Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Clasificación de la insatisfacción de estudiantes utilizando el algoritmo SMOTE en una regresión logística

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Natali Salome RUBINA CAMARGO

ASESOR

Mg. Carlos Alberto JAIMES VELASQUEZ

Lima, Perú

2021

Referencia bibliográfica

Rubina, N. (2021). *Clasificación de la insatisfacción de estudiantes utilizando el algoritmo SMOTE en una regresión logística*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Natali Salome Rubina Camargo
Tipo de documento de identidad	DNI
Número de documento de identidad	09951346
URL de ORCID	No aplica
Datos de asesor	
Nombres y apellidos	Mg. Carlos Alberto Jaimes Velasquez
Tipo de documento de identidad	DNI
Número de documento de identidad	42762905
URL de ORCID	https://orcid.org/0000-0002-8794-0972
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Zoraida Judith Huamán Gutierrez
Tipo de documento	DNI
Número de documento de identidad	09890094
Miembro del jurado 1	
Nombres y apellidos	Ricardo Luis Pomalaya Verastegui
Tipo de documento	DNI
Número de documento de identidad	10460674
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad (Empresa, Instituciones, Poblaciones locales, regionales y nacionales)

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.0560257 Longitud: -77.0844226
Año o rango de años en que se realizó la investigación	Agosto 2021 - noviembre 2021
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL EN LA MODALIDAD VIRTUAL PARA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA

En Lima, siendo las 18:30 horas del domingo 03 de octubre del 2021, se reunieron los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional: Mg. Zoraida Judith Huamán Gutierrez (PRESIDENTA), Mg. Ricardo Luis Pomalaya Verastegui (MIEMBRO) y el Mg. Carlos Alberto Jaimes Velasquez (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**CLASIFICACIÓN DE LA INSATISFACCIÓN DE ESTUDIANTES UTILIZANDO EL ALGORITMO SMOTE EN UNA REGRESIÓN LOGÍSTICA**", presentado por la señorita **Bachiller Natali Salome Rubina Camargo**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del trabajo de suficiencia, la Presidenta invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los miembros del Jurado Evaluador, la expositora mereció la aprobación de **SOBRESALIENTE,** con un calificativo promedio de **DIECISIETE (17).**

A continuación, los miembros del Jurado dan manifiesto que la participante **Bachiller Natali Salome Rubina Camargo** en vista de haber aprobado la sustentación del Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 19:00 horas se levantó la sesión firmando para constancia la presente Acta.

Mg. Zoraida Judith Huamán Gutierrez
PRESIDENTA

Mg. Ricardo Luis Pomalaya Verastegui
MIEMBRO

Mg. Carlos Alberto Jaimes Velasquez
MIEMBRO ASESOR

La Vicedecana de la Facultad de Ciencias Matemáticas, Mg. Zoraida Judith Huamán Gutiérrez, certifica virtualmente la participación del Jurado Evaluador, el titulando, el acto de instalación y el inicio, desarrollo y término del acto académico de sustentación, dejando constancia en el acta respectiva.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADÍSTICA

INFORME DE EVALUACIÓN DE ORIGINALIDAD

El Director de la Escuela Profesional de Estadística, Dr. Roger Pedro Norabuena Figueroa, informa lo siguiente:

1. Operador del programa informático de similitudes: Dr. Roger Pedro Norabuena Figueroa
2. Documento evaluado: Trabajo de Suficiencia Profesional para optar el Título Profesional de Licenciada en Estadística, titulado: CLASIFICACIÓN DE LA INSATISFACCIÓN DE ESTUDIANTES UTILIZANDO EL ALGORITMO SMOTE EN UNA REGRESIÓN LOGÍSTICA
3. Autor de la tesis: NATALI SALOME RUBINA CAMARGO
4. Fecha de recepción de la tesis: 11/01/2023
5. Fecha de aplicación del programa informático de similitudes: 11/01/2023
   - Software utilizado: Turnitin
6. Configuración del programa detector de similitudes:
   - Excluye textos entrecomillados
   - Excluye bibliografía
   - Excluye cadenas menores a 40 palabras
7. Porcentaje de similitudes según programa detector de similitudes: seis por ciento (6%)
8. Fuentes originales de las similitudes encontradas:
   Fuentes de internet: 6 %
   Publicaciones: 0 %
9. Calificación de originalidad:
   - Documento cumple criterios de originalidad, sin observaciones

Lima, 11 de enero del 2023

UNMSM
Firmado digitalmente por NORABUENA FIGUEROA Roger Pedro FAU 20148092282 soft
Motivo: Soy el autor del documento
Fecha: 11.01.2023 12:23:04 -05:00

Dr. Roger Pedro Norabuena Figueroa
Director

RESUMEN

Uno de los aspectos de mayor importancia en nuestra institución es evaluar la satisfacción de nuestros alumnos para con nuestros servicios. Medir este nivel de satisfacción de los estudiantes permite acercarnos a un objetivo nuestro. Por lo tanto, una tarea fundamental dentro de nuestra institución es desarrollar una estrategia para la satisfacción de nuestros alumnos. La insatisfacción es uno de los mayores retos a superar por parte de los negocios modernos. Algunas investigaciones anteriores señalan que el 72% de los clientes comparte una experiencia positiva con otras personas (aproximadamente con 6 conocidos). Sin embargo, aunque esa cifra parezca positiva, el 13% de los clientes insatisfechos comparte su opinión negativa con 15 personas o más. Por esto y otros motivos la importancia de esta investigación. El objetivo principal de este trabajo es clasificar a nuestros alumnos respecto a su insatisfacción con el ciclo académico, ya que poder clasificar a un estudiante como insatisfecho antes que realmente lo sea sería un gran aporte para nuestra institución. Para cumplir este objetivo se planteó abordar la clasificación de los alumnos mediante la regresión logística aplicando el algoritmo SMOTE que es una técnica que se utiliza cuando la variable respuesta presenta el número de observaciones dentro de sus categorías muy desbalanceadas, el comportamiento de este modelo fue evaluado por métricas como la Sensibilidad, especificidad, Curva ROC y clasificación global.

Palabras Claves: Regresión logística, datos desbalanceados, SMOTE, satisfacción del estudiante, sensibilidad.

ABSTRACT

One of the most important aspects in our institution is to evaluate the satisfaction of our students with our services. Measuring this level of student satisfaction allows us to get closer to our goal. Therefore, a fundamental task within our institution is to develop a strategy for the satisfaction of our students. Dissatisfaction is one of the biggest challenges for modern businesses to overcome. Some previous research indicates that 72% of customers share a positive experience with other people (approximately 6 acquaintances). However, even if that number seems positive, 13% of dissatisfied customers share their negative opinion with 15 or more people. For this and other reasons the importance of this research. The main objective of this work is to classify our students regarding their dissatisfaction with the academic cycle, since being able to classify a student as dissatisfied before they really is would be a great contribution to our institution. To meet this objective, it was proposed to address the classification of students through logistic regression applying the SMOTE algorithm, which is a technique used when the response variable presents the number of observations within its highly unbalanced categories, the behavior of this model was evaluated by metrics such as sensitivity, specificity, ROC curve and global classification.

Keywords: Logistic regression, unbalanced data, SMOTE, student satisfaction, sensitivity.

ÍNDICE

Lista de Tablas

Lista de Figuras

CAPÍTULO I. INTRODUCCIÓN

En el sector educativo peruano, durante los últimos años las instituciones han centrado casi todos su empeño en idear tacticas de captación y detener el aumento de estudiantes desertores, lo cual también se debe a que el mercado educativo actual se ha vuelto muy competitivo y con diversas opciones para los alumnos, es por esto que para las instituciones educativas se ha vuelto una tarea fundamental conocer la satisfacción de los estudiantes con los servicios que brindan, ya que la satisfacción actualmente es más que un indicador diremos que es una estrategia de captación ya que por medio de la recomendación se podrá atraer nuevos clientes en este caso estudiantes y también porque la insatisfacción en muchos casos es motivo de deserción. Frente a esta situación y problemática los modelos estadísticos son una alternativa como instrumentos de clasificación y también de predicción en este trabajo para la insatisfacción de estudiantes es decir mediante la investigación de patrones frecuentes de comportamiento se puede conocer el perfil de este tipo de estudiantes.

Este estudio se realizará para una institución educativa sin fines de lucro que practica y promueve la investigación para brindar una formación y educación de calidad a los diversos sectores de la población principalmente a la juventud estudiantil de nuestro país. Debido a la coyuntura de la pandemia se ha dado mayor fuerza y realce a la educación virtual, ofreciendo ciclos completamente virtuales. Nuestra institución cuenta con el área de Subdirección académica en la cual se encuentra el departamento de estadística al cual pertenezco, cumpliendo las funciones de analista de datos.

Realizar encuestas académicas es una práctica que ofrece varias ventajas entre ellas recabar información, por ejemplo, sobre las opiniones y percepciones de los estudiantes de una institución, acerca de los cursos, talleres, seminarios, evaluaciones escolares y otros servicios que se brindan

que tienen cierto impacto en los alumnos. Si nos referimos específicamente a las encuestas de satisfacción estas se aplican con el objetivo de saber la opinión de los alumnos acerca de la calidad de nuestros servicios como el material académico que se les brinda, las evaluaciones, el desempeño de sus docentes entre otros. Para nosotros como institución es muy importante la realización de las encuestas de satisfacción las cuales realizamos a todos nuestros ciclos.

Ahora bien, las técnicas de clasificación forman parte de la minería de datos y se encuentran dentro del aprendizaje supervisado, en este trabajo estas técnicas son usadas principalmente para crear un clasificador que diferencie y determine el perfil de nuestros estudiantes para poder conocer y predecir la insatisfacción al inicio del ciclo de estudio. Existen muchas técnicas de clasificación entre ellas la regresión logística la cual usaremos en este estudio.

La Regresión Logística binaria es una técnica estadística para predecir clases binarias, esto quiere decir que nuestra variable respuesta tendrá solo dos categorías posibles por ejemplo sano enfermo, éxito fracaso y en nuestro caso satisfecho frente a insatisfecho. Es una de los métodos más usados por su simplicidad y sus ventajas entre ellas el no cumplimiento del supuesto de la normalidad multivariada.

Un punto muy importante a considerar es la presentación de nuestra data, la cual se encuentra desbalanceada esto quiere decir que en nuestra variable dependiente una categoría es proporcionalmente mayoritaria respecto a la otra clase y esto resulta un problema al evaluar el modelo construido ya que los indicadores de exactitud casi siempre resultan muy altos debido a que la clase mayor o clase de éxito tiene mucho más número de observaciones.

Cuando los datos presentan esta particularidad Moreno, Rodríguez, Sicilia, Riquelme, y Ruiz (2009) señalan que los modelos clasificadores muestran una inclinación hacia la categoría mayor, es así que se minimiza el error de clasificación y clasifica acertadamente las observaciones

de categoría mayor, todo lo contario sucede con la categoría minoritaria. Además, señalan que frente a este problema se puede aplicar varios tratamientos entre ellos el oversampling (sobre muestreo) cuya técnica radica en balancear las proporciones de las clases incorporando observaciones de la categoría menor, uno de los algoritmos más representativos de esta técnica es el SMOTE (Syntetic Minority Over-sampling Technique).

Kunal (2016) señala que SMOTE es un algoritmo relativamente nuevo de sobre muestreo de minorías sintéticas ya que crea e incluye observaciones artificiales mediante la regla del vecino más cercano y así modificar la tendencia de aprendizaje del clasificador hacia la clase minoritaria.

El objetivo principal de esta investigación es clasificar a los estudiantes respecto a su insatisfacción mediante Regresión Logística Binaria aplicando el algoritmo SMOTE para datos desproporcionados.

CAPÍTULO II. DESCRIPCIÓN DE LA ACTIVIDAD

2.1 Información del lugar donde se desarrolló la actividad

La institución educativa es una Asociación Civil sin fines de lucro. Es un consorcio educativo y académico orientado a la investigación a través de una aprendizaje humanista, científico e integral, con una verdadera conciencia de nuestro país como diversidad multicultural. Se tiene como prioridad favorecer y desarrollar la formación integral de sus estudiantes con un elevado nivel académico e investigativo, desarrollando un profundo sentido crítico que profundiza en el conocimiento de nuestra realidad nacional y mundial, por lo tanto, es importante promover en los alumnos el desarrollo de un pensamiento fraterno y solidario para que actúen con responsabilidad social.

Se cuenta con una brillante trayectoria como institución educativa formando estudiantes con un alto nivel académico y cultural, también orientamos al estudiante en la elección de su carrera, contemplando su disposición y aptitudes.

Misión

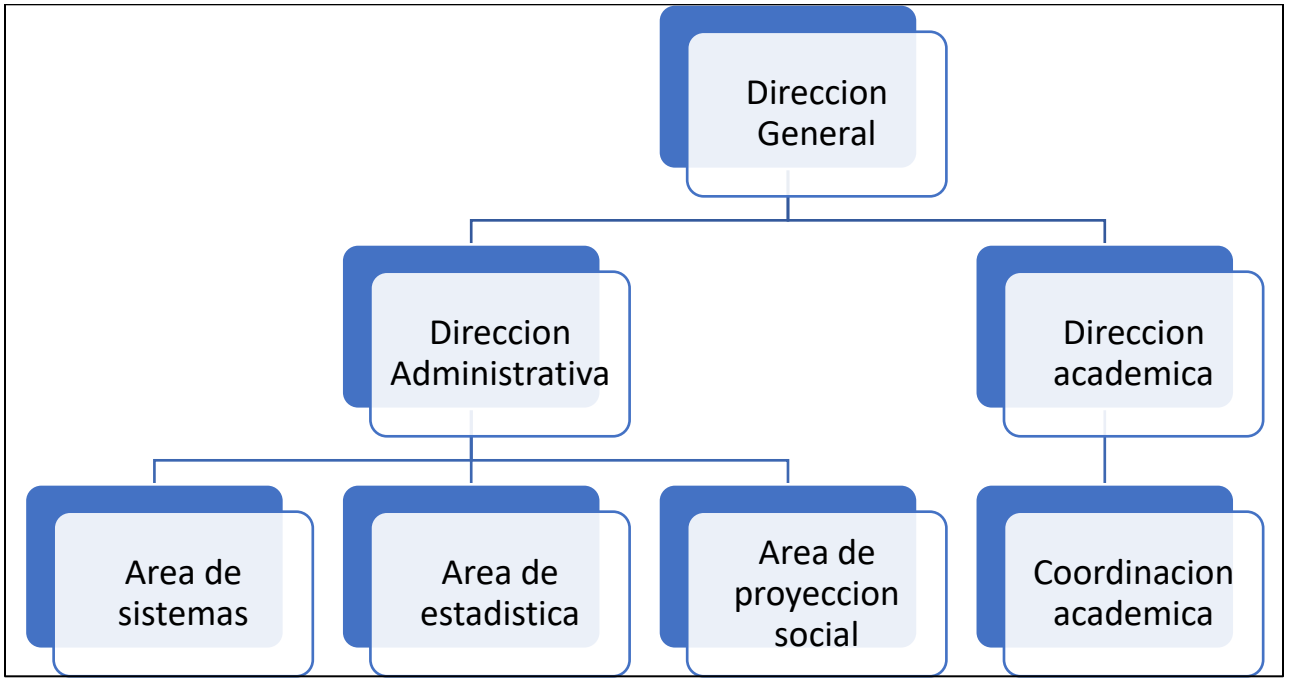
La institución educativa tiene la misión de brindar una educación completa e integral luego de proceder de la educación secundaria, mediante una enseñanza de calidad a los estudiantes que anhelan seguir estudios universitarios o institutos profesionales de nuestro país. Nuestra Institución garantiza una educación de calidad a estudiantes, Capacitándolos en valores, para que puedan responder a las exigencias actuales de un mundo cada vez más competitivo a través de herramientas con los últimos avances en todas las áreas y facilitar y capacitar la creación y aplicación de proyectos innovadores. Pero lo más importante es el desarrollo de la formación integral en todos los aspectos humanos para desarrollar sus habilidades.

Visión

Ser una institución educativa líder en el campo de la educación del país, donde el aprendizaje que ofrece nuestra institución en valores y a la vez cumpla con todos los estándares de calidad, a través de metodologías que se encuentran a la vanguardia de este siglo XXI, donde el alumno desempeña un papel de liderazgo en su proceso de aprendizaje y la tecnología, sea una herramienta importante para su crecimiento**.** El deseo de aprender, explorar y valorar los conocimientos y valores no solo ayuda a los estudiantes a tener un alto nivel de conciencia, sino también a una conciencia crítica y una conciencia democrática. Estamos comprometidos con el desarrollo social, ambiental y económico de su entorno.

Figura 1

Organigrama de la empresa



Nota. Elaboración propia

2.2 Descripción de la actividad

El área de estadística se encuentra dentro del departamento de subdirección académica cuya función principal es la elaboración de los planes de ciclo donde establece de forma precisa las actividades o tareas que se realizarán en un periodo lectivo, algunas de estas tareas son la organización de horarios de acuerdo a los cursos y docentes por materia, las fechas programadas para las evaluaciones, implementación de seminarios y asesorías, preparación del material académico por curso que se le brindara al estudiante.

Objetivos

- Formar estudiantes con un alto nivel académico y cultural
- Lograr un óptimo desarrollo de las actividades académicas y culturales programadas.
- Brindar herramientas adecuadas al estudiante para su ingreso y posterior formación profesional en la universidad o centro superior.

El área de estadística es la encargada de la planificación, la organización, el diseño de la toma de información para luego analizar e interpretar los resultados de las diferentes encuestas que se realiza para las diferentes áreas dentro de la institución. Toda esta información se muestra por medio de informes que son presentados y expuestos para luego ser derivado a las diferentes áreas para la toma de decisiones y lineamientos en base a estos indicadores.

Una de las tareas específicas es planificar, elaborar y la realización de las encuestas de satisfacción de nuestros servicios para con los alumnos, con el fin de identificar las necesidades de los estudiantes y cualquier problema con el producto o servicio. Otra tarea específica es la elaboración del cuestionario para conocer el perfil de nuestros estudiantes y así poder direccionar la implementación de servicios de acuerdo a las características de nuestros alumnos.

2.3 Finalidad del trabajo TSP

Una tarea fundamental dentro de nuestra institución es desarrollar una estrategia para la satisfacción de nuestros estudiantes, ya que la insatisfacción genera un malestar y podría traer una mala recomendación de nuestros servicios, además es una de las posibles causas de abandono en el ciclo de estudio, por lo tanto poder clasificar un estudiante como insatisfecho antes que realmente lo sea sería un gran aporte para así poder mejorar los servicios con los que ellos no se sienten bien y evitar su posible mala recomendación y fuga o deserción.

La investigación realizada contribuirá a la institución educativa a una mejora en cuanto a la antelación de estudiantes en la satisfacción de un servicio brindado asimismo se verá la importancia de aplicar técnicas de balanceo de datos en una regresión logística cuando se presenta este tipo de datos.

2.4 Objetivos del Trabajo TSP

Objetivo General

- Diseñar un modelo de clasificación a partir de los datos de ingreso que identifique las variables que inciden en la insatisfacción del estudiante.

Objetivos Específicos

- Determinar el perfil del estudiante insatisfecho.
- Describir la técnica y el algoritmo, para predecir la insatisfacción de los estudiantes con el ciclo académico a partir de sus datos de ingreso.

2.5 Metodología

Método

La presente investigación se refiere al uso de la regresión logística que es una técnica de modelamiento para clasificación, el cual se usará mediante el software R, el cual tiene como una de sus grandes virtudes ser un software libre.

La presente investigación es un trabajo importante porque permite conocer el perfil de los estudiantes insatisfechos, identificando las características significativas de los alumnos que no se sienten satisfechos con el servicio académico del ciclo de estudio propuesto por el área de Subdirección académica y de esta manera clasificar a esta clase de estudiantes para una mejor toma de decisiones.

CAPÍTULO III. MARCO TEÓRICO

3.1 Minería de Datos

La minería de datos está dirigida al hallazgo de conocimiento en base de datos o también llamado KDD, en ella se encuentra patrones y se formulan relaciones mediante diversas tecnicas. Para que luego desde estos patrones y relaciones se tomen decisiones y se pueda predecir el comportamiento a futuro.

Gutiérrez y Molina (2016) señalan que el desarrollo de extracción de patrones a partir de datos se le denomina minería de datos, además indican que es un instrumento fundamental y esencial de los negocios modernos ya que es capaz de convertir los datos en inteligencia de negocios.

Dentro de las diversas metodologías que presenta la minería de datos se encuentra el aprendizaje supervisado de clasificación, es ahí donde se encuentra la regresión logística que se aplica para clasificación binaria.

Las técnicas y/o metodologías de la minería de datos se fijan en función de diferentes métodos, entre ellos se encuentran las metodologías SEMMA y CRISP-DM, para este trabajo de investigación usaremos la metodología CRISP.

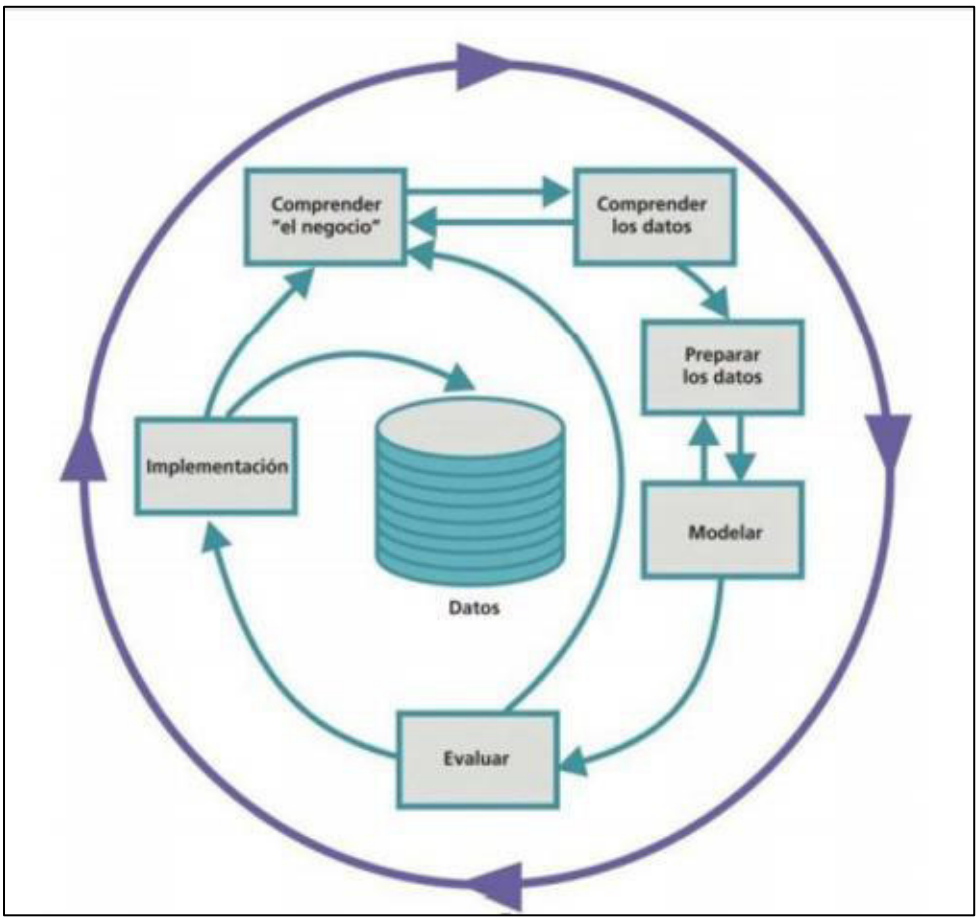
Metodología CRISP-DM (Cross Industry Standard Process for Data Mining)

Se trata de una metodología estándar abierto para minería de datos que incluye un manual de los procedimientos estándar conocidos de los expertos en minería de datos. Según el Manual CRISP-DM de IBM SPSS Modeler, “Es un método probado para orientar sus trabajos de minería de datos. Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos”.

En la figura 2 se aprecia el ciclo del modelo, la cual se estructura en seis fases.

Figura 2

Etapas de la metodología CRISP



Nota: Adaptado de "CRISP-DM 1.0: Step-by-step data mining guide" Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Semantic Scholar.

Las fases son:

1. Comprensión del Negocio: es la fase inicial que se centra en la comprensión de los requisitos y objetivos del proyecto, esta fase nos proporcionara un pre modelo para lograr nuestros objetivos.

2. Comprensión de los Datos: es la fase de entendimiento de los datos, es aquí donde se realizan tareas como la recolección de datos, descubriendo correlaciones, concordancias importantes para reconocer y formular las hipótesis.

3. Preparación de los Datos: luego de la fase 2 se inicia la tarea de la preparación, estas tareas incluyen la limpieza de los datos, y su transformación preparándolos para el procesamiento, incluye tareas de normalización, discretización, tratamiento de valores perdidos.

4. Modelamiento: es aquí donde se eligen las técnicas que se utilizaran para modelar los datos del estudio a realizar. Para elegir el algoritmo o la técnica se deben seguir requerimientos que sean apropiados y de acuerdo a la forma de nuestros datos.

5. Evaluación: en esta etapa se procede a la evaluación del modelo o modelos para verificar que se cumpla con los objetivos del proyecto.

6. Implementación o Despliegue: luego de que el modelo ha sido diseñado y evaluado, se procede a generar un informe que visualice transformar el conocimiento dentro de la organización.

3.1.1 Regresión logística

Introducción

Los métodos de regresión son una parte importante de muchos análisis de datos relacionados con la explicación de la relación causal entre una variable respuesta y una o más variables predictivas o independientes. La regresión lineal se usa para el modelamiento de una relación causal entre una variable respuesta continua y un grupo de variables explicativas continuas, esta relación es lineal.

Pero lo que comúnmente ocurre es que, la variable dependiente no es continua sino más bien categórica y toma sólo dos o más posibles valores. En 1972, Nelder y Wedderburn,

publicaron un artículo que marcó un hito en el análisis de regresión, demostrando que la familia de los modelos lineales podía ser generalizada de manera que introduzca miembros que habían sido analizados como modelos particulares. Ellos nombraron como modelos lineales generalizados a esta nueva familia compuesta por los modelos clásicos más los modelos en los cuales la variable respuesta no necesariamente tiene una distribución normal en los cuales se incluye una variedad de distribuciones de probabilidad denominada Familia Exponencial. El análisis de regresión logística se enmarca en esta nueva familia de los modelos lineales generalizados.

La Regresión Logít o modelo logístico es una técnica o algoritmo de clasificación que actualmente se usa bastante en todos los campos. En estos casos como su nombre lo dice lo que se quiere es clasificar por lo tanto la probabilidad se mostrara como una partición de dos. Este clasificador analiza la relación entre múltiples variables independientes y una variable dependiente categórica. Este clasificador muestra la significación de las variables independientes en términos de probabilidades.

Nieto (2010), señala que para la aplicación de la regresión logit no se necesita que cumpla con los supuestos que se presentan en una típica regresión lineal. Además, indica que esta técnica de regresión, será apropiado a datos donde la variable respuesta o dependiente puede tomar dos posibles valores, es así que esta técnica se utiliza cuando la variable dependiente “y” no tiene una distribución normal y el predictor lineal compuesto por las variables independientes pueden ser categóricos ya sea binarias o politómicas, ordinales y también combinación de numéricas y categóricas.

El Modelo de regresión Logístico

Sea la variable respuesta Y, la cual sólo puede tomar los valores Y=1 indica la presencia de la característica de interés, con probabilidad de ocurrencia igual a π (P(Y=1) = π) y Y=0 que significa la ausencia con probabilidad 1-π (P(Y=0) =1-π).

