UNIVERSIDAD DE CUENCA

Facultad de Ciencias Químicas

Carrera de Ingeniería Industrial

Comparativa de modelos de clasificación para inferir la probabilidad de deserción estudiantil en la Facultad de Ciencias Químicas de la Universidad de Cuenca

Trabajo de titulación previo a la obtención del título de Ingeniera Industrial

Autora:

Karla Rafaela Palacios Alvear

CI: 0105503189

Correo electrónico: karafapalacios@gmail.com

Director:

Ing. Carlos Mauricio Sánchez Alvarracin

CI: 0102367653

Cuenca, Ecuador

17-febrero-2021

Resumen: En el presente trabajo se muestra la aplicación de modelos de clasificación comparativos, a través de variables específicas, para determinar la deserción universitaria respecto de los estudiantes de la Facultad de Ciencias Químicas de la Universidad de Cuenca. En este contexto, a través de la minería de datos se aplicaron dos modelos de clasificación: K- vecinos más próximos (knn) y regresión logística (rl) a fin de catalogar al alumnado de primer año en dos poblaciones, a saber: deserción o permanencia. Los datos fueron obtenidos de la ficha socioeconómica, presentada por los referidos estudiantes, desde el año 2014 hasta el 2018, además se identificaron los grupos poblacionales correspondientes a quienes abandonaron la carrera en el primer año y a quienes continuaron con sus estudios. Con base a esto, fue posible interrelacionar las variables para agrupar las mismas mediante el análisis de componentes principales (ACP). Los datos fueron separados para entrenamiento y validación de los modelos. Los sistemas fueron modelados en RapidMiner generando una matriz de confusión, lo que permitió determinar que el modelo knn presenta mejor exactitud de 73,30% frente a un 54,67% del modelo de Regresión Logística. Finalmente, se concluye que las principales causas de deserción son: el total ingreso, total egreso, mensual pago de arriendo, avalúo acumulado de vehículos, tipo de colegio. A través de la matriz de confusión se evaluaron los modelos (knn y rl) seleccionando al modelo knn como mejor opción. Por últimos se verificó que el modelo knn tiene un error del 20% respecto la realidad.

Palabras claves: Deserción Universitaria. Regresión Logística. K Vecinos más Cercanos (knn).

Abstract: This degree work shows an application of comparative classification models, through specific variables, to determine the university dropout of students from the Faculty of Chemical Sciences of the University of Cuenca. In this context, through data mining, two classification models were applied: K- nearest neighbors (knn) and logistic regression to classify first-year students into two populations: dropout or permanence. The data was obtained from the socio-economic record of the students from 2014 to 2018, in addition, the population groups corresponding to those who dropped out in the first year and those who continued with their studies were identified. Based on this, it was possible to interrelate the variables to group them through principal component analysis (PCA). The data were separated for training and validation of the models. The systems were modeled in RapidMiner generating a confusion matrix, which allowed determining that the knn model presents a better current of 73.30% compared to 54.67% of the Logistic Regression model. Additionally, it was concluded that the most relevant variables are those that make up the main component 1: total income, total expenses, monthly rent payment, type of high school, cumulative valuation of vehicles. Through the confusion matrix, the models (knn and rl) were evaluated, selecting the knn model as the best option.

Finally, it was verified that the knn model has an error of 20% with respect to reality.

Keywords: **Dropout. Logistic regression. K- Nearest Neighbors (knn).**

Indice del Trabajo

Cláusula de Propiedad Intelectual

Karla Rafaela Palacios Alvear autor/a del trabajo de titulación "Comparativa de modelos de clasificación para inferir la probabilidad de deserción estudiantil en la Facultad de Ciencias Químicas de la Universidad de Cuenca", certifico que todas las ideas, opiniones y contenidos expuestos en la presente investigación son de exclusiva responsabilidad de su autor/a.

Cuenca, 17 de Febrero del 2021

Karla Rafaela Palacios Alvear

C.I : 0105503189

Cláusula de licencia y autorización para publicación en el Repositorio Institucional

Karla Rafaela Palacios Alvear en calidad de autor/a y titular de los derechos morales y patrimoniales del trabajo de titulación "Comparativa de modelos de clasificación para inferir la probabilidad de deserción estudiantil en la Facultad de Ciencias Químicas de la Universidad de Cuenca", de conformidad con el Art. 114 del CÓDIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN reconozco a favor de la Universidad de Cuenca una licencia gratuita, intransferible y no exclusiva para el uso no comercial de la obra, con fines estrictamente académicos.

Asimismo, autorizo a la Universidad de Cuenca para que realice la publicación de este trabajo de titulación en el repositorio institucional, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Cuenca, 17 de febrero del 2021

Karla Rafaela Palacios Alvear

C.I : 0105503189

1. Introducción

El proceso de ingreso a las instituciones de educación superior en el Ecuador ha cambiado durante los últimos años. Hasta el año 2013, se regía según el proceso de admisión establecido internamente en cada institución. Posteriormente, por disposición normativa, las universidades debieron alinearse a la Ley Orgánica de Educación Superior (LOES). En el año 2014 se implementó el examen SER BACHILLER, así pues; en este examen los estudiantes obtenían la nota del examen de grado y el puntaje para acceder a la educación superior pública o privada. Para el año 2019 el puntaje de ingreso a la universidad se modificó a 70% de la nota del resultado del examen de SER BACHILLER y el 30% de la nota del examen de ingreso de cada universidad (Ministerio de Educación, 2018).

Diferentes estudios analizan la deserición universitario, por ejemplo, Sinchi y Gómez (2018), realizaron una investigación con 6854 estudiantes que ingresaron a las diferentes universidades de la ciudad de Cuenca en el periodo 2015-2016, siendo la Universidad de Cuenca la que presentó mayor porcentaje de abandono de sus estudiantes.

Con el objetivo de realizar un seguimiento a los posibles estudiantes desertores, resulta necesario analizar las variables que influyen en la decisión de retirarse. La deserción estudiantil debe ser analizada con grandes volúmenes de datos, para obtener modelos predictivos o descriptivos, según sea necesario. Bajo este concepto se encuentra la minería de datos, que se caracteriza por la existencia de técnicas matemáticas no tan simples de clasificación, basándose en el análisis de correlación de variables, a fin de determinar si un estudiante es un posible desertor o no.

En la Figura 1 se aprecian algunos algoritmos de la minería de datos enfocados a la clasificación, como son: knn (k vecinos más cercanos), árboles de clasificación, regresión logística, support vector machines, naive bayes, entre otras. A su vez, se pueden encontrar softwares libres generales enfocados a la minería de datos, por ejemplo, RapidMiner permite en un entorno gráfico el modelado y análisis de datos, aplicando en cadena los operadores disponibles. Adicionalmente, RapidMiner proporciona más de 500 operadores, incluyendo los necesarios para realizar distintas operaciones, brindando seguridad y exactitud en los resultados (Beltran y Poveda, 2010).

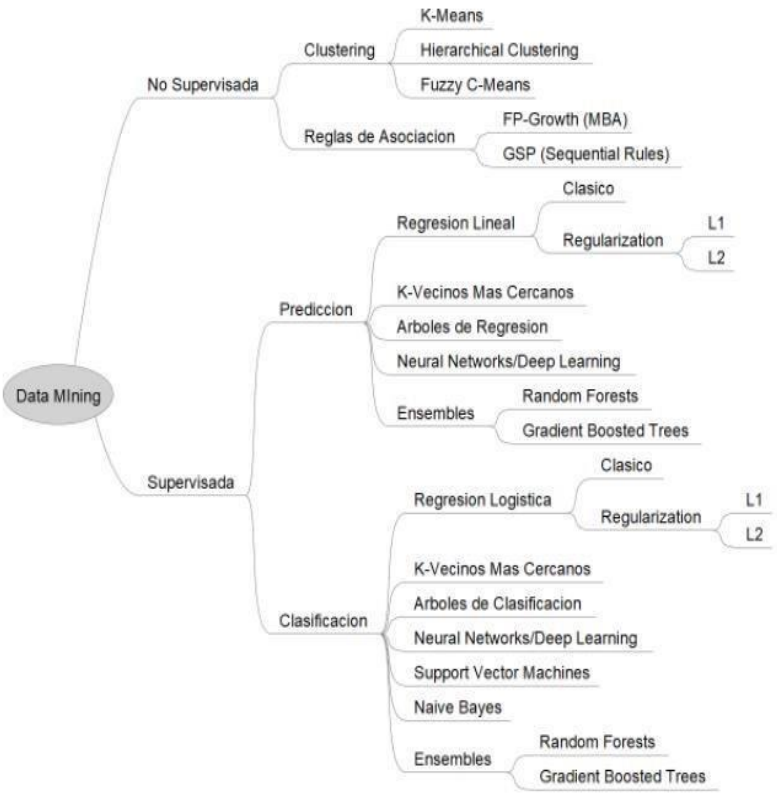

Figura 1. Tipos de Minería de datos (Quintanilla, 2013)

Respecto a los algoritmos y deserción estudiantil, existen varias investigaciones que demuestran que a través de algoritmos y una big data es posible predecir el abandono estudiantil, como es un estudio enfocado a la deserción de los estudios en línea, donde los algoritmos aplicados son: Naive Bayes, Random Forest, Regresión Logística y K vecinos más cercanos; y algunas variables empleadas son: fecha de inicio, fecha de fin del curso, categoría del curso, acceso al fórum del curso, videos del cuso visualizados, módulo, etc. concluyendo que regresión logística obtuvo la máxima exactitud al momento de predecir (Umer, Susnjak, Mathrani, & Suriadi, 2017).

Valero, Salvador y García (2010) menciona que las técnicas árbol de decisión y knn son idóneas para análisis de deserción estudiantil, debido a los óptimos resultados obtenidos en la Universidad Tecnológica de Izúcar de Matamoros (México). En estos determinaron que los alumnos desertan por tres causas principales: edad, ingresos familiares y nivel de inglés.

Fernández, Solís, Hernández y Moreira, (2019), publicaron un estudio con modelos explicativos y predictivos de la deserción estudiantil, en programas académicos de grado en el Tecnológico de Costa Rica. Los referidos autores realizaron análisis con modelos antes de la primera matrícula y después del primer ciclo lectivo, concluyendo que aumenta la capacidad de precisión en la predicción al analizar a los estudiantes después del primer ciclo. Además,

determinaron que existen dos variables que son altamente significativas para determinar la deserción en los ciclos siguientes: cantidad de créditos cursados y nota media de créditos cursados. En esta investigación se realizó la predicción correspondiente con algunos modelos de algoritmo, dando como conclusión que la regresión logística se encuentra entre los que tienen mejor precisión.

Arismendy y Morales (2018) realizaron una tesis en la Universidad de los Llanos (Colombia) donde aplicaron únicamente el modelo de regresión logística para comparar la deserción en todas las facultades, concluyendo que este modelo es eficiente para este tipo de estudio. Adicionalmente, los resultados muestran que, las facultades con mayor riesgo de deserción temprana son la de Ciencias Básicas e Ingeniería y la Facultad de Ciencias de la Salud. En estas, aumenta la probabilidad de deserción el haber estudiado en un colegio particular.

En la Escuela Superior Politécnica del Litoral, se realizó un estudio para predecir los estudiantes que son posibles desertores basados en su demografía y características académicas, como por ejemplo, tipo de residencia, indicador del nivel socioeconómico, promedio general de notas, numero de materias reprobadas, número de materias aprobadas, etc. Para ello se aplicó regresión logística, árboles de decisión y knn, Naives Bayes, finiquitando que el modelo con mayor porcentaje de detección es regresión logística y que la deserción es más común al inicio de los estudios (Noboa, Ordóñez, & Magallanes, 2018).

En las investigaciones citadas se puede observar que se han aplicado diversos algoritmos y diferentes variables, según el objetivo de cada investigación, concluyendo que la deserción es un hecho preocupante en varias universidades a nivel nacional y en varios países. Fernández et al. (2019) señalan que los datos de estudiantes que culminaron primer ciclo presentan mejor precisión de predicción. Paralelamente Noboa et al., (2018) determina que mientras más avanza el estudiante en sus estudios menos probable es su deserción. Bajo estas circunstancias, en la presente investigación se analizó únicamente la deserción en el primer año universitario.

Algunas variables adicionales encontradas en la revisión bibliográfica para analizar la deserción estudiantil en la educación superior y que se aplican en este trabajo de investigación, son: problemas de movilidad (personas que vienen de otra ciudad); tipo de bachillerato (bachillerato en ciencias básicas, bachillerato en ciencias con especialidad en físico matemático, bachillerato en contabilidad y bachillerato técnico); y, tipo de colegio

(particular, fiscal o fiscomisional); por tanto estas variables junto con la ficha socioeconómica, son primordiales para analizar la probabilidad de éxito de los estudiantes. Además, sobre técnicas de deserción y variables empleadas, se observó que las técnicas de k-vecinos y regresión logística, son las más utilizadas y con buenos resultados, sin embargo; no se han comparado su efectividad entre estas. En algunos casos se han aplicado de manera independiente y en otros, se han comparado con diferentes modelos de minería de datos. El algoritmo knn es idóneo para tareas de clasificación donde las relaciones entre las características y las clases son numerosas o difíciles de entender. Su función es asignar los datos no etiquetados a la clase de los ejemplos etiquetados más similares, para ello, empieza con datos de entrenamiento que son clasificados en varias categorías, etiquetadas por una variable nominal, adicionalmente tiene un conjunto de prueba que contiene datos sin etiquetar pero con similares características a los datos de entrenamiento, por último, a la instancia de prueba sin etiqueta se le asigna la clase de la mayoría de los k vecinos más cercanos. El algoritmo knn identifica *k* registros en los datos de entrenamiento que son los más cercanos en similitud, donde *k* es un número entero especificado de antemano (Witten, Frank, & Hall, 2011). (Ver sección 2.5)

El algoritmo de regresión logística nos permite modelar un resultado binario, es por ello que es una buena opción para ser aplicada en este estudio. Un parámetro necesario en este para este modelo es definir un umbral de selección, de esta manera si sobrepasa el umbral, se etiqueta a la clase como 1, caso contrario como 0 (Witten et al., 2011).

Con base a lo mencionado anteriormente, el objetivo de esta investigación fue determinar las variables que influyen significativamente sobre la deserción estudiantil en la Facultad de Ciencias Químicas de la Universidad de Cuenca, aplicando los modelos de knn y regresión logística con la ayuda del software de RapidMiner y finalmente inferir en estudiantes matriculados recientemente.

2. Materiales y Métodos

Como caso de estudio, se buscó determinar la probabilidad de deserción estudiantil en los primeros ciclos según la información y variables que influyen en el ingreso a la universidad pública ecuatoriana, usando las herramientas de minería de datos mediante el software denominado RapidMiner. Esta investigación se realizó en ocho fases, estas son:

2.1. Definición de variables

En esta fase se procedió a identificar las variables capaces de ser empleadas en los algoritmos de clasificación y que, adicionalmente, se encuentren disponibles para la investigación. Con este antecedente, las mismas fueron extraídas de la ficha socioeconómica que los estudiantes llenan al momento de matricularse. Cabe destacar que al ser la ficha socioeconómica común para todas las carreras, permite tener las mismas variables. Estas se clasificaron en: factores personales, vivienda familiar e integrantes. Por último, fueron transformadas de variables nominales a categóricas numéricas en RapidMiner, para poder ser aplicadas en los algoritmos mediante el operador "Nominal to Numerical".

2.1.1. Datos personales

Género. Se asignó un valor a cada variable de la siguiente manera: 0 para hombres, 1 para mujeres y 2 para las diversidades sexogenéricas o género GLBTI.

Fecha de nacimiento. Es una variable polinómica que asignó a cada año de nacimiento un valor correspondiente.

Estado civil. Se identificaron cuatro categorías: 0 para soltero, 1 para casado, 2 para unión libre y 3 para divorciado.

Etnia. Existen seis clasificaciones. Se asignaron valores de la siguiente manera: 1 para mestiza, 2 para indígena, 3 para mulata, 4 para afrodescendiente, 5 para blanca y 6 para otras. **Régimen de educación secundaria.** Hace referencia al sostenimiento del colegio en el que se graduó el estudiante. Se identificaron tres categorías: 0 para fiscales, 1 para fiscomisionales y 2 para particulares.

Título de bachiller. Esta variable se refiere al título que se le otorga al estudiante al graduarse del colegio; entre los títulos más comunes están: Bachiller en Ciencias Básicas, Bachiller en Ciencias con especialidad en Físico Matemático, Bachiller en Ciencias con especialidad en Químico Biólogo, Bachiller en Contabilidad y Bachiller Técnico. Se le asignó un valor a cada título de bachiller, por ejemplo, 0 para Bachiller en Ciencias con especialidad en Físico Matemático, 1 para Bachiller Técnico en Explotaciones Agropecuarias, 2 para Bachiller Técnico en Comercio, 3 para Bachiller Técnico Industrial Instalaciones, Equipos y Maquinarias Eléctricas, 4 para Bachiller Técnico Industrial Esp. Mecanizado y Construcciones Metálicas, 5 para Bachiller en Ciencias con especialidad en Químico Biólogo, etc.

Año de graduación. Esta variable nos indica en qué año el estudiante terminó sus estudios de bachillerato.

¿Alguna vez estudio en otra universidad? Es una variable binaria de sí o no. Se asignó el valor 0 para No y 1 para Sí.

Otra universidad. En esta variable se coloca el nombre de la universidad anterior, en el caso que el estudiante hubiera estudiado en otra universidad.

¿Ha cursado una carrera diferente en la Universidad De Cuenca? Es una variable binaria que se le ha asignado 1 para Si y 0 para No.

En la Tabla 1, se muestran los valores numéricos asignados a las diferentes categorías de algunas variables nominales. Tabla 1

Valor numérico asignado a la clasificación de las variables nominales

Variable Numérica	Variables Nominales					
	Género	Estado civil	Etnia	Régimen de Educación Secundaria	¿Alguna vez estudio en otra universidad?	¿Ha cursado una carrera diferente en la Universidad De Cuenca?
0	Hombres	Soltero	Mestizo	Fiscales	No	No
1	Mujeres	Casado	Mulato	Fiscomisionales	Si	Si
2	GLBTI	Unión Libre	Afrodescendiente	Particulares		
3		Divorciado	Blanco			
4			Indígena			
5			Otros			

2.1.2. Vivienda familiar

Tenencia de la vivienda. Existen cuatro categorías para esta variable, asignando 0 a vivienda propia, 1 a vivienda propia con hipoteca, 2 a vivienda arrendada y 3 a vivienda cedida. **Zona de la vivienda.** Es una variable binaria, asignando 0 para rural y 1 para urbana. **Provincia donde se ubica la vivienda.** Lugar donde se ubica la vivienda. Se fijó un valor numérico a cada provincia según el orden como se presentaban en la base de datos de elaboración propia. Por ejemplo, Cañar estaba primero en la base de datos se le asignó el número 1, Azuay que fue el segundo en la base de datos, se le asignó el número 2, etc. **Cantón donde se ubica la vivienda.** Lugar dentro de la provincia donde se encuentra la vivienda. Se estableció un valor numérico a los diferentes cantones según se encontraban en la base de datos elaborada. Por ejemplo, 1 para

a Azogues, 2 para Gualaceo, 3 para Piñas, 4 para Zaruma, 5 para Cuenca; así sucesivamente hasta asignar un número a cada cantón. **Mensual pago Arriendo**. En caso de que la vivienda sea arrendada, valor del pago mensual. **Avalúo acumulado de vehículos.** Esta variable nos indica el valor acumulado del costo de los vehículos disponibles en la vivienda.

Como se puede observar en las líneas posteriores, en la Tabla 2 se muestran algunas variables nominales referentes a vivienda familiar junto a los valores numéricos asignados.

Tabla 2

Valor numérico asignado a la clasificación de las variables nominales referentes a viviendas

Variable Numérica	**Variables Nominales**	
	Tenencia de la vivienda	**Zona de la vivienda**
0	Propia	Rural
1	Propia con Hipoteca	Urbana
2	Vivienda Arrendada	
3	Vivienda Cedida	

2.1.3. Integrantes

Número de integrantes de la familia. Esta variable nos indica cuántas personas forman parte del hogar.

Número de integrantes estudiantes. Hacen referencia a cuántas personas que forman parte de la familia son estudiantes.

Número de hijos del estudiante. Es una variable polinómica, que nos indica la cantidad de hijos que tiene el estudiante.

Ocupación del jefe de familia. Puede ser bien el padre o la madre.

Total de egresos. Valor de gastos relevantes mensuales.

Total de Ingresos. Ingresos mensuales en el hogar.

2.2. Compresión de datos

En esta fase se recolectaron los datos para conocer su estructura, identificar la calidad y convertirlos en un formato idóneo para analizarlos. Por citar un ejemplo, en la ficha socioeconómica se registra el nombre del colegio donde el estudiante culminó su bachillerato, por lo que, mediante el nombre del colegio se investigó el sostenimiento (fiscal, fiscomisional o particular). Otro factor importante en la comprensión de datos fue verificar que el estudiante no se haya matriculado por tres semestres seguidos, para asegurar su deserción. Adicionalmente, los datos mencionados en las líneas precedentes, que reposan en los archivos de las carreras de la Facultad de Ciencias Químicas, entre los años 2014 – 2018, fueron proporcionados por la el Departamento de Admisión y por la Facultad de Ciencias Químicas de la Universidad de Cuenca. Se obtuvo una base de datos de 1015 estudiantes.

2.3. Preparación de los datos

Se normalizaron las variables para interrelacionarlas estadísticamente en el software RapidMiner. Este programa es completo, flexible y con una interfaz gráfica fácil de entender por lo que permite realiza la transformación de datos en menor tiempo (Zainal, Sulaiman, & Jali, 2015).

En el programa SPSS se analizó que variables no aportan con información para el estudio, estas variables son identificadas mediante la comunalidad de cada variable, que hace referencia a la proporción de varianza explicada por el conjunto de factores comunes resultantes. Las comunalidades son representadas con valor entre 0 y 1, mientras más cercanas al cero menor aporte tienen en el estudio, siendo posible eliminarlas (Rodríguez y Mora, 2001).

Con las variables que quedan, en el software SPSS se realizaron pruebas conexas (tests) para indicar la posibilidad de realizar el ACP, desde el punto de vista estadístico (ver sección 3). Arancibia (2005) menciona algunas pruebas para ver la factibilidad de aplicar un análisis multivariante como es ACP; las pruebas aplicadas en este estudio son:

- *El Test de Esfericidad de Bartlett:* Permite comprobar la existencia de correlaciones entre las variables, mediante la oposición de la hipótesis nula que asegura que la matriz de correlaciones es una matriz identidad. Si p-valor < 0.05 se rechaza Ho (hipótesis nula) y se continúa con el análisis.

- *El Indice de Kaiser-Meyer Olkim:* Cuanto más cerca de 1 sea el valor obtenido, implica que la relación entre variables es alta. Los valores menores que 0.5 indican que no se permite realizar ACP.

A continuación se realizó una rotación VARIAMAX de los factores iniciales, igualmente en el software SPSS. Según Montoya (2007) el método VARIMAX es utilizado para que cada componente rotado obtenga altas correlaciones con unas cuantas variables, permitiendo identificar claramente los factores en cada componente.

Según León, Solano, y Tilano (2008), para estos tipos de estudios el porcentaje de varianza debe ser mayor o igual al 70%. En este estudio se logró disminuir el número de variables aplicando el análisis ACP trabajando con un porcentaje de varianza del 71%, que hace referencia a 11 componentes principales (ver sección 3). Con la rotación VARIAMAX y varianza de 71% se identificó fácilmente las variables principales de cada componente, en caso de no aplicar la rotación VARIAMAX, es oportuno usar un porcentaje de varianza mayor para una correcta asignación de variables a cada componente.

Para finalizar se formó una nueva base de datos para ser utilizada más adelante, en los algoritmos de clasificación (knn y rl). Para los valores actuales de la nueva base de datos Fernández, (2011) indica que se calcula aplicando la ecuación (1) para cada componente principal en cada dato o elemento de la base de datos:

$$CP_i = U_{i1}X_{i1} + U_{i2}X_{i2} + U_{i3}X_{i3} + \cdots + U_{n1}X_{n1} \quad (1)$$

Donde

U: es el coeficiente de correlación de cada componente principal con la variable asignada.

X: es el valor propio de cada ejemplo o dato referente a cada variable.

2.4. Partición de datos

En el algoritmo knn para determinar la eficiