Investigación Sobre La Deserción Escolar En Colombia Mediante El Uso De Técnicas De Machine Learning

Corporación Universitaria Remington.

Facultad de Ingeniería

Seminario en Machine Learning

Humberto Diosa Guiral

Gustavo López Bedoya

Juan Esteban Ospina López

Hernán Darío Reina Morales

John Fredy Mira Mejía

Seminario - Diplomado Como Opción de Trabajo de Grado

2024

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

1. Resumen

De Acuerdo con los datos del Ministerio de Educación Nacional, alrededor de unos 400.000 niños y jóvenes han desertado de sus instituciones educativas entre finales de 2022 e inicios del 2023, Esta cifra muestra un notable incremento en comparación con el año anterior, cuando el promedio de deserción escolar fue de 330,000 estudiantes entre 2021 y 2022 (Moreno, 2023)

Con lo anteriormente dicho estas cifras son alarmantes ya que tienen una influencia significativa y negativa en el desarrollo social e individual de los niños, niñas y jóvenes que deciden tomar esta decisión y que eventualmente se verán impedidos de acceder a un trabajo que les proporcione una vida digna, mejorar su nivel de vida y contribuir al progreso del país.

La falta de infraestructura en algunas zonas del país se convierten también en un obstáculo bastante grande a la hora de acceder a una buena educación para los niños y jóvenes, la falta de transporte, la distancia entre su hogar y el colegio, la falta de maestros debido al difícil acceso y las malas instalaciones son retos por los cuales tienen que atravesar los jóvenes que residen en zonas rurales, en donde la mejor salida para seguir con su formación como persona es el trabajo de campo o colaborar incluso con grupos armados ilegales, de acuerdo con esto los municipios con mayor índice de deserción en el país son Putumayo con una tasa del 8,11%, después a este se encuentran Arauca, Guainía y Caquetá (Rodríguez, 2023) lugares donde la violencia a raíz del conflicto armado el índice es bastante alto.

De acuerdo con toda la información anterior se busca utilizar las herramientas del machine learning, para predecir la cantidad de niños y jóvenes que podrían terminar su vida educativa en un futuro y sobre esta manera lograr que estas personas cambien su estilo de vida y tengan la

educación que por diferentes motivos ya anteriormente mencionados se vieron en la obligación de abandonar y en muchos casos no volver a retomar por diferentes razones o circunstancias en las cuales se tienen que ver enfrentados día a día en donde algunos lo logran superar y otros se ven derrotados.

2. Palabras clave

Deserción escolar, herramientas, inteligencia artificial, predicción, educación.

3. Objetivos

3.1 Objetivo general

- Analizar la deserción escolar en Colombia por medio de la aplicación de técnicas de machine learning.

3.2 Objetivos específicos

- Realizar una revisión de literatura relacionada con la deserción escolar en Colombia.

- Identificar los factores clave que contribuyen a la deserción escolar según los resultados obtenidos.

- Proponer intervenciones basadas en los hallazgos del análisis.

4. Marco conceptual

4.1 Generalidades

En este apartado, se construye los conceptos que sustentan la investigación. Este marco construye un fundamento para la comprensión y análisis del fenómeno que está bajo estudio, estableciendo vínculos entre ideas y teorías relevantes. Por lo tanto, el marco conceptual sirve para ofrecer al lector las definiciones necesarias para la comprensión del proyecto junto con la revisión bibliográfica para establecer un acuerdo entre los investigadores y sus lectores (Equipo editorial, 2023).

4.2 Definición de conceptos

4.2.1 Deserción Escolar

Este fenómeno se refiere al abandono prematuro de la educación formal, la cual involucra la interrupción de un proceso esencial para el crecimiento individual y colectivo. Según el Ministerio de Educación Colombiano, se puede interpretar a la deserción escolar como la renuncia de estudiantes a la educación, debido a una serie de factores que pueden surgir por problemas sociales, familiares o individuales (Ministerio de Educación Nacional, s.f.).

4.2.2 Machine Learning

Proveniente de la inteligencia artificial es una técnica que suele utilizarse para instruir a máquinas a prepararse por su propia cuenta. Lo anterior se realiza por medio de algoritmos la cual son instrucciones que se usan con el objetivo de que la máquina analice los datos, aprenda de ellos y pueda realizar predicciones. El machine learning puede emplearse con diferentes formatos y tipos de datos, ya sean imágenes, sonidos o textos. Por lo tanto, con el transcurso del tiempo y después de que las máquinas aprenden, estas fortalecen sus decisiones y predicciones de manera autónoma sin la necesidad de que estén programadas. (IMPULSO_06 Formación y futuro, s.f.).

