Modelo predictivo de la deserción académica en la UPTC durante el semestre de 2019 Aplicación de Support Vector Machines y Metodología CRISP-DM.

Andrés Rivera y Daniel Portilla
Facultad de Ciencias Exactas y Naturales, programa de Licenciatura en informática.
Pasto, Nariño, Colombia.
andress101013@gmail.com
danielportilla30@gmail.com

Resumen— Este trabajo presenta un estudio que utiliza CRISP-DM y máquina de vectores de soporte (SVM) sobre deserción académica en los semestres del 2019 en el la Universidad Pedagógica y Tecnológica de Colombia. Se Analizan datos sociodemográficos y académicos de los estudiantes para predecir la deserción académica. Se hace énfasis en las fases de entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue. Se identifican variables influyentes como la edad, el sexo y la clase socioeconómica entre otras. Los resultados muestran que estimamos con precisión la deserción estudiantil con una precisión promedio del 80%. Este modelo se entrega a través de una interfaz de usuario intuitiva y visualmente atractiva. Este estudio contribuye a mejorar el conocimiento sobre la deserción académica y proporciona una herramienta útil para la identificación e intervención temprana del riesgo de abandono estudiantil.

Palabras claves: Desempeño académico, investigación pedagógica, estudio académico, variables sociodemográficas, Education Data Mining.

I. Introducción

La disponibilidad de datos en el campo educativo brinda oportunidades para comprender la deserción académica, la minería de datos se ha convertido en una herramienta invaluable para analizar e identificar tendencias y patrones para predecir en este caso la deserción académica. [1] Según Han, Pei y Kamber (2006), la minería de datos es esencial para extraer conocimiento valioso de grandes cantidades de datos, como los que se encuentran en el sector educativo.

El alcance de este trabajo es aplicar la metodología CRISP-DM para abordar el problema de la deserción estudiantil. [2] Enfoque CRISP-DM propuesto por Chapman. (2000) proporcionan una estructura metodológica sólida que guía el proceso de minería de datos a través de seis etapas: entendimiento del negocio, entendimiento de los datos, preparación de datos, modelado, evaluación e implementación.

Cada etapa del modelo CRISP-DM implica el uso de diferentes datos e información, así como métodos específicos para lograr los objetivos establecidos. [3] Witten y Frank (2016) destacan la importancia de comprender profundamente los datos antes de aplicar técnicas de minería de datos, subrayando la relevancia de la fase de Entendimiento de los datos en el proceso.

II. Materiales y Métodos

A. Materiales y métodos

Para cada fase del modelo CRISP-DM, se utiliza materiales y documentos respectivos que contemplen el buen desarrollo de la minería de datos enfocándose en el tema principal que es la deserción académica, para ello, a continuación, se encuentra una breve descripción en cada etapa en donde se verá reflejado lo que se va a utilizar.

- Entendimiento del negocio

Para la primera etapa se tendrá en cuenta el objetivo general de este trabajo, basándose primeramente en el título, el cual es comprender y analizar el dataset que contiene una serie de datos que abarca la deserción académica durante el semestre de 2019 en la UPTC y observar la información predispuesta de cada alumno.

- Entendimiento de los datos

Se realizará una recopilación de datos relevantes de cada uno de los estudiantes, en donde se vean reflejados por una serie de gráficos que permitan contemplar a simple vista y de manera más fácil lo que ocurre con estos datos.

- Preparación de los datos

Se realizará un análisis de los datos relevantes de cada uno de los estudiantes, con la ayuda de gráficos y estadísticas entender la naturaleza de los datos reflejados.

- Modelado

En la parte de modelado, se eligen técnicas como SVM (Support Vector Machine) para predecir el estado final de los estudiantes en función de los atributos proporcionados, para agrupar a los estudiantes en segmentos homogéneos según sus características. Esto transforma el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, y finalmente entrena un modelo predictivo; evaluando su desempeño para predecir el desempeño de los estudiantes.

- Evaluación

Esta etapa implica evaluar y ajustar, si es necesario, varios factores para mejorar el desempeño. Además, se realiza un análisis de sensibilidad para comprender la influencia de diferentes variables en las predicciones.

- Despliegue

Para la última etapa se implementa el modelo, como también una serie de retroalimentación y seguimiento para el éxito del proyecto.

III. Entendimiento del negocio

A. ¿De qué trata el conjunto de datos?

Casos de deserción académica presentados por algunos estudiantes presentados por la UPTC en el primer y segundo semestre de 2019, entre los atributos del dataset están características sociales y demográficas.

Es importante entender el objetivo principal que dio paso a realizar este trabajo, en este caso hablamos de la deserción estudiantil que hace énfasis en el área de pregrado, la cual está enfocada la base de datos que se utilizó para realizar la respectiva minería de datos (data mining) o tratamiento de los mismos.

