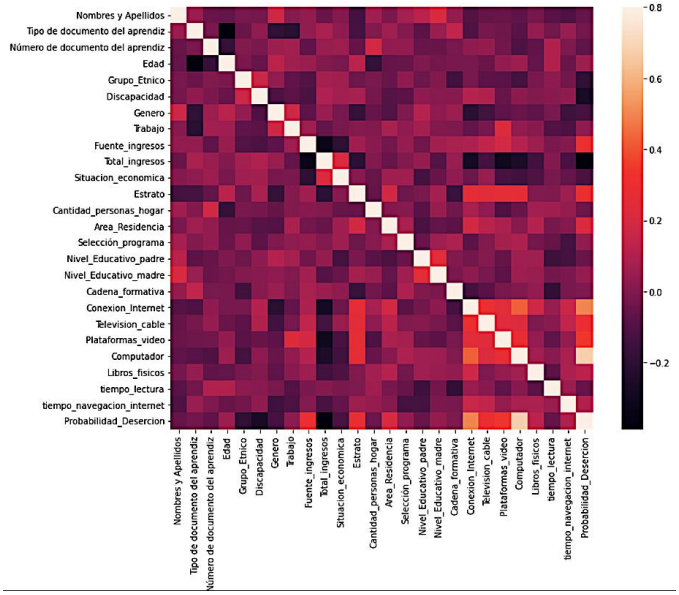

Figura 4. Matriz de correlación entre las variables del dataset.

Además, se ha observado que las cantidades de datos correspondientes a la etiqueta "Probabilidad_Deserción" están equilibradas. Una distribución equilibrada es esencial para evitar el sesgo hacia un valor específico debido a la sobrecarga de datos. En consecuencia, el recuento total de los datos de valor alto es de 158, mientras que el de los datos de valor bajo asciende a 130. Este equilibrio contribuye a la solidez del conjunto de datos para el posterior entrenamiento del modelo.

En particular, la variable "Ordenador" muestra una relación directa positiva con una marcada tendencia contraria a la variable predicha, mientras que "Ingresos_totales" muestra una relación inversa negativa. Por el contrario, la variable "Cadena_formativa" muestra una tendencia frágil, cercana a 0, lo que sugiere una correlación menos sólida con la variable a predecir. Este análisis matizado permite comprender mejor la intensidad de las relaciones entre las distintas características y la variable pronosticada en el gráfico 5.

	Probabilidad_Desercion
Probabilidad_Desercion	1.000000
Computador	0.676869
Conexión_Internet	0.499351
Plataformas_video	0.344952
Fuente_ingresos	0.306035
Television_cable	0.298406
Estrato	0.297625
Area_Residencia	0.211992
Libros_fisicos	0.126341
tiempo_navegacion_internet	0.102464
Edad	0.054479
Nivel_Educativo_madre	0.039872
Selección_programa	0.038524
Trabajo	0.034674
Cadena_formativa	-0.006072
Tipo de documento del aprendiz	-0.026372
Cantidad_personas_hogar	-0.038593
Nivel_Educativo_padre	-0.048406
tiempo_lectura	-0.056430
Número de documento del aprendiz	-0.066115
Situacion_economica	-0.110789
Género	-0.124250
Grupo_Etnico	-0.184308
Discapacidad	-0.258417
Total_ingresos	-0.387009

Figura 5. Correlación entre las variables del dataset variables con "Probabilidad_Desercion".

El análisis reveló una distribución equilibrada en las cantidades de datos correspondientes a la etiqueta "Probabilidad_Deserción". Garantizar un conjunto de datos

equilibrado evita el sesgo hacia un valor concreto debido a la sobrecarga de datos. En consecuencia, el recuento total de datos de valor alto es de 158, mientras que el de datos de valor bajo asciende a 130 en la Figura 6. Este equilibrio mejora la fiabilidad del conjunto de datos para el posterior entrenamiento del modelo, garantizando una representación justa de los casos de alta y baja probabilidad.

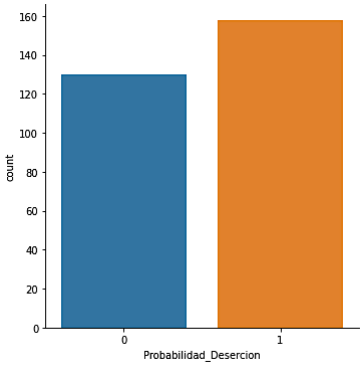

Figura 6. Histograma que respresenta la distribuación de las variables "Probabilidad_Desercion"

En nuestra fase de preprocesamiento de datos, implementamos un LabelEncoder que transforma los datos categóricos del conjunto de datos en valores numéricos. Esta conversión mejora la interpretabilidad de los datos cuando se utilizan en un modelo de predicción. A continuación, los datos se dividen en conjuntos de entrenamiento y de prueba. Por último, se lleva a cabo un proceso de selección de características para identificar y retener las características más influyentes para su aplicación en el modelo de clasificación. Este enfoque integral garantiza que el conjunto de datos se prepara y optimiza adecuadamente para las fases posteriores de desarrollo y evaluación del modelo.

El proceso de selección de características se ejecutó utilizando sklearn [28,29] para identificar las características que contribuyen más eficazmente a la variable de predicción. El resultado de este proceso muestra las características que obtuvieron puntuaciones más altas mediante la función mutual_info_regression. Esta función estima la información entre dos variables aleatorias, despreciando los valores negativos y midiendo su dependencia. Por ejemplo, un valor de dependencia de cero indica independencia, mientras que los valores más altos sugieren una mayor dependencia [30].

Se realizó una comparación de las variables del conjunto de datos con la variable "Probabilidad_Deserción", considerando aquellas con una dependencia superior a 0,04 (véase la Tabla 1). Descuidar este proceso crucial podría comprometer la precisión de muchos modelos, especialmente en algoritmos lineales como la regresión lineal y logística. Las ventajas de este proceso de selección de características abarcan la mitigación del sobreajuste, la mejora de la precisión y la reducción del tiempo de entrenamiento del algoritmo en la Figura 7.

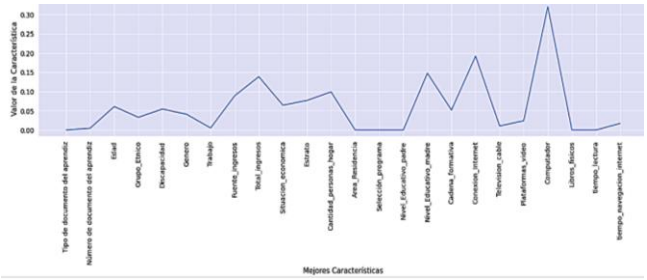

Figura 7. Gráfico de las mejores características de la variable "Probabilidad_Desercion" para el entrenamiento del modelo

El conjunto de datos de la Tabla 1 presentado se sometió a un proceso de selección de características, empleando sklearn [28, 29] para identificar las variables que contribuyen significativamente a la predicción de la "Probabilidad_Deserción". Utilizando la función mutual_info_regression, que mide la información y la dependencia entre variables, se consideraron las características con puntuaciones superiores a 0,04 para su inclusión. La tabla resultante presenta una lista de variables del conjunto de datos que influyen sustancialmente en la predicción de la probabilidad de abandono. Este proceso de selección no sólo refina el conjunto de datos para el desarrollo del modelo, sino que también tiene el potencial de mejorar la precisión del modelo, mitigar el exceso de ajuste y reducir el tiempo de entrenamiento del algoritmo, lo que es particularmente crucial para los algoritmos lineales como la regresión lineal y logística.

Tabla I.
Dataset variables concerning the variable "Probabilidad_Desercion"

Nombre Variable	Relación con la variable 'Probability_desertion.'
Edad (age)	0.07070558555477158
Discapacidad (disability)	0.06229595224645257
Fuente_ingreso (source of income)	0.11081585774576297
Total_ingresos (total income)	0.16694237796367295
Genero (gender)	0.09648723369634071
Conexion_Internet (Internet connection)	0.16856403341616444
tiempo_navegacion_internet (Internet browsing time)	0.10168724007805485
Plataformas_video (Video platforms)	0.10292620632059357
Computador (computer)	0.3767926110486908
Nivel_Educativo_madre (Mother educational level)	0.054086599691605564

6) Etapa de modelado: En esta fase, se exploraron diversas técnicas de aprendizaje automático para abordar el problema modelado, teniendo en cuenta el conjunto de datos específico del que se disponía:

--Kmeans: se empleó por su capacidad para agrupar variables basándose en similitudes con la variable a predecir. Sin embargo, los resultados no arrojaron una precisión satisfactoria, como se indica en la Tabla 2.

--Isolation forest: este método demostró un rendimiento subóptimo con el conjunto de datos considerado, como se indica en la Tabla 2.

--Lineal regressions: también se probó una técnica ampliamente utilizada para escenarios de predicción. Sin embargo, no

alcanzó una precisión aceptable para el conjunto de datos, como se refleja en la Tabla 2.

--Decision tree: Obtuvo los resultados más prometedores, como se muestra en la Tabla 2. En consecuencia, la sección siguiente se centrará en la evaluación del modelo a partir de los resultados obtenidos con esta técnica.

Tabla II.
Precisión de las técnicas usadas de the Machine Learning.

Técnica Aplicada	Precisión
K-means	0.44292682926829274
Insolation Forest	0.30341463414634146
Decision Tree	0.9051724137931034
Regression	0.35065757367560246

7) Modelo de evaluación del árbol de decisión: se implementó utilizando la biblioteca sci-kit-learn de Python [30], aplicando la clase DecisionTreeClassifier. La precisión que se obtuvo fue de 0,91 Tabla 4, lo que indica que el modelo puede predecir correctamente el 91% de las observaciones del conjunto de prueba. Las variables más influyentes que contribuyen a la variable predicha se consideraron predictores para mejorar el rendimiento del modelo.

La matriz de confusión de la Tabla 3 es una herramienta fundamental para evaluar el rendimiento del algoritmo de clasificación cuantificando aciertos y errores. En este modelo, los valores correspondientes a Positivos (predicciones positivas correctas) y Negativos (predicciones negativas correctas) en la matriz de la Tabla 3 reflejan las estimaciones precisas del modelo. Mientras que otros valores representan instancias en las que el modelo cometió errores, es crucial señalar que las predicciones correctas superan a las incorrectas, lo que da como resultado una evaluación positiva del modelo aplicado al conjunto de datos.

Tabla III.
Valores y distribución del modelo de árbol de decisión con matriz de confusión para el conjunto de datos trabajado.

		Predicción	
		positivos	**negativos**
Observación	**positivos**	True Positives (VP) 45	False Negatives (FN) 9
	negativos	false positives (FP) 3	True Negatives (VN) 59

Tras el entrenamiento del modelo de clasificación del árbol de decisión, calculamos las mediciones de precisión e índice de error, que dieron como resultado 0,91 y 0,09, respectivamente. Los resultados indican un alto nivel de precisión en las predicciones del modelo. De ahí que el algoritmo de clasificación de árbol de decisión surja como la solución más recomendable para este problema. Este algoritmo facilita la clasificación de los usuarios en categorías específicas basadas en características distintivas, representando visualmente las distintas vías de decisión.

El árbol de decisión está estructurado con ramas y nodos, como se muestra en la Figura 8. Los nodos internos representan cada característica considerada durante la toma de decisiones, mientras que las ramas representan decisiones basadas en condiciones específicas. Los nodos finales encapsulan los resultados del proceso de decisión [30]. Esta representación gráfica mejora la interpretabilidad y proporciona información sobre los factores críticos que influyen en las predicciones del modelo.

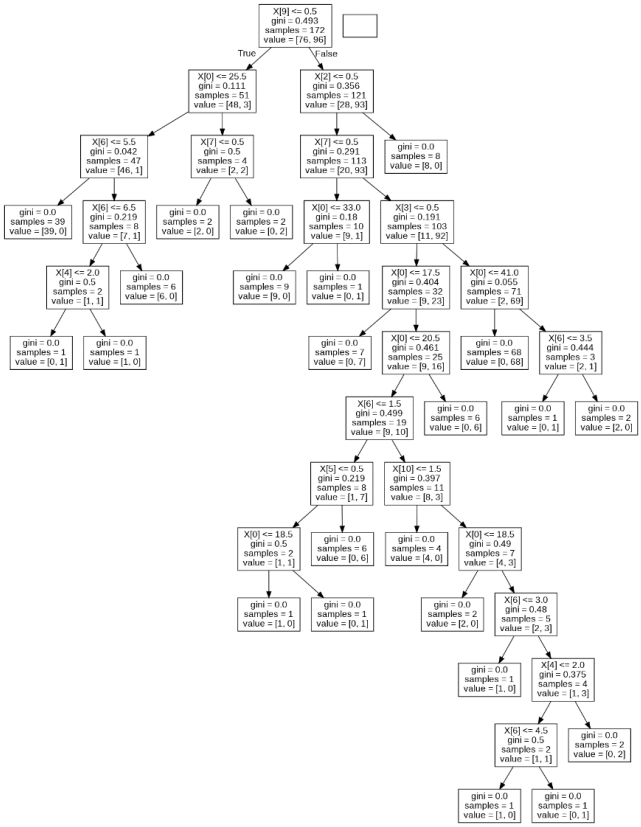

Figura 8. Árbol de decisión obtenido del dataset de entrenamiento. Profundidad del arbol: 12. Nodos terminal nodes: 23.

IV. CONTRIBUCIONES

A partir de los resultados y análisis de las fases anteriores, se ha desarrollado con éxito una plataforma de software predictivo. El software consta de un backend (del lado del servidor) creado con Django, que permite una integración web perfecta con Python, y un componente frontend (del lado del usuario final) construido con HTML, CSS y JS. Para la gestión de la base de datos se utilizó MongoDB, un sistema de bases de datos NoSQL. La estructura basada en documentos de MongoDB resultó ventajosa, simplificando la integración de datos y eliminando las complejidades asociadas a las bases de datos relacionales tradicionales.

Esta plataforma de software ofrece un entorno de desarrollo escalable y robusto, ampliamente reconocido por su eficacia en el desarrollo de aplicaciones web. La construcción de la aplicación se facilitó a través de la consola de comandos, y la herramienta Sublime Text desempeñó un papel fundamental en la gestión del diseño y los códigos de programación. Esta cadena de herramientas garantiza un proceso de desarrollo eficiente y ágil, que da como resultado una herramienta de predicción eficaz y fácil de usar.

A. Tecnológias

El proyecto utilizó el lenguaje de programación Python, las librerías Numpy, Pandas, Matploplib, Pymongo y Scikit-learn

para la gestión de datos y el uso de algoritmos de aprendizaje automático, una base de datos MongoDB para almacenar los datos y el framework Django, que utiliza el patrón Modelo Vista Controlador (MVC) [33] para el desarrollo de la interfaz de obtención de datos del usuario.

B. Descripción functional de interfaces

Para diseñar y desarrollar las interfaces de la aplicación se utilizaron patrones de diseño, que son soluciones que evitan que se cometan errores comunes o minimizan sus impactos, además de permitir la implementación de un desarrollo ágil (menos costes, menos tiempo). Estos patrones responden a preguntas como: ¿Cómo puede navegar el usuario por la aplicación? ¿Qué espera encontrar el usuario en una página web? ¿Cómo mostrar la información? ¿Incluir un motor de búsqueda?

Como resultado de los procesos llevados a cabo anteriormente, una de las principales interfaces del Software es la interfaz de registro de usuario, donde el alumno introduce la información sociodemográfica que se analizará posteriormente. Esta información se divide en varias categorías, como la información sobre el programa de entrenamiento y sus datos de identificación, donde se encuentran características como la edad, el género y el nivel socioeconómico, que son variables que se consideran para la ejecución del algoritmo de aprendizaje automático seleccionado.

Los pasos en el proceso de esta interfaz se encuentran en la Figura 9:

1) Se muestra el formulario (información sociodemográfica) que el usuario solicita. Este formulario contiene algunos datos en blanco y otros con valores por defecto, como es el caso de los campos Regional, Centro de Formación y Programa de Formación, entre otros.

2) Se reciben los datos de la solicitud de envío (petición) y se vinculan al formulario. La vinculación significa que todos los datos introducidos por el usuario y los errores están disponibles cuando se vuelve a mostrar el formulario.

3) Se realiza la limpieza y validación de los datos. En este proceso se comprueba que los datos son adecuados para los campos del formulario y no tienen caracteres inválidos que puedan poner en riesgo la seguridad del servidor.

4) Si algún dato presenta algún error de los mencionados en el punto anterior, el formulario devuelve los respectivos mensajes de error por cada campo que presente un conflicto.

5) Si todos los datos son correctos, se almacenan en la base de datos, y se aplica el algoritmo de Machine Learning, devolviendo la probabilidad de abandono con valores altos (0) y bajos (1).

6) Una vez completadas todas las acciones, el usuario es redirigido a la página final, donde se muestran los resultados de la aplicación del algoritmo en el apartado anterior.

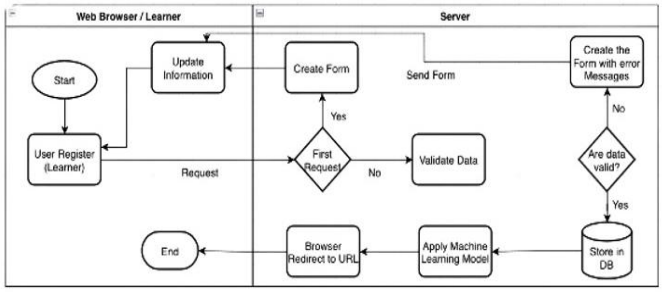

Figura 9. Diagrama de flujo de los datos de registro de los estudiantes.

La interfaz de ingreso de usuarios de bienestar permite al usuario ingresar a la interfaz donde puede consultar la información de caracterización realizada a los estudiantes que previamente diligenciaron la encuesta sociodemográfica. En esta interfaz se habilita el perfil de bienestar del aprendiz, que permite al usuario consultar la información de los estudiantes que han sido caracterizados y si tienen alta o baja probabilidad de deserción.

El comportamiento del sistema se basa en el registro del usuario (aprendiz) a través del ingreso de información sociodemográfica. Luego, los datos son registrados, evaluados y transformados en una predicción dentro del servidor para que el usuario (bienestar del alumno) pueda consultar la información del alumno y asignarla a una posible ayuda disponible.

C. Integración del modelo de ML con el Software

En el proceso de integración del modelo de árbol de decisión seleccionado con el Software, se utilizó la librería joblib de Python. Esta librería permite exportar el modelo en formato .plk para su posterior evaluación y uso dentro del Software desarrollado. A continuación, se importó en la capa views.py de Django llamada a través de las páginas .html que contienen si información del alumno.

Para la conexión con la base de datos MongoDB se utilizó la librería pymongo de Python. De esta librería se importó la clase MongoClient, que funciona como intermediario para entrar en la consola Client y poder manipular todos los datos.

V. EVALUACIÓN Y RESULTADOS

En cuanto a la evaluación de los modelos de aprendizaje automático propuestos, medimos la métrica Accuracy Tabla 2 para identificar el de mayor porcentaje de acierto y el que se implementaría en el Software. En esta propuesta inicial, tras seleccionar las características para la predicción, se toman los diferentes valores de la base de datos inicial para mejorarla Figura 10. Precisión. Finalmente, se verificó la existencia de una tasa de abandono alta (0) o baja (1) según la base de datos inicial.

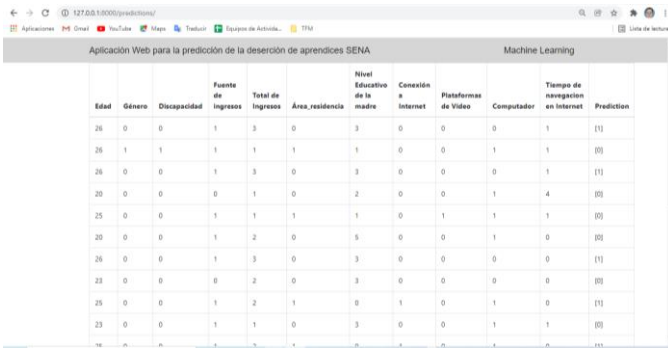

Figura 10. Resultados de la predicción a partir de los datos introducidos en las características principales del algoritmo de aprendizaje automático.

Por otro parte, las tareas solicitadas a los participantes en la evaluación en las que la calidad de uso del producto de software se refiere a la visión de calidad del usuario final [34]. Por lo tanto, para este análisis, se consideran diferentes aspectos, como:

--Eficacia se refiere a la capacidad del producto software para permitir a los usuarios alcanzar el objetivo de su desarrollo en un contexto de uso determinado. De nuevo, se consideran variables como la flexibilidad en los datos de entrada, el multilenguaje y el número de variables.

-La productividad se refiere a la capacidad del software para permitir a los usuarios utilizar los recursos adecuados teniendo en cuenta la eficiencia. Este aspecto tiene en cuenta el tiempo necesario para completar la tarea, el esfuerzo del usuario y el coste económico.

--La satisfacción se refiere a la respuesta del usuario a la interacción con el software, teniendo en cuenta la facilidad de uso y la aplicabilidad.

--La seguridad se refiere a la capacidad del software para alcanzar niveles aceptables de riesgo en un contexto de uso específico. Se tienen en cuenta características como las licencias y los contratos de uso.

Diversas técnicas e instrumentos permiten llevar a cabo este proceso de evaluación. Por ejemplo, a la hora de probar la usabilidad de este Software, una de las técnicas de evaluación utilizadas fue el llamado Paseo Cognitivo, que no requiere mucha inversión para su ejecución y nos permite evaluar aspectos relevantes de la navegación. Además, este método nos permitirá medir la usabilidad del Software desde las primeras etapas de desarrollo hasta su implementación dentro de un público objetivo, teniendo en cuenta características como la Facilidad de aprendizaje, la eficiencia de uso, la retención en el tiempo, las tasas de error y la Satisfacción [35].

Para la evaluación se define un conjunto de tareas representativas a ejecutar por el participante donde intervienen el mayor número de elementos de la interfaz de usuario y que permiten identificar los factores críticos que pueden presentar inconvenientes. Las tareas que se solicitaron a los participantes de la evaluación fueron:

Una vez finalizada la sesión de caminata cognitiva, se realizaron comentarios grupales que permitieron concluir que teniendo en cuenta el perfil de los usuarios, las interfaces son usables, cumpliendo con las condiciones especificadas y con capacidad de comprensión, aprendizaje, operatividad y atracción [34].

Otra técnica utilizada fue la Evaluación Heurística, un método de inspección llevado a cabo por un grupo de expertos en usabilidad que, tras revisar la interfaz en varias ocasiones, midieron el grado de cumplimiento de la heurística [36]. Para este proceso se contó con el apoyo de 5 instructores de la red en análisis y desarrollo de sistemas de información que desempeñaron el papel de evaluadores de software.

VI. Conclusiones, Discusiones, y trabajo futuro

En conclusión, este proyecto abordó con éxito, el reto imperativo de predecir el abandono de los aprendices incorporando variables socioeconómicas cruciales y empleando técnicas avanzadas de aprendizaje automático. La técnica de clasificación de árbol de decisión resultó ser la más eficaz, al alcanzar una precisión del 91% y una tasa de error mínima del 9%, lo que corrobora su posición de vanguardia. Esta técnica, especialmente adecuada para usuarios con características específicas, ofrece un marco sólido para una clasificación precisa dentro de categorías específicas.

El proyecto puso de relieve la importancia de los factores socioeconómicos en el abandono de los aprendices, revelando un notable impacto en las tasas de abandono, especialmente en las Instituciones de Educación Superior (IES). La pronta identificación de estas características a través de algoritmos de clasificación de Machine Learning tiene el potencial de reducir sustancialmente las altas tasas de deserción en programas de formación de nivel técnico y profesional facilitados por la tecnología SENA.

Nuestro trabajo destacó la correlación entre la disponibilidad de equipos informáticos y la deserción de aprendices, subrayando el profundo impacto del acceso a la información en los procesos educativos. Los estudiantes de zonas remotas, particularmente en la Colombia profunda, enfrentan menores tasas de escolaridad debido a la desigualdad en el acceso a la tecnología y la información.

El algoritmo de Árbol de Decisión, identificado como la solución más efectiva, está ahora implementado en la aplicación, contribuyendo significativamente al cumplimiento del objetivo general y permitiendo la aplicación de estrategias de retención temprana por parte de la unidad de Bienestar del Aprendiz.

La aplicación web desarrollada, basada en aprendizaje automático, no solo ayuda a predecir la deserción de los estudiantes del SENA, sino que también contribuye a la reducción de la tasa, la optimización de los tiempos de atención de los usuarios y la agilización de la gestión de ayudas por parte de las unidades pertinentes. Adicionalmente, propone soluciones basadas en IA que mejoran el bienestar general de los estudiantes del SENA.

Todos los objetivos específicos fueron alcanzados con una tasa de éxito del 100%. El análisis integral de variables socioeconómicas y académicas frente a la deserción escolar, la caracterización de la población objetivo y la aplicación de diversas técnicas de aprendizaje automático demostraron la efectividad del enfoque.

A medida que avance el proyecto, los trabajos futuros

profundizarán en un análisis más matizado, que incluya aspectos personales, psicológicos, económicos, laborales, de ubicación y parentales. Se incorporarán nuevas variables académicas para mejorar la precisión predictiva y se tendrán en cuenta factores institucionales como el origen, la situación pública o privada y el apoyo financiero para mejorar la identificación de estrategias de retención.

El sistema no sólo identifica las posibles causas de la deserción estudiantil a partir de los datos analizados, sino que también propone soluciones para la retención, aprovechando algoritmos de procesamiento del lenguaje natural. Se integrarán mecanismos de aprendizaje continuo, lo que garantizará que el sistema evolucione y mejore su éxito en la toma de decisiones.

En el futuro, se probarán diferentes algoritmos para obtener métricas superiores, se incorporarán tecnologías como Big Data y wearables para la identificación de emociones y se presentará la información mediante gráficos estadísticos para mejorar la experiencia del usuario. Además, se desarrollará una aplicación móvil para mejorar la portabilidad y accesibilidad del sistema.

En esencia, este proyecto representa un avance significativo en la modelización predictiva del abandono de los aprendices, demostrando el potencial de las técnicas avanzadas de aprendizaje automático para abordar retos críticos en los sistemas educativos.

AGRADECIMIENTOS

Agradecimientos al Servicio Nacional de Aprendizaje (SENA) Colombia por apoyar y contribuir al desarrollo de este proceso piloto de análisis de información para la permanencia de los estudiantes en los procesos de aprendizaje de la institución.

REFERENCIAS

[1] C. Vasco, C. Cardona, D. Caro, M.F. Molano, M. Pinzón, S. Gómez. C. Estrategias para la permanencia en educación superior: experiencias significativas. Ministerio de Educación Nacional 2015.

[2] C.M. Lopera. Consejo Nacional de Educación Superior (Colombia). (n.d.). Acuerdo por lo Superior 2034: propuesta de política pública para la excelencia de la educación superior en Colombia en el escenario de la paz.

[3] SPADIES - Sistemas información. (2021). SPADIES - Sistemas información. Available online: https://www.mineducacion.gov.co/sistemasinfo/spadies/ (accessed on 9-12-2021).

[4] G.J. Paramo, C.A. Correa Maya. Deserción estudiantil universitaria. Conceptualización. Revista Universidad EAFIT 2012, 35(114), 65–78.

[5] E. Himmel. Modelo de análisis de la deserción estudiantil en la educación superior. Calidad En La Educación 2002, 17, 91. https://doi.org/10.31619/caledu.n17.409

[6] D.M. Moreno, A.M. González. Deserción escolar. Revista internacional de Psicología 2005, 6(1), 1-3.

[7] J.R. Lagunas, M.A.L. Piña. La deserción escolar universitaria. La experiencia de la UAM. Entre el déficit de la oferta educativa superior y las dificultades de la retención escolar. El cotidiano 2007, 22(142), 98.

[8] C.G. Ruiz, D.M.D. Muriel, J.F. Gallego, E.C. Vélez. S.G. Gómez, K.G. Portilla. Deserción Estudiantil en la Educación Superior Colombiana: Metodología de seguimiento, diagnóstico y elementos para su prevención. Ministerio de educación nacional, 2009.

[9] S. Christenson, A.L. Reschly, C. Wylie. Handbook of research on student engagement 2012, 840, New York: Springer.

[10] G. Fonseca, F. García. Permanencia y abandono de estudios en estudiantes universitarios: un análisis desde la teoría organizacional. Revista de la educación superior **2016**, *45*(179), 25-39.

[11] M. Rodríguez Urrego. La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados. Pedagogía y Saberes 2019, 51, 49–66.

[12] R. Timaran. J. Jiménez. Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM. Formacion Universitaria. 2014, 1–19.

[13] A.A. Oñate Bowen. Análisis de la deserción y permanencia académica en la educación superior aplicando minería de datos. Ingeniería de Sistemas, 2016. Available online: https://repositorio.unal.edu.co/bitstream/handle/unal/57387/alvaroagustino%C5%84atebowen.2016.pdf?sequence=1 (accessed on 14-2-2022).

[14] B. Cuji, W. Gavilanes, R. Sanchez. Modelo predictivo de deserción estudiantil basado en arboles de decisión. Espacios 2017, 38(55), 17.

[15] G. González Díaz. Metodología para el diseño generativo de un sólido de revolución. OPENAIRE 2020.

[16] P.E. Ramírez, E.E. Grandón. Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados. Formación universitaria 2018, 11(3), 3-10.

[17] V.X. Guerrero. Aplicación de la técnica de minería de datos para la predicción de la deserción estudiantil universitaria (Bachelor's thesis, Universidad Técnica de Ambato. Facultad de Ciencias Humanas y de la Educación. Carrera de Docencia en Informática), 2020.

[18] E. Romero Sánchez, M. Hernández Pedreño. Análisis de las causas endógenas y exógenas del abandono escolar temprano: una investigación cualitativa. Educación XX1: revista de la Facultad de Educación 2019.

[19] L. Breiman. Statistical Modeling: The Two Cultures. In Statistical Science 2001, 16(3).

[20] P. Elger, E, Shanaghy. AI as a Service. Serverless machine learning with AWS. Available online: www.manning.com (accessed on 26-01-2022).

[21] M.E. Fenner. Machine Learning with Python for Everyone, 2020.

[22] R. Hernández, C. Fernández, M. Baptista. Metodología de la Investigación. 6th Edición, 2014, 1–736.

[23] V. Galán. Aplicación de la Metodología CRISP-DM a un proyecto de Minería de Datos en el entorno Universitario. (Tesis de grado. Escuela Politecnica Superior, Ingeniería en Informática, Universidad Carlos III de Madrid). Madrid. Available online: http://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf?sequence=1 (accessed on 10-2-2022).

[24] Conceptos básicos de ayuda de CRISP-DM. Conceptos básicos de ayuda de CRISP-DM. Available online: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview (accessed on 29-01-2021).

[25] N.L.J. Benavides, G.E.O. Sabogal. Factores que inciden en la deserción de los aprendices del CGTS. SENA Regional Valle 2019, 4(3), 1–21.

[26] SNIES. National Higher Educations Information System, 2021 Available online: https://snies.mineducacion.gov.co/portal/

[27] A. Meneses. Factores asociados a la deserción de estudiantes que ingresaron por condición de excepción indígena a la Universidad. Available online: http://www.scielo.org.co/pdf/soec/n20/n20a03.pdf (accessed on 12-10-2021).

[28] Kaggle. Feature selection using SelectKBest. Available online: https://www.kaggle.com/jepsds/feature-selection-using-selectkbest (accessed on 14-2-2022).

[29] Decision Trees. Scikit-Learn. Available online: https://scikit-learn.org/stable/modules/tree.html (accessed on 10-12-2021).

[30] Scikit-learn: machine learning in Python. Scikit-Learn 1.0.2 Documentation. Available online: https://scikit-learn.org/stable/ (accessed on 10-12-2021).

[31] W. McKinney. Python data analysis library. 2017.

[32] S. Dauzon, A. Bendoraitis, A. Ravindran. Django: Web Development with Python. 2016.

[33] Y.M. Rivero, M.V.G. Sánchez, Y.M. Suárez. Evaluation model for the Software using metric indicators to science and technology surveillance. Revista Cubana de Información en Ciencias de la Salud (ACIMED). 2009, 20(6), 125-140.

[34] W.O. Sánchez. La usabilidad en Ingeniería de Software: definición y características. 2015. Available online: http://www.redicces.org.sv/jspui/bitstream/10972/1937/1/2.%20La%20usabilidad%20en%20Ingenieria%20de%20Software-%20definicion%20y%20caracteristicas.pdf (accessed on 15-3-2022).

[35] M.P. González, J. Lorés, A. Pascual, T. Granollers. Evaluación Heurística de Sitios Web Académicos Latinoamericanos dentro de la Iniciativa UsabAIPO. In Proceedings of the VII INTERACTION 2006. Congreso Internacional Interacción Persona Ordenador, 2006, 1, 145–157.

[36] Ley 581 de 2012, “Ley de protección de datos de Colombia,” 2012.
[37] A. Llauro´ et al., "Identification and comparison of the main variables af fecting early university dropout rates according to knowledge area and i nstitution," Heliyon, 2023.Heliyon, 2023.
[38] "Improvement of Academic Analytics Processes Through the Identification of the Main Variables Affecting Early Dropout of First-Year Students in Technical Degrees. A Case Study," Int. J. Interact. Multimedia. Artif. Intell., 2023.Int. J. Interact. Multimedia. Artif. Intell., 2023.