Según (Iberdrola, s.f.) existen tres categorías en este campo del Machine Learning: *El aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo*:

- En el *aprendizaje supervisado*, los algoritmos reciben instrucciones mediante un sistema que va asociado a los datos, lo que les permite tomar decisiones o hacer predicciones. Un ejemplo práctico sería un detector de spam que analiza correos electrónicos y los clasifica como spam o no spam, utilizando patrones extraídos del historial de correo electrónico, como el remitente, la proporción de texto/imagen, las palabras clave del asunto, entre otros aspectos
- En contraste, el *aprendizaje no supervisado* implica algoritmos que carecen de previo conocimiento y se encaran a datos desorganizados, buscando identificar patrones que puedan organizar de alguna manera los datos. Por ejemplo, en el marketing, estos algoritmos se emplean para extraer patrones a partir de grandes cantidades de datos en redes sociales, facilitando la creación de campañas publicitarias altamente segmentadas.

- El *aprendizaje por refuerzo* tiene como función que el algoritmo pueda aprender de sus propias experiencias. Esto implica la capacidad de tomar decisiones óptimas en diversas circunstancias mediante un proceso de ensayo y error, otorgando recompensas por las decisiones correctas. En la actualidad, se aplica en áreas como reconocimiento facial, diagnóstico médico y clasificación de secuencias de ADN.

Dentro de la disciplina de machine learning, podemos encontrar numerosos modelos y técnicas capaces de resolver una tarea determinada. Por lo tanto, primero se inicia realizando un resumen de los modelos y técnicas de esta disciplina. De acuerdo con el curso realizado en la plataforma Crehana y complementando de acuerdo con (IAT, s.f.) existen tres tipos de modelos: lógicos, geométricos y probabilísticos. Los modelos lógicos, como los árboles de decisión, transforman datos y probabilidades en reglas de actuación. Los modelos geométricos operan en espacios de instancias utilizando líneas o planos para clasificar, y pueden basarse en la noción de distancia. Y los modelos probabilísticos, que emplean estadística bayesiana, clasifican características y variables objetivo como variables aleatorias, manipulando la incertidumbre en el proceso de modelado.

Los algoritmos de Machine Learning más comunes de acuerdo con (Redacción APD, 2019) son:

1. **Algoritmos de regresión:** Acá se estima y comprende las conexión que hay entre variables. Este algoritmo de regresión se apoya de una variable que es dependiente y una sucesión de variables cambiantes, siendo así un algoritmo que es eficaz para la predicción.
2. **Algoritmos bayesianos:** Estos se basan del teorema de Bayes la cual clasifica cada valor como independiente de otro. Este predice una clase en función de un conjunto de características, por medio de la probabilidad.
3. **Algoritmos de agrupación:** Se usa en aprendizaje la cual no es supervisado. Funciona para categorizar datos etiquetados. Su función es por medio de la indagación de conjuntos dentro de los datos y son representados por la variable K el número de grupos.
4. **Algoritmos de árbol de decisión:** Los algoritmos de árbol de decisión son herramientas utilizadas en el ámbito de la inteligencia artificial y el aprendizaje automático. Estas herramientas toman su nombre de su estructura visual, que se asemeja a un árbol con ramificaciones. La idea principal detrás de estos algoritmos es simular el proceso de toma de decisiones humano de una manera lógica y estructurada.

   En un árbol de decisión, cada nodo representa una pregunta o prueba sobre una característica específica. Estas características podrían ser variables como el precio, la edad, la ubicación, entre otras, dependiendo del contexto del problema que se esté abordando. Las ramas que se desprenden de cada nodo representan las posibles respuestas a esa pregunta o prueba, dividiendo el conjunto de datos en subconjuntos más pequeños en función de la condición evaluada.

   A medida que se avanza por el árbol, se llega a nodos terminales llamados hojas, que representan las decisiones finales o las predicciones del modelo. Cada hoja del árbol se asocia con una clase o un valor específico, dependiendo de la naturaleza del problema.

Este enfoque de bifurcación y ramificación permite que los algoritmos de árbol de decisión capturen de manera efectiva patrones complejos y relaciones en los datos, facilitando la toma de decisiones automatizada en diversas aplicaciones, desde la clasificación de datos hasta la predicción de resultados..

5. **Algoritmos de redes neuronales:** Las redes neuronales artificiales tiene unidades dispuestas de series de capas, donde cada una de ellas se conecta a las capas anexas. Estas redes se basan en los sistemas biológicos, como lo es el cerebro y sus procesos de información. De esta manera, sirven para resolver problemas por medio de un gran número de elementos de procesamiento. Este algoritmo aprende también por medio del ejemplo y experiencia la cual lo hace útil para modelar relaciones no lineales en datos de alta dimensión.
6. **Algoritmos de reducción de dimensión:** La noción fundamental detrás de los algoritmos de reducción de dimensión radica en la idea de disminuir la cantidad de variables en un conjunto de datos, con el propósito de identificar de manera más eficiente la información esencial que se busca. Estos algoritmos se revelan como herramientas valiosas en el ámbito del análisis de datos y el aprendizaje automático, ya que permiten simplificar y condensar la información manteniendo, al mismo tiempo, su relevancia y utilidad.
7. **Algoritmos de aprendizaje profundo:** Los algoritmos de aprendizaje profundo constituyen una categoría avanzada en el campo del aprendizaje automático, destacando por su capacidad para procesar datos a través de múltiples capas de redes neuronales. En lugar de depender de un único estrato de análisis, estos algoritmos se caracterizan por la presencia

de múltiples capas o niveles de procesamiento, lo que otorga a la red una capacidad de aprendizaje y extracción de características más sofisticadas.

Este enfoque en capas múltiples facilita la identificación y comprensión de patrones complejos en los datos, permitiendo que la red capture relaciones más profundas y sutiles entre las variables.

Cuando se introduce el aprendizaje automático en un contexto de investigación, una de las estrategias que se adecua especialmente al tema es la metodología basada en el modelo Random Forest. Este enfoque representa una técnica de aprendizaje supervisado que destaca por generar múltiples árboles de decisión a partir de un conjunto de datos de entrenamiento. La singularidad de esta metodología radica en la combinación de los resultados de estos múltiples árboles, con el objetivo de crear un modelo consolidado y más robusto en comparación con los resultados individuales de cada árbol (Lizares, 2017).

El proceso de construcción de cada árbol en un Random Forest se lleva a cabo mediante un proceso de dos etapas. En la primera etapa, se realiza una selección aleatoria de subconjuntos de datos del conjunto de entrenamiento. Esto significa que cada árbol se entrena con una porción diferente del conjunto de datos, lo que promueve la diversidad en la información que cada árbol aprende. En la segunda etapa, se construye un árbol de decisión con cada uno de estos subconjuntos de datos, utilizando un proceso de bifurcación y toma de decisiones.

La combinación de los resultados de estos árboles individuales se logra mediante un proceso de promedio o votación, dependiendo de la naturaleza del problema. Esta agregación de resultados contribuye a reducir el impacto de posibles sobreajustes y mejora la capacidad del modelo para generalizar patrones en datos nuevos y no vistos.

5. Marco contextual

5.1 Generalidades

Se define al marco contextual como la parte escrita del estudio. Incluye información con relación a los escenarios físicos y temporales de una situación particular. Esta sección puede incluir datos e información sobre los antecedentes sociales, históricos, culturales y económicos de los temas que está investigando. Esto es relevante para introducir un escenario que contextualice y defina el tema en estudio (Tesis y Másters, s.f.).

Este trabajo se enfoca en el método exploratorio y de investigación, siendo así una investigación de tipo no experimental, que busca la recolección de estadísticas y datos provenientes de estudios e investigaciones confiables ya realizados anteriormente. Esto, con el fin de analizar las diferentes técnicas de machine learning que se podrían aplicar para predecir cuando los estudiantes podrían abandonar la educación escolar en el contexto colombiano.

5.2 Contexto Educativo en Colombia

Colombia es uno de los países más grandes de América Latina, ocupando el tercer puesto con más población después de Brasil y México, contando con una población joven y diversa. Mayoritariamente los jóvenes residen en zonas urbanas con un porcentaje del 76% y el resto de población en zonas rurales (Revisión de políticas nacionales de educación, 2016). En Colombia de acuerdo con la constitución política, todos los colombianos tienen derecho al estudio o educación. Las personas pueden tener ingreso a las instituciones educativas públicas dependiendo de la región,

las cuales están estructuradas en varias etapas: la educación inicial, la educación preescolar, la educación básica que está formada por cinco grados en primaria y cuatro grados en secundaria, la educación media que consta de dos grados más y la finalización con un título de bachiller, y por último la educación superior en la universidad (Colombia Potencia De La Vida, 2022).

Dentro de los horarios establecidos, la mayoría de las escuelas y colegios manejan un horario de 5 a 6 horas. Sin embargo, el gobierno también ha realizado esfuerzos en implementar una jornada única para toda institución con 7 horas de jornada de conformidad con la Ley General de Educación de 1994. Cuando los estudiantes culminan con la educación básica secundaria, obtienen el certificado de estudios de Bachillerato que sirve como prerrequisito para estudios de pregrado o educación superior en la universidad.