Las razones detrás de la deserción son diversas y complejas, es esencial abordar este fenómeno desde múltiples perspectivas para desarrollar estrategias efectivas de prevención y retención, para ello se utilizará datos que brinda él data-set, en los cuales se puede ver reflejado, además de utilizar como ejemplo, el estrato, aunque suene algo clasista es imposible no tener en cuenta los ingresos familiares, o incluso del estudiante, debido a que si no posee los recursos esenciales puede dar paso a ser una de las causas o la causa principal para desertar.

Es importante reconocer que el abandono de la escuela no es sólo un fracaso individual, sino un reflejo de deficiencias sistémicas en la educación. Las instituciones educativas deben asumir la responsabilidad de identificar y abordar los factores que conducen al abandono escolar, como los problemas financieros, las dificultades académicas y la falta de apoyo social y emocional.

Además, es fundamental reconocer que la deserción estudiantil no solo representa una pérdida personal para los individuos, sino también una pérdida para la sociedad en su conjunto. La falta de educación superior puede limitar las oportunidades de empleo y el desarrollo económico, perpetuando así el ciclo de desigualdad y marginación. Por lo tanto, invertir en programas de retención estudiantil no solo beneficia a los individuos, sino que también contribuye al bienestar y progreso de la sociedad en su conjunto.

De acuerdo con [4] Quiroz (2022), Las Tecnologías de la información y la comunicación TIC se ha convertido en un instrumento fundamental para que los países aceleren su crecimiento y desarrollo. Incluso ha sido considerada por el Programa de las Naciones Unidas como un indicador del desarrollo humano. Las TIC han avanzado significativamente hasta lograr permear en las distintas esferas de la sociedad: educativas, cultural, económicas, política, social, entre otras. De tal importancia que se ha convertido en el objeto de estudio para diferentes disciplinas, analizando el comportamiento, la influencia, el impacto, los efectos y la relación que estas tienen con diferentes variables socio-económicas. En muchos casos, la deserción estudiantil está estrechamente ligada a desigualdades socioeconómicas y culturales. Los estudiantes de bajos ingresos, minorías étnicas y grupos marginados enfrentan desafíos adicionales que pueden obstaculizar su éxito académico, es por ello que el auge de las nuevas tecnologías y la transformación que estas han provocado a varios ámbitos, como lo afirma la autora, ha logrado solucionar ciertas dificultades que generaban la deserción de los estudiantes, cabe recalcar que el uso de un dispositivo o un acceso de red a internet, también contempla en gran parte a la economía, pero que gracias a los diferentes proyectos que ha implementado el estado, se ha podido observar una baja en la estadística de deserción en las diferentes instituciones, llegando a ser útil para las personas que más lo necesitan.

Es importante tener en cuenta un respaldo o antecedente al momento de trabajar un tema en específico que permita la elaboración de un trabajo o investigación de manera correcta, por esta razón, para la elaboración de este trabajo el cual se enfoca en la deserción estudiantil, que es el indicador principal del data-set, se hace referencia al Sistema para Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) el cual hace parte indispensable al momento de extraer información de estudiantes que desertan o hacen cambios entre programas en las instituciones de educación superior, permitiendo mejoras al momento de monitorear estos movimientos.

Por otra parte, hay que tener en cuenta cuando un estudiante se clasifica como desertor, [5] según el gobierno nacional de Colombia, es importante resaltar que un estudiante se clasifica como desertor del sistema de educación superior cuando no se ha matriculado por dos o más periodos consecutivos en algún programa académico, y para el caso de los cierres estadísticos oficiales correspondientes a los indicadores tasa de deserción anual y tasa de ausencia inter semestral siempre se toma el primer semestre de cada año como el indicador de referencia para dicha vigencia. Es decir, el indicador de tasa de deserción anual para el año 2021, corresponde a los estudiantes

identificados como desertores, que cumplieron dos semestres consecutivos sin matricularse, y que estaban matriculados en el primer semestre del año 2020.

A continuación, se mostrará un ejemplo de la extracción estadística de la tasa de deserción anual, en donde se ve involucrados diferentes niveles de formación, en los cuales están presentes tres años, a los cuales corresponden desde el 2019 hasta el 2021, en donde se puede observar un alza y baja entre el transcurso de los periodos.