En cuanto a la variable independiente o predictora definida como X pueden ser cuantitativas o cualitativas las cuales tienen el papel de variables explicativas. La variable respuesta o dependiente definida como Y, es una variable aleatoria la cual tiene distribución Bernoulli con probabilidad condicional Y=1 dado que X=x:

$$P(Y=1/X=x) = E(Y/X=x) = \pi$$

$$V(Y/X=x) = \pi\,(1-\pi).$$

Esta probabilidad condicional puede ser calculada a partir de un modelo que tiene la forma de una curva sigmoidea, en particular esta curva sigmoidea puede ser la función logística:

$$P(Y / X = x) = \pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Donde P(Y/X=x) = π que corresponde a la probabilidad de éxito de la variable aleatoria discreta llamada distribución Bernoulli, por lo tanto, la combinación lineal de las variables independientes es el predictor lineal $\eta = \beta_0 + \beta_1 X$

Modelo Logit

Si ahora aplicamos el logaritmo natural, obtenemos una ecuación lineal que permite un manejo matemático aún más fácil y de mayor comprensión:

$$f^{-1}(\eta) = Ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x = \eta$$

El término a la derecha de la igualdad es la expresión de una recta, idéntica a la del modelo general de regresión lineal y a la derecha tenemos la esperanza de la variable respuesta a la cual se le ha aplicado una transformación logit**.**

Podemos extender el modelo para el caso de varias variables explicativas, entonces el modelo será:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}$$

Recordemos que, π es la probabilidad de que el individuo tenga la característica de interés (insatisfacción del estudiante). Lo importante de esta transformación es que, ahora, la variable respuesta es el logit (π) y está relacionada linealmente con la combinación lineal de variables predictivas (predictor lineal). Luego el modelo de regresión logística se reduce a un modelo de regresión lineal múltiple, donde las variables independientes pueden ser cualitativas y cuantitativas.

$$\log it(\pi(x)) = ln\left(\frac{P(Y/X=x)}{1-P(Y/X=x)}\right) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

El predictor lineal del modelo de regresión logística puede contener variables categóricas, (binarias/politómicas), numéricas (intervalo/razón) o combinaciones de variables numéricas y categóricas.

Estimación de parámetros

La diferencia fundamental entre el análisis de regresión lineal y la regresión logística es que en el primero se considera que los errores son variables aleatorias continuas no observables distribuidas normalmente con media cero y varianza constante (σ^2), en cambio en la regresión logística binaria la target o variable respuesta no es continua sino es discreta y sólo puede tomar dos valores (dicotómica) esto quiere decir la ocurrencia o no de un evento, por lo tanto su distribución obedece a una Bernoulli cuya media es π y su varianza π(1-π). En cuanto a la distribución y a los valores que pueden tomar los errores decimos:

$$\varepsilon_i = \begin{cases} 1-\pi_i & y_i = 1 \\ -\pi_i & y_i = 0 \end{cases} \quad \text{i= 1, 2…n}$$

En una regresión logística los errores se distribuyen de la misma forma que la variable dependiente Y, por lo tanto, su distribución del error aleatorio tiene distribución Bernoulli

$$\varepsilon_i \rightarrow B(1,\pi) \quad \text{i=1,2,...,n}$$

En cuanto a la esperanza o media de los errores es $E(\varepsilon_i) = \pi_i$

y la varianza es $V(\varepsilon_i) = \pi_i(1-\pi_i)$ (no constante) i=1,2,…,n..

Dado que, se conoce la distribución de probabilidades de los errores, lo natural es utilizar el método de estimación de máxima verosimilitud.

En el caso del modelo logístico (modelo logit), tenemos dos problemas:

1) La función objetivo $S(\beta)$ a maximizar para obtener las estimaciones de los parámetros del modelo, β, no es lineal en los parámetros, por lo que se deberán utilizar métodos numéricos para la estimación.

2) La varianza de los errores no es constante por lo que se deberá transformar la función objetivo, incorporando la estructura de covarianzas.

Por lo tanto, la ecuación del modelo logístico múltiple luego del ajuste quedaría de la siguiente forma:

$$Logit(\pi_i(x)) = \ln\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{iK} \quad \text{i=1, 2….. n}$$

Es importante recordar que al ajustar el modelo este no da como resultado la probabilidad de que la característica o particularidad esté presente (Insatisfacción en estudiantes) sino el logit ($\pi(x)$), o también el logaritmo por lo que se tendría que hacer la transformación, entonces para

poder interpretar mejor los resultados podemos calcular el antilogaritmo. Esta transformación no modifica la forma de interpretar el signo del coeficiente tal como se hace en una regresión lineal, signo positivo aumenta la probabilidad de éxito mientras que un signo negativo disminuye la probabilidad.

$$\left(\frac{\pi_i(x)}{1-\pi_i(x)} \right) = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} \ldots e^{\hat{\beta}_{k1} x_k}$$

El lado izquierdo de esta ecuación es conocido como logaritmo de ODDS RATIO o la razón de probabilidades y se define como:

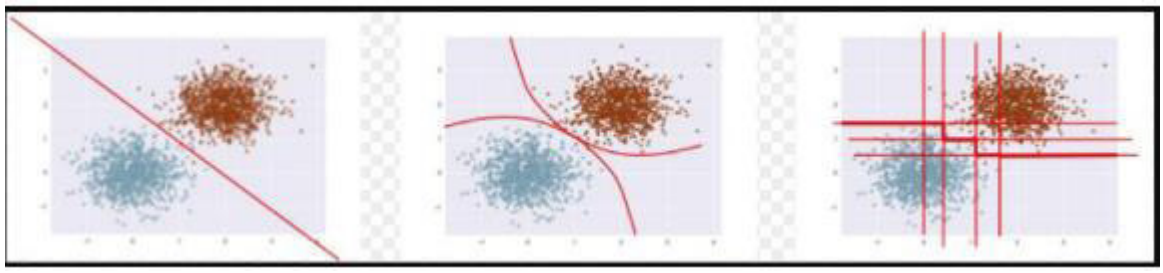
$$OR_{j,ajustado} = e^{\beta_j}$$

3.1.2 Datos desbalanceados

La finalidad de un algoritmo de una técnica de clasificación es aprender a separar o clasificar, las dos categorías de la target o variable respuesta. Hay diversas formas para realizar esta tarea, que se basan en muchos supuestos estadísticos, algunas de estas se observan en la figura 3.

Figura 3

Posibles Clasificadores para separar 2 categorías de la target

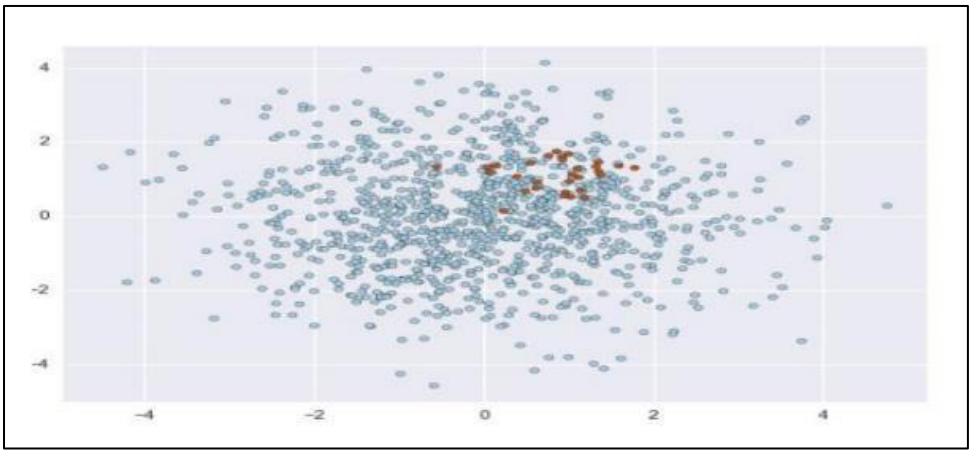


Nota: Tomado de Datos Desbalanceados *por Fawcett, 2016*

Actualmente en muchos estudios se presentan datos donde las dos categorías de la variable respuesta presentan una diferencia notable en el número de sus observaciones es decir una desigualdad de las proporciones, como se observa en la figura 4.

Figura 4

Clases desproporcionada de la target



Nota: Tomado de *Datos Desbalanceados* por Fawcett, 2016

¿Que hacer frente a estas situaciones?

En la realidad actual se presentan sucesos pocos frecuentes como por ejemplo el número de casos de una enfermedad rara en una población, como consecuencia de esto se ha dado una desproporción considerable en el número de casos dentro de una categoría, a esta problemática la llaman clases desbalanceadas. El aprendizaje automático tiene la difícil tarea de dar solución a este tipo de datos no balanceados, ya que los algoritmos no funcionan correctamente frente a este conjunto de datos.

La mayoría de clasificadores convencionales logran excelentes precisiones para la clase mayoritaria es decir donde se encuentra la mayor cantidad de observaciones, mientras que para la clase minoritaria no sucede lo mismo ya que no se clasifica bien a esta categoría. En este tipo de datos casi siempre el interés suele estar en la categoría menos representada, sin embargo, muchos

clasificadores los tratan como valores atípicos y se concentran en resultados de los indicadores globales.

En Fawcett (2016), se presenta un bosquejo de técnicas y enfoques útiles cuando se presenta este tipo de datos:

1) Dar una armonía de las categorías en la data de entrenamiento con alguna técnica:

- En la categoría minoritaria aplicar sobre muestreo.
- En la clase de mayor proporción eliminar observaciones.
- Crear nuevas observaciones sintetiza en la clase minoritaria.

2) Replicar las observaciones de la clase menor y modificar a una lista de localización de outlayers.

3) Cuando se aplica el algoritmo, o después de la aplicación:

- Adaptar las proporciones de las clases.
- Ajusta el punto de decisión.
- Cambiar, adecuar algoritmos ya existentes para que pueda reconocer este tipo de datos donde las clases son desproporcionadas.

4) Crear mediante un programa un algoritmo nuevo que pueda dar buenos indicadores de clasificación para este tipo de datos.

Técnicas para datos desbalanceados

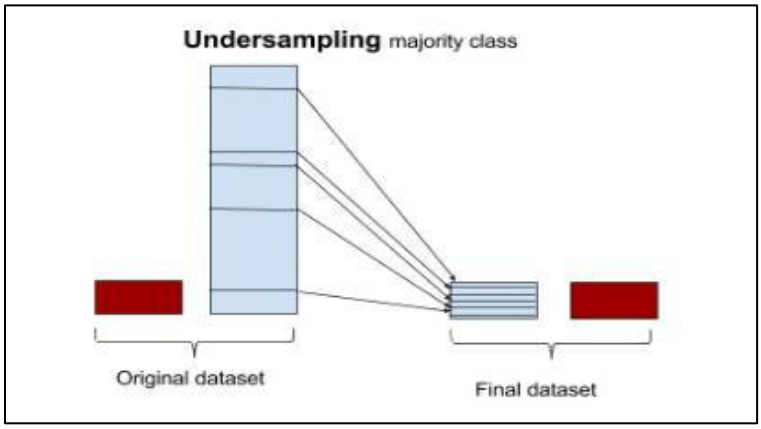
La finalidad de estas técnicas es transformar los datos desbalanceados en una data con proporciones parecidas aplicando un algoritmo o mecanismo. Este cambio se da modificando el número de observaciones de la base original y se trata de dar una proporción similar en ambas categorías o clases. Las técnicas más utilizadas son:

- **Submuestreo o Undersampling**

Esta metodología reduce al azar observaciones de la clase mayor y así igualar al tamaño de la clase menor. todo esto para nivelar las muestras y que el modelo aprenda en la data de entrenamiento, pero con datos balanceados. Por esta técnica de sub-muestreo, se cabe la posibilidad de suprimir ciertas observaciones de la categoría mayor que pueden ser más caracteristicos, por lo tanto, se estaría eliminando información útil.

Figura 5

Funcionamiento del submuestreo



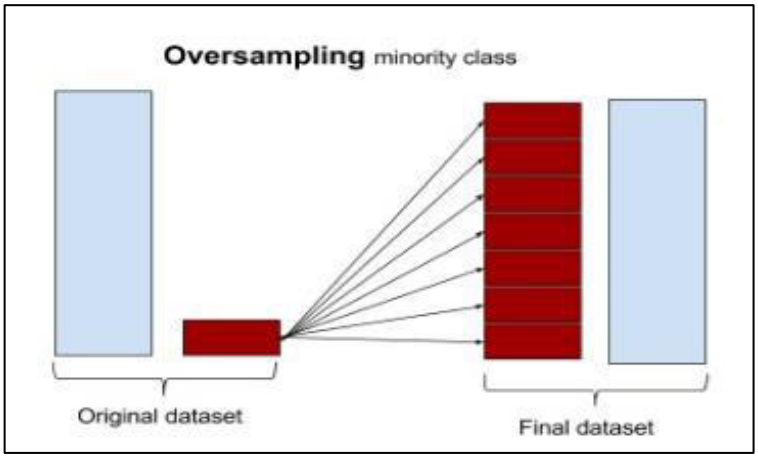
Nota: Tomado de *Datos Desbalanceados* por Fawcett, 2016

- **Sobre muestreo Oversampling**

Este método es contrario al sub muestreo ya que se basa en replicar observaciones de la clase menor al azar hasta llegar a tener tantas observaciones como la otra categoría o clase esto se realiza mediante muestreo con reemplazo. Una de sus principales ventajas es que de la base original no se pierde información. En cuanto a los inconvenientes de trabajar con esta técnica, es que el solo a trabajar la clase menor podría traer consecuencias como de sobreajuste del clasificador.

Figura 6

Funcionamiento del sobre muestreo



Nota: *Tomado de* Datos Desbalanceados *por Fawcett, 2016*

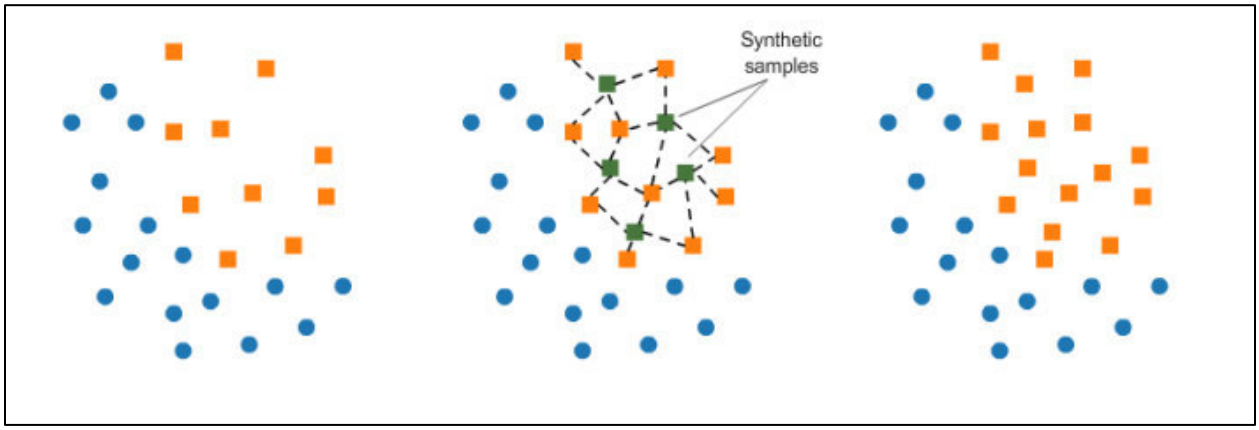
Existen otras técnicas para el problema de desbalanceo de datos, pero en este trabajo se utilizará el algoritmo SMOTE el cual explicaremos a continuación.

SMOTE: Técnica de sobre-muestreo de Minorías sintéticas

Es un algoritmo o mecanismo de sobre-muestreo con reemplazo para la categoría menor. En este estudio utilizaremos esta técnica para balancear nuestros datos. Esta técnica funciona creando nuevas observaciones sintéticas minoritarias al azar calculando los K vecinos más cercanos para estos nuevos ejemplos Estas nuevas observaciones se añaden entre la observación escogida y sus vecinos.

Figura 7

Esquema de Algoritmo de SMOTE



Nota: *Tomado de* Datos Desbalanceados *por Fawcett, 2016*

En este trabajo utilizaremos el algoritmo SMOTE para nuestro problema de datos desbalanceados que como hemos indicado consiste en la producción de observaciones sintéticos, esta aplicación opera en "el espacio característico" en vez de en "el espacio de datos". El sobre muestreo se aplica en la clase minoritaria creando nuevas observaciones de esta categoría, esto lo realiza insertando muestras sintéticas entre los k vecinos más próximos de manera aleatoria de la categoría minoritaria.

Pariona (2017) explica cómo se generan estas muestras sintéticas, lo primero que se realiza es coger la diferencia que existe entre el vector de la muestra y su más próximo k vecino para luego multiplicarla por un número escogido al azar entre 0 y 1, y lo adiciona al vector característico de la muestra.

Funcionamiento paso a paso del algoritmo SMOTE

Pariona (2017) describe este mecanismo de sobre muestreo de la siguiente manera:

- Obtiene la proporción de observaciones originales como parámetro que serán sebremuestreados.

- Estima el número de observaciones para nivelar ambas categorías mediante interpolación.
- Determina los k vecinos que se encuentran más próximos de las observaciones de la categoría menor.
- Los k vecinos se eligen y deciden generando N distancias entre la observación primaria y sus contiguos.
- En cuanto a la regla de distancia a utilizar se escogerá de acuerdo a la naturaleza de nuestros datos.
- Para cada característica de la observación elegida para ser sobre muestreada, se estima la diferencia entre el vecino escogido aleatoriamente y el vector de atributos de la muestra.
- Al final el algoritmo retorna el conjunto de ejemplos sintéticos

3.1.3 Evaluación de Clasificación

En la evaluación de los modelos de clasificación y predicción existen diversas métricas, la que comúnmente se utiliza es la métrica de exactitud, pero en los casos donde la variable respuesta se encuentra muy desbalanceada esta métrica resulta engañosa y no valida. Las medidas o indicadores de desempeño más sugeridas o recomendadas para datos no balanceados son la sensibilidad, F-Measure, Precisión, la ROC y AUC (Wu et al. 2018).

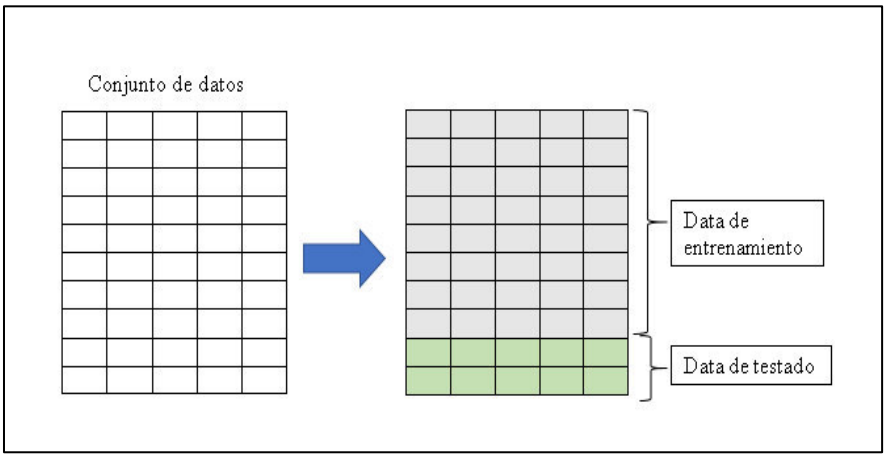
En cuanto a la estimación al fin de evaluar las técnicas usadas para clasificar, diremos que es una de las partes más importantes del trabajo y en minería de datos se plantea utilizar la matriz de confusión para definir de acuerdo a ella indicadores como la sensibilidad y la precisión ya que nos permite validar el modelo sobre el conjunto de entrenamiento. Además de la evaluación también nos permite comparar técnicas de clasificación como es el caso de este trabajo que realizara el modelo de clasificación utilizando la regresión logit con y sin el algoritmo SMOTE y así escoger el modelo con los indicadores destacados.

División de la data original en muestra de entrenamiento y de testeo

Esta técnica de minería de datos se basa en dividir el conjunto total de datos en dos partes, uno de ellos será la muestra de entrenamiento y la otra la muestra de prueba o testado, esto se realiza para obtener una buena predicción. La parte de entrenamiento lo usa para entrenar el modelo de clasificación y suele ser más grande (aproximadamente de 70% a 80%), es aquí donde el modelo aprende para luego predecir. En cuanto a la parte de testado se usa para comprobar si efectivamente el modelo ha aprendido de la predicción.

Figura 8

Esquema de división de la muestra



Nota: *Elaboración propia*

Tabla de clasificación

La tabla de clasificación también llamada matriz de confusión es un instrumento o herramienta que tiene por finalidad evaluar el rendimiento de la técnica de aprendizaje supervisado. De esta manera podemos verificar la capacidad clasificatoria del modelo.

Fawcett (2004) señala que la matriz de confusión o también llamada tabla de clasificación, está compuesta por datos que contienen las predicciones halladas por una técnica de clasificación,

en este caso la regresión logística es así que realiza la comparación para todas las observaciones en la tabla de aprendizaje, la predicción resultante frente a la categoría que pertenecen realmente.

La Matriz de Confusión nos posibilita visualizar a través de una tabla de contingencia los valores observados que se encuentran en las filas, mientras que las predicciones se representan en las columnas de acuerdo a cada categoría, por lo tanto lo que realmente nos permite ver son los aciertos (verdaderos positivos y verdaderos negativo) y errores (falsos positivos y falsos negativos) de los resultados de la clasificación que ha realizado nuestro modelo.

Recordemos que en la regresión logística la respuesta observada toma valores cero o uno, es decir, los individuos pertenecen a uno de dos grupos.

Una vez ajustado el modelo se obtienen las respuestas estimadas que corresponden a valores en el intervalo [0, 1]. (Probabilidades estimadas). Por lo tanto, es necesario establecer una regla de clasificación (basada en una combinación lineal de las variables explicativas) de modo que se pueda tener un criterio (punto de corte), de modo que, a partir de las probabilidades estimadas, se pueda clasificar a un individuo en uno de los dos grupos, pero tratando de minimizar el error de clasificación (falsos positivos y/o falsos negativos), o equivalentemente tratando de maximizar la sensibilidad y/o especificidad.

Una vez establecido el criterio de clasificación, se construye la tabla de clasificación o matriz de confusión.

Tabla 1

Esquema de Matriz de confusión

	Predicho	
Observado	Positivos	Negativos
Positivos	a	b
Negativos	c	d

Nota: elaboración propia

Podemos observar en la tabla 1 los valores a y d, los cuales son los que están acertadamente clasificados (para ambas clases) y se encuentran en la diagonal. Los valores c y b son las clasificaciones incorrectas. En esta matriz se muestran las predicciones correctas tanto para los casos positivos (verdaderos positivos VP) y negativos (verdaderos negativos VN). Pero también se visualiza las predicciones incorrectas.

En resumen, podemos determinar para nuestro trabajo los siguientes términos:

Verdadero positivo (a=VP): Un estudiante que se encuentra insatisfecho y la prueba lo predice insatisfecho.

Verdadero Negativo (d=VN): Un estudiante que se encuentra satisfecho y la prueba lo predice satisfecho.

Falso Positivo (c=FP): Un estudiante que se encuentra insatisfecho, pero la prueba lo predice erróneamente como satisfecho conocido como el error tipo II.

Falso negativo (b=FN): Un estudiante que se encuentra satisfecho, pero la prueba lo predice erróneamente como insatisfecho.

Métricas para la evaluación de una clasificación

Precisión: Es una medida de precisión de los positivos predichos en contraste con la tasa de positivos reales, es decir representa la proporción de casos positivos pronosticados que son

correctamente positivos observados y se calcula como la división de los verdaderos positivos entre la suma de estos más los falsos positivos. (Powers, 2008).

$$\text{Precisión} = \frac{VP}{VP+FP}$$

Sensibilidad: Es una medida que nos muestra la proporción de observaciones clasificadas como positivas, en los que se confirma que ciertamente son positivos en las observaciones reales es decir se predicen correctamente.

$$Sensibilidad = \frac{VP}{VP+FN}$$

F-measure (medida F): Es una medida que está compuesta por la precisión y la sensibilidad. Es la media armónica entre la sensibilidad (Recall) y la precisión.

$$\text{F} = \frac{Precision*Sensibilidad}{Precision+Sensibilidad}$$

CURVAS ROC

(Curvas de operación característica del receptor): Este es uno de los principales indicadores para valorar modelos cuyos datos son de naturaleza desproporcionados, se determina delineando a los verdaderos positivos frente a la tasa de falsos positivos. Por lo tanto, diremos que representa la sensibilidad frente a la especificidad.

Tharwat (2018) afirma que un indicador para contrastar las diversas curvas ROC de los modelos es el AUC (área bajo la curva), este valor varía entre cero y uno, el que obtenga el valor superior probablemente será el de mejor rendimiento. Este valor es un buen indicador para evaluar que también funciona el modelo construido, cuando este valor es uno diremos que esta prueba es perfectamente capaz de diferenciar ambas clases o categorías sería una prueba perfecta, muy por el contrario si el AUC es 0.5 el modelo no es capaz de diferenciar entre la categoría positiva y la categoría negativa. Una forma de interpretar este valor es como una

probabilidad, por ejemplo si el AUC es 0.65, diremos que hay 65% de probabilidad de que el modelo pueda discriminar entre ambas clases.

La tabla de contingencia puede proporcionar varias medidas. El comportamiento de estos indicadores depende de donde se ponga el punto de corte, si este se desplaza a la derecha se reducen los falsos positivos, pero a la vez se incrementan los falsos negativos. Por lo tanto si se reduce la sensibilidad se acrecienta la especificidad, entonces la elección del punto de corte es importante y su elección depende mucho del objetivo a cumplir.

Por lo tanto, la curva ROC es útil en las siguientes situaciones:

• Conocer el desempeño general de la prueba observando el área debajo de la curva.

• Comparar dos pruebas o dos umbrales (puntos de corte). Comparacion de dos curvas o dos puntos en una curva.

• Seleccionar el punto de corte adecuado según el objetivo del estudio.

3.2 Satisfacción del estudiante

La satisfacción estudiantil se ocupa de ilustrar y definir aspectos relacionados con las actividades educativas, procedimientos, de pedagogía aprendizaje, como también de los diversos asuntos que incurren en las actividades de los alumnos y profesores (Questión Pro. 2020).

Eyzaguirre (2015) indica que la satisfacción estudiantil es un indicador que mide la eficacia de los múltiples aspectos académicos, por lo tanto señala que esto lo convierte en un indicador de calidad.

Por otro lado Gento (2012) como se citó en Nobario (2018) señala que “la satisfacción estudiantil se enfoca en todo aquello que el estudiante percibe y lo considera importante, por cada servicio educativo que la institución educativa brinda y el estudiante lo toma en cuenta en el

momento preciso de valorar su nivel de satisfacción estudiantil por la calidad del servicio educativo".

Las Instituciones Educativas de Formación que acogen un planteamiento de satisfacción al cliente educativo, es decir al estudiante, tienen como responsabilidad satisfacer en todo el conjunto, al alumno y a la sociedad en general, por lo tanto, frente a la expectativa de la oferta educativa tienen que alcanzar esta o superarla. Por eso, casi todas las estrategias de estas instituciones se centran en reforzar la credibilidad, mejorar sus servicios académicos y no académicos, creación de nuevos productos, cumplir con las metas educativas institucionales y sobretodo asegurar que se cumpla con la satisfacción de toda la sociedad educativa.

Actualmente en nuestro país la mayoría de instituciones educativas han implantado sistemas de gestión de aseguramiento de la calidad, ya que las certificaciones y acreditaciones se han vuelto un requisito indispensable en el marco de la ley educativa todo esto asegurara y garantizara un buen funcionamiento de la institución en cuanto a la calidad e innovación de sus servicios. La norma ISO 9001 es una en la cual se han fijado muchas instituciones para su tarea de mejoramiento de la imagen y así atraer nuevos clientes y aumentar la confianza en la institución.

La gestión educativa, es un proceso orientado al desarrollo educativo con la finalidad de mejorar la didáctica pedagógica, la función de los directivos y del personal administrativo, para conservar la imagen institucional y así poder superar las necesidades educativas. Es así que en nuestra institución la gestión es un proceso sistemático que tiene pasos establecidos en donde se debe de empezar con una planeación y evaluación, estos pasos a seguir son:

a) La primera fase o etapa es la autoevaluación, es aquí donde se recopila y analiza toda la información, es decir la realización de la encuesta que evaluara todos los servicios brindados,

eso nos permite reconocer las fortalezas y también nuestros puntos débiles para así mejorar y poder realizar un plan de mejoramiento, la autoevaluación es uno de los puntos esenciales durante la aplicación de los proyectos y programas.

b) La segunda fase es la implementación para el mejoramiento, luego de identificar fortalezas y puntos débiles, se establecen metas y estrategias claras dirigidas al cumplimiento delas metas institucionales.

c) La tercera fase es la realización y el seguimiento, es aquí donde luego de la implementación se ejecutan el plan de mejoramiento para una mejor toma de decisiones.

Estas fases constituyen el camino y la marcha que se necesita para el desarrollar e implementar una gestión educativa de calidad.

Para el cumplimiento de la primera fase nuestro compromiso consiste en la aplicación de una encuesta a estudiantes, donde se recogerá la percepción sobre los principales componentes del servicio educativo.

En nuestra institución educativa la medición de la satisfacción del estudiante está a cargo del departamento de Subdirección académica, el cual tiene como uno de sus principales brazos al área de estadística quien se encarga específicamente de esta tarea.

El instrumento que usamos para medir el grado de satisfacción con el ciclo de estudio del estudiante es un cuestionario validado por juicio de expertos, es una escala Likert de cinco puntos.

3.3 Marco Histórico

3.3.1 Antecedentes Nacionales

Pariona (2017) construye en su tesis “Clasificación de fuga de clientes en una entidad financiera utilizando el algoritmo SMOTE para datos desbalanceados en una regresión logística”, un modelo con datos desproporcionados, realiza una comparación aplicando y sin aplicar el algoritmo SMOTE, con la finalidad de predecir la disminución de clientes en instituciones financieras. En esta investigación se evalúa la capacidad clasificatoria de ambos modelos utilizando métricas como sensibilidad, especificidad y curca ROC. Los resultados de este trabajo fueron que al comparar ambos modelos el primero con los datos originales sin balancear y el segundo utilizando la técnica SMOTE , este obtuvo resultados superiores dado que obtuvo los indicadores sensibilidad más altos comparado con la técnica de submuestreo simple y sin ninguna técnica para balancear datos.

Meza (2018), en su investigación “Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo Adaboost desbalanceado y la regresión logística asimétrica” realiza una comparación para datos asimétricos formulando modelos usando el algoritmo Adaboost y la regresión logística para predecir el comportamiento de usuarios de una compañia de operadores móviles. El modelo logístico lo realiza desde el enfoque de machine learning como es nuestro caso ya que su objetivo también fue la clasificación. Los indicadores de desempeño para escoger el mejor modelo fueron la sensibilidad métrica que es utilizada para este tipo de datos, el F-measure y el área bajo la curva ROC refiere que utiliza estas medidas ya que trabaja con datos desbalanceados. Sus principales resultados mostraron que al comparar los modelos de la regresión logística con los modelos del algoritmo Adaboost este último obtuvo el mejor desempeño en cuanto al AUC.

Tarazona (2016), en su investigación en la Universidad Nacional de Ingeniería "Identification of Factors and variables describing the quantily and distribution of fatal vehicular accidents in metropolitan city of Lima using data Mining techniques: random forest, boosting, decisión tres", aborda la problemática de los siniestros viales fatales ocurridos en Lima Metropolitana, mediante metodologías de minería de datos identifico los factores asociados a esta problema, al igual que nuestra investigación su data era asimétrica por lo tanto uso SMOTE para balancear, en sus resultados se logró identificar que las variables tipo de vehículo, la vía de ocurrencia, invasión de carril y desobediencia de los conductores a la señalización de tránsito son los factores relevantes que están asociados en el desenlace fatídico de los siniestros viales.

3.3.2 Antecedentes Internacionales

Martínez (2019) España**,** en su tesis de maestría "Modelos de clasificación para incidencias en entornos industriales con datos no balanceados" realiza la creación de un modelo de clasificación para optimizar el rendimiento de las máquinas y predecir fallas en estas. En este trabajo se obtuvo muy buenos resultados en los indicadores para todos los modelos estimados, lo que si encontró fue un mal desempeño en los modelos de regresión logística y Naive Bayes. Trabajo con las métricas de sensibilidad, el F-measure y las curvas ROC.

Arnejo (2017) en su investigación "Métodos para la mejora de predicciones en clases desbalanceadas en el estudio de bajas de clientes (CHURN)" en España, detalla como objetivo estudiar las técnicas que permitan hacer predicciones de la forma más fiable posible atendiendo al problema del desbalanceo ya que explica que el clasificador utilizado siempre aprende más de la clase o categoría que tiene más observaciones en la variable respuesta y así da como resultados predicciones equivocadas. Este estudio se centra en la problemática de abandono de clientes y comparación de técnicas de machine learning.

De Juan (2017), en su tesis realizada en la universidad Internacional de la Rioja, Madrid, "Análisis y optimización de algoritmos de clasificación supervisada sobre operaciones impagadas en tarjetas de créditos" realizo un algoritmo que obtuvo mejores indicadores para la identificación de características para los clientes que no efectúan el pago de sus tarjetas de créditos, este trabajo se enfocó en la técnicas de minería de datos mediante el aprendizaje supervisado y no supervisado, en cuanto al desbalanceo de datos uso el algoritmo SMOTE para tratar este problema. Usó las métricas precisión o accuracy, sensibilidad, AUC.

CAPITULO IV METODOLOGÍA

4.1 Comprensión y entendimiento del negocio

En referencia a la situación de negocio en la institución educativa, se cuenta con una base de datos que se recopila al inicio del ciclo académico cuyo objetivo es conocer el perfil de nuestros estudiantes y también de los datos de la encuesta satisfacción de los alumnos con el ciclo académico. Sin embargo, hasta el momento no existe ningún estudio sobre el perfil de los estudiantes insatisfechos, de los que se puedan sacar conclusiones para clasificar a estos estudiantes o para hacer predicciones.

El objetivo principal de este trabajo es clasificar a los estudiantes insatisfechos antes que efectivamente lo sea.

- **Población y Muestra**

El universo de este trabajo fue la población de estudiantes de los ciclos de estudio dados en el periodo 2019 II de la institución educativa ubicada en Lima Metropolitana. La unidad de análisis es un estudiante de los ciclos de estudio dados en el periodo 2019 II de la institución educativa ubicada en Lima Metropolitana.

La muestra final fue de 5683 estudiantes.

• Muestra de entrenamiento: para este trabajo se decidió aplicar el 80 por ciento del total de la muestra, lo cual equivale a 4546 estudiantes.

• Muestra de validación: en cuanto a la muestra de testeo o validación resulto el 20 por ciento de la muestra equivalente a 1137 estudiantes.

4.2 Comprensión y Recolección de datos:

- **Instrumentos**

Para poder realizar este trabajo se utilizaron los siguientes instrumentos:

- Cuestionario del perfil del estudiante, este cuestionario es tomado al inicio del ciclo de estudio mediante la plataforma o intranet. Con este instrumento nuestro objetivo es conocer e identificar características demográficas, académicas de nuestros alumnos para brindarles un mejor proceso de enseñanza y aprendizaje de acuerdo a sus necesidades y expectativas ayudándoles en su formación integral.
- Cuestionario de satisfacción del estudiante con el ciclo de estudio; A diferencia del instrumento del perfil, este se toma casi al finalizar el ciclo de estudio para conocer su satisfacción con los diversos servicios que se le brinda al alumno tales como material de estudio, evaluaciones, desempeño del docente entre otros pero principalmente y de manera general conocer su satisfacción con el ciclo de estudio.

- **Recolección de datos**

La institución educativa brinda al estudiante una intranet la cual es una plataforma dinámica, interactiva y fácil de navegar, donde tiene toda la información de su ciclo académico y de sus cursos de una manera ordenada, a la cual puede acceder desde cualquier dispositivo. Los cuestionarios de perfil y de satisfacción del estudiante son aplicados desde esta intranet según un cronograma establecido, luego son invocados y motivados por los auxiliares en sus aulas para el correcto llenado de la encuesta. Toda esta información se almacena en una base de datos administrada por el área de sistemas de la institución para luego ser enviada al área de estadística en donde se realiza su procesamiento.

- **Descripción de los datos**

Variables

Variable Dependiente: Es la satisfacción del estudiante al final del ciclo. Esta codificado como 0: Satisfecho y 1: Insatisfecho.

Variables Independientes: Para definir y determinar las variables explicativas o predictoras se usaron variables sociodemográficas y académicas, entre las cuales tenemos:

Ciclo: en la institución se brindan a los estudiantes diferentes ciclos académicos, de los cuales ellos eligen de acuerdo a sus objetivos. Variable cualitativa nominal

Tipo de colegio: se refiere a la naturaleza del colegio en el cual el estudiante ha estudiado o estudia actualmente puede ser nacional o particular.

Horas de estudio: Se le consultó al alumno si dedicaba horas de repaso luego de las clases, esta variable es dicotómica ya sea positiva o negativa.

Preparación académica: se refiere a si el estudiante tiene alguna experiencia en preparación antes de este ciclo.

Materia con mayor dificultad: se le pregunto al estudiante en que materia tiene mayor dificultad académica, esta variable esta recodificada en materias de letras y materias de matemáticas.

Motivo de estudio: se le consultó al estudiante si se había matriculado por voluntad propia o por obligación de sus padres o algún familiar.

Edad: como vemos esta variable esta recodificada en estudiantes menores o que tienen 17 años y alumnos mayores de 17 años.

Tabla 2

Variables seleccionadas

Variable	Naturaleza	Tipo	Criterio de medición	Escala
Y: Satisfacción	Cualitativa	Dicotómica	0= Satisfecho 1= Insatisfecho	Nominal
X1: Ciclo	Cualitativa	Dicotómica	1= Ciclo A 2= Ciclo B 3= Ciclo C	Nominal
X2: Tipo de colegio	Cualitativa	Dicotómica	0: Privado 1: Estatal	Nominal
X3: ¿Dedicas horas de estudio?	Cualitativa	Dicotómica	0: Si 1: No	Nominal
X4: ¿Tiene preparación académica?	Cualitativa	Dicotómica	0: Si 1: No	Nominal
X5: Materia con mayor dificultad	Cualitativa	Dicotómica	0: letras 1: Matemáticas y ciencias	Nominal
X6: ¿Postulo a la universidad?	Cualitativa	Dicotómica	0: Si 1: No	Nominal
X7: Motivo de estudio	Cualitativa	Dicotómica	0: Obligado 1: Voluntad propia	Nominal
X8: ¿Actualmente labora?	Cualitativa	Dicotómica	0: No 1: Si	Nominal
X9: ¿Con quién vive?	Cualitativa	Dicotómica	0= Familia nuclear 1= Otro	Nominal
X10: Edad recodificada	Cualitativa	Dicotómica	0: mayor 17 años 1: menor igual a 17 años	Nominal

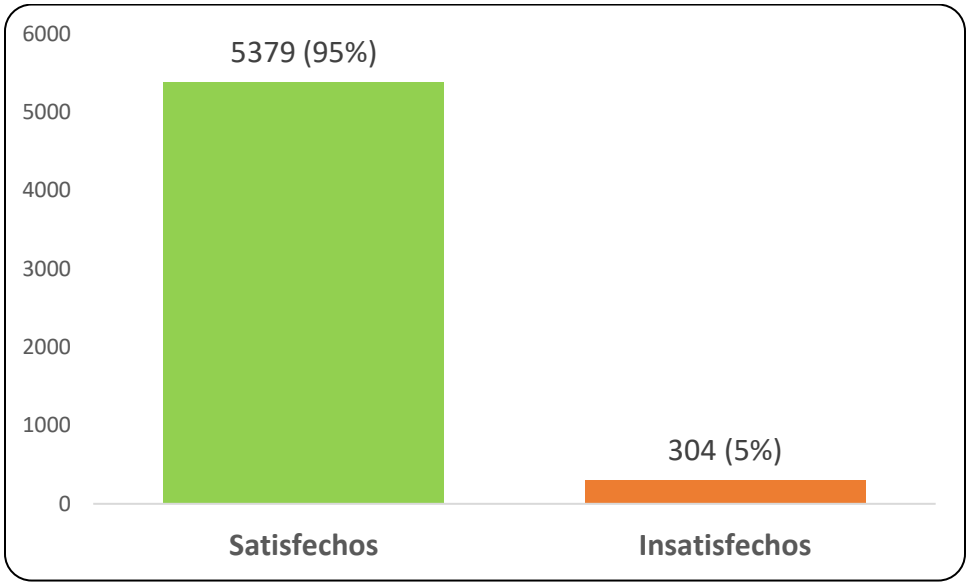
Nota. Elaboración propia

- **Exploración de los datos**

El gráfico 5 muestra la distribución de la variable de interés que como sabemos es una variable dicotómica nos referimos a la satisfacción del estudiante en la institución, 1= insatisfecho y 0= satisfecho, se puede observar claramente la diferencia notoria entre ambas categorías, este es un claro ejemplo de datos desbalanceados. Podemos decir que el 95% de estudiantes se encuentran satisfechos con el ciclo académico mientras que el 5% de alumnos indican estar insatisfechos. Nuestros datos son bastantes desproporcionados donde la clase minoritaria es insatisfecha.

Gráfico 1

Distribución de la satisfacción del estudiante con el ciclo académico



Nota. Elaboración propia

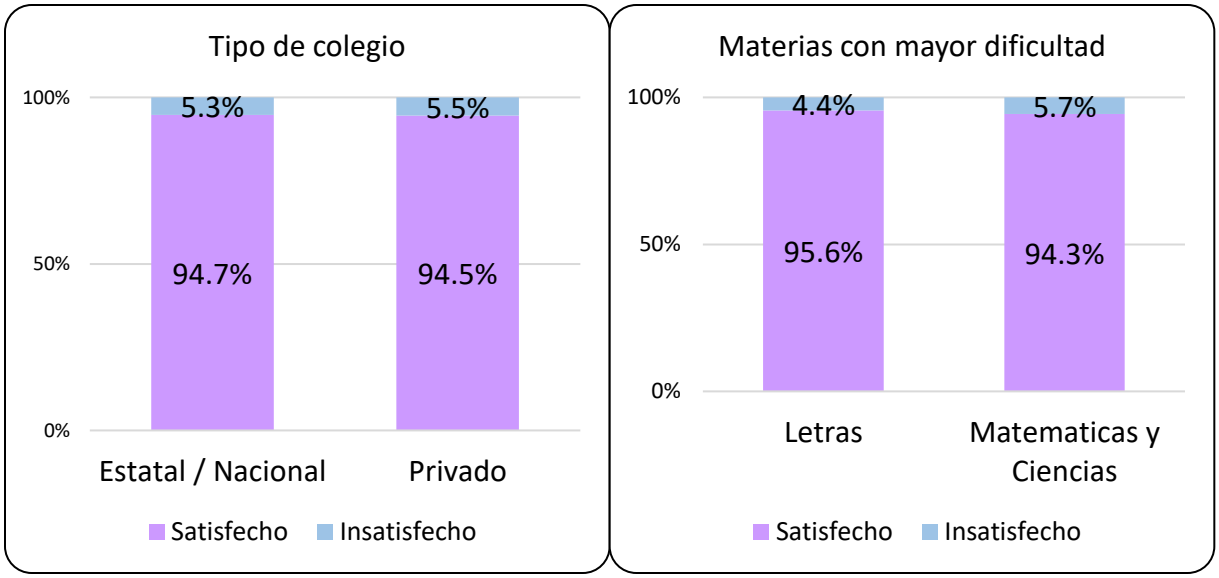
Al revisar los datos podemos concluir que estos están completos. En cuanto a los valores nulos, no hay ya que desde que la encuesta se pone en la intranet del estudiante se dispone que el alumno tiene que responder todas las preguntas sino no puede continuar con la encuesta.