5.3 Deserción Escolar y desafíos en el Sistema Educativo Colombiano

Se han reportado grandes avances en el sistema educativo colombiano en los últimos años. Sin embargo, Colombia cuenta con ciertos desafíos en el sistema educativo. Dentro de dichos desafíos se evidencia que Colombia cuenta con un sector público pequeño, limitando así la capacidad de ofrecer estos servicios (Revisión de políticas nacionales de educación, 2016). La corrupción es otro motivo la cual afecta la educación, y los demás motivos incluyen la pobreza, desigualdad, el conflicto interno y embarazos a temprana edad. Específicamente hablando de la pobreza, una persona de cada tres vive en estado de pobreza en Colombia, la cual nos lleva al 33% que es

superior al promedio de OCDE del 11% (OCDE, 2015a, 2014c) y donde se evidencia en mayor parte es en zonas rurales, en la Guajira, Cauca y Chocó (DNP, 2015) (Revisión de políticas nacionales de educación, 2016).

Las cifras recientes demuestran que Colombia aun enfrenta la deserción escolar debido a los desafíos anteriormente mencionados. Esto indica que aún quedan ciertos objetivos que cumplir con el fin de evitar dicha deserción. De acuerdo con el Ministerio de Educación Nacional, son más de 400 mil niños, niñas y jóvenes que desertaron el colegio en un rango comprendido de finales del 2022 e inicios del 2023. Lo anterior, revela que ha habido un incremento de deserción a comparación del periodo anterior. (Portafolio, 2023) (United Way, 2023).

5.4 Técnicas y modelos de Machine Learning para la predicción de la deserción escolar

Tomando como referencia a investigaciones previas, (Valencia, 2017) nos expone en su trabajo de investigación la aplicación de una técnica de machine learning para predecir posibles estudiantes a abandonar una institución. Como primer paso, se ingresa a la base de datos de la institución con los datos de cada estudiante, con el fin de obtener la información necesaria de los estudiantes. Luego se implementó la técnica de redes neuronales bajo el uso de mapas autoorganizados utilizando la herramienta Matlab y configurando ciertos scripts para poder personalizar la configuración y el diseño de los mapas. Por último, se analizaron los resultados, la cual se concluyó que es posible obtener las predicciones necesarias para saber qué tipos de estudiantes pueden

desertar la educación bajo la técnica de redes neuronales de machine learning. Dentro de sus resultados se obtuvo que los estudiantes más probables de desertar son los que tienen problemas de índole social, económico, familiar y situaciones con experiencias previas.

Por otra parte, se realizó una investigación en la Escuela ECBTI de la UNAD en Colombia en el año 2021, donde se implementó un modelo de predicción con tecnologías de minería de datos. (Pérez, 2021) expone que las instituciones cuentan con la información necesaria de los estudiantes la cual sirven para la aplicación de diferentes técnicas de "Data mining", la cual el árbol de decisión se adecua como una buena opción para la predicción.

El investigador desarrolló la implementación con la metodología CRISP-DM, y se usó el entorno de minería de datos WEKA la cual cuenta ya con ciertos algoritmos. Dentro de los algoritmos escogidos por el investigador, se encuentran "Bayes", "Functions", "Lazy" y "Árboles de decisión (Trees)". Posteriormente, los resultados fueron positivos ya que arrojaron estadísticas similares a los que la institución tenía con relación a las personas que abandonan sus estudios. Además, los porcentajes de predicción fueron altos, dejando en evidencia que es posible implementar estas herramientas de machine learning bajo la técnica de "Data mining".

Con base en los estudios presentados por Villamarín (2017) y Ávila (2021) sobre predicción de la tasa de deserción escolar mediante métodos de machine learning, se puede

concluir las redes neuronales y también los árboles de decisión pueden resolver eficazmente este problema. Estos son métodos eficaces para identificar y predecir el abandono.

6. Desarrollo e implementación del aprendizaje

Este trabajo de grado se lleva a cabo para analizar las causas y ver la cantidad de niños y jóvenes que toman la opción de continuar con su vida académica académicos una vez toman la decisión de desertar del mismo, debido al motivo que sea que lo haya llevado a tomar esta decisión.

6.1 Recopilación y Análisis de datos

6.1.1 Recolección de datos

Se realiza una búsqueda tomando como referencia fuentes de entidades cuyo fin es obtener información eficaz y verídica sobre el número de jóvenes desertores en el país.

El procedimiento de búsqueda de la data se lleva a cabo a través de la información suministrada por el ministerio de educación y entidades asociadas a esta misma al igual que el SIMPADE.

6.1.2 Análisis de datos

Una vez recopilada la información se analiza cada uno de los campos que incluye la data.

6.2 Recopilación de datos

Esta fase inició con la búsqueda de fuentes de información que contarán con un historial de jóvenes que desertaron en el país y sus motivos, búsqueda que condujo hasta el ministerio de educación y el SIMPADE.

El SIMPADE, aborda información del estudiante, su entorno familiar, el contexto educativo y municipal con la finalidad de ser sometido a un análisis por diversos niveles de administración

del sistema educativo con la finalidad de tomar decisiones que mejoren la estadía escolar (Ministerio De Educación Nacional República De Colombia).

De Acuerdo con la información del ministerio de educación nacional se observa en la tabla 1 la cantidad de jóvenes en total que salieron del EPBM (Matrícula estadística educación preescolar, básica y media) durante los periodos 2020, 2021 y 2022.

Tabla 1. Deserción en (Nivel De Educación Preescolar, Básica y Media)

Año	Alumnos Desertores	Alumnos Caracterizados	% Caracterización
2020	181.009	26.121	14,43%
2021	277.792	44.909	16,17%
2022	337.104	76.525	22,70%

Tabla 1. Deserción en Nivel De Educación Preescolar, Básica y Media. Tomado de CGR, Oficio MEN 2023EE132218

Para el 2022 observamos un aumento del porcentual de alumnos caracterizados con un 22.7% y una disminución en la deserción, estando en 53.4% en 2021-2020 y pasar a 21.3% en 2022-2021.

Figura 1. Evolución de embarazos a pronta edad

Figura 1. Evolución de embarazos a pronta edad. Tomado de CGR, datos del SIMPADE (MEN)

Este tema genera retos muy difíciles que intervienen el proceso de aprendizaje de los jóvenes, como se puede ver en el año 2020 se registraron 2.115 embarazos los cuales se representa el 8% de desertores, lo cual incrementó 2.220 embarazos en 2021, disminuyendo a 5% de desertores, en 2022 contaron 2.865 embarazos una subida absoluta que representa el 3.7% de desertores.

Tabla 2. Cantidad de víctimas por discriminación en entorno escolar

Información individual	2020	2021	2022	% en 2022
Discriminación LGBTI	17.321	16.440	15.487	28%
Discriminación razón apariencia fisica	7.274	6.652	14.048	25%
Discriminación razón a su pertenencia étnica	1.646	1.153	3.852	7%
Discriminación razón discapacidad	3.502	2.601	4.880	9%
Discriminación razón religión	15.587	16.236	5.166	9%
Discriminación otras razones	3.060	2.961	12.215	22%
Total	**48.390**	**46.043**	**55.648**	**100%**

Tabla 2. Cantidad de víctimas por discriminación en entorno escolar. Tomado de CGR, datos del SIMPADE oficio MEN 2023EE132218

Entre las víctimas de discriminación, la comunidad LGBTI es una de las que más se destaca, que cuenta con 28% de casos registrados de acuerdo con este tipo de discriminación contados en el año 2022.

Adicionalmente se observa una cantidad significativa de víctimas por su apariencia física 25% en el periodo de 2022 contando con 7.274 en 2020 luego 6.652 en 2021 y por último aumentando a 14.048 en el año 2022.

Tabla 3. Deserción escolar, ámbito individual

Concepto	2020	2021	2022
Bajo rendimiento escolar	2.742	4.594	5.361
Poco gusto por el estudio	4.162	7.423	8.417
Dificultades académicas	1.355	1.642	3.029

Tabla 3. Deserción escolar, ámbito individual. Tomado de CGR, datos del SIMPADE oficio MEN 2023EE132218

Este tipo de deserción de manera individual escolar por factores individuales o personales nos refleja unas tendencias alarmantes pero que al total de deserciones cada vez va disminuyendo, estando en 31.6% en 2020 a 30.4% en 2021 y al 22% en 2022, la deserción en su totalidad es 16.807alumnos en 2022 teniendo en cuenta estos registros que se tomaron y que fueron evidenciados.

Tabla 4. Deserción escolar en el ámbito familiar

Concepto	2020	2021	2022	Peso relativo 2020	Peso relativo 2021	Peso relativo 2022
Cambio de residencia	11.444	20.778	36.953	82,6%	85,0%	73,9%
Desempleo de los padres o acudientes	1.379	1.581	2.380	10,0%	6,5%	4,8%
Desplazamiento forzado	129	80	181	0,9%	0,3%	0,4%
Cambio de país	901	2.012	10.493	6,5%	8,2%	21,0%

Tabla 4. Deserción escolar en el ámbito familiar. Tomado de CGR, datos del SIMPADE oficio MEN 2023EE132218

Cambiar de lugar de residencia es motivo de deserción que reflejó una disminución proporcional durante el año analizado, de todos modos, este es un factor que representa sobre todo para 2022.

Otro factor que se muestra en la tabla es el desempleo de los padres que se ve reflejado con una representatividad del 4.8% de casos familiares en 2022, la falta de empleo llega a generar un problema en la economía de la familia haciendo difícil acceder a servicios y recursos educativos. Factores como el desplazamiento forzado y cambio de país son motivos bastante relevantes, que cuentan con una representación de 0.4% y 21%.

Tabla 5. Deserción por causas de la institución educativa

INSTITUCIONAL	2.020	2.021	2022	Peso relativo 2020	Peso relativo 2021	Peso relativo 2022
ESTABLECIMIENTO EN ZONA LEJANA	149	430	491	49,3%	59,1%	49,6%
CONFLICTOS ENTRE ESTUDIANTES	35	11	234	11,6%	1,5%	23,6%
HACINAMIENTO	31	81	94	10,3%	11,1%	9,5%
PROBLEMAS INFRAESTRUCTURA	44	165	86	14,6%	22,7%	8,7%
JORNADAS NO ADECUADAS AL TIEMPO DEL ESTUDIANTE	43	41	85	14,2%	5,6%	8,6%

Tabla 5. Deserción por causas de la institución educativa. Tomado de CGR, datos del SIMPADE oficio MEN 2023EE132218

Las instituciones educativas ubicadas en zonas lejanas representan más o menos el mismo ritmo durante el periodo, lo cual refleja que los alumnos que asisten a instituciones que se encuentran en áreas muy alejadas se ven sometidos a dificultades para obtener su derecho a la educación. En el 2022 el conflicto entre estudiantes tuvo un importante incremento pasó de 1.5% representando a esto al 23%. Adicionalmente, el hacinamiento se ve representado en 9.5% en 2022, este factor afectará negativamente el aprendizaje debido a una falta de espacio suficiente en las aulas.

Por otro lado, las jornadas no adecuadas al tiempo del estudiante muestran un incremento del 107% en el año 2022 frente al 2021 influidos por las muestras de representación con un aumento relativo de tres puntos porcentuales.

Tabla 6. Deserción escolar por causas de contexto

CONTEXTO	2.020	2.021	2022	Peso relativo 2020	Peso relativo 2021	Peso relativo 2022
Drogadicción	134	150	206	12,0%	11,3%	9,6%
Presencia de grupos armados	207	123	219	18,5%	9,3%	10,2%
Inseguridad	153	149	328	13,7%	11,2%	15,2%
Matoneo escolar bullying	4	6	121	0,4%	0,5%	5,6%
Riesgo de reclutamiento	60	73	85	5,4%	5,5%	3,9%
Distancia hogar al establecimiento	562	826	1.195	50,2%	62,2%	55,5%

Tabla 6. Deserción escolar por causas de contexto. Tomado de CGR, datos del SIMPADE oficio MEN 2023EE132218

La drogadicción es una de las causas que más influencia tiene en la deserción escolar, teniendo en cuenta que las cifras a priori son bajas en la proporción del contexto.

Los grupos armados, teniendo en cuenta que los valores varían de un año al otro, se puede observar un registro tal de incidencia considerable.

El factor de la inseguridad es bastante relevante al momento de tomar la decisión desertar de los estudios y de este modo los datos lo muestran en el año 2020, se observaron 153 casos de alumnos que desertaron y su decisión estuvo relacionada con la inseguridad, y en los años posteriores fue aumentando

Las distancias que puede llegar a ver entre el establecimiento educativo y el hogar del

estudiante es un factor que dificulta que los estudiantes asistan a los colegios, la información nos muestra datos significativos en los periodos en los cuales se llevó a cabo la investigación en proporción de contextos superan el 50%. En 2020, se registraron 562 casos, por otro lado en el año 2021 y 2022 las cifras aumentaron.

7. Metodología

Basados en la investigación de diferentes proyectos de deserción escolar en el país encontramos esta metodología ya aplicada y basada en machine learning de la cual tomaremos como ejemplo y analizaremos su proyección.

Esta metodología se basa en concretar la información recolectada con lo que se forma una base de datos con diferentes estudiantes, como lo es los periodos académicos, información como la ciudad, programa o materia, género y estrato social. Lo anterior, como principal idea de encontrar un modelo de Machine Learning la cual pronostique los desertores posibles dentro de la base de los datos disponibles.

Primero, se realiza un estudio que se basa en explorar los datos, con el fin de evidenciar características antes de realizar cualquier tipo de modelo.

Con el fin de realizar este método se debe aplicar Sweetviz; la cual se trata de una biblioteca que utiliza la visualización automática. El reporte para generar debe incluir: descripción general de los conjuntos de datos, propiedades de la variable asociaciones categóricas, numéricas, más frecuentes, más grandes y pequeños.

La variable de ciudad comparte unas categorías tomando en cuenta las principales ciudades del país como lo es (Bogotá, Medellín, Cali y Eje cafetero)

En programas o materias, se tienen en cuenta las principales materias que presentan las instituciones educativas.

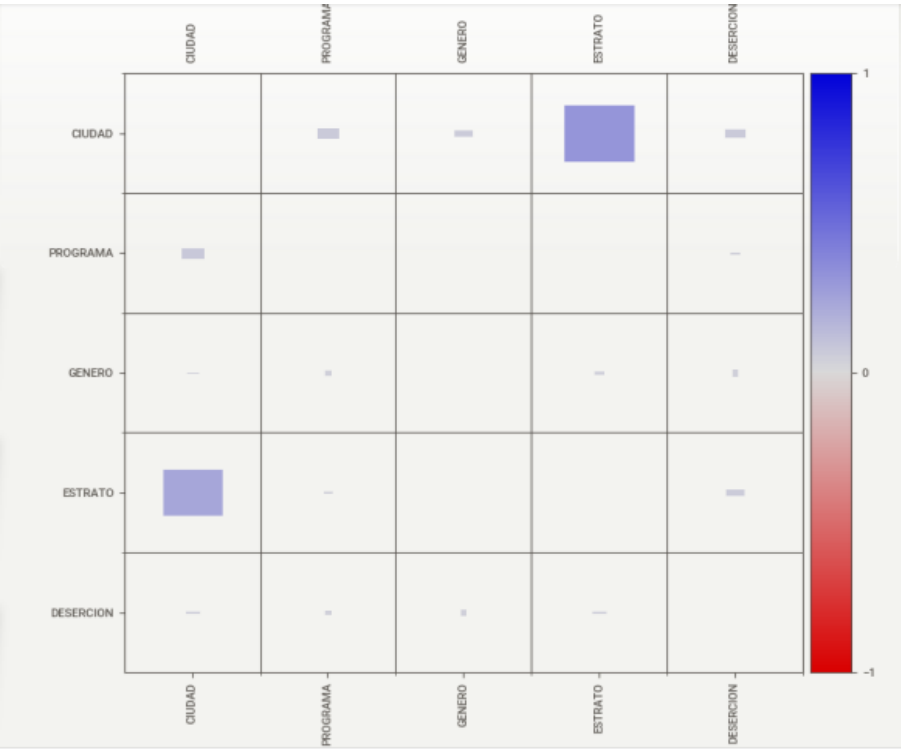
El estrato socioeconómico se distribuye en los tres principales estratos en los que se presenta esta problemática como lo son los estratos 1, 2 y 3.

Los estudiantes se clasifican como Femenino y Masculino.

En la deserción, se considera probables desertores y él no se encuentra en amenaza de deserción.

En la figura 2 se muestra una matriz de correlación de las variables analizadas.

Ilustración. Matriz de correlación de las variables en una base de datos



En la matriz se verá evidenciado un bajo paralelismo en las variables que son estrato socioeconómico y ciudad puesto que las otras no deberían mostrar paralelismo en este análisis, como se muestra en la posible imagen que se podrá obtener a partir de los datos.

Con la investigación de esta metodología podemos tomar como propuesta los pasos mencionados para llevar a cabo un proyecto y desarrollo basado en Machine Learning, ya que son fuentes confiables que se llevaron a cabo y funcionaron en su momento y así aportar al tema tratado de prevención de la deserción escolar.

8. Resultados

La metodología nos indica que utilizando la matriz de confusión y realizando el análisis de los 3 modelos en funcionamiento, encontraremos que los valores de dan exactitud realizó la clasificación, será próximo a 0,6 para esos tres modelos. Los mejores análisis pertenecen al modelo Random Forest.

Como se muestra en la siguiente tabla de un ejemplo de resultados comparativos de los modelos.

Tabla 7. Resultados de los modelos por medio de la matriz de confusión

	Árbol de decisión	Random Forest	Regresión logística
Verdaderos positivos	0.43	0.26	0.33
Verdaderos negativos	0.59	0.79	0.68
Falsos negativos	0.57	0.74	0.67
Falsos positivos	0.41	0.21	0.32
Accuracy (Exactitud)	0.52	0.59	0.54

A los que desertan se les asignó el 1 y a los no desertores el 0. Los verdaderos negativos fueron mayoritariamente clasificados de manera correcta por medio del modelo

9. Conclusiones

La deserción escolar es un problema que afecta a muchos estudiantes en Colombia. A través de este proyecto, se puede lograr evidenciar este problema con el fin de identificar y prevenir dicha deserción en la educación.

La predicción se lleva a cabo gracias a las diferentes herramientas de Machine Learning explicadas, la cual son base principal para la prevención de estudiantes en vulnerabilidad. Esto gracias a que también se cuenta con las estadísticas que reflejan los motivos del abandono estudiantil en Colombia actualmente.

Es de aclarar, que ya se han realizado previas implementaciones e investigaciones de este tema en Colombia, donde se han demostrado resultados positivos en base a las predicciones arrojadas por Machine Learning. Por lo tanto, la tarea es la de continuar implementando estas herramientas para ir disminuyendo la deserción a medida que pasa el tiempo.

Referencias

Bastidas, L.R. (2007). *El inicio del siglo XXI*. Planeta. Sitio web: http://www.rbastidasl.com/libro-inicio-del-sigloxxi.

Borges, J.L. (2013). *Ficciones*. Buenos Aires, Argentina: Debolsillo.

Colombia Potencia De La Vida. (15 de 03 de 2022). *Colombia Potencia De La Vida*. Obtenido de Colombia Potencia De La Vida: https://www.mineducacion.gov.co/portal/Preescolar-basica-y-media/Sistema-de-educacion-basica-y-media/233839:Sistema-educativo-colombiano

Equipo editorial, E. (18 de 02 de 2023). *Concepto*. Obtenido de Concepto: https://concepto.de/marco-conceptual/

IAT. (s.f.). *IAT*. Obtenido de IAT: https://iat.es/tecnologias/inteligencia-artificial/machine-learning/

Iberdrola. (s.f.). *Iberdrola*. Obtenido de Iberdrola: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

IMPULSO_06 Formación y futuro. (s.f.). *IMPULSO_06 Formación y futuro*. Obtenido de IMPULSO_06 Formación y futuro: https://impulso06.com/como-machine-learning-esta-transformando-la-educacion/#:~:text=Los%20algoritmos%20de%20Machine%20Learning,los%20estudiantes%20a%20tener%20%C3%A9xito.

Ministerio de Educación Nacional. (s.f.). *Ministerio de Educación Nacional*. Obtenido de Ministerio de Educación Nacional: https://www.mineducacion.gov.co/1621/article-82745.html

Moreno, Y. (27 de June de 2023). *Aumenta deserción escolar: más de 300.000 niños abandonaron el colegio.* Recuperado el 27 de December de 2023, de Portafolio: https://www.portafolio.co/economia/finanzas/aumenta-desercion-escolar-mas-de-300-000-ninos-abandonaron-el-colegio-585052

Pérez, M. L. (2021). *Modelo de Predición de Deserción Estudiantil, Apoyado en Tecnologías de Data Mining, En un Curso de Primera Matrícula De La Escuela ECBTI De La Unad.* Colombia.

Redacción APD. (04 de 04 de 2019). *APD*. Obtenido de APD: https://www.apd.es/algoritmos-del-machine-learning/

Rodríguez, J. C. (26 de June de 2023). *Más de 300 mil niños y adolescentes abandonaron el sistema educativo en lo corrido de 2023.* Recuperado el 27 de December de 2023, de Infobae: https://www.infobae.com/colombia/2023/06/27/mas-de-300-mil-ninos-y-adolescentes-abandonaron-el-sistema-educativo-en-lo-corrido-de-2023/

Tesis y Másters. (s.f.). *Tesis y Másters*. Obtenido de Tesis y Másters: https://tesisymasters.com.co/marco-contextual/

United Way. (19 de 10 de 2023). *United Way*. Obtenido de United Way: https://unitedwaycolombia.org/2023/10/19/desercion-escolar-en-colombia-un-desafio-que-se-agrava/#:~:text=Entre%202021%20y%202022%2C%20el,y%20social%20de%20los%20j%C3%B3venes

Valencia, J. H. (2017). *Análisis de la deserción estuiantil en la FCECEP utilizando Machine Learning específicamente mapas auto organizados de Kohonen.* Santiago de Cali