TABLA I
Tasa de deserción anual

Nivel de formación	2019	2020	2021
Universitario	8,25%	8,02%	8,89%
TyT	14,84%	13,39%	16,51%
Técnico profesional	18,05%	13,65%	18,79%
Tecnológico	13,20%	13,26%	15,32%
Total general	**9,29%**	**8,85%**	**10,08%**

En la siguiente figura se muestra la tasa de deserción anual pero desde el año 2010 hasta el 2021 y de igual manera que en los porcentajes de la anterior tabla se puede concluir que uno de los motivos de deserción puede ser el transcurso de cada año, por ejemplo, la aparición del Covid-19 a finales del año 2019 e inicios del 2020, que según la gráfica muestra un bajo porcentaje de deserción, esto debido que se implementaron las clases virtuales o remotas, en donde más de un estudiante salió beneficiado gracias a un nivel de exigencia menor al que normalmente se ha llevado a cabo, todo por la aparición de esta circunstancia Fig. 1.

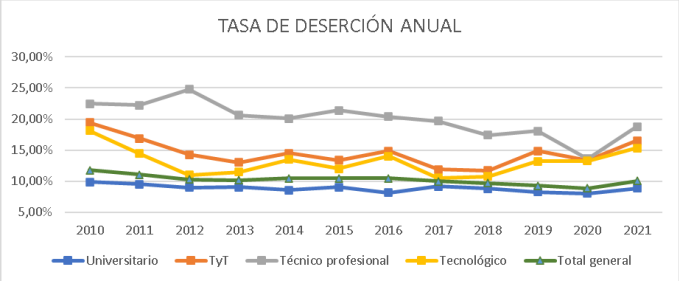

Fig. 1 Corte de los datos: cierre estadístico de 2022

Para concluir, este sistema implementado por el ministerio de educación nacional tiene en cuenta otras variables aparte del nivel de formación, las cuales son sector, sexo, metodología, área de conocimiento, entre otras. Todas las anteriores pueden dar paso para llegar al factor que influye en la deserción estudiantil.

Cabe recalcar que este trabajo intenta aplicar algo similar al objetivo del SPADIES, en donde al realizar el tratamiento de datos, se logra monitorear de mejor manera la información brindada y permite encontrar el producto influyente de la deserción, una vez se realice lo anterior, daría paso para implementar estrategias que permitan bajar el índice de deserción.

IV. Entendimiento de los Datos

A. *Descripción de variables*

El conjunto de datos consta de 13 variables:

TABLA II
Descripción de variables

No	VARIABLE	DESCRIPCIÓN
1	Descripción	Semestre actual a la toma de datos. Nominal: primer semestre de 2019 – segundo semestre de 2019.
2	Nombre_Facultad	Facultad al cual pertenece el estudiante. Nominal: Contiene 11 valores únicos.
3	Nombre Programa	Programa al cual pertenece el estudiante. Nominal: Contiene 41 valores únicos.
4	Jornada	Jornada a la cual pertenece el estudiante. Nominal: "Nocturno", "Diurno", "Otra".
5	Modalidad	Modalidad de estudio a la cual pertenece el estudiante. Nominal: “Distancia”, “Presencial”.
6	Nombre_Sede	Sede a la cual pertenece el estudiante. Nominal: Contiene 4 valores únicos.
7	Tipo_Iden_Est	Tipo de documento de identidad del estudiante. Nominal: C - Cedula, T – Tarjeta de identidad, E – Cedula de Extranjería.
8	Lug_Expedicion	Lugar de expedición del documento Nominal: Contiene 49 valores únicos.
9	Lugar_Nacimiento	Lugar de nacimiento del estudiante. Nominal: Contiene 49 valores únicos.
10	Fecha_Nacimiento	Fecha de nacimiento del estudiante. Numérica: 237 valores únicos.
11	Sexo	Sexo del estudiante. Nominal: F - Femenino, M – Masculino.
12	Estrato	Estrato del estudiante. Numérico: Rango del 1 al 5.
13	Nombre Estado	Estado académico del estudiante. Nominal: “Retirado con cupo reservado” – “No matriculado”.

El data-set contiene un total de mil cuarenta y seis (1046) registros, trece atributos (13), novecientos treinta y seis (936) datos faltantes repartidos de la siguiente manera, ver [9] punto 3, 3.1, 3.2 y 3.3.

TABLA III
Descripción de datos faltantes

CATEGORIA	NUMERO DE DATOS FALTANTES
'DESCRIPCION'	0
'NOMBRE_FACULTAD'	0
'NOMBRE_PROGRAMA'	0
'JORNADA'	47
'MODALIDAD'	83
'NOMBRE_SEDE'	0
'TIPO_IDEN_EST'	0
'LUG_EXPEDICION'	315
'LUGAR_NACIMIENTO'	315
'SEXO'	0
'NOMBRE_ESTADO'	0
'ESTRATO'	176
'FECHA_NACIMIENTO'	0
'TOTAL'	936

Se analizo el conjunto de datos con la finalidad de entender cómo se distribuye el conjunto objetivo, en este caso denominado NOMBRE_ESTADO, esta variable indica el estado actual de los estudiantes en relación con la institución educativa. El análisis evidencia la siguiente distribución de clases.

- Retirado con cupo reservado: Representa el porcentaje de estudiantes en el estado " Retirado con cupo reservado ".
- No matriculado: Indica el porcentaje de estudiantes en el estado " No matriculado ".

Es importante la distribución del conjunto objetivo para entender cómo se representan las diferentes clases en el conjunto de datos; esto brinda información relevante para la toma de decisiones y el desarrollo de estrategias en el campo educativo.

El porcentaje de cada clase en el conjunto objetivo se muestra a continuación, Fig. 2:

- Retirado con cupo reservado: 75.9%
- No matriculado: 24%

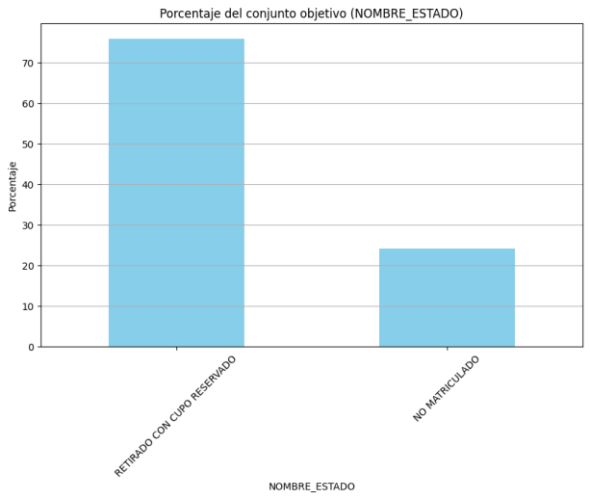

Fig. 2 Distribución conjunto objetivo

En cuanto a la distribución contamos que el 75.9 % de los estudiantes que figuran en la data-set tienden a Retirarse de la Universidad con reserva de cupo y el 24% de los estudiantes tienden a no matricularse.

B. Estadística De Los Datos

Calcular las estadísticas de los datos es una parte fundamental en el análisis de datos, puesto que proporcionan información valiosa para resumir, comprender y tomar decisiones basadas en los datos que tenemos.

Existen medidas de tendencia central que proporcionan un resumen de un conjunto de datos, por ejemplo, de los datos proporcionados, realicemos una suma de datos y el resultado lo dividimos por el número total de valores de ese conjunto, es la definición a lo que comúnmente llamamos MEDIA.

El conjunto de datos a trabajar contiene variables categóricas y numéricas, por lo cual se debió asignar un valor numérico único a cada categoría y luego calcular la media de estos valores, lo que nos llevó a los siguientes resultados ver [9] punto 3, 3.5.

TABLA IV
Estadística de los datos – Medias Categóricas

CATEGORIA	MEDIA NUMÉRICA	MEDIA CATEGÓRICA
'DESCRIPCION'	0.36	PRIMER SEMESTRE DE 2019
'NOMBRE_FACULTAD'	5.18	DERECHO Y CIENCIAS SOCIALES
'NOMBRE_PROGRAMA'	40.8	ADMINISTRACION DE SERVICIOS DE SALUD
'JORNADA'	0.50	DIURNO
'MODALIDAD'	0.052	PRESENCIAL
'NOMBRE_SEDE'	2.36	SANTA FE DE BOGOTA
'TIPO_IDEN_EST'	0.13	C
'LUG_EXPEDICION'	10.27	AGUAZUL
'LUGAR_NACIMIENTO'	10.38	AGUAZUL
'SEXO'	0.40	M
'NOMBRE_ESTADO'	0.24	RETIRADO CON CUPO RESERVADO

Por otra parte, las medias de las categorías numéricas nos dieron el siguiente resultado ver [9] punto 3, 3.6.

TABLA V
Estadística de los datos – Medias Numéricas

VARIABLE	MEDIA
'ESTRATO'	1
'EDAD'	24

En cuanto a la variable Fecha_Nacimiento , se tuvo en cuenta la edad que tenían los estudiantes a la fecha de la toma de datos, es decir 2019, por lo cual se transformó el dato de Fecha_Nacimiento por EDAD.

Por otra parte, identificamos esos valores que aparecen con mayor frecuencia en el conjunto de datos, lo que normalmente denominamos MODA, ver [9] punto 3, 3.7.

TABLA VI
Estadística de los datos – Moda

VARIABLE	MODA
'ESTRATO'	2
'EDAD'	23
'DESCRIPCION'	PRIMER SEMESTRE DE 2019
'NOMBRE_FACULTAD'	CIENCIAS DE LA EDUCACION
'NOMBRE_PROGRAMA'	ADMINISTRACION DE EMPRESAS
'JORNADA'	DIURNO
'MODALIDAD'	PRESENCIAL
'NOMBRE_SEDE'	TUNJA
'TIPO_IDEN_EST'	TUNJA
'LUG_EXPEDICION'	C
'LUGAR_NACIMIENTO'	TUNJA
'SEXO'	M
'NOMBRE_ESTADO'	RETIRADO CON CUPO RESERVADO

En cuanto a la desviación estándar de los datos numéricos, en este caso EDAD y SEXO, obtuvimos, ver [9] punto 3, 3.8.