Antes de iniciar con el modelamiento, se hará un análisis descriptivo de algunas variables independientes o predictoras, debemos indicar que todas son categóricas.

En cuanto al tipo del colegio del estudiante podemos ver que este tiene 2 categorías colegio estatal y colegio privado, si observamos la proporción de satisfacción dentro de cada categoría se visualiza que es muy homogénea y se encuentra alrededor del 5%. En lo que se refiere a las materias con mayor dificultad para el estudiante este se re categorizó en cursos de letras y en cursos de matemáticas y ciencias también se observa una proporción muy similar de estudiantes satisfechos e insatisfechos en ambas categorías. (Figura 9)

Figura 9

Distribución de las variables Tipo de colegio y Materias que se les dificulta vs la Satisfacción del estudiante



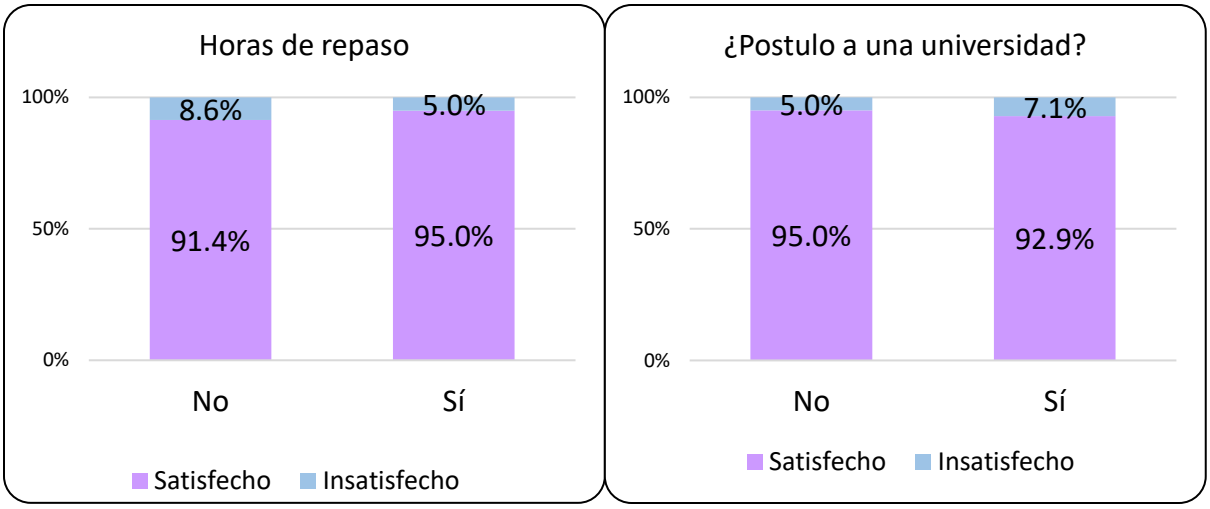
Nota. Elaboracion propia

En cuanto a si repasa o no luego de clases podemos observar que la distribución de la insatisfacción del estudiante dentro cada categoría es muy similar, dentro de la clase no repasa la insatisfacción es un poco mayor (8.6%) comparada con la clase si repasa (5%). En lo que se refiere a si postulo a la universidad se visualiza que dentro de la categoría no postulo la

insatisfacción del estudiante con el ciclo de estudio es 5% mientras que para la categoría si postulo la insatisfacción es de 7%. (Figura 10)

Figura 10

Distribución de las variables horas de repaso y si postulo a la universidad versus Satisfacción del estudiante

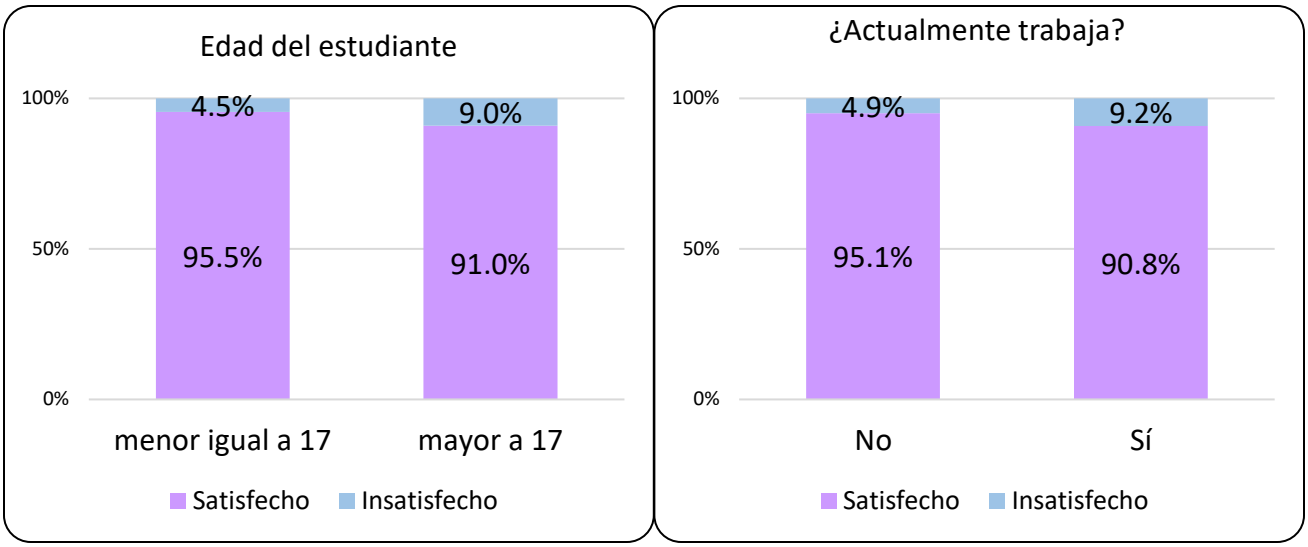


Nota: Elaboración propia

Respecto a la edad del estudiante la cual ha sido recodificada en menor e igual a 17 años y mayor a 17 años se observa que la distribución de la insatisfacción del estudiante dentro cada categoría menor e igual de 17 años es del 5%, mientras que en la categoría mayor de 17 años es un poco mayor (9%). En cuanto a si trabaja el alumno se visualiza que dentro de la categoría no trabaja la insatisfacción del estudiante con el ciclo de estudio es 4.9% mientras que para la categoría si trabaja la insatisfacción es de 9.2%. (Figura 11)

Figura 11

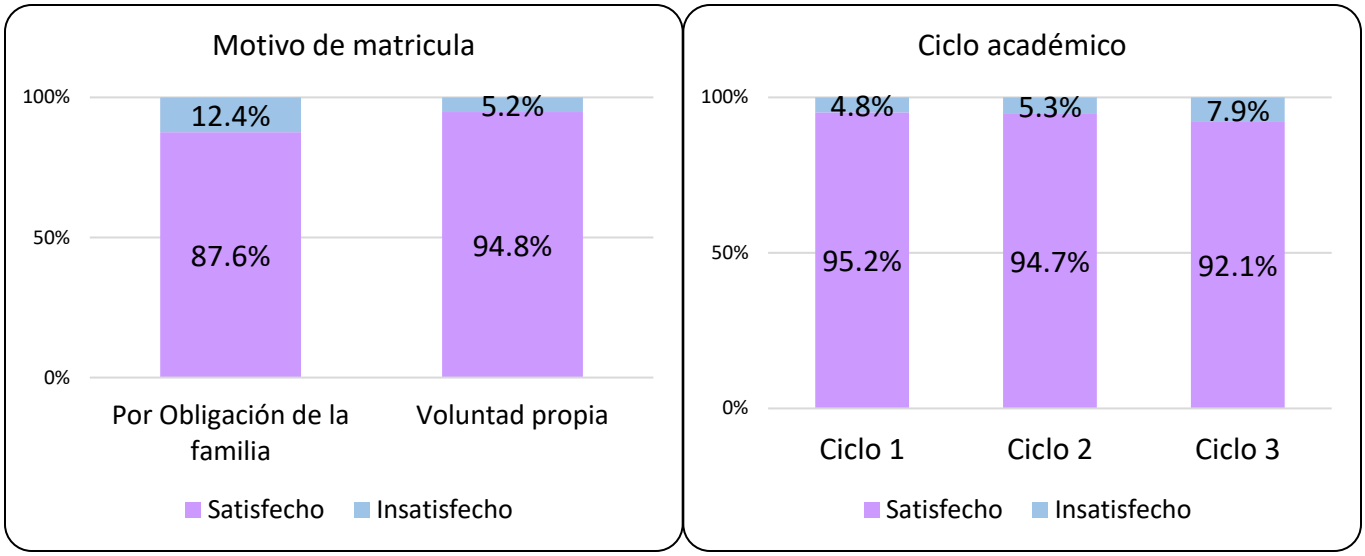
Distribución de las variables edad y si trabaja el estudiante versus Satisfacción del estudiante



Nota: Elaboración propia

En cuanto al motivo de matrícula si fue por voluntad propia o por obligación de la familia se observa que la distribución de la insatisfacción del estudiante es un poco mayor en la categoría por obligación de la familia (12.4%) comparada con la insatisfacción dentro de la categoría por voluntad propia (5.2%). En lo que se refiere al ciclo académico vemos que son tres ciclos diferentes, la distribución de la insatisfacción del estudiante con el ciclo de estudio es similar en todos los ciclos siendo un poco mayor en el ciclo 3 (7.9%). (Figura 12)

Figura 12 *Distribución de las variables motivo de matrícula y ciclo académico versus Satisfacción del estudiante*



Nota: Elaboración propia

4.3 Preparación y tratamiento de datos

- Tratamiento de datos

En esta fase se prepara los datos para aplicar la técnica de minería de datos que se va a emplear sobre ellos es decir regresión logística. En términos de registros y variables, se van a usar todos los que compone la base de datos, sin embargo, hay variables que no tienen asociación con la insatisfacción del estudiante, por lo tanto, no son necesarios para nuestro modelo de clasificación, por lo que se puede prescindir de algunas de ellas.

Se realizó el test de Chi cuadrado para ver la relación que existe entre las variables explicativas con respecto a la variable respuesta Y: Insatisfacción. De acuerdo a estos resultados se excluyeron algunas variables que no resultaron significativas con la insatisfacción del alumno. Al final se quedó con las siguientes variables (Tabla 3):

Tabla 3

Variables cualitativas respecto a la Insatisfacción del estudiante

Variable	Chi cuadrado	P valor
X1: Ciclo	6.5352	0.0381
X2: Dedicas horas de estudio	10.913	0.0009
X3: ¿Postuló a la universidad?	7.3534	0.0067
X4: Motivo de estudio	9.0798	0.0026
X5: ¿Actualmente labora?	16.594	0.000
X6: ¿Con quién vive?	10.743	0.001
X7: Edad recodificada	34.358	0.000

Fuente: Elaboración propia

- Selección de muestras: Training y Testing

Primero debemos seleccionar una muestra de entrenamiento y una muestra para validar nuestro modelo, en nuestro caso se decidió 80% y 20% para cada una de las muestras respectivamente de nuestra base original. La construcción y el aprendizaje del modelo se desarrollará con el 80%, luego el modelo será testeado con el 20%, es decir con la data testing. Se observa en la tabla 4 que la proporción de cada categoría de la variable respuesta en la base total es del 95% frente a un 5%.

Tabla 4

Tabla de contingencia de la base original

Variable	Cantidad	% del total
0: Satisfacción	5379	94.7%
1: Insatisfacción	304	5.3%

Nota: Elaboración propia

El sostener y verificar que se cumpla la proporción dentro de cada categoría de la variable respuesta dada en la base total es muy importante y fundamental, por lo tanto, esa misma proporción que es aproximadamente 95% vs 5% se debe cumplir en ambas muestras tanto trainnig y testing.

Tabla 5

Tabla cruzada de la base para el train y el test

Base	Variable	Cantidad	% del total
Base training	0: Satisfacción	4295	94.5%
	1: Insatisfacción	251	5.5%
Base testing	0: Satisfacción	1084	95.3%
	1: Insatisfacción	53	4.7%

Nota: elaboración propia

4.4 Modelado: Técnica de clasificación regresión logística

- **Regresión logística sin balanceo de datos**

Se realizó un modelo logístico (tabla 6) con el propósito de establecer que variables influyen en la insatisfacción del estudiante con el ciclo académico. Para el ajuste del modelo solo se utilizo la muestra de entrenamiento, no se realizó el balanceo de las categorías de la variable respuesta, como sabemos nuestra categoría minoritaria es insatisfacción=1. De acuerdo a la tabla, se puede observar que las variables Ciclo y la edad del estudiante resultaron significativas (p-valor < 0,05).

Tabla 6

Modelo de Regresión logística con data original

Variables	Coeficientes	Pr(>|z|)	
(Intercepto)	-1.69326	0.000	(*)
Ciclo2	0.02422	0.870	
Ciclo3	0.61050	0.006	(*)
Horas repaso (si)	-0.36712	0.072	
Postulo (si)	0.19847	0.278	
Motivo estudio (voluntad propia)	-0.56214	0.1482	
Labora (si)	0.3615	0.100	
¿con quién vive? (otro familiar)	0.2649	0.069	
Edad (menor e igual a 17 años)	-0.6916	0.000	(*)

Fuente: Elaboración propia con R

En cuanto a la tabla de clasificación o matriz de confusión resultante de la regresión logística sin aplicar el algoritmo SMOTE (Tabla N°7) para el problema del desbalance de datos, observamos una tendencia de clasificación hacia la categoría mayor, es decir clasifica correctamente a los estudiantes satisfechos con lo cual el error de clasificación disminuye y clasifica incorrectamente a los estudiantes insatisfechos, no clasifica correctamente a ningún estudiante insatisfecho, con lo cual estos resultados no cumplen el objetivo principal de este trabajo.

Tabla 7

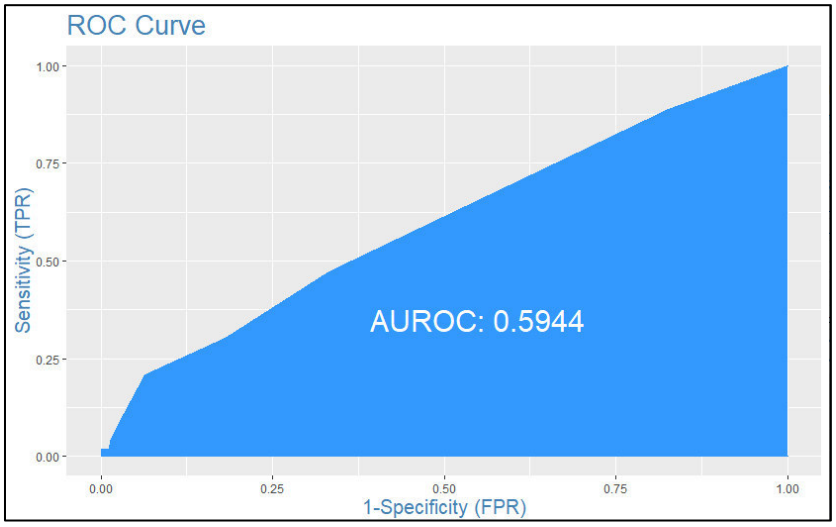
Tabla de clasificación para datos desbalanceados al realizar la regresión logística

	Predicho		
Real	Insatisfecho (+)	Satisfecho (-)	Total
Insatisfecho	0	65	65
Satisfecho	0	1356	1356
Total	0	1421	1421

Fuente: Elaboración propia con R

Figura 13

Curva ROC Regresión logística sin balancear los datos



Nota. Elaboración propia

En la tabla 8 se presentan los indicadores propuestos para la evaluación del modelo, se observa que en cuanto a la exactitud o accuracy es un valor alto pero como ya lo hemos comentado esta métrica no es adecuada para cuando nuestros datos se encuentran desbalanceados como es nuestro caso, la sensibilidad es cero ya que nuestro modelo no clasifico ningún estudiante insatisfecho correctamente, caso contrario sucede con la especificidad el cual es 1 ya que todos los estudiantes satisfechos fueron clasificados correctamente, por lo tanto nuestro modelo no clasifica correctamente a los casos positivos (insatisfechos) y por último una curva AUC del 59%.

Tabla 8

Métricas para evaluar el modelo de la data original

	Regresión Logística
Sensibilidad	0
Especificidad	1
Accuracy	0.954
Precisión	----
AUC	0.594

Fuente: Elaboración propia

- **Modelo logístico usando algoritmo SMOTE**

Para realizar la simetría de nuestra data se utilizó el algoritmo SMOTE que incluye las técnicas de sobre muestreo, quedando de la siguiente manera la proporción (Tabla 9)

Tabla 9

Sobre muestreo de datos para aplicar SMOTE

Categoría	Estudiantes	% estudiantes
Satisfecho	12885	42.9 %
Insatisfecho	17180	57.1 %
Total	30065	100 %

Fuente: Elaboración propia

Este modelo con datos balanceados se muestra en la Tabla 10. Para la estimación del modelo se trabajó únicamente con la muestra de entrenamiento, se realizó el balanceo de las categorías de la variable respuesta, de acuerdo a la tabla 10 se puede observar que las variables Ciclo, horas de repaso, si postulo a la universidad, motivo por el cual estudian, si laboran actualmente y la edad del estudiante resultaron significativas (p-valor < 0,05).

Tabla 10

Regresión logística utilizando el algoritmo SMOTE

Variables	Coeficientes	Pr(>|z|)
(Intercepto)	1.14735	0.000 (*)
Ciclo2	0.01603	0.549
Ciclo3	0.82052	0.000 (*)
Horas repaso (si)	-0.38077	0.000 (*)
Postulo (si)	0.18751	0.000 (*)
Motivo estudio (voluntad propia)	-0.62486	0.000 (*)
Labora (si)	0.38687	0.000 (*)
¿con quién vive? (otro familiar)	0.16322	0.000 (*)
Edad (menor 17 años)	-0.60667	0.000 (*)

Fuente: Elaboración propia con R

En cuanto a la matriz de confusión para la regresión logística utilizando el algoritmo SMOTE (Tabla N 11) se observa que el modelo presenta una tendencia de clasificación hacia la categoría minoritaria, es decir clasifica correctamente a los estudiantes insatisfechos, lo cual es el objetivo principal de este estudio.

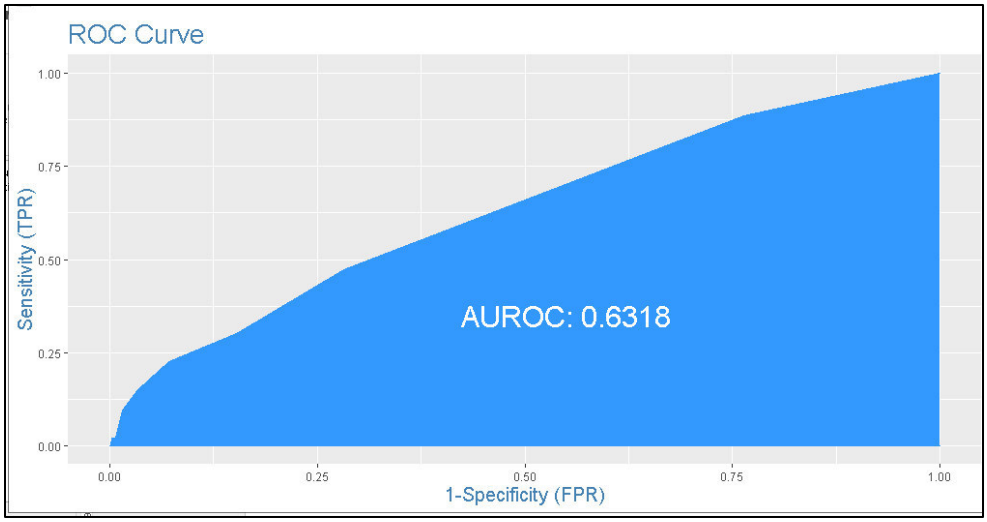
Tabla 11

Tabla de clasificación con SMOTE

	Predicho		
Real	Insatisfecho	Satisfecho	Total
Insatisfecho	55	10	65
Satisfecho	553	803	1356
Total	608	813	1421

Nota: Elaboración propia

Figura 14 *Curva ROC con SMOTE*



Nota: Elaboración propia

En la tabla 12 se presentan los indicadores para la evaluación del modelo ajustado, se observa que en cuanto a la sensibilidad resulto un valor alto con lo cual podemos decir que este modelo clasifica correctamente a los estudiantes insatisfechos, la sensibilidad es 0.41 ya que no

clasifica correctamente a los casos negativos (estudiantes satisfechos), la exactitud del modelo es .60, la precisión es 0.09 y una curva AUC del 63%.

Tabla 12

Metricas para evaluación del modelo aplicando SMOTE

	Regresión Logística
Sensibilidad	0.85
Especificidad	0.41
Accuracy	0.60
Precisión	0.09
AUC	0.632

Fuente: Elaboración propia

- **Comparación de clasificadores e indicadores**

En la tabla 13 se observa los indicadores para los modelos obtenidos con la regresión logística sin balanceo y con balanceo de datos. Para poder identificar qué modelo es el mejor para nuestro TSP, debemos analizar las métricas de clasificación propuestas para datos desbalanceados, sensibilidad, la precisión y la curva ROC. En cuanto a la sensibilidad podemos observar que cuando aplicamos el algoritmo SMOTE este indicador es alto, caso contrario sucede cuando no aplicamos este algoritmo, debemos prestar mucha atención a este indicador ya que nuestro objetivo principal es clasificar a la clase minoritaria. En cuanto a la precisión vemos que es baja al realizar la regresión aplicando el algoritmo SMOTE esto podemos interpretarlo como que nuestro modelo identifica bien a los estudiantes satisfechos, pero también incluye muestras de la clase satisfecha.

La exactitud baja cuando aplicamos este algoritmo, pero recordemos que esta medida no es confiable para datos desbalanceados, en cuanto a la precisión resulta ser baja con SMOTE, el AUC del modelo balanceado tiene un superior valor. Por lo tanto, al evaluar estos indicadores

podríamos decir que el modelo ajustado aplicando el sobre muestreo tiene mejor rendimiento al momento de clasificar a los estudiantes insatisfechos (clase minoritaria).

Tabla 13

Comparación para evaluación de modelos

	Regresión Logística	Regresión logística SMOTE
Sensibilidad	0	0.85
Especificidad	1	0.41
Exactitud (Accuracy)	0.954	0.60
Precisión	----	0.09
Curva AUC	0.594	0.632

Fuente: Elaboración propia con R

4.5 Evaluación de los Objetivos del TSP

El objetivo principal de este TSP es diseñar un modelo de clasificación a partir de los datos de ingreso de los estudiantes para poder identificar las variables que inciden en la insatisfacción con el ciclo. Como vimos la regresión logística con SMOTE es el mejor clasificador comparado con el modelo de la regresión logística sin SMOTE. En cuanto a la interpretación de los coeficientes obtenidos tenemos que:

- La chance o probabilidad de sentirse insatisfecho del estudiante que se encuentra matriculado en el ciclo integral con respecto al ciclo 1 es 2.27 veces más, esto es manteniendo las demás variables constantes.
- El dedicar horas de repaso luego de las clases disminuye la probabilidad que el estudiante se sienta insatisfecho a que se sienta satisfecho, en un 68% respecto a cuándo no repasa luego de clases, manteniendo las demás variables constantes.

- La chance o probabilidad de sentirse insatisfecho del estudiante que ha postulado a alguna universidad con respecto al que no ha postulado es 1.2 veces más, esto es manteniendo las demás variables constantes.
- El haberse matriculado por voluntad propia disminuye la chance que el estudiante se sienta insatisfecho a que se sienta satisfecho, en un 54% respecto a cuándo se matricula por obligación de sus padres, manteniendo las demás variables constantes.
- La chance o probabilidad de sentirse insatisfecho del estudiante que si labora con respecto al que no labora es 1.47 veces más, esto es manteniendo las demás variables constantes.
- El vivir con otro familiar y no con su familia nuclear aumenta la probabilidad de sentirse insatisfecho del estudiante en 1.2 veces más, permaneciendo las demás variables constantes.
- El tener menos o 17 años reduce la probabilidad que el estudiante se sienta insatisfecho a que se sienta satisfecho, específicamente en un 55% respecto a cuándo tiene más de 17 años, permaneciendo las otras variables constantes.

- **Variables que están asociadas e influyen en la insatisfacción del estudiante con el ciclo**

  Las variables que resultaron significativas con la insatisfacción del estudiante son: el ciclo de estudio, las horas de repaso, si postulo a la universidad, motivo por el cual estudian, si laboran actualmente, con quien vive y la edad del estudiante. (p-valor < 0,05).

- **Predicción.**

Se desarrollo la predicción de nuevos estudiantes con el modelo que obtuvo el mejor rendimiento. La ecuación final es:

$$Y = 1/(1 + \exp(-(1.14+0.016(\text{ciclo2}) +0.82(\text{ciclo3}) - 0.38(\text{horas que repasa}) + 0.19(\text{si postulo}) + 0.16(\text{con quien vive}) - 0.62(\text{motivo de matrícula}) +0.39(\text{si labora})-0.61(\text{edad})).$$

4.6 Implantación

La implementación del presente proyecto es una fase importante ya que gracias a este estudio se puede conocer desde el ingreso del estudiante a la institución variables que inciden en la insatisfacción con el ciclo en general. Como se dijo al inicio del trabajo es muy importante conocer y clasificar al estudiante insatisfecho antes que realmente lo sea, esto permitirá a la institución estar atento a este perfil de alumno para brindar mejores servicios y no caer de repente en la fuga o deserción de este estudiante.

El volumen de los datos en proceso es grande motivo por el cual se seleccionarán muestras, pero también sabemos que los alumnos insatisfechos son una proporción pequeña del total de estudiantes. La minería de datos se realizará al inicio de los ciclos de estudio como un plan de supervisión.

CONCLUSIONES

Las conclusiones de este trabajo son:

- El modelo de regresión logística aplicado en este TSP es una buena técnica cuando nuestro objetivo es clasificar y predecir, pero se ha podido evidenciar que los resultados son diferentes al aplicar y no aplicar el algoritmo SMOTE para balancear nuestra data.
- Según los indicadores de desempeño que se utilizaron en este estudio para evaluar la capacidad clasificatoria del modelo, es la regresión logística con SMOTE quien tuvo el mejor rendimiento para clasificar a los estudiantes insatisfechos con su ciclo de estudio.
- Al comparar la sensibilidad (indicador de clasificación parcial) de ambos modelos es la regresión logística con algoritmo SMOTE quien resulto superior; se concluye que el segundo es el mejor modelo ya que obtiene la sensibilidad más alta (85%).
- El perfil del estudiante que indica estar insatisfecho con el ciclo de estudio se caracteriza por pertenecer al ciclo 3, no repasar después clase, tener más de 17 años, que actualmente se encuentre laborando que viva con otro familiar y que estudie por obligación de su familia.

RECOMENDACIONES

- Se ha evidenciado que el problema de datos desbalanceados es muy importante al momento de construir un modelo con este tipo de datos ya que si trabajamos sin ninguna técnica para solucionar esta dificultad podríamos obtener resultados que nos son válidos y pensar que el modelo clasifica bien cuando no lo hace.
- Se recomienda realizar mayores investigaciones sobre técnicas para los datos desbalanceados, no solo SMOTE, también técnicas de minería de datos que son útiles cuando nuestros datos presentan esta característica, en la actualidad se ha adaptado diferentes modelos de Machine Learning.
- En cuanto a la insatisfacción del estudiante con su ciclo de estudio se pudo identificar el perfil del estudiante insatisfecho, otro estudio muy interesante seria elaborar el perfil del estudiante desertor o también el perfil del estudiante de acuerdo a su rendimiento académico.

REFERENCIAS

Bramer, M. (2007). *Principles of data mining* (Vol. 131). London: Springer. DOI: 10.1007/978-1-84628-766-4.

Candia Oviedo, D. I. (2019). *Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático.*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0:* Guía de minería de datos paso a paso. SPSS Inc., (2000)

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Researchers. [Gráficos ROC: notas y consideraciones prácticas para investigadores] http://web.archive.org/web/20151002215629/http://home.comcast.net/~tom.fawcett/public _html/papers/ROC101.pdf.

Fawcett, T. (25 de agosto de 2016). Learning from Imbalanced Classes. [Aprender de clases desequilibradas]. *Silicon Valley Data Science*. Obtenido de https://svds.com/learning-imbalanced-classes/

Gutiérrez, J. y Molina, B. (2016). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. Revista Ontare. 3. 33. 10.21158/23823399.v3.n2.2015.1440.

Haibo, H.; Yunqian, M. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications.* Hoboken, New Jersey, John Wiley & Sons.

Hosmer, D. W., & Lemeshow, S. (2000). "Introduction to the logistic regression model. Applied Logistic Regression, Second Edition, pags 1-30.

Jaramillo, A., & Paz Arias, H. P. (2015). *Aplicación de técnicas de minería de datos para determinar las interacciones de los estudiantes en un entorno virtual de*

aprendizaje. Revista Tecnológica - ESPOL, 28(1). Recuperado a partir de http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351

Kunal J. (2016). Practical Guide to deal with Imbalanced Classification Problems in R. Analytics Vidhya. Learn Everything About Analytics. Disponible en: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-dealimbalanced-classification-problems/

Lizares Castillo, M. (2017). *Comparación de modelos de clasificación: Regresión logística y árboles de clasificación para evaluar el rendimiento académico.* [Tesis de Licenciatura, Universidad Nacional Mayor de San Marcos] oai:cybertesis.unmsm.edu.pe:20.500.12672/7122

Martínez Raya, J. M. (2019) *Modelos de clasificación para incidencias en entornos industriales con datos no balanceados.* [Tesis de Maestría, Universidad Oberta de Catalunya]. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/99386/7/jmartinezrayTFM0619memoria.pdf

Meza Rodríguez, A. R. (2018) *Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo Adaboost desbalanceado y la regresión logística asimétrica.* [Tesis de Maestría, Universidad Nacional Agraria La Molina] https://repositorio.lamolina.edu.pe/handle/UNALM/3245/

Moreno, J & Rodriguez, Daniel & Sicilia, M. & Riquelme, José & Ruiz, Y. (2009). *SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias.* Recuperado de: https://www.researchgate.net/publication/229045207

Nieto, S. (2010). *Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio.* Universidad Autónoma Metropolitana. https://doi.org/10.4067/S0071-17132000003500023

Obregón Sandoval, J. (2016) *Desarrollo de una Herramienta de Diagnóstico de Fallos en Motores de Inducción Mediante la técnica Adaboost*. [Tesis de Maestría, Universidad de Valladolid]. http://uvadoc.uva.es/handle/10324/18912/

Pariona Huarhuachi, J. C. (2017). *Clasificación de fuga de clientes en una entidad financiera utilizando el algoritmo Smote para datos desbalanceados en una regresión logística*. [Tesis de Licenciatura, Universidad Nacional Agraria La Molina] https://repositorio.lamolina.edu.pe/handle/UNALM/3329/

Powers, D. 2008. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. [Evaluación: desde la precisión, la recuperación y el factor F hasta la ROC , la información, el marcado y la correlación] https://arxiv.org/abs/2010.16061

Rouse M. (2008). What type of data mining has your organization embraced? AWS analytics tools help make sense of big data. Disponible en: http://searchsqlserver.techtarget.com/definition/data-mining

Serna Pineda, S. (2009). *Comparación de árboles de regresión y clasificación y regresión logística.* [Tesis de Maestría, Universidad Nacional de Colombia] https://repositorio.unal.edu.co/handle/unal/2421

Sierra, B. (2006). Aprendizaje Automático: conceptos básicos y avanzados, Aspectos prácticos utilizando el software WEKA, Madrid, España, Editorial PEARSON PRENTICE HALL.

Tharwat, A. (2018). Classification assessment methods. Applied Computing and Informatics. doi: 10.1016/j. aci.2018.08.003

Wu, Z., Lin, W. y Ji, Y. (2018). An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics. in IEEE Access 6: 8394-8402. doi: 10.1109/ ACCESS.2018.2807121

ANEXO

Programa en R utilizado

```
#paquetes necesarios para este modelo
rm(list=ls())
library(caret)
library(tidyverse)
library(broom)
library(ISLR)
library(GGally)
library(modelr)
library(pROC)
library(cowplot)
library(OneR)
library(rlang)
library(dplyr)
library(readxl)
######lectura de data

library(readxl)

reg <- read_excel("base1.xlsx")
reg
head(reg)
str(reg)
#dependencia
chisq.test(reg1$sat,reg1$col)
chisq.test(reg1$sat,reg1$tercol)
chisq.test(reg1$sat,reg1$mat)
chisq.test(reg1$sat,reg1$lugar)
chisq.test(reg1$sat,reg1$CICLO)
chisq.test(reg1$sat,reg1$hor)
```

```
chisq.test(reg1$sat,reg1$horep)
chisq.test(reg1$sat,reg1$prep)
chisq.test(reg1$sat,reg1$pos)
chisq.test(reg1$sat,reg1$edadrec)
chisq.test(reg1$sat,reg1$estu)
chisq.test(reg1$sat,reg1$viv)
chisq.test(reg1$sat,reg1$lab)

###seleccion de variables de la base de datos final
table(reg1$sat)
prop.table(table(reg1$sat))*100
reg1[,1:ncol(reg1)] <- lapply(reg1[,1:ncol(reg1)],as.factor)
reg1[,1:ncol(reg1)] <- lapply(reg1[,1:ncol(reg1)],as.numeric)
reg1[,1:ncol(reg1)] <- lapply(reg1[,1:ncol(reg1)],as.factor)
summary(reg1)
#####Particion de la data
train_size <- floor(0.75*nrow(reg1))
set.seed(123)
train_reg1_MAS <- sample(seq_len(nrow(reg1)), size =   train_size)
ytrain <- reg1$sat[train_reg1_MAS]
ytest <- reg1$sat[-train_reg1_MAS]
### se escoge el 80% de la muestra total para training
train_MAS <- reg1[train_reg1_MAS,]
train_MAS
### se escoge el complemento: 20% para la muestra de testing
test_MAS <- reg1[-train_reg1_MAS,]
prop.table(table(train_MAS$sat))*100
table(train_MAS$sat)
prop.table(table(test_MAS$sat))*100
table(test_MAS$sat)
```

```
table(train_MAS$sat)
prop.table(table(train_MAS$sat))


######Modelo logit

train <- glm(sat ~ . , family = binomial, data= train_MAS)
summary(train)
####EVALUACION DEL MODELO
predicted_value <- predict(train,test_MAS,type = "response")
predicted_class <- ifelse(predicted_value>0.5, "2","1")
performance_data<-data.frame(observed=test_MAS$sat,
                             predicted= predicted_class)
positive <- sum(performance_data$observed=="2")
negative <- sum(performance_data$observed=="1")
predicted_positive <- sum(performance_data$predicted=="2")
predicted_negative <- sum(performance_data$predicted=="1")
total <- nrow(performance_data)
data.frame(positive, negative,predicted_positive,predicted_negative)

tp<-sum(performance_data$observed=="2" & performance_data$predicted=="2")
tn<-sum(performance_data$observed=="1" & performance_data$predicted=="1")
fp<-sum(performance_data$observed=="1" & performance_data$predicted=="2")
fn<-sum(performance_data$observed=="2" & performance_data$predicted=="1")
data.frame(tp,tn,fp,fn)
accuracy <- (tp+tn)/total
error_rate <- (fp+fn)/total
sensitivity <- tp/positive
especificity <- tn/negative
precision <- tp/predicted_positive
npv <- tn / predicted_negative
data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)
```

```
#####CURVA ROC
roc_data <- function(model, test_MAS, step=0.01) {
  out<-data.frame()
  cut <- seq(step, 1, by = step)
  for (i in cut) {
    predicted_value <- predict(train,test_MAS,type = "response")
    predicted_class <- ifelse(predicted_value> i, "2","1")
    performance_data<-data.frame(observed=test_MAS$sat,
                                 predicted= predicted_class)
    predicted_positive <- sum(performance_data$predicted=="2")
    predicted_negative <- sum(performance_data$predicted=="1")
    positive <- sum(performance_data$observed=="2")
    negative <- sum(performance_data$observed=="1")
    total <- nrow(performance_data)
       tp<-sum(performance_data$observed=="2" & performance_data$predicted=="2")
    tn<-sum(performance_data$observed=="1" & performance_data$predicted=="1")
    fn<-sum(performance_data$observed=="2" & performance_data$predicted=="1")
       sensitivity <- tp/positive
    especificity <- tn/negative
    precision <- tp/predicted_positive
    npv <- tn / predicted_negative
    fpr<- fn/negative
       out<-rbind(out,c(i,1-especificity,sensitivity,especificity,precision,npv,fpr))
  }
  names(out)<-c("cut","1-Especificity","Sensivity","Especificity","Precision","npv","fpr")
  return(out)
}
roc_graph <- roc_data(train,test_MAS,step = 0.01)
plot(roc_graph$`1-Especificity`,roc_graph$Sensivity,xlab = "1-especificity", ylab =
"Sensitivity",type = "l")
```

```
index<-seq(1,nrow(roc_graph), by=nrow(roc_graph)*0.05)
text(roc_graph$`1-Especificity`[index],roc_graph$Sensivity[index], labels = roc_graph$cut[index], cex=0.6, pos=4)

############REGRESION LOGISTICA CON BALANCEO DE DATOS
#library(DMwR)
library(dplyr)
library(unbalanced)
table(train_MAS$sat)
X <- train_MAS %>% dplyr::select(-sat)
#X[,1:ncol(X)] <- lapply(X[,1:ncol(X)],as.factor)
X[,1:ncol(X)] <- lapply(X[,1:ncol(X)],as.numeric)
Y <- train_MAS$sat
proceso_smoote <- ubSMOTE(X, Y,k=5)
trainS<-cbind(proceso_smoote$X, sat=proceso_smoote$Y)
trainS[,1:ncol(trainS)] <- lapply(trainS[,1:ncol(trainS)],as.integer)
trainS[,1:ncol(trainS)] <- lapply(trainS[,1:ncol(trainS)],as.factor)
table(trainS$sat)
#####modelo

modeloS<-glm(sat ~ . , data=trainS, family=binomial(link="logit"))
summary(modeloS)
#####Evaluacion del modelo con SMOTE
predicted_value <- predict(modeloS,test_MAS,type = "response")
predicted_class <- ifelse(predicted_value>0.5, "2","1")
performance_data<-data.frame(observed=test_MAS$sat,
                             predicted= predicted_class)
positive <- sum(performance_data$observed=="2")
negative <- sum(performance_data$observed=="1")
predicted_positive <- sum(performance_data$predicted=="2")
```

```
predicted_negative <- sum(performance_data$predicted=="1")
total <- nrow(performance_data)
data.frame(positive, negative,predicted_positive,predicted_negative)
tp<-sum(performance_data$observed=="2" & performance_data$predicted=="2")
tn<-sum(performance_data$observed=="1" & performance_data$predicted=="1")
fp<-sum(performance_data$observed=="1" & performance_data$predicted=="2")
fn<-sum(performance_data$observed=="2" & performance_data$predicted=="1")
data.frame(tp,tn,fp,fn)
accuracy <- (tp+tn)/total
error_rate <- (fp+fn)/total
sensitivity <- tp/positive
especificity <- tn/negative
precision <- tp/predicted_positive
npv <- tn / predicted_negative
data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)

###### CAMBIANDO PUNTO DE CORTE

predicted_value <- predict(modeloS,test_MAS,type = "response")
predicted_class <- ifelse(predicted_value>0.75, "2","1")
performance_data<-data.frame(observed=test_MAS$sat,
                             predicted= predicted_class)
positive <- sum(performance_data$observed=="2")
negative <- sum(performance_data$observed=="1")
predicted_positive <- sum(performance_data$predicted=="2")
predicted_negative <- sum(performance_data$predicted=="1")
total <- nrow(performance_data)
data.frame(positive, negative,predicted_positive,predicted_negative)
tp<-sum(performance_data$observed=="2" & performance_data$predicted=="2")
tn<-sum(performance_data$observed=="1" & performance_data$predicted=="1")
fp<-sum(performance_data$observed=="1" & performance_data$predicted=="2")
```

```
fn<-sum(performance_data$observed=="2" & performance_data$predicted=="1")
data.frame(tp,tn,fp,fn)
accuracy <- (tp+tn)/total
error_rate <- (fp+fn)/total
sensitivity <- tp/positive
especificity <- tn/negative
precision <- tp/predicted_positive
npv <- tn / predicted_negative
data.frame(accuracy,error_rate,sensitivity,especificity,precision,npv)

####CURVA ROC
roc_data <- function(model, test_MAS, step=0.01) {
  out<-data.frame()
  cut <- seq(step, 1, by = step)
  for (i in cut) {
    predicted_value <- predict(modeloS,test_MAS,type = "response")
    predicted_class <- ifelse(predicted_value> i, "2","1")
    performance_data<-data.frame(observed=test_MAS$sat,
                                 predicted= predicted_class)
    predicted_positive <- sum(performance_data$predicted=="2")
    predicted_negative <- sum(performance_data$predicted=="1")
    positive <- sum(performance_data$observed=="2")
    negative <- sum(performance_data$observed=="1")
    total <- nrow(performance_data)
       tp<-sum(performance_data$observed=="2" & performance_data$predicted=="2")
    tn<-sum(performance_data$observed=="1" & performance_data$predicted=="1")
    fn<-sum(performance_data$observed=="2" & performance_data$predicted=="1")
      sensitivity <- tp/positive
    especificity <- tn/negative
    precision <- tp/predicted_positive
    npv <- tn / predicted_negative
```

```
      fpr<- fn/negative
          out<-rbind(out,c(i,1-especificity,sensitivity,especificity,precision,npv,fpr))
      }
      names(out)<-c("cut","1-Especificity","Sensivity","Especificity","Precision","npv","fpr")
      return(out)
    }
    roc_graph <- roc_data(modeloS,test_MAS,step = 0.01)
    plot(roc_graph$`1-Especificity`,roc_graph$Sensivity,xlab = "1-especificity", ylab =
"Sensitivity",type = "l")
    index<-seq(1,nrow(roc_graph), by=nrow(roc_graph)*0.05)
    text(roc_graph$`1-Especificity`[index],roc_graph$Sensivity[index], labels =
roc_graph$cut[index], cex=0.6, pos=4)
```