TRABAJO DE GRADO
Opción Seminario-Diplomado.

SISTEMA DE PREDICCIÓN DE LA DESERCIÓN ESCOLAR, UTILIZANDO ESTRATEGIAS DE MACHINE LEARNING

Corporación Universitaria Remington
Facultad de Ingeniería
Ingeniería en sistemas

Estudiantes:
Carlos Andres Mendoza Garcia

Tutor: Juan Carlos Briñez de León

Opción de Trabajo de grado Seminario-Diplomado.
2024

Dedicatoria

Principalmente a Dios, por darme la fuerza necesaria para culminar esta meta.
A mis padres, a pesar de las adversidades son mi motor para seguir adelante.
A mis amigos principalmente a Jessica, por brindarme su apoyo moral en esas noches que tocaba investigar.
Y, finalmente, a los que no creyeron en mí, con su actitud lograron que tomará más impulso.

Contenido

Resumen

Este trabajo presenta el desarrollo de un sistema de predicción para la deserción escolar en escuelas de educación superior, basado en el análisis de datos utilizando técnicas de machine learning. El sistema tiene como objetivo proporcionarnos predicciones para evitar la deserción temprana en la educación superior y así crear ayudas educativas para apoyarlos.

Este dataset recompila diversas bases de datos en una institución de educación superior, contiene información sobre estudiantes de varias carreras de pregrado. Los datos incluyen características demográficas, socioeconómicas y trayectoria académica al momento de inscripción, así como el rendimiento académico al final del primer y segundo semestre. La idea de implementación es construir un modelo de clasificación que predice la deserción escolar y el éxito académico.

Esta información se somete a un proceso de limpieza y normalización para asegurar la calidad y consistencia de los datos.

A partir de este análisis, se propone utilizar algoritmos de Aprendizaje supervisado (Clasificación) como KNeighborsClassifier y LinearDiscriminantAnalysis con el objetivo de clasificar a los estudiantes en diferentes grupos a categorías específicas.

Palabras clave

Sistemas de predicción, Rendimiento académico, deserción escolar, Clasificación.

Marco conceptual

Deserción Escolar: La deserción escolar es una problemática social ya que es cuando los estudiantes se alejan del sistema educativo. Este fenómeno tiene múltiples causas, que incluyen factores económicos, familiares, sociales y académicos. La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) y otras entidades internacionales han advertido que la deserción escolar limita las oportunidades de los jóvenes (UNESCO, Deserción escolar: una revisión de los factores y soluciones., 2021), disminuye el desarrollo económico y aumenta las desigualdades sociales.

Éxito Académico: El éxito académico se refiere al cumplimiento de los objetivos de aprendizaje establecidos en los programas educativos. Este concepto es complejo, ya que no se limita a la obtención de buenas calificaciones, sino que también implica la adquisición de competencias, habilidades, y el desarrollo integral del estudiante. Factores como la motivación, la metodología de enseñanza y el apoyo familiar suelen jugar un papel importante.

Machine Learning (ML): Machine learning, o aprendizaje automático, es un subcampo de la inteligencia artificial que permite que las computadoras aprendan patrones a partir de datos sin ser explícitamente programadas para ello. En el contexto educativo, el ML permite analizar grandes cantidades de datos históricos de estudiantes para identificar patrones y hacer predicciones. Las técnicas de ML como los árboles de decisión, redes neuronales y modelos de regresión son comúnmente empleadas para resolver problemas de clasificación y regresión, como los que se presentan al predecir la deserción y el éxito académico.

Predicción y Modelado Predictivo: La predicción en machine learning se basa en el modelado predictivo, que es el uso de un modelo matemático para prever el comportamiento futuro de un sistema. En este caso, se usan datos históricos y características de los estudiantes para prever si un estudiante pudiera desertar o tener éxito en sus estudios (Cruz, 2022). Los algoritmos de clasificación, como los modelos de regresión logística y los modelos de aprendizaje supervisado, son comunes para estas aplicaciones.

Factores de Riesgo en la Deserción Escolar y el Rendimiento Académico: Los factores que influyen en la deserción escolar y en el éxito académico suelen agruparse en varias categorías, entre ellas:

- **Factores personales**: incluyen el historial académico del estudiante, la motivación, la salud mental y física.
- **Factores familiares**: nivel socioeconómico, apoyo emocional, y nivel educativo de los padres.
- **Factores institucionales**: calidad de la enseñanza, infraestructura educativa, métodos de evaluación y relaciones con los docentes.
- **Factores sociales**: ambiente de la comunidad, influencia de pares, y accesibilidad al transporte escolar.

Marco contextual

Contexto Educativo Global: A nivel mundial, el sistema educativo se enfrenta a altos índices de deserción escolar, especialmente en regiones de bajos ingresos o con limitaciones en el acceso a recursos educativos (Moreno Bernal, 2014). La pandemia de COVID-19 exacerbó estos problemas, haciendo que la tasa de abandono escolar aumentara en diversas partes del mundo (UNESCO, Interrupción educativa y respuesta al Covid-19., 2020). Esto ha impulsado a las instituciones a buscar soluciones tecnológicas para monitorear y predecir el rendimiento y la permanencia de los estudiantes.

Situación en América Latina: En América latina según el reporte del Sistema de Información de Tendencias Educativas en América Latina, SITEAL (2010), a partir de los 13 años comienza a observarse un incremento sostenido en el porcentaje de los adolescentes que abandona la escuela a nivel regional (Román, 2013). Las instituciones educativas locales han comenzado a adoptar tecnologías avanzadas, incluyendo machine learning, para abordar este problema y mejorar el apoyo académico. Esto refleja una tendencia hacia la implementación de soluciones tecnológicas en los sectores educativos y de administración escolar, para mejorar los niveles de retención.

Desarrollo de Tecnología Educativa: El uso de herramientas de inteligencia artificial en la educación está en crecimiento. Actualmente, existen plataformas y programas de análisis de datos enfocados en la mejora del rendimiento educativo, aunque muchas de ellas aún están en fase experimental. Esto plantea una oportunidad para innovar y desarrollar soluciones específicas que puedan adaptarse al contexto y necesidades de cada institución educativa.

Relevancia Social del Proyecto: La deserción escolar y el éxito académico tienen efectos a largo plazo no solo en la vida de los estudiantes, sino también en el desarrollo económico y social de una región. Un sistema que pueda predecir y mitigar estos problemas tendría un impacto positivo en la sociedad, al mejorar la retención educativa y proporcionar una educación de calidad. Asimismo, permitiría a las instituciones educativas ser más proactivas, tomando decisiones basadas en datos que puedan ayudar a personalizar el apoyo a estudiantes en riesgo.

Descripción de caso de estudio.

El proyecto se trata sobre realizar un sistema de predicción de la deserción escolar y se centra en desarrollar un modelo utilizando ML para identificar ciertos factores que provocan la deserción escolar y interferir a tiempo para aportar ayudas educativas o psicológicas.
Los principales factores que podemos medir es la trayectoria académica, demografía y factores socioeconómicos de los estudiantes y el rendimiento académico de los estudiantes al final del primer y segundo semestre para poder determinar cuál es la mayor afectación.

El proyecto se base en un modelo de datos que incluye las siguientes variables traducidas al español (M.V.Martins, 2021):

- **Estado civil**: Estado civil del estudiante.
- **Curso**: Carrera técnica, tecnológica o universitaria del estudiante.
- **Asistencia diurna/noche:** Horario de asistencia del estudiante.
- **Titulación anterior:** Titulación obtenida anteriormente.
- **Titulación previa(grado)**: Titulación previa al grado (título académico).
- **Nacionalidad**: Nacionalidad del estudiante.
- **Titulación de la madre**: Carrera cursada por la madre del estudiante.
- **Titulación del padre**: Carrera cursada por el padre del estudiante.
- **Profesión de la madre**: Profesión de la madre.
- **Profesión del padre**: Profesión del padre.
- **Grado de admisión**: Nota de admisión al curso.
- **Desplazados**: Desplazado por la violencia.
- **Necesidades educativas especiales**: Enfermedad o necesidad especial para poder tomar el curso.
- **Deudor:** Deudas académicas del estudiante.
- **Gastos de matrícula al día**: Gastos en matriculas al día del estudiante.
- **Género**: Genero del estudiante.
- **Becario**: Estudiante con privilegio de beca.
- **Edad**: Edad del estudiante en el momento de la matricula.
- **Internacional**: Estudiante de intercambio.

- **Unidades curriculares 1er sem (acreditadas):** Materias acreditadas del primer semestre.
- **Unidades curriculares 1er sem (matriculados):** Materias matriculadas del primer semestre.
- **Unidades curriculares 1er sem (evaluaciones):** Materias evaluadas del primer semestre.
- **Unidades curriculares 1er sem (aprobadas):** Materias aprobadas del primer semestre.
- **Unidades curriculares 1er sem (calificación):** Materias calificadas del primer semestre.
- **Unidades curriculares 1er sem (sin evaluaciones):** Materias sin evaluación del primer semestre.
- **Unidades curriculares 2º sem (acreditadas):** Materias acreditadas del segundo semestre.
- **Unidades curriculares 2º sem (matriculados):** Materias matriculadas del segundo semestre.
- **Unidades curriculares 2º sem (evaluaciones):** Materias evaluadas del segundo semestre.
- **Unidades curriculares 2º sem (aprobadas):** Materias aprobadas del segundo semestre.
- **Unidades curriculares 2º sem (calificación):** Materias calificadas del segundo semestre.
- **Unidades curriculares 2º sem (sin evaluaciones):** Materias evaluadas del segundo semestre.
- **Objetivo:** Clasificación de los estudiantes (deserción, matriculado o graduado).

Pregunta problema

¿Cómo desarrollar un sistema basado en machine learning para predecir la deserción escolar y el éxito académico en estudiantes basado en datos experimentales, permitiendo intervenciones tempranas para mejorar la retención y el rendimiento?

Hipótesis

El análisis de la información de los estudiantes a la hora de matricularse junto con el sistema de machine learning de clasificación de aprendizaje supervisado basado en el análisis de factores académicos, socioeconómicos y personales puede predecir con alta precisión la probabilidad de deserción escolar y el éxito académico de los estudiantes, permitiendo intervenciones oportunas que mejoren la retención y el rendimiento.

Objetivos

Objetivo general.

Implementar estrategias computacionales para la clasificación de datos experimentales de los estudiantes para predecir la probabilidad de deserción escolar y el éxito académico, haciendo uso de algoritmos de Machine Learning.

Objetivos específicos.

- Caracterizar y procesar los factores clave (académicos, personales, socioeconómicos e institucionales) que influyen en la deserción escolar y el éxito académico de los estudiantes, con miras a la toma de decisiones informadas.
- Implementar un algoritmo de Machine learning para el aprendizaje supervisado (Clasificación) de los datos con miras al sistema de predicción.
- Evaluar y analizar el desempeño de los algoritmos implementados para la toma de decisiones.
- Validar el funcionamiento de toma de decisiones a partir de datos nuevos.

Desarrollo e implementación del aprendizaje

El dataset que analizaremos son datos recopilados para muestra y pruebas para lograr un sistema preciso para predecir la deserción escolar en tempranas etapas de estudio ya que se cuenta con un déficit de rendimiento académico.
El dataset cuenta con 4424 registros y 36 columnas que ayuda a nutrir el modelo de entrenamiento.
La intención es desarrollar el entrenamiento primero identificando los valores frecuentes entre los datos, identificar columnas que no nos sirve para nuestro estudio y por último entrenar el modelo de predicción para así lograr una implementación final.

Los datos son producto de una investigación del Instituto Politécnico de Portalegre (M.V.Martins, 2021)

Preparación y análisis de los datos

Para realizar un buen análisis haremos los primeros pasos con el dataset:

Tabla 1. Variables de Dataset

```
[ ] <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 4424 entries, 0 to 4423
    Data columns (total 31 columns):
     #   Column                                          Non-Null Count  Dtype
    ---  ------                                          --------------  -----
     0   Marital status                                  4424 non-null   int64
     1   Course                                          4424 non-null   int64
     2   Daytime/evening attendance                      4424 non-null   int64
     3   Previous qualification                          4424 non-null   int64
     4   Previous qualification (grade)                  4424 non-null   float64
     5   Nacionality                                     4424 non-null   int64
     6   Mother's qualification                          4424 non-null   int64
     7   Father's qualification                          4424 non-null   int64
     8   Mother's occupation                             4424 non-null   int64
     9   Father's occupation                             4424 non-null   int64
     10  Admission grade                                 4424 non-null   float64
     11  Educational special needs                       4424 non-null   int64
     12  Debtor                                          4424 non-null   int64
     13  Tuition fees up to date                         4424 non-null   int64
     14  Gender                                          4424 non-null   int64
     15  Scholarship holder                              4424 non-null   int64
     16  Age at enrollment                               4424 non-null   int64
     17  International                                   4424 non-null   int64
     18  Curricular units 1st sem (credited)             4424 non-null   int64
     19  Curricular units 1st sem (enrolled)             4424 non-null   int64
     20  Curricular units 1st sem (evaluations)          4424 non-null   int64
     21  Curricular units 1st sem (approved)             4424 non-null   int64
     22  Curricular units 1st sem (grade)                4424 non-null   float64
     23  Curricular units 1st sem (without evaluations)  4424 non-null   int64
     24  Curricular units 2nd sem (credited)             4424 non-null   int64
     25  Curricular units 2nd sem (enrolled)             4424 non-null   int64
     26  Curricular units 2nd sem (evaluations)          4424 non-null   int64
     27  Curricular units 2nd sem (approved)             4424 non-null   int64
     28  Curricular units 2nd sem (grade)                4424 non-null   float64
     29  Curricular units 2nd sem (without evaluations)  4424 non-null   int64
     30  Target                                          4424 non-null   int64
    dtypes: float64(4), int64(27)
    memory usage: 1.0 MB
```

Como se muestra en la tabla anterior el dataset tiene 31 columnas que se describen a continuación:

- **Estado civil** (1 – soltero 2 – casado 3 – viudo 4 – divorciado 5 – unión libre 6 – separados legalmente).
- **Curso** (33 - Tecnologías de Producción de Biocombustibles 171 - Animación y Diseño Multimedia 8014 - Servicio Social (asistencia nocturna) 9003 - Agronomía 9070 - Diseño de Comunicación 9085 - Enfermería Veterinaria 9119 - Ingeniería Informática 9130 - Equincultura 9147 - Gestión 9238 - Servicio Social 9254 - Turismo 9500 - Enfermería 9556 - Higiene Bucal 9670 - Gestión de Publicidad y Marketing 9773 - Periodismo y Comunicación 9853 - Educación Básica 9991 - Gestión (asistencia nocturna)).
- **Asistencia diurna/noche** (1 – diurno 0 - nocturno).
- **Titulación anterior** (1 - Educación secundaria 2 - Educación superior - licenciatura 3 - Educación superior - grado 4 - Educación superior - maestría 5 - Educación superior - doctorado 9 - 12º año de escolaridad - no completado 10 - 11º año de escolaridad - no completado 12 - Otros - 11º año de escolaridad 14 - 10º año de escolaridad 15 - 10º año de escolaridad - no completado 19 - Educación básica 3º ciclo (9º/10º/11º año) o equivalente 38 - Educación básica 2º ciclo (6º/7º/ 8º año) o equivalente. 39 - Curso de especialización tecnológica 40 - Educación superior - Licenciatura (1er ciclo) 42 - Curso técnico superior profesional 43 - Educación superior - maestría (2º ciclo)).
- **Titulación previa(grado)** (Grado de titulación previa (entre 0 y 200)).
- **Nacionalidad** (1 - portugués; 2 - Alemán; 6 - Español; 11 - Italiano; 13 - neerlandés; 14 - Inglés; 17 - lituano; 21 - angoleño; 22 - Cabo Verde; 24 - guineano; 25 - mozambiqueño; 26 - Santome; 32 - Turco; 41 - brasileño; 62 - rumano; 100 - Moldova (República de); 101 - Mexicano; 103 - ucraniano; 105 - ruso; 108 - Cubano; 109 - Colombiana).
- **Titulación de la madre** (1 - Educación Secundaria - 12º Año de Escolaridad o Ec. 2 - Educación Superior - Licenciatura 3 - Educación Superior - Grado 4 - Educación

Superior - Maestría 5 - Educación Superior - Doctorado 6 - Frecuencia de la Educación Superior 9 - 12º Año de Escolaridad - No Completado 10 - 11º Año de Escolaridad - No Completado).

- **Titulación del padre** (1 - Educación Secundaria - 12º Año de Escolaridad o Ec. 2 - Educación Superior - Licenciatura 3 - Educación Superior - Grado 4 - Educación Superior - Maestría 5 - Educación Superior - Doctorado 6 - Frecuencia de la Educación Superior 9 - 12º Año de Escolaridad - No Completado 10 - 11º Año de Escolaridad - No Completado).
- **Profesión de la madre** (0 - Estudiante 3 - Técnicos y Profesiones de Nivel Medio 4 - Personal Administrativo 5 - Servicios Personales, Trabajadores y Vendedores de Seguridad y Protección 6 - Agricultores y Trabajadores Calificados en la Agricultura, Pesca y Silvicultura 7 - Trabajadores Calificados en la Industria, Constructores y artesanos 8 - Operadores de instalación y máquinas y trabajadores de montaje 9 - Trabajadores no calificados 10 - Profesiones de las Fuerzas Armadas 90 - Otra situación 99 - (en blanco)).
- **Profesión del padre** (0 - Estudiante 3 - Técnicos y Profesiones de Nivel Medio 4 - Personal Administrativo 5 - Servicios Personales, Trabajadores y Vendedores de Seguridad y Protección 6 - Agricultores y Trabajadores Calificados en la Agricultura, Pesca y Silvicultura 7 - Trabajadores Calificados en la Industria, Constructores y artesanos 8 - Operadores de instalación y máquinas y trabajadores de montaje 9 - Trabajadores no calificados 10 - Profesiones de las Fuerzas Armadas 90 - Otra situación 99 - (en blanco)).
- **Grado de admisión** (Nota de admisión (entre 0 y 200)).
- **Desplazados** (1 – Sí 0 – No).
- **Necesidades educativas especiales** (1 – Sí 0 – No).
- **Deudor** (1 – Sí 0 – No).
- **Gastos de matrícula al día** (1 – Sí 0 – No).
- **Género** (1 – Hombre 0 – Mujer).
- **Becario** (1 – Sí 0 – No).

- **Edad** en el momento de la matrícula.
- **Internacional** (1 – Sí 0 – No).
- Unidades curriculares 1er sem (acreditadas).
- Unidades curriculares 1er sem (matriculados).
- Unidades curriculares 1er sem (evaluaciones).
- Unidades curriculares 1er sem (aprobadas).
- Unidades curriculares 1er sem (calificación).
- Unidades curriculares 1er sem (sin evaluaciones).
- Unidades curriculares 2º sem (acreditadas).
- Unidades curriculares 2º sem (matriculados).
- Unidades curriculares 2º sem (evaluaciones).
- Unidades curriculares 2º sem (aprobadas).
- Unidades curriculares 2º sem (calificación).
- Unidades curriculares 2º sem (sin evaluaciones).
- Objetivo (deserción, matriculado o graduado).

Las variables están orientadas en obtener un registro de un estudiante para observar si alguna de estas condiciones puede afectar en su deserción.

Vamos a realizar un análisis de datos:

Tabla 2. Análisis variables

```
[ ] #Análisis de los datos
    Conjunto_Datos.describe()
```

	Marital status	Course	Daytime/evening attendance\t	Previous qualification	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	...	Curricular units 1st sem (evaluations)	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	...	4424.000000	4424.000000	4.424000e+03	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4.424000e+03	4424.000000
mean	1.178571	8856.642631	0.890823	4.577758	132.613314	1.873192	19.561935	22.275316	10.960895	11.032324	...	8.299051	4.706600	4.458091e+15	0.137658	0.541817	6.232143	8.063291	4.435805	3.928634e+15	0.150316
std	0.605747	2063.566416	0.311897	10.216592	13.188332	6.914514	15.603186	15.343108	26.418253	25.263040	...	4.179106	3.094288	6.172473e+15	0.690880	1.918546	2.195951	3.947951	3.014764	5.970124e+15	0.753774
min	1.000000	33.000000	0.000000	1.000000	95.000000	1.000000	1.000000	1.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000
25%	1.000000	9085.000000	1.000000	1.000000	125.000000	1.000000	2.000000	3.000000	4.000000	4.000000	...	6.000000	3.000000	1.125000e+01	0.000000	0.000000	5.000000	6.000000	2.000000	1.100000e+01	0.000000
50%	1.000000	9238.000000	1.000000	1.000000	133.100000	1.000000	19.000000	19.000000	5.000000	7.000000	...	8.000000	5.000000	1.350000e+01	0.000000	0.000000	6.000000	8.000000	5.000000	1.300000e+01	0.000000
75%	1.000000	9556.000000	1.000000	1.000000	140.000000	1.000000	37.000000	37.000000	9.000000	9.000000	...	10.000000	6.000000	1.216667e+16	0.000000	0.000000	7.000000	10.000000	6.000000	1.166667e+16	0.000000
max	6.000000	9991.000000	1.000000	43.000000	190.000000	109.000000	44.000000	44.000000	194.000000	195.000000	...	45.000000	26.000000	1.733333e+16	12.000000	19.000000	23.000000	33.000000	20.000000	1.857143e+16	12.000000

8 rows × 30 columns

En el análisis de datos que nos proporciona en la consulta nos da lo siguiente:

Notamos que el estado civil de los estudiantes puede afectar en la toma de decisiones ya que la mayoría de los estudiantes según el análisis de datos son tipo 6 (separados legalmente) y esto puede influir a la tendencia de deserción escolar; mientras el 50% son solteros.

En la nacionalidad de los estudiantes que participaron en este estudio vemos que son nacionalidad 109 (colombiana), otro punto que puede influir en la deserción escolar; por otra parte, el 50% son portugueses.

Tabla 3. Ajustes valores variable Target

```
# Revisando opciones de una variable
Conjunto_Datos['Target'].unique() #Muestra las opciones de la variable
```

```
array(['Dropout', 'Graduate', 'Enrolled'], dtype=object)
```

```
[9] # Cambiando valores en la variable
    Opciones = {'Dropout':0,'Graduate':1,'Enrolled':2}
    Conjunto_Datos['Target'] = Conjunto_Datos['Target'].map(Opciones)
    Conjunto_Datos.head()
```

ous ion	Previous qualification (grade)	Nacionality	Mother's qualification	Father's qualification	Mother's occupation	Father's occupation	...	Curricular units 1st sem (approved)	Curricular units 1st sem (grade)	Curricular units 1st sem (without evaluations)	Curricular units 2nd sem (credited)	Curricular units 2nd sem (enrolled)	Curricular units 2nd sem (evaluations)	Curricular units 2nd sem (approved)	Curricular units 2nd sem (grade)	Curricular units 2nd sem (without evaluations)	Target
1	122.0	1	19	12	5	9	...	0	0.000000e+00	0	0	0	0	0	0.000000e+00	0	0
1	160.0	1	1	3	3	3	...	6	1.400000e+01	0	0	6	6	6	1.366667e+16	0	1
1	122.0	1	37	37	9	9	...	0	0.000000e+00	0	0	6	0	0	0.000000e+00	0	0
1	122.0	1	38	37	5	3	...	6	1.342857e+16	0	0	6	10	5	1.240000e+01	0	1
1	100.0	1	37	38	9	9	...	5	1.233333e+16	0	0	6	6	6	1.300000e+01	0	1

Realizamos un ajuste a la variable Target (Objetivo), cambiando sus valores a valores algorítmicos para poder realizar operaciones.
'Dropout':0,'Graduate':1,'Enrolled':2

Con estos valores modificados procedemos a realizar evaluación de las variables a través de tablas de frecuencia:

Tabla 4. Tabla de frecuencia para variable Target

Tabla de Frecuencia:

	Intervalo	Frecuencia Absoluta	Frecuencia Acumulada
0	[1.0, 1.2)	2209	2209
1	[0.0, 0.2)	1421	3630
2	[1.8, 2.002)	794	4424
3	[0.2, 0.4)	0	4424
4	[0.4, 0.6)	0	4424
5	[0.6, 0.8)	0	4424
6	[0.8, 1.0)	0	4424
7	[1.2, 1.4)	0	4424
8	[1.4, 1.6)	0	4424
9	[1.6, 1.8)	0	4424

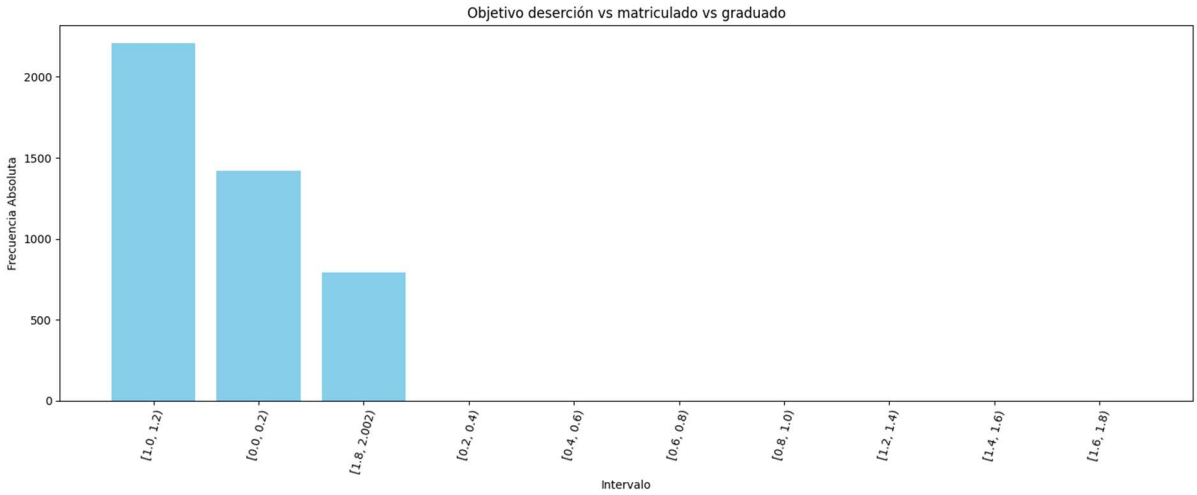

Figura 1. Objetivo deserción vs matriculado vs graduado

En este grafico podemos observar como el sistema esta escalonado donde la mayoría de los estudiantes son graduados, seguido de los estudiantes con deserción escolar y por ultimo los estudiantes matriculados.

Esto nos da a entender que la gran mayoría de estudiantes si se logran graduar teniendo un cambio de visión en el análisis.

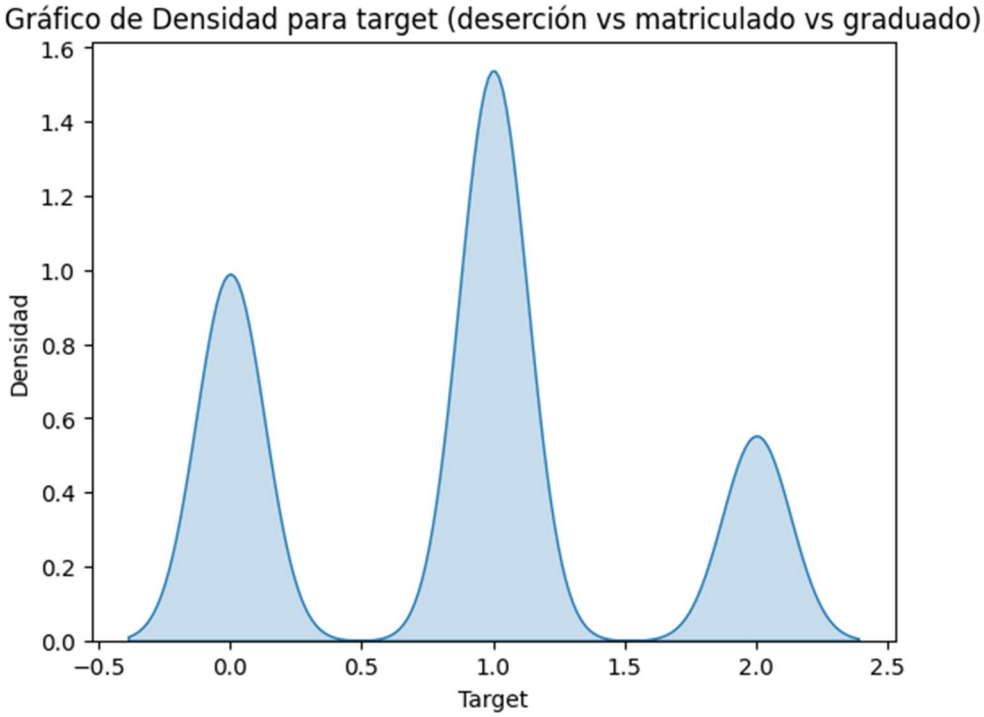

Figura 2. Gráfico de densidad por target

En el grafico de densidad lo podemos ver más gráficamente como los estudiantes graduados superan ampliamente al resto.

Tabla 5. Tabla de frecuencia para variable nacionalidad

Tabla de Frecuencia:

	Intervalo	Frecuencia Absoluta	Frecuencia Acumulada
0	[1.0, 11.8)	4332	4332
1	[33.4, 44.2)	38	4370
2	[22.6, 33.4)	22	4392
3	[11.8, 22.6)	18	4410
4	[98.2, 109.108)	12	4422
5	[55.0, 65.8)	2	4424
6	[44.2, 55.0)	0	4424
7	[65.8, 76.6)	0	4424
8	[76.6, 87.4)	0	4424
9	[87.4, 98.2)	0	4424

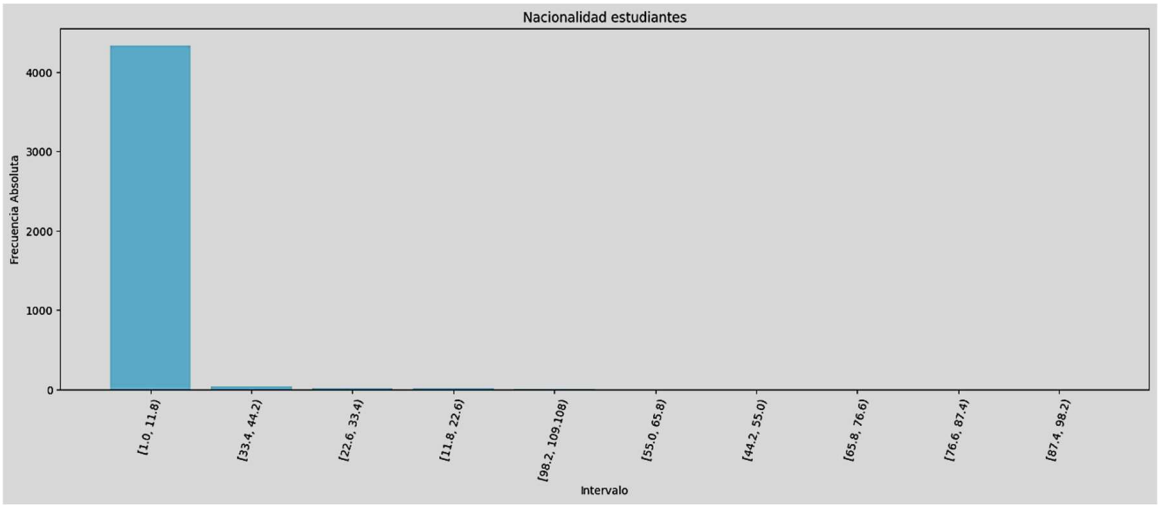

Figura 3. Tabla de frecuencia Nacionalidad

Aquí podemos observar que casi en su totalidad todos los estudiantes tienen nacionalidad entre 1 y 12 que se refiere a las siguientes nacionalidades: 1 - Portugués; 2 - Alemán; 6 - Español; 11 – Italiano.

Mientras el intervalo mas bajo son las nacionalidades comprendidas entre 55 y 66 que son: 41 - brasileño; 62 – rumano.

Con esto podemos concluir que no hay una relación que la nacionalidad afecte a la deserción escolar ya que como analizamos anteriormente hay un nivel alto de graduados

junto con un nivel alto de un grupo de nacionalidades, si la nacionalidad afectara los grupos de datos estuvieran más dispersos.

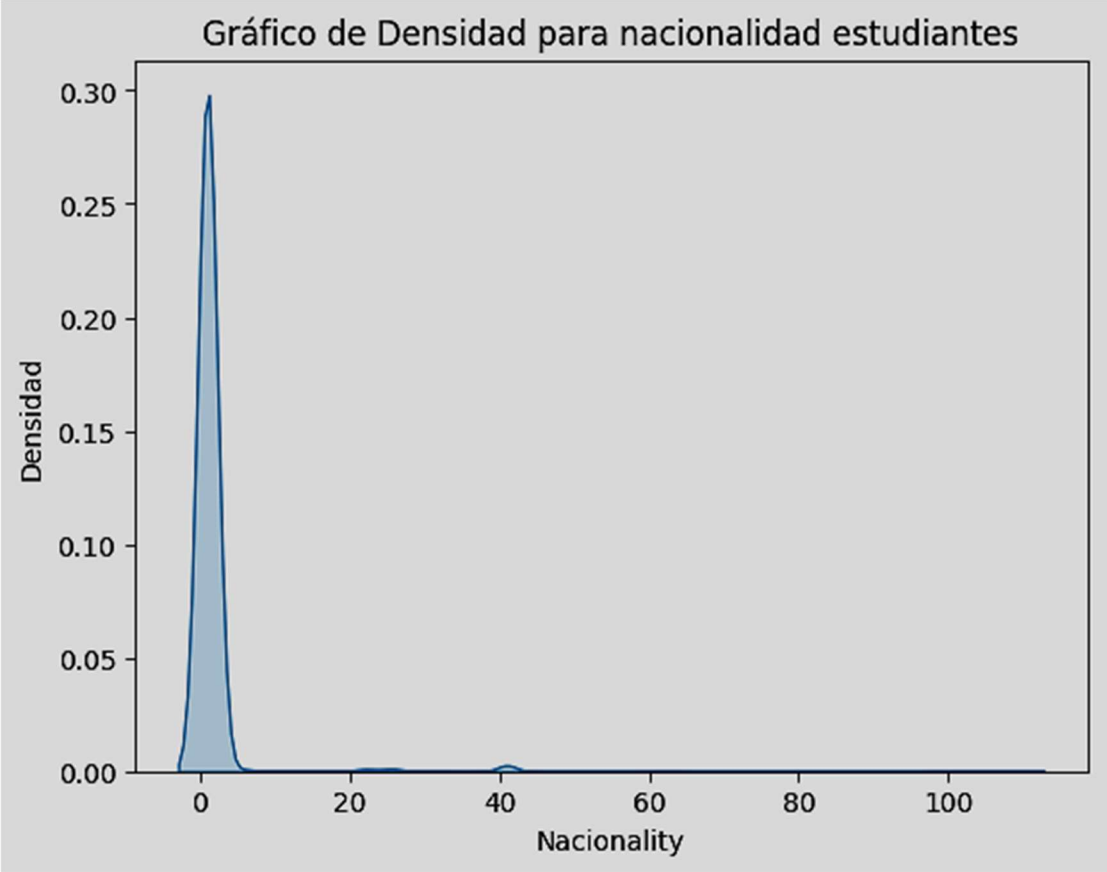

Figura 4. Gráfico densidad nacionalidad

Podemos concluir con el grafico de densidad que la nacionalidad de la gran mayoría de estudiantes se encuentra dentro del mismo grupo de nacionalidades.

Modelo de toma de decisiones

Aprendizaje supervisado (Clasificación)

Después de ver los datos, el comportamiento de estos y de analizar que los datos se encuentran dividido en grupos no muy correlacionados se toma la decisión de usar como

modelo de decisiones el aprendizaje supervisado para poder obtener un modelo predictivo de clasificación para poder tener una respuesta oportuna para evitar la deserción escolar.

Asignamos los datos para entrada y salida y dividimos los datos para realizar el entrenamiento y testeo.

```
#Divide datos en entradas y salidas
import numpy as np
Datos_matriz=np.array(Conjunto_Datos)
X = Datos_matriz[:,0:-1] #datos de entrada (Todas las variables del estudiante)
Y = Datos_matriz[:,-1] #Datos de salida (La decisiòn)

# Divide datos en Entrenamiento y testeo
import sklearn
from sklearn.model_selection import train_test_split
X_train, X_test,Y_train, Y_test= train_test_split(X,Y,test_size=0.1,random_state=751)

#Para mejorar la escala de los datos se hace normalization (Ignorar)
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figura 5. Código para entrenamiento y división del modelo

Realizamos la división de los datos por matrices de datos de entrada al modelo y datos de salida que es la toma de la decisión.

Análisis de desempeño

Realizamos el entrenamiento del modelo usando diferentes clasificadores de inteligencia artificial que nos darán unos porcentajes de precisión después de analizar los datos de entrada.

```
# Evaluando casos mediante todos los clasificadores
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,precision_score

Modelo_0 = KNeighborsClassifier(5)
Modelo_0.fit(X_train, Y_train)
Y_pred_0 =Modelo_0.predict (X_test)
print("Accuracy KNN",accuracy_score(Y_test, Y_pred_0))

Modelo_1 = GaussianNB()
Modelo_1.fit(X_train, Y_train)
Y_pred =Modelo_1.predict (X_test)
print("Accuracy Bayes",accuracy_score(Y_test, Y_pred))

Modelo_2 = LinearDiscriminantAnalysis()
Modelo_2.fit(X_train, Y_train)
Y_pred_2 =Modelo_2.predict (X_test)
print("Accuracy LDA",accuracy_score(Y_test, Y_pred_2))

Modelo_3 = QuadraticDiscriminantAnalysis()
Modelo_3.fit(X_train, Y_train)
Y_pred_3 =Modelo_3.predict (X_test)
print("Accuracy QDA",accuracy_score(Y_test, Y_pred_3))

Modelo_4 = DecisionTreeClassifier()
Modelo_4.fit(X_train, Y_train)
Y_pred_4 =Modelo_4.predict (X_test)
print("Accuracy Tree",accuracy_score(Y_test, Y_pred_4))

Modelo_5 = SVC()
Modelo_5.fit(X_train, Y_train)
Y_pred_5 =Modelo_5.predict (X_test)
print("Accuracy SVM",accuracy_score(Y_test, Y_pred_5))
```

```
Accuracy KNN 0.6478555384740487
Accuracy Bayes 0.6478555384740487
Accuracy LDA 0.7426636568848759
Accuracy QDA 0.7110600488812641
Accuracy Tree 0.672686230248307
Accuracy SVM 0.7358916478555305
```

Figura 6. Código de evaluación casos

Después de realizar la evaluación de los diferentes casos teniendo como entrada el dataset suministrado los diferentes clasificadores nos dieron buenos resultados a la hora de realizar predicciones con nuestro modelo.

El modelo que nos arrogo el porcentaje más alto de predicción fue el LDA con un 74%.

Validación del modelo

Para dar uso del modelo de predicción el usuario tendrá que ingresar unos datos de entrada del estudiante tales como el estado civil, curso, asistencia, titulación anterior y previa, nacionalidad, Titulación tanto de la madre como del padre, gatos de matrícula, género y las unidades curriculares.

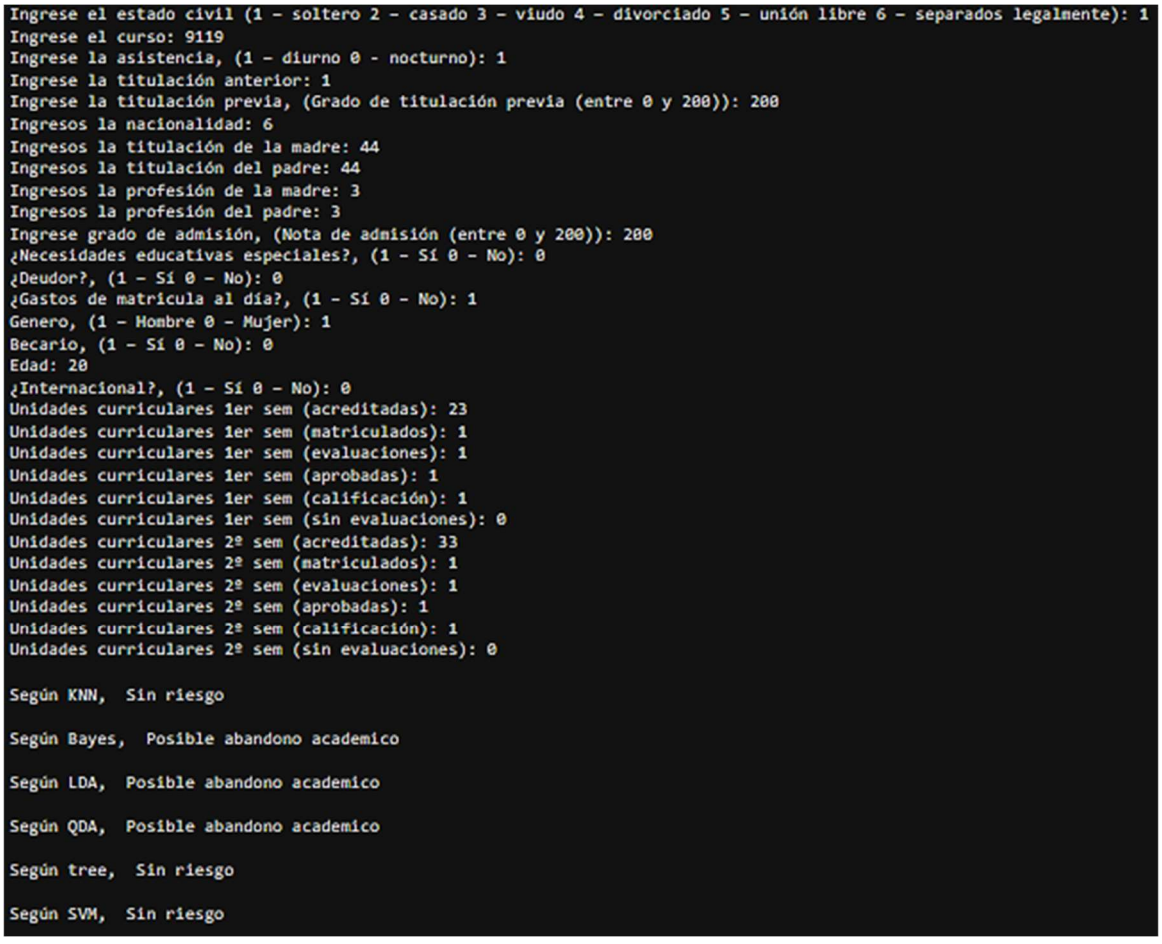

Figura 7. Resultado validación modelo predicción

Conclusiones y trabajos futuros

El desarrollo de un sistema de predicción ha permitido analizar los parámetros que pueden influir en la deserción escolar, como los problemas sociales y socioeconómicos afecta considerablemente a los jóvenes que a pesar de querer estudiar se ven obligados al abandono escolar.

Este trabajo ha permitido identificar los factores clave que influyen en el rendimiento y la permanencia de los estudiantes, integrándolos en un modelo predictivo que demuestra su capacidad para analizar grandes volúmenes de datos y detectar patrones relevantes con alta precisión.

Con la implementación del modelo de predicción a través del aprendizaje automatizado por clasificación hemos encontrado una manera de prevenir a tiempo un abandono prematuro de los estudiantes, identificamos que factores económicos y la edad son los principales ocasionales de deserción escolar.

Trabajos futuros:

Este modelo de ML puede ser explotado para hacerlo automático y enviar alertas educativas cada vez que se determina riesgo de deserción escolar en los estudiantes, es un caso de uso muy potente para desarrollar en colegios y universidades.

Este proyecto sienta las bases para futuras investigaciones, invitando a una exploración más profunda de los factores socioemocionales y culturales en el modelado predictivo. Asimismo, abre la posibilidad de implementar el sistema en entornos educativos reales, evaluando su impacto en la reducción de la deserción y el mejoramiento del rendimiento estudiantil.

Referencias bibliográficas

Cruz, E. G. (2022). *Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios.* Obtenido de https://doi.org/10.33412/pri.v13.1.3039

M.V.Martins, D. T. (12 de 12 de 2021). *"Predicción temprana del rendimiento de los estudiantes en la educación superior: un estudio de caso" Tendencias y aplicaciones en sistemas y tecnologías de información, vol.1, en la serie Avances en Sistemas Inteligentes y Computación. .* Obtenido de https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Moreno Bernal, D. M. (2014). *La Deserción Escolar: Un problema de Carácter Social. In Vestigium Ire, 6(1).* Obtenido de http://revistas.ustatunja.edu.co/index.php/ivestigium/article/view/795

Román, M. (2013). *Factores asociados al abandono y la deserción escolar en América Latina: una mirada en conjunto. REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 11(2), 33-59.* Obtenido de https://www.redalyc.org/pdf/551/55127024002.pdf

UNESCO. (01 de 09 de 2020). *Interrupción educativa y respuesta al Covid-19.* Obtenido de https://es.unesco.org/covid19/educationresponse

UNESCO. (2021). *Deserción escolar: una revisión de los factores y soluciones.* Obtenido de https://unesdoc.unesco.org/