Revista CIENCIA Y TECNOLOGÍA
ISSN 1810-6781 Rev. Cienc. Tecnol. 20(4): 59-71, (2024)

Modelo predictivo basado en Naive Bayes a través de Machine Learning Supervised y la deserción estudiantil, en centros de Educación Tecnológicos públicos de la región La Libertad

Predictive model based on Naive Bayes through Supervised Machine Learning and student dropout, in public Technological Education centers in the La Libertad region

Víctor Jaime Polo Romero*

Facultad de Ingeniería, Universidad Nacional del Santa, Av. Universitaria S/N–Ciudad Universitaria, Nuevo Chimbote, Perú.

*Autor correspondiente: vpolo@unitru.edu.pe (V. Polo)

DOI: 10.17268/rev.cyt.2024.04.05

RESUMEN

La presente investigación, expresa un modelo predictivo, para estimar estudiantes con riesgo de abandonar los estudios en los centros de educación superior tecnológicos públicos de la región La Libertad. El modelo, se fundamenta en el algoritmo de clasificación, Naive Bayes en máquinas de aprendizaje supervisado guiado por la metodología CRISP DM. La investigación es aplicada, descriptiva, no experimental y diseño transversal. Los datos se obtienen de fichas socioeconómicas, matriculas y notas históricas, para obtener el Set de datos inicial, luego del procesamiento, se obtiene el set de datos definitivo. En la implementación, se usó Python, a través de júpiter notebook, de Google Colaboratory. Una parte del set de datos definitivo, se usó para entrenar, validar y otra para evaluar la confiabilidad del modelo. Se entrena un objeto del algoritmo, con el set definitivo, y obtiene el modelo predictivo. Generado el modelo, con el set de datos de prueba se realiza una predicción y evalúa la confiabilidad de los resultados. Con los resultados esperados del set de datos de prueba, se verifica un grado de confiabilidad del modelo obtenido del 93%. Para visualizar el número de patrones correctos e incorrectos reconocidos por el modelo, se usó la Matriz de Confusión.

Palabras clave: Algoritmo naive bayes; modelo predictivo; deserción estudiantil; algoritmos supervisados; matriz de confusión; aprendizaje automático.

ABSTRACT

This research expresses a predictive model to estimate students at risk of dropping out of school in public technological higher education centers in the La Libertad region. The model is based on the Naive Bayes classification algorithm in supervised learning machines guided by the CRISP DM methodology. The research is applied, descriptive, non-experimental and cross-sectional in design. The data is obtained from socioeconomic records, enrollments and historical notes to obtain the initial data set, after processing, the final data set is obtained. In the implementation, Python was used through Jupiter Notebook from Google Colaboratory. A part of the final data set was used to train, validate and another to evaluate the reliability of the model. An object of the algorithm is trained with the final set, and the predictive model is obtained. Once the model is generated, a prediction is made with the test data set and the reliability of the results is evaluated. With the expected results of the test data set, a degree of reliability of the obtained model of 93% is verified. To visualize the number of correct and incorrect patterns recognized by the model, the Confusion Matrix was used.

Keywords: Naive Bayes algorithm; predictive model; student dropout; supervised algorithms; confusion matrix; machine learning.

1. INTRODUCCIÓN

El aprendizaje automático, o machine learning, es una sub disciplina de la inteligencia artificial; cada día adquiere mayor importancia al proporcionar a los algoritmos, la capacidad de aprender, a partir de datos históricos recolectados, analizados, y establecer patrones de comportamiento. Estos patrones son utilizados en el desarrollo de muchas aplicaciones a nivel empresarial, en ciencias de la salud, ingeniería y educación entre

Fecha de envío:02-08-2024 Fecha de aceptación: 29-11-2024 Fecha de publicación: 28-12-2024

otros, permitiendo tomar mejores decisiones y enfrentar la incertidumbre a través de la elaboración de modelos predictivos como la presente investigación.

El algoritmo de aprendizaje usado en esta investigación, es el aprendizaje supervisado; en términos de Gironés y otros, lo definen como algoritmos que, a partir de un juego de datos de entrenamiento su objetivo será clasificar correctamente todas las instancias nuevas de acuerdo a un solo objetivo, en este caso la deserción estudiantil el seria la variable de clase (Gironés et al., 2017).

El nivel de deserción estudiantil en el ámbito educativo, es un problema que las instituciones de nivel superior en américa latina vienen atravesando. Investigadores del tema, afirman que el índice anual, oscila alrededor del 57% (Claudio, 2007). Esta problemática ahora es un tema de mucha preocupación para las autoridades de los Institutos Superiores Tecnológicos y otros especialistas. Para la UNESCO, de los ingresantes a estos centros de educación superior, en periodos normales, se gradúan cerca del 43% de los que ingresan en cada programa (Sánchez, 2015). El conocimiento y uso de técnicas minería de datos, permiten identificar patrones y predecir eventos, con un porcentaje considerable de confiabilidad (Timar & Jim, 2015). La presente investigación, busca determinar un modelo inteligente que permita predecir estudiantes que desertarían su centro de formación, dado características de su entorno. El ámbito elegido como referencia para la presente investigación, fue el Instituto Superior Tecnológico Publico Trujillo, de la provincia de Trujillo, departamento La Libertad, sin embargo, la misma problemática se observa en los institutos superiores tecnológicos públicos de las diferentes regiones, pudiendo dicho proyecto extenderse a todo el ámbito nacional. Esta problemática, motivó pensar en la posibilidad de crear un modelo predictivo, para predecir, con cierto grado de confianza, estudiantes que podrían ser candidatos a desertores o en su defecto determinar a los estudiantes que tienen riesgo de no continuar estudios el próximo periodo académico y de ser así poder informar al área de tutoría, para una atención personalizada a través de su departamento; llevando esta situación a formular la presente interrogante: ¿Es posible, a través de un modelo predictivo basado en algoritmos de aprendizaje supervisados, determinar estudiantes que abandonaran sus estudios de los programas que se imparten en los centros de formación tecnológicos públicos de la libertad?.