TABLA VII
Estadística de los datos – Desviación estándar

VARIABLE	DESVIACIÓN ESTÁNDAR
'ESTRATO'	0.78
'EDAD'	7.1

En la variable ESTRATO se puede inferir que en general los valores adoptados en esta variable están relativamente cerca de la media, teniendo una dispersión moderada de 0.78.

Por otra parte, en cuanto a la edad su desviación estándar es de 7.1 años lo cual no indica que hay una cantidad significativa de variabilidad en las edades de las personas en el conjunto de datos, lo que sugiere una distribución más dispersa alrededor de la media.

C. Tarea ML a realizar

Para determinar la tarea de Machine Learning que mejor se adapte al objetivo del trabajo actual, es importante definir primero que es el Machine Learning.

Machine Learning es un campo interdisciplinar que combina varios conceptos tales como la informática, las matemáticas, las ciencias de la computación entre otras, con el objetivo de desarrollar algoritmos y modelos que permitan realizar tareas específicas de una manera automatizada. Según [6] T. Mitchell en su libro "Machine Learning" (1997) un programa de computadora se dice que aprende de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su rendimiento en las tareas T, medida por P, mejora con la experiencia E, lo que implica que los algoritmos pueden mejorar su rendimiento en tareas específicas mientras más experiencia o datos tengan.

Una parte esencial al momento de elegir la mejor tarea ML a aplicar, es entender algunos autores dividen el aprendizaje en supervisado y no supervisado, un ejemplo de ellos es [7] Bishop en su libro "Pattern Recognition and Machine Learning" (2006), el cual distingue entre el aprendizaje supervisado y el no supervisado. El aprendizaje supervisado es aquel que se construye a partir de un conjunto de datos que contiene ejemplos de entrada y salida deseados, mientras que el aprendizaje no supervisado implica la construcción de un modelo a partir de un conjunto de datos no etiquetado para descubrir patrones intrínsecos o estructuras subyacentes.

Dentro de la praxis al intentar encontrar una tarea ML para lograr el objetivo del trabajo que es llegar a ese modelo predictivo que ayude a detectar la deserción estudiantil en programas de pregrado tomando como ejemplo la Universidad Pedagógica y Tecnológica de Colombia, nos encontramos con varias tareas que al momento de evaluar su acurracy (desempeño o exactitud) pues no era el apropiado para su implementación. Encontramos en la Support Vector Machines (SVM) una tarea de ML apropiada para el complimiento del objetivo del presente trabajo.

SVM nos da la capacidad de manejar datos complejos [8] Según Cortes y Vapnik en su trabajo seminal "Support-Vector Networks" (1995), SVM puede encontrar hiperplanos óptimos en espacios de alta dimensión para separar datos de diferentes clases, lo que nos indica que una de las fortalezas de SMV es su capacidad de encontrar la mejor separación posible entre diferentes clases de datos. Especialmente útil cuando se trabaja con conjuntos de datos complejos en los que las relaciones entre las características pueden no ser lineales como el que tenemos.

D. Preparación de datos

La preparación de datos es un proceso crucial en el trabajo planteado, [1] según Han, Pei y Kamber (2006), es fundamental la calidad de los datos para tener confiabilidad y precisión en cualquier modelo predictivo o de análisis de datos, este proceso permite corregir errores inconsistencias, identificar valores atípicos y datos faltantes, garantizando que el modelo final se base en datos de alta calidad.

En cuanto a la preparación de datos en primera instancia se decidió verificar los datos faltantes con un total de novecientos treinta y seis (936) datos faltantes repartidos de la siguiente manera, ver [9] punto 3, 3.3.

TABLA VIII
Datos faltantes

VARIABLE	DATOS FALTANTES
'JORNADA'	47
'MODALIDAD'	83
'LUG_EXPEDICION'	315
'LUGAR_NACIMIENTO'	315
'EXTRATO'	176
TOTAL	936

Teniendo en cuenta el valor anterior se decidió eliminar los datos faltantes quedando con un total de 582 registros, un total de 8148 datos y 0 datos faltantes.

Teniendo el conjunto limpio, se procede a verificar si en los datos consolidados hay valores duplicados, teniendo como resultado un total de cero (0) valores duplicados.

Por otra parte, fue necesario transformar la variable Fecha_Nacimiento en la variable EDAD, esto nos facilita mejorar la interpretación de los datos e interpretar patrones o tendencias relacionadas con la edad en el conjunto de datos, esta transformación se realizó obteniendo las edades de los estudiantes a la fecha de la toma de datos.

Para aplicar SVM (Support Vector Machines) fue necesario codificar las variables categóricas en este caso se utilizo la técnica de codificación one-hot, técnica que convierte las variables categóricas a variables que puedan se procesadas de mejor manera por el algoritmo de aprendizaje automático en este caso SVM para un mejor procesamiento.

Para cada categoría, la codificación one-hot crea una nueva columna en el conjunto de datos, esta columna tendrá un valor de 1 si la observación pertenece a esa categoría y un valor de 0 si no.

La implementación de la codificación one-hot represento una manera efectiva de representar las variables categóricas lo que contribuye a mejorar la precisión y el rendimiento general del modelo de aprendizaje automático.

V. MODELADO

A. *Modelo*

Teniendo en cuenta que la tarea de aprendizaje automático (ML) que se seguirá en el modelo es de Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés "Support Vector Machines"), se llevó a cabo un experimento utilizando los datos de deserción académica de la UPTC. Para aplicar SVM en primera instancia, se convirtió el dato de FECHA_NACIMIENTO a EDAD. Dado que los datos fueron obtenidos en el año 2019, se calculó la edad hasta el 31 de diciembre de 2019.

Una vez que el dataset estuvo limpio y transformado, se procedió a preparar los datos. Teniendo en cuenta que el objetivo del modelo es la predicción de deserción académica en estudiantes de la UPTC, se seleccionaron los registros más influyentes al realizar un análisis de deserción. Las categorías elegidas para el entrenamiento son de esencial importancia al generar un análisis de deserción académica, datos como la 'edad', 'sexo', 'estrato', 'jornada', 'modalidad', 'nombre_sede', 'tipo_iden_est', 'nombre_facultad', 'lugar_nacimiento', ofrecen una visión más amplia para comprender quiénes son más propensos a desertar.

Teniendo listo los datos de entrenamiento se experimenta sin ningún tipo de métricas e hiperparametro que mejoren el modelo obteniendo como resultado un acurracy de 0.77 ver [9] punto 5, 5.1, por ende fue necesario ajustar métricas e hiperparametros que nos lleven a un modelo más eficiente, por tal motivo se emplea un kernel lineal, una de las razones por la cual se emplea es porque computacionalmente es más eficientes que los kernels no lineales, lo que conlleva a entrenarse más rápido y ser más escalable para conjuntos de datos más grande, por otra parte Los modelos con kernels lineales tienden a ser menos propensos al sobreajuste en comparación con modelos con kernels no lineales más flexibles. Esto es especialmente relevante cuando se trabaja con conjuntos de datos ruidosos o con una cantidad limitada de datos de entrenamiento como en este caso.

Otro parámetro que se decidió utilizar es el de regularización C controla el equilibrio entre la maximización del margen y la minimización de errores de clasificación en SVM. Un valor más alto de C penaliza más fuertemente los errores de clasificación, lo que lleva a un límite de decisión más ajustado y potencialmente a un sobreajuste en conjuntos de datos más ruidosos. Por tal motivo, en el caso de la predicción de deserción académica, la precisión en la clasificación de los estudiantes en riesgo es crucial.

B. *Evaluación*

Teniendo claro los hiperparametros a utilizar se ejecuta el modelo teniendo en cuenta el Kernel lineal y los valores de regularización C [0.1, 1, 10, 100], experimentando con ellos se logro identificar que el mejor valor de regularización fue el de 10 permitiendo un margen más estrecho para el modelo siendo más efectiva en cuanto a las relaciones entre los atributos y la deserción, sin comprometer significativamente la generalización del modelo.

Por otra parte, se utilizo la validación cruzada, la cual nos arrojo una estimación mas precisa del rendimiento del modelo, al realizar cinco (5) divisiones al conjunto de datos, se logra observar de una manera mas robusta el rendimiento promedio del modelo, así como su variabilidad, teniendo como resultado los siguientes valores [0.80645161 0.75268817 0.84946237 0.78494624 0.84946237] Fig. 3.

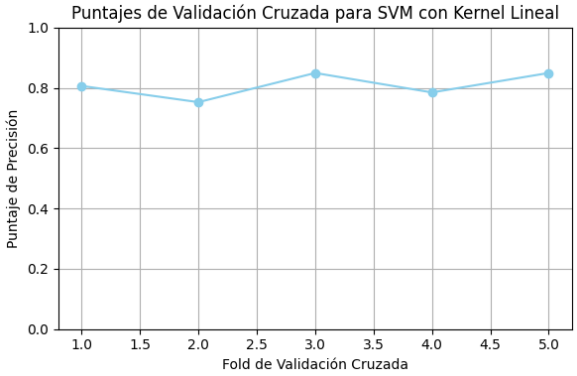

Fig. 3 Puntajes validación cruzada.

C. *Despliegue*

Una vez el modelo fue evaluado, se decide desplegarlo utilizando Flask, por lo cual fue necesario configurar una aplicación Flask y sus rutas para manejar las solicitudes de predicción, Fig. 4.

```python
# Ruta para la página principal
@app.route('/')
def home():
    return render_template('index2.html', categorias_sexo=categorias_sexo,
                           categorias_jornada=categorias_jornada,
                           categorias_modalidad=categorias_modalidad,
                           categorias_sede=categorias_sede,
                           categorias_tipo_iden=categorias_tipo_iden,
                           categorias_nombre_facultad=categorias_nombre_facultad,
                           categorias_lugar_nacimiento=categorias_lugar_nacimiento,
                           accuracy=scores.mean())

# Ruta para la predicción
@app.route('/predict', methods=['POST'])
def predict():
    if request.method == 'POST':
        # Obtener los datos del formulario
        edad = float(request.form['EDAD'])
        sexo = request.form['SEXO']
        estrato = float(request.form['ESTRATO'])
        jornada = request.form['JORNADA']
        modalidad = request.form['MODALIDAD']
        nombre_sede = request.form['NOMBRE_SEDE']
        tipo_iden_est = request.form['TIPO_IDEN_EST']
        nombre_facultad = request.form['NOMBRE_FACULTAD']
        lugar_nacimiento = request.form['LUGAR_NACIMIENTO']

        # Crear un DataFrame con los nuevos datos
        new_data = pd.DataFrame({'EDAD': [edad], 'SEXO': [sexo], 'ESTRATO': [estrato],
                                 'JORNADA': [jornada], 'MODALIDAD': [modalidad],
                                 'NOMBRE_SEDE': [nombre_sede], 'TIPO_IDEN_EST': [tipo_iden_est],
                                 'NOMBRE_FACULTAD': [nombre_facultad],'LUGAR_NACIMIENTO': [lugar_nacimiento]})

        # Aplicar codificación one-hot
        new_data_encoded = encoder.transform(new_data)

        # Realizar la predicción con el clasificador SVM entrenado
        prediction = svm_classifier.predict(new_data_encoded)[0]

        # Renderizar la plantilla de resultados
        return render_template('result2.html', prediction=prediction)
```

Fig. 4 Creación rutas de predicción.

Cuando se recibe una solicitud de predicción, el modelo SVM se utiliza para hacer predicciones sobre los datos proporcionados. Las predicciones resultantes pueden ser devueltas al cliente como una respuesta HTTP, permitiendo así la integración del modelo en aplicaciones web en tiempo real. Fig. 5.

```
udenar@DESKTOP-8MISQQ9 MINGW64 ~/Desktop/MODELOSVM
$ python app.py
C:\Users\Udenar\Desktop\MODELOSVM\app.py:16: UserWarning: Could not infer format
, so each element will be parsed individually, falling back to `dateutil`. To en
sure parsing is consistent and as-expected, please specify a format.
  df['FECHA_NACIMIENTO'] = pd.to_datetime(df['FECHA_NACIMIENTO'])
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment.
 Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
C:\Users\Udenar\Desktop\MODELOSVM\app.py:16: UserWarning: Could not infer format
, so each element will be parsed individually, falling back to `dateutil`. To en
sure parsing is consistent and as-expected, please specify a format.
  df['FECHA_NACIMIENTO'] = pd.to_datetime(df['FECHA_NACIMIENTO'])
 * Debugger is active!
 * Debugger PIN: 640-230-788
```

Fig. 5 Funcionamiento.

En cuanto a la parte visual del despliegue, se desarrollo con el objetivo de reflejar la esencia de la Universidad Pedagógica y Tecnológica de Colombia, tanto por los colores como por la disposición de elemento, cada elemento del despliegue fue añadido con el fin de crear una experiencia coherente y atractiva que refleje la identidad de la institución.

Al ingresar al modelo, los usuarios se encuentran con una interfaz de usuario intuitiva, los campos que deben completarse para realizar la predicción están claramente etiquetados y organizados, lo que facilita el proceso para los usuarios.

Entre los campos que se solicitan para realizar la predicción se encuentran: edad, sexo, estrato, jornada, modalidad, nombre de la sede, tipo de identificación del estudiante nombre de la facultad y lugar de nacimiento. Estos datos proporcionan información clave que se utilizará para generar recomendaciones personalizadas para cada usuario, Fig. 6.

Fig. 6 Apartado inicial del Modelo.

Una vez se diligencia los campos establecidos, el resultado de la predicción es crucial, ya que para cada predicción se genera un texto con las recomendaciones correspondientes según sea el caso. Estas predicciones guiaran las acciones a futuro, desde la detección temprana hasta la implementación de medidas preventivas. Fig. 7.

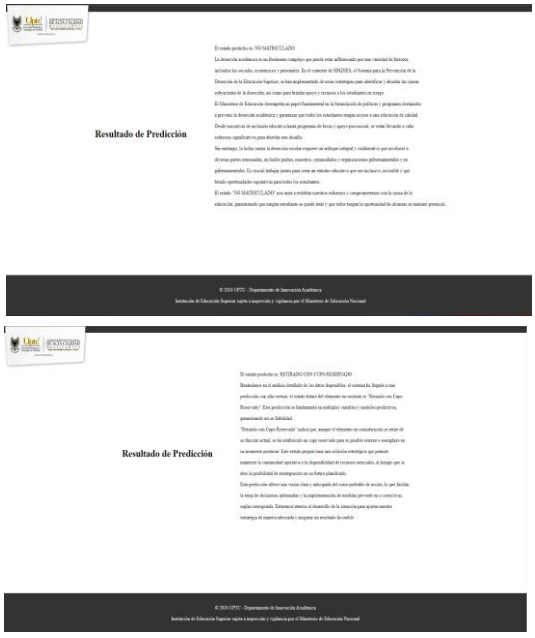

Fig. 7 Apartado predicciones

En este sentido, la precisión y confiabilidad de nuestro modelo no solo son esenciales, sino que también son la clave para impulsar decisiones informadas y efectivas en el ámbito que sea necesario.

CONCLUSIONES

Es importante destacar que esta herramienta proporciona una identificación temprana de la deserción académica, este estudio contribuye a la generación de conocimiento en el campo de la educación y análisis de datos. La aplicación de modelos de aprendizaje automático como las Support Vector Machines (SVM) y la aplicación de la metodología CRISP-DM, arrojan resultados útiles para nuevas líneas de investigación, así como análisis de factores de riesgo que pueden llevar a una comprensión de los procesos que conducen a la deserción académica, permitiendo desarrollo de actividades y políticas de permanecía y titulación exitosa.

Esta herramienta proporciona una perspectiva solida para abordar el problema de la deserción académica, no solo permitiendo identificar a estudiantes con riesgo de desertar, sino también proporcionar información valiosa sobre los factores que contribuyen a este fenómeno. Este trabajo sirve como base para la implementación de políticas educativas y programas de apoyo y permanencia que fortalezcan una titulación exitosa en los estudiantes, contribuyendo en el campo de la educación y análisis de datos.

RECONOCIMIENTO

Queremos expresar nuestro sincero agradecimiento a la Universidad Nariño y a la Universidad CESMAG junto a los respectivos docentes y directivos docentes que nos acompañaron en este proceso académico en el que se vio reflejado un crecimiento personal y profesional. También queremos agradecer a nuestras familias y amigos que han estado acompañándonos a lo largo de esta travesía y que aportaron en nuestro desarrollo como futuros licenciados en informática, muchas gracias.

Referencias

[1] J. Han, J. Pei, y M. Kamber, *Data mining, southeast Asia edition*, 2a ed. Oxford, Inglaterra: Morgan Kaufmann, 2006.

[2] Chapman, P., "*CRISP-DM 1.0: Step-by-step data mining guide*," CRISP-DM Consortium, 2000.

[3] I. H. Witten y E. Frank, *Data mining: Practical machine learning tools and techniques*, second edition, 2a ed. Oxford, Inglaterra: Morgan Kaufmann, 2005.

[4] A. Q. Ruiz, "*Análisis de la deserción estudiantil en la educación superior en Colombia asociada al uso de las tecnologías de la información y la comunicación*", EAFIT Escuela de Economía y Finanzas, Colombia, 2022.

[5] "SPADIES - *Estadísticas de deserción*". SPADIES. Accedido el 18 de marzo de 2024. [En línea]. Disponible: https://www.mineducacion.gov.co/sistemasinfo/spadies/secciones/Estadisticas-de-desercion/

[6] T. Mitchell, *Machine Learning*. Nueva York, NY, Estados Unidos de América: McGraw-Hill Professional, 1997

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1a ed. Nueva York, NY, Estados Unidos de América: Springer, 2006.

[8] C. Cortes y V. Vapnik, "*Support-vector networks*", Mach. Learn., vol. 20, núm. 3, pp. 273–297, 1995.

[9] A. Rivera y D. Portilla, "*Modelo predictivo de la deserción académica en la UPTC durante el semestre de 2019 Aplicación de Support Vector Machines*," [En Línea], 2019. Disponible en: https://colab.research.google.com/drive/1G-jFCC1XHrbw_Wh1cCTCW1LVo3x-cgjf?usp=sharing