La problemática de la deserción, no es nueva, ya ha sido estudiada por otros investigadores tales como: Masabanda & Zapata (2019), estudiaron esta problemática estudiantil en la Universidad Técnica de Cotopaxi. A través de la aplicación de técnicas de minería de datos, guiados por la metodología KDD para la confección de su modelo predictivo, llegaron a determinar factores tales como: conducta en el aula, bullying, motivación del docente – estudiante, bajo conocimiento de la asignatura, entre otras, son los factores que tienen mayor influencia en la problemática de la deserción estudiantil, dimensiones que usaron para su set de datos y luego de entrenar algoritmos de aprendizaje como J48, Random Forest y otros, obtuvieron un modelo que luego de ser evaluado, obtuvo 92% de confiablidad.

A su vez, Pérez & Nieto (2020). Diseñaron de un sistema para enfrentar la problemática de la deserción estudiantil en la Universidad Tecnológica del Perú y poder predecir que alumnos abanderarían sus programas de estudios. Usaron el algoritmo denominado máquinas de vectores de soporte, para entrenar su set de datos y obtener su modelo que luego fue implementado. Al evaluar el modelo predictivo obtuvo un grado de certeza superior a 90%, superando el grado base que se trazaron en la hipótesis de su investigación.

La presente investigación, considera como hipótesis, que la creación de un modelo predictivo basado en el algoritmo naive bayes, a través del aprendizaje automatico, permitirá determinar, estudiantes que desertarían el programa que estudian, con un grado de confianza superior al 80%, en los institutos de educación superior tecnológicos públicos de la libertad. La herramienta de programación usada para el proyecto es Python por su versatilidad, amplia variedad de librerías para machine learnig y facilidad de uso (Pérez 2016).

Phyton, ofrece librerías potentes, para trabajar con modelos predictivos, basados métodos bayesianos (Hinojosa, 2016).

El objetivo de la investigación es encontrar un modelo predictivo que permita, determinar si un estudiante abandonara sus estudios profesionales, con un buen porcentaje de confianza, teniendo en cuenta, características personales y de su entorno cercano. En términos de Jordi Gironés, respecto a un modelo, “Podemos entender el modelo como la habilidad de aplicar una técnica a un juego de datos con el fin de predecir una variable objetivo o encontrar un patrón desconocido” (Gironés et al., 2017).

El modelo predictivo, esta basado en el método estadístico naive bayes como algoritmo de clasificación supervisado en machine learning. Para la evaluación usara como métrica, el coeficiente de determinación y para identificar patrones correctos e incorrectos, usara la matriz de confusión.

Los modelos predictivos de deserción estudiantil, buscan determinar la probabilidad de que un estudiante abandone la universidad, teniendo en cuenta las reglas de conducta y el entorno del estudiante (Cuji, Gavilanes, & Sánchez, 2017).

El marco metodológico que guiará el desarrollo de la presente investigación, es la metodología denominada CRISP -DM, muy utilizada en proyectos de minería de datos por su versatilidad (Mamani, 2019).

2. METODOLOGÍA

2.1 Objeto de estudio

El objeto de estudio de la presente investigación es la obtención de un modelo predictivo que, permita estimar con cierto grado de confiabilidad si un estudiante, culminara sus estudios académicos.

El tipo de investigación, por su finalidad, es aplicada, porque propone una solución del mundo real; es decir dado un conjunto de característica de un estudiante, el modelo responde generando conocimiento, indicando si el estudiante continuará o no sus estudios el próximo semestre. Por su naturaleza es no experimental; es decir no manipula variables. Por su profundidad es descriptiva porque caracteriza hechos concretos de la vida real. Por su alcance, es transversal, capturando datos en un solo momento.

2.2 Instrumentos de recolección de datos

Siendo que los institutos superiores tecnológicos públicos de la región, poseen características semejantes entre ellos: periodos académicos, programas de estudios, fechas de admisión, sin costo de estudios, docentes contratados y nombrados por el estado etc. Las problemáticas también son muy semejantes, por esta razón se tomó como muestra, el IESTP Trujillo, de la provincia de Trujillo, región la Libertad. La muestra fue tomada por conveniencia y se limitó a los estudiantes del programa: Computación e Informática.

La técnica empleada para recabar información, fue la revisión de documentos, entre ellos fichas socioeconómicas, notas de registros académicos de 10 periodos académicos anteriores. Dichos datos, luego de un preprocesamiento, permitió confeccionar el primer set de datos inicial con el cual luego de un preprocesamiento se convertirá en set de datos definitivo.

El set de datos que representa la muestra del estudio, está conformado por los valores sobre las variables independientes, denominadas dimensiones: carga familiar, edad del estudiante, promedio ponderado, situación laboral, ingreso familiar, vivienda, servicio de internet, seguro de salud, régimen alimentario, discapacidad, con quien vive y la variable dependiente Desertor, formando el set de datos inicial. Luego de un análisis de correlación basado en Pearson, se pudo reducir a siete dimensiones, siendo estas las más relevantes, porque se pudo constatar que tenían una correlación mayor a 0.5.

2.3 Métodos y técnicas

El desarrollo de la investigación, estuvo guiada por la aplicación de las seis etapas de la metodología denominada CRISP DM, las cuales se describen a continuación:

Etapa1: Comprensión del negocio, permite tener un conocimiento de la situación problemática, para el caso de estudio se revisaron datos, basados en las fichas de matrículas de alumnos y los reportes de matrículas, notas de los periodos académicos y fichas socio económicas del programa del programa de estudios en mención, considerando una totalidad de 10 periodos y una data aproximada de 500 registros.

Esta etapa, la metodología, consiste en: evaluar la situación actual, establecer los objetivos del negocio, y establecer los objetivos a nivel de minería de datos.

Etapa 2: Comprensión de los datos, aquí, con los datos de campo revisados y recopilados en diferentes formatos tal cual se encontraron; datos de matrículas de alumnos y los reportes de matrículas, notas de los periodos académicos y fichas socio económicas del programa de estudios, considerando una totalidad de 10 periodos y una data aproximada de 500 registros. Se extraen las características más relevantes, según enfoques teóricos y se construye el Set de Datos primario en formato XLMS, con los registros históricos encontrados.

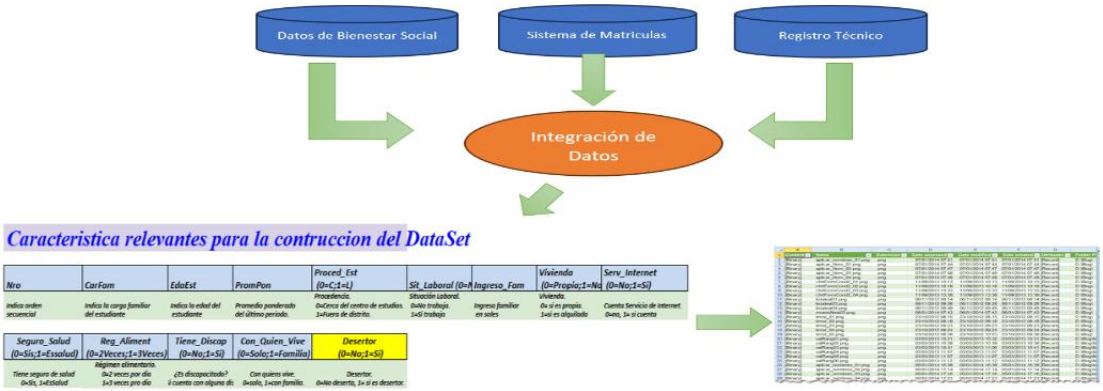

Figura 1: Comprensión de los datos

Etapa 3: Preparación de datos, en esta etapa el set de datos inicial, formulado en la etapa anterior, pasa por un procesamiento de limpieza de datos, para evitar datos perdidos o nulos en el set de datos, se discretiza el set de datos, para dar mayor rendimiento al algoritmo que será entrenado, se trata el problema de la dimensionalidad basados en el análisis de correlación de Karl Pearson.

Esta etapa concluye con la obtención del set de datos definitivo.
Etapa 4: Modelado, en esta etapa, se construye el modelo propuesto, creando una instancia del algoritmo naive bayes, el cual es entrenado con el 80% del set de datos definitivo y el 20% restante del set de datos definitivo, se reserva para la siguiente etapa de modelo, donde será evaluado su confiabilidad, a través de la generación de una predicción donde el vector resultante, se compara con el vector de prueba del set de prueba y estima su confiabilidad.

Etapa 5: Evaluación, en esta etapa, la metodología CRISP DM, evalúa el nivel de confiabilidad de cada modelo en el bloque entrenamiento para su validación y en el bloque de pruebas para la confiablidad del modelo propuesto. En esta etapa también se detalla la cantidad de estudiantes desertores y no desertores que el modelo reconoció correctamente y en cuantos se equivocó a través de la métrica denominada Matriz de Confusión.

Etapa 6: Implementación, en esta etapa el modelo implementado y probado es llevado a producción, construyéndose el sistema de información para la toma de decisiones tanto en escritorio, web o en aplicaciones móviles. La implementación de estas soluciones no es el propósito del presente trabajo de investigación.

3. RESULTADOS Y DISCUSIÓN

En esta parte, se muestra las evidencias encontradas en la investigación, que permiten establecer el cumplimiento de cada objetivo propuesto y general. A su vez comprobar la hipótesis planteada, al evaluar la confiabilidad del modelo, mediante el uso de la estadística descriptiva e inferencial.

Objetivo específico 1: **Obtener el juego de datos inicial de la situación problemática**. Durante el desarrollo de este objetivo, se procedió a aplicar las etapas 1 y 2 de la metodología CRISP DM.
La metodología CRISP-DM es la más utilizada para proyectos de minería de datos (Espinosa, 2020).
En primer lugar, comprender el negocio, evaluando la situación actual acerca de la deserción estudiantil en el Iestp de referencia, donde se estima que el promedio de estudiantes que concluyen sus estudios oscila entre 40 y 45 % del total de ingresantes y de allí los que llegan a titularse, están entre el 25 y 30 % de los que terminaron sus estudios, situación problemática con semejantes proporciones en los institutos tecnológicos de la región. Luego, se identifican los objetivos del negocio, tales como: la permanencia de los estudiantes en los diferentes programas académicos durante los 3 periodos de duración, brindar una asesoría personalizada a los estudiantes con riesgo de deserción y cumplir con los estándares de calidad estipulados por el SINEASE para la acreditación de los programas académicos que se imparte. En esta etapa también se establece los objetivos a nivel de minería de datos tales como: establecer un modelo que permita predecir estudiantes que abandonarían los estudios durante los periodos académicos que consta sus programas y encontrar patrones de características de estudiantes con probabilidades de desertar de sus programas de estudios.

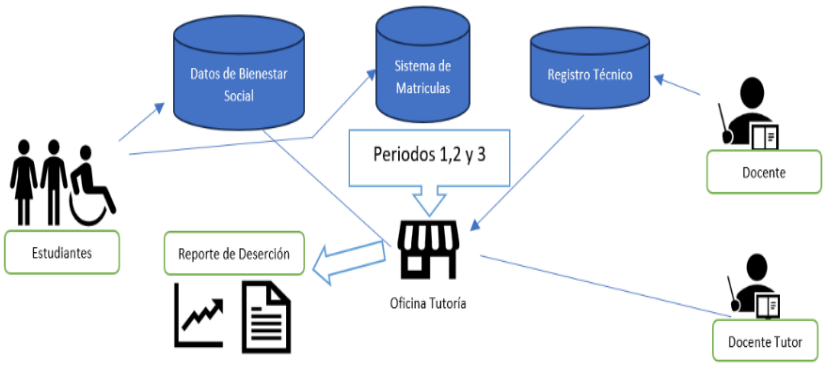

Figura 2: *Comprensión del negocio*

En segundo lugar, comprender los datos, considerando los datos recopilados, basados en las fichas de matrículas de estudiantes, reportes de matrículas, notas de los periodos académicos anteriores y fichas socio económicas del programa computación e Informática del Iestp Trujillo desde el año 2010 y considerando una totalidad de 10 periodos y una data aproximada de 500 registros. Se extraen las características más relevantes, según enfoques teóricos y se construye el Set de Datos primario en formato XLSX con los datos de registros históricos encontrados como se indican en la siguiente tabla.

Tabla 1: Característica identificadas inicialmente para el dataset:

Nro	CarFam	EdaEst	PromPon	Proced_Est (0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda (0=Propia;1=No)	Serv_Internet (0=No;1=Si)
Indica orden secuencial	Indica la carga familiar del estudiante	Indica la edad del estudiante	Promedio ponderado del último periodo.	Procedencia. 0=Cerca del centro de estudios. 1=Fuera de distrito.	Situación Laboral. 0=No trabaja. 1=Si trabaja	Ingreso familiar en soles	Vivienda. 0= si es propia. 1=si es alquilada	Cuenta Servicio de Internet. 0=no, 1= si cuenta

Seguro_Salud (0=Sis;1=Essalud)	Reg_Aliment (0=2Veces;1=3Veces)	Tiene_Discap (0=No;1=Si)	Con_Quien_Vive (0=Solo;1=Familia)	Desertor (0=No;1=Si)
Tiene seguro de salud 0=Sis, 1=EsSalud	Régimen alimentario. 0=2 veces por día 1=3 veces pro día	¿Es discapacitado? 0= no, 1=Si cuenta con alguna discapacidad	Con quiens vive. 0=solo, 1=con familia.	Desertor. 0=No deserta, 1= si es desertor.

Objetivo específico 2*: Preparar el juego de datos para el entrenamiento del algoritmo naive bayes y prueba del modelo*. Para el cumplimiento de este objetivo, el set de datos inicial pasa por un procesamiento de limpieza de datos, discretización, normalización y atender la dimensionalidad.

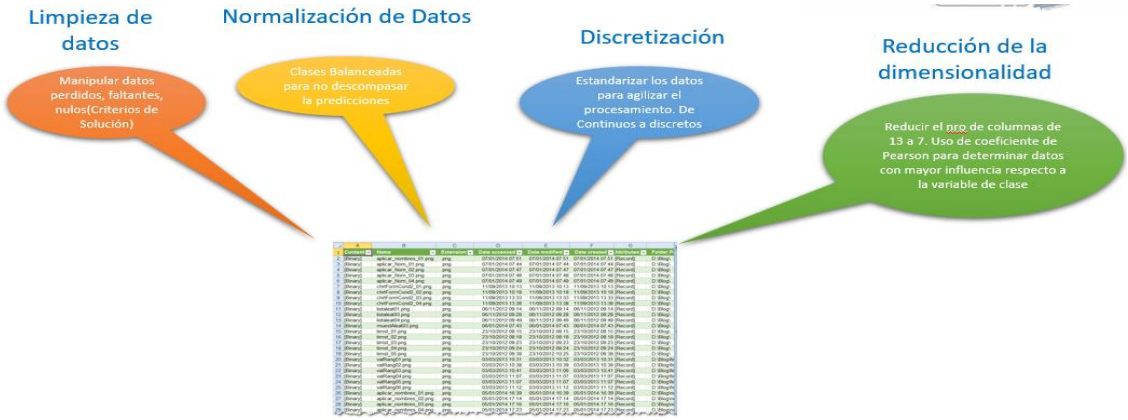

Figura 3: *Preparación de datos*

El set de datos inicial, se obtuvo de la recopilación de datos que se encontraban en varios formatos, tales como fichas físicas y hojas de cálculo, estableciendo un único formato en Excel con 12 características de mayor impacto en la deserción, como sostienen autores citados anteriormente.

Posteriormente se importó la data al cuaderno de Jupiter Notebook, donde se realizó el trabajo de la limpieza de datos, identificando columnas sin datos, valores nulos o algún dato perdido. Este proceso no fue muy complicado, porque las columnas en su mayoría, tenían datos completos y algunos pocos no correspondían con el tipo de dato de la columna, situación que se pudo solucionar en algunos casos asignando el promedio de los datos restantes para completar el dato erróneo. Depurada la data, otro aspecto a tomar en cuenta en la preparación de datos, fue el problema de la dimensionalidad, muy común en proyectos de esta naturaleza; es decir dimensiones que muchas veces tienen poco o nada que ver con el problema; es decir poca significancia. En ese caso se tenía 12 dimensiones y había que establecer cuáles de ellas tienen mayor relación con la variable predictora y obtener un set de datos solo con las dimensiones que son relevantes para el modelo. Esto se logró con el análisis de correlación de Karl Pearson, el cual establece que los coeficientes de correlación son la expresión numérica que indica el grado de relación existente entre 2 variables.

Sus valores varían entre los límites -1 y +1.

Si r = 0, no existe relación entre las variables.

Si r= 1, existe correlación perfecta.

Si r=-1, existe una correlación perfecta negativa.

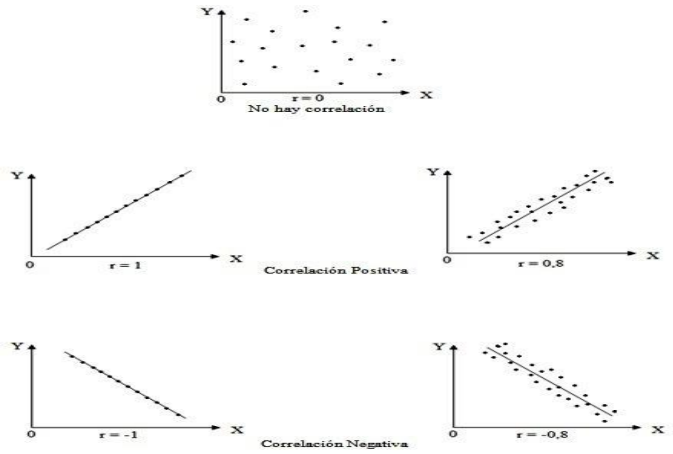

Figura 4: *Correlación de Karl Pearson*
Fuente: Nel, 2017.

Tabla 2: Tabla de correlación Pearson.

Valor	Significado de Correlación
-1	Negativa grande y perfecta
-0,9 a -0,99	Negativa muy alta
-0,7 a -0,89	Negativa alta
-0,4 a -0,69	Negativa moderada
-0,2 a -0,39	Negativa baja
-0,01 a -0,19	Negativa muy baja
0	Correlación nula
0,01 a 0,19	Positiva muy baja
0,2 a 0,39	Positiva baja
0,4 a 0,69	Positiva moderada
0,7 a 0,89	Positiva alta
0,9 a 0,99	Positiva muy alta
1	Positiva grande y perfecta

En la presente investigación, se buscó dimensiones que tengan como mínimo una correlación moderada, estableciendo el índice de medición a 0.5, dado que este indicador va de 0.4 a 0.69.

A continuación, el procedimiento realizado.

a.- Discretización de datos. – Estandarizar los datos, de continuos a discretos. Las variables categóricas se transforman a ceros (No) y Unos (Si), obteniendo el siguiente Set de Datos de la figura adjunta.

b.- Normalización de datos. Se busca que la variable de clase desertor, esté balanceada para garantizar la robustes del modelo. Para este caso, se verifica y transforma algunos registros con el propósito que las clases no estén sesgadas para ningún extremo, esta situación podría llevar a producir un modelo que no refleje la realidad. A continuación, el data set discretizado y normalizado y discretizado.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Nro	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Far	Vivienda(0=Propia;1=No)	Serv_Internet(0=No;1=Si)	Seguro_Salud (0=Sis;1=Essalud)	Reg_Aliment(0=2Veces;1=3Veces)	Tiene_Discap(0=No;1=Si)	Con_Quien_Vive (0=Solo;1=Familia)	Desertor(0=No;1=Si)
2	1	0	22	18	0	0	1800	0	1	1	1	0	1	0
3	2	2	28	12	1	1	1025	1	0	0	2	0	1	1
4	3	1	21	15	0	0	1300	0	1	1	1	0	1	0
5	4	3	31	13	1	1	1500	1	1	1	1	0	1	1
6	5	1	20	15	0	0	1800	0	1	1	1	0	0	0
7	6	0	18	14	0	0	1560	0	1	0	1	0	0	0
8	7	0	23	15	0	0	1800	0	1	1	1	0	0	0
9	8	3	42	12	1	1	1350	0	0	1	1	0	1	1
10	9	0	17	18	0	0	1300	0	0	0	0	0	0	0
11	10	1	19	12	0	1	1280	1	1	0	1	0	1	1
12	11	0	22	15	0	0	1650	0	1	1	1	0	0	0
13	12	3	28	13	1	1	1025	1	0	0	0	0	1	1
14	13	0	24	15	1	0	2000	0	1	1	1	0	1	0
15	14	2	25	12	0	1	1100	1	0	0	0	0	0	1
16	15	0	18	16	0	0	1600	1	1	1	1	0	0	0
17	16	0	17	18	0	0	1500	0	1	1	1	0	0	0
18	17	2	29	12	1	1	1025	1	1	0	0	1	1	1
19	18	0	20	16	0	0	1750	0	1	1	1	0	0	0
20	19	0	22	15	0	0	2500	0	1	1	1	0	1	0
21	20	2	48	13	0	1	1350	0	0	1	1	0	1	1
22	21	0	17	18	0	0	1800	1	1	1	1	0	0	0
23	22	0	23	16	0	0	1350	0	0	0	1	0	0	0
24	23	1	19	17	0	0	1600	0	1	1	1	0	0	0
25	24	0	18	18	0	0	1400	0	1	1	1	0	0	0
26	25	1	18	16	0	0	1850	0	1	1	1	0	0	0
27	26	0	22	14	0	0	1850	0	1	1	1	0	0	0
28	27	3	34	13	1	1	1450	0	0	1	1	0	1	1
29	28	0	20	15	0	0	1600	0	1	1	1	0	0	0
30	29	0	18	17	0	0	1900	0	1	1	1	0	0	0
31	30	0	17	18	0	0	1750	0	1	1	1	0	0	0
32	31	0	16	17	0	0	1500	0	1	1	1	0	0	0
33	32	0	19	15	0	0	1600	0	1	1	1	0	0	0
34	33	2	36	12	1	1	1025	1	0	0	0	0	1	1
35	34	0	18	17	0	0	1800	0	0	1	1	0	0	0
36	35	1	23	15	1	0	2000	0	0	1	1	0	1	0
37	36	0	22	17	0	1	1800	1	0	1	1	0	0	0
38	37	3	28	13	0	1	1300	1	0	1	0	0	1	1

Figura 5: *Data Set discretizada y normalizada*

c.- Limpieza de datos. – Manipular datos perdidos, faltantes y nulos.

1.- Importando los datos del archivo DataSet_Tesis24.xlsx

```
[1] import pandas as pd
    URL='/content/drive/MyDrive/Colab Notebooks/Caso1Tesis/DataSet_Tesis24.xlsx'
    df = pd.read_excel(URL)
    df
```

	Nro	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)	Serv_Internet(0=No;1=Si)	Seguro_Salud(0=Sis;1=Essalud)	Reg_Aliment(0=2Veces;1=3Veces)	Tiene_Discap(0=No;1=Si)	Con_Quien_Vive(0=Solo;1=Familia)	Desertor(0=No;1=Si)
0	1	0	22	18	0	0	1800	0	1	1	1	0	1	0
1	2	2	28	12	1	1	1025	1	0	0	2	0	1	1
2	3	1	21	15	0	0	1300	0	1	1	1	0	1	0
3	4	3	31	13	1	1	1500	1	1	1	1	0	1	1
4	5	1	20	15	0	0	1800	0	1	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	496	0	25	14	0	0	2000	0	0	0	1	0	1	1
496	497	0	20	17	0	0	1950	0	0	0	1	0	1	0
497	498	3	27	13	1	1	1200	1	0	0	0	0	1	1
498	499	0	18	17	0	0	1550	0	0	0	1	0	0	0
499	500	0	19	18	0	0	1600	0	0	0	1	0	0	0

500 rows × 14 columns

Figura 6: *Set de datos importado de una tabla de Excel.*

2.- Verificando que el set de datos, no tenga valores nulos, vacíos o perdidos.

```
#Columnas sin datos
df.isna().sum()
```

```
Nro                                  0
CarFam                               0
EdaEst                               0
PromPon                              0
Proced_Est(0=C;1=L)                  0
Sit_Laboral (0=No;1=Si)              0
Ingreso_Fam                          0
Vivienda(0=Propia;1=No)              0
Serv_Internet(0=No;1=Si)             0
Seguro_Salud(0=Sis;1=Essalud)        0
Reg_Aliment(0=2Veces;1=3Veces)       0
Tiene_Discap(0=No;1=Si)              0
Con_Quien_Vive(0=Solo;1=Familia)     0
Desertor(0=No;1=Si)                  0
dtype: int64
```

Figura 7: *Muestra datos completos*

3.- Eliminando el campo Nro., que no contribuye al modelo, debido a que es un valor de orden. Verificando.

```
# Remover Columna "Nro" que no contribuye al modelo

df = df.drop('Nro', axis=1)
df.columns
```

```
# Verificando las dimensiones y visualizando el nuevo df
rows=df.shape[0]
columns=df.shape[1]
print ( rows,columns )
df
```

500 13

	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)	Serv_Internet(0=No;1=Si)	Seguro_Salud(0=Sis;1=Essalud)	Reg_Aliment(0=2Veces;1=3Veces)	Tiene_Discap(0=No;1=Si)	Con_Quien_Vive(0=Solo;1=Familia)	Desertor(0=No;1=Si)
0	0	22	18	0	0	1800	0	1	1	1	0	1	0
1	2	28	12	1	1	1025	1	0	0	2	0	1	1
2	1	21	15	0	0	1300	0	1	1	1	0	1	0
3	3	31	13	1	1	1500	1	1	1	1	0	1	1
4	1	20	15	0	0	1800	0	1	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	0	25	14	0	0	2000	0	0	0	1	0	1	1
496	0	20	17	0	0	1950	0	0	0	1	0	1	0
497	3	27	13	1	1	1200	1	0	0	0	0	1	1
498	0	18	17	0	0	1550	0	0	0	1	0	0	0
499	0	19	18	0	0	1600	0	0	0	1	0	0	0

500 rows × 13 columns

Figura 8: *Muestra y verifica eliminación del campo Nro.*

d.- Reduciendo la dimensionalidad. – Reducir el número de columnas de 13 a 8.

Uso de coeficiente de Pearson para determinar datos con mayor influencia respecto a la variable de clase.

```
[ ] # Seleccionar las columnas con mayor correlacion  respecto a la Desercion.
    corr[abs(corr['Desertor(0=No;1=Si)']) > 0.5]
```

	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)	Desertor(0=No;1=Si)
CarFam	1.000000	0.712805	-0.749520	0.733252	0.821657	-0.642793	0.590623	0.780689
EdaEst	0.712805	1.000000	-0.700666	0.578953	0.653693	-0.405781	0.270281	0.627928
PromPon	-0.749520	-0.700666	1.000000	-0.607012	-0.740259	0.481199	-0.512677	-0.683085
Proced_Est(0=C;1=L)	0.733252	0.578953	-0.607012	1.000000	0.600533	-0.431427	0.455690	0.575684
Sit_Laboral (0=No;1=Si)	0.821657	0.653693	-0.740259	0.600533	1.000000	-0.599748	0.706441	0.772507
Ingreso_Fam	-0.642793	-0.405781	0.481199	-0.431427	-0.599748	1.000000	-0.565717	-0.586454
Vivienda(0=Propia;1=No)	0.590623	0.270281	-0.512677	0.455690	0.706441	-0.565717	1.000000	0.527800
Desertor(0=No;1=Si)	0.780689	0.627928	-0.683085	0.575684	0.772507	-0.586454	0.527800	1.000000

Figura 9: *Reducción de dimensiones de 13 a 8.*

Como se puede apreciar en el gráfico, en la columna izquierda tenemos 7 dimensiones de la variable X y una para la variable Y. Aquellas dimensiones son las que tienen correlación respecto a la variable Y mayor a 0.5, ligeramente moderada según la tabla de Pearson.
Por lo tanto, la dimensionalidad se redujo de 13 a 8 en nuestro Set de Datos.

```
# El Set de Datos con el que se trabajara
df1
```

	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)	Desertor(0=No;1=Si)
0	0	22	18	0	0	1800	0	0
1	2	28	12	1	1	1025	1	1
2	1	21	15	0	0	1300	0	0
3	3	31	13	1	1	1500	1	1
4	1	20	15	0	0	1800	0	0
...	...	...	...	...	...	...	...	...
495	0	25	14	0	0	2000	0	1
496	0	20	17	0	0	1950	0	0
497	3	27	13	1	1	1200	1	1
498	0	18	17	0	0	1550	0	0
499	0	19	18	0	0	1600	0	0

500 rows × 8 columns

Figura 10: *Set de Datos definitivo*

Objetivo específico 3: Crear el modelo predictivo propuesto. En relación a la problemática del caso de estudio, en esta etapa se realizó la implementación del modelo basado en el algoritmo, Naive Bayes en máquinas de aprendizaje supervisadas usando python, pandas, numpy y scikit-learn.
A través de la librería scikit-learn, procedemos a importar el algoritmo de clasificación que usaremos para crear una instancia, la cual será entrenada con el set de datos definitivo.
3.1.- Se importa los algoritmos de aprendizaje, métricas y libreras graficas para la implementación de los modelos.

```
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.naive_bayes import GaussianNB
```

Figura 11a. Importación del algoritmo naive bayes, métricas y libreras gráficas para el modelo.

3.2.- Se verifica el DataSet que participará en el entrenamiento y prueba

```
# El Set de Datos con el que se trabajara
m
```

	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)	Desertor(0=No;1=Si)
0	0	22	18	0	0	1800	0	0
1	2	28	12	1	1	1025	1	1
2	1	21	15	0	0	1300	0	0
3	3	31	13	1	1	1500	1	1
4	1	20	15	0	0	1800	0	0
...	...	...	...	...	...	...	...	...
495	0	25	14	0	0	2000	0	1
496	0	20	17	0	0	1950	0	0
497	3	27	13	1	1	1200	1	1
498	0	18	17	0	0	1550	0	0
499	0	19	18	0	0	1600	0	0

500 rows × 8 columns

Figura 11b. Importación del algoritmo naive bayes, métricas y libreras gráficas para el modelo.

3.3.- Se separa el data set en las variables X y la variable Y.

```
X = df[caracteristicas]

y = df["Desertor(0=No;1=Si)"]

X
```

	CarFam	EdaEst	PromPon	Proced_Est(0=C;1=L)	Sit_Laboral (0=No;1=Si)	Ingreso_Fam	Vivienda(0=Propia;1=No)
0	0	22	18	0	0	1800	0
1	2	28	12	1	1	1025	1
2	1	21	15	0	0	1300	0
3	3	31	13	1	1	1500	1
4	1	20	15	0	0	1800	0
...	...	...	...	...	...	...	...
495	0	25	14	0	0	2000	0
496	0	20	17	0	0	1950	0
497	3	27	13	1	1	1200	1
498	0	18	17	0	0	1550	0
499	0	19	18	0	0	1600	0

500 rows × 7 columns

Figura 12: *DataSet* seleccionado

```
#Visualizando y
y
```

```
0      0
1      1
2      0
3      1
4      0
      ..
495    1
496    0
497    1
498    0
499    0
Name: Desertor(0=No;1=Si), Length: 500, dtype: int64
```

Figura 13: *Separación del Set de Datos*

3.4.- Luego se procede a separar la data en entrenamiento (80%) y prueba (20%)

```
## Separando nuestra data en entrenamiento y prueba

X_train,X_test,y_train, y_test = train_test_split( x,y,test_size=.20,random_state=42323232)
```

Figura 14*: Separación del Data Set en entrenamiento y prueba (filas)*

3.5.- Se genera una instancia del algoritmo para ser entrenado con el Set de Datos.

```
## Instanciando el modelo
Algoritmo1 = GaussianNB()
```

Figura 15*: Define una instancia del algoritmo naive bayes.*

3.6.- Se entrena el objeto instanciado con el Set de Datos de entrenamiento para obtener el modelo propuesto.

```
#Entrenando el Algoritmo
Modelo=Algoritmo1.fit(X_train, y_train)
```

Figura 16*: Modelo Generado*

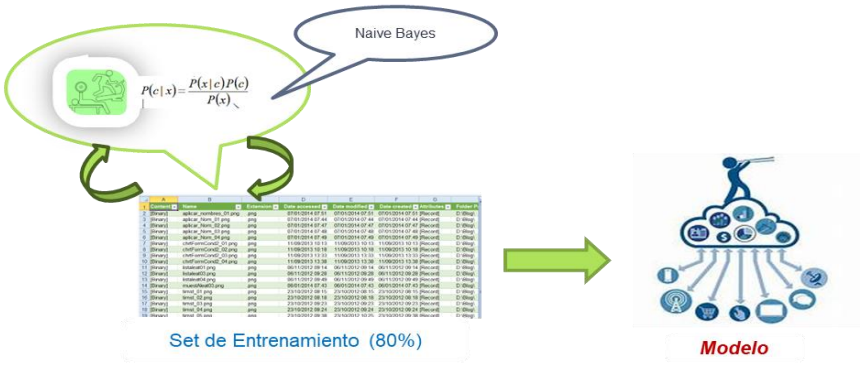

Figura 17: *Modelo Generado en forma grafica.*

Objetivo específico 4: Determinar la confiabilidad del modelo obtenido. Para el presente objetivo, procedemos a realizar la evaluación del modelo determinado su validación y confiabilidad.
En esta etapa también detallamos la cantidad de estudiantes desertores y no desertores que el modelo reconoció correctamente y en cuantos se equivocó a través de la métrica denominada Matriz de Confusión.

```
# Validando el Modelo1 con una prediccion de validacion
y_predV = Modelo.predict(X_train)
print("El Coeficiente de Validacion del Modelo basado en ", Algoritmo1, "es : ",accuracy_score(y_train, y_predV)*100 ,"%")
# Evaluando la Confiabilidad del Modelo con una prediccion de evaluacion
y_predC = Modelo1.predict(X_test)
print("El Coeficiente de Confiabilidad del Modelo basado en ", Algoritmo1, " es : ",accuracy_score(y_test, y_predC)*100 ,"%")

print("Datos del Vector Prueba \n",y_test)
print("Datos del Vector Prediccion\n",y_predC)
# Matriz en modo texto
matriz = confusion_matrix(y_test,y_predC)
plot_confusion_matrix(conf_mat=matriz, figsize=(4,4), show_normed=False)
plt.tight_layout()
```

Figura 18: *Validación y confiabilidad del Modelo.*

El coeficiente de validación del Modelo basado en GaussianNB() es : 93.5 %
El coeficiente de confiabilidad del Modelo basado en GaussianNB() es : 93.0 %
Se puede ver que en el bloque de datos de entrenamiento nos ofrece un 93.5%, esto representa la validación o robustez del modelo y en el bloque de datos de prueba 93 % de confiabilidad.
En la métrica denominado matriz de confusión, analizaremos la cantidad de elementos que reconoce correctamente; es decir estudiantes que no desertan (0) y estudiantes que si desertor (1).

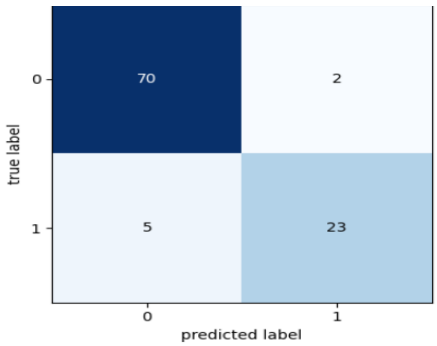

Figura 19*: Reporte de resultados.*

Análisis de resultados de la Matriz de confusión:
De 72 estudiantes que no desertan en la prueba de estudio, el modelo, reconoce a 70 y se equivoca en 2 estudiantes. Sostiene que 2 si desertan. De 28 estudiantes que, si desertan en la prueba de estudio, el modelo, reconoce a 23 y no reconoce a 5. Sostiene que 5 no desertan.

Error en la predicción=7 estudiantes.

Puede verse que, en el bloque de datos de entrenamiento un valor de 93.5%, esto representa la validación o robustez del modelo y en el bloque de datos de prueba 93 %, esto representa la confiabilidad del modelo.
En la métrica matriz de confusión, se representan a los estudiantes que no desertan (0) y estudiantes que si desertan (1).
Respecto al objetivo general, podemos decir que se implementó un modelo basado en naive bayes a través de máquinas de aprendizaje asociado a la deserción estudiantil en los institutos de educación superior tecnológicos públicos de la libertad, con muy buenos resultados.

Con respecto a la contrastación de la hipótesis tenemos:
Los resultados de la evaluación del modelo obtenido, superan ampliamente el supuesto de que un modelo predictivo basado en naive bayes a través de máquinas de aprendizaje, determinaría si un estudiante desertaría sus estudios profesionales con más de 80% de confiabilidad. Como pudo demostrarse, el modelo predijo con 93% de confianza y con una tasa de errores en el reconocimiento de patrones de 7/100 (7%) de modo general entre desertores y no desertores, superando ampliamente los esperado en la hipótesis que fue de 80%.
Con Masabanda & Zapata (2019), quien también utilizo algoritmos de aprendizaje supervisados para la misma problemática, sus resultados fueron del 92% de confiabilidad, mientras que en el modelo de la presente investigación basada en naive bayes, se obtuvo 93% de confiabilidad. Los resultados fueron muy cercanos.
Con Pérez & Nieto (2020), quienes también investigaron la problemática de la deserción estudiantil, diseñaron un sistema para predecir la deserción de estudiantes, usando el clasificador de Soporte Vectorial, y obtuvieron un modelo predictivo con buenos resultados. Sin embargo, en el presente trabajo de investigación, implementamos un modelo basado en naive bayes y la confiabilidad mejoro un poco más que el estudiado por Pérez y Nieto.

4. CONCLUSIONES

Comprendida la situación problemática, se logró establecer el juego de datos inicial, basado en 12 características, que se recopilaron de las fichas socioeconómicas.
Se logró procesar el juego de datos inicial, eliminando características con valores nulos, normalizado, discretizando y redimensionando para finalmente obtener el juego de datos definitivo.
Se logró obtener el modelo predictivo, a partir del entrenamiento de una instancia del algoritmo naive bayes, sobre el set de datos definitivo logrado.
Se midió el grado de confiabilidad del modelo predictivo, obteniendo un 93% de confiabilidad.

REFERENCIAS BIBLIOGRÁFICAS

Claudio, R. (2007). *Informe sobre la educación superior en América Latina y el Caribe 2000-2005*. https://www.ses.unam.mx/curso2013/pdf/informe_educacion_superiorAL2007.pdf

Cuji, B.; Gavilanes, W.; & Sánchez, R. (2017). *Modelo predictivo de deserción estudiantil basado en árboles de decisión*. Revista Espacios, 38(55). https://www.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

Espinosa, J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, Investigación y Tecnología*, 21(1), 1-10. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432020000100008

Gironés, J.; Casas, J.; Minguillón, J. & Caihuelas, R. (2017). *Minería de datos: Modelos y algoritmos*. Universidad Oberta de Catalunya. https://books.google.com.pe/books/about/Miner%C3%ADa_de_datos.html?id=sOn-swEACAAJ&redir_esc=y

Hinojosa, Á. (2016). *Python paso a paso*. RA-MA Editorial. https://www.ra-ma.es/libro/python-paso-a-paso_47942/

Mamani, D. (2019). *Modelo de minería de datos basado en factores asociados para la predicción de deserción estudiantil universitaria* [Tesis de licenciatura, Universidad Nacional Mayor de San Marcos]. Repositorio UNMSM. https://repositorio.unam.edu.pe/bitstreams/9706dd46-07a2-4ba4-8491-527dbefdb3e1/download

Masabanda, J. & Zapata, C. (2019). *Modelo basado en minería de datos para determinar factores de deserción estudiantil en la Facultad de Ciencias de la Ingeniería y Aplicadas de la Universidad Técnica de Cotopaxi* [Tesis de maestría, Universidad Técnica de Cotopaxi]. Repositorio UTC. https://repositorio.utc.edu.ec/bitstreams/000d2451-4aa5-4340-87f3-9228f3d9ada3/download

Pérez, A. (2016). *Python fácil*. Marcombo. https://www.alpha-editorial.com/E-book/9786076227800/Python+F%C3%A1cil

Pérez, J. & Nieto, J. (2020). *Reflexiones metodológicas de investigación educativa: Perspectivas sociales*. Universidad Nacional Abierta y a Distancia (UNAD). http://hdl.handle.net/11634/31292

Sánchez, D. (2015). *La tendencia del abandono escolar en Ecuador: Período 1994-2014*. http://hdl.handle.net/2078/255557

Timar, R. & Jim, J. (2015). Extracción de perfiles de deserción estudiantil en la institución universitaria CESMAG. *Investigium IRE*, 3(2), 5-15. https://investigiumire.unicesmag.edu.co/index.php/ire/article/view/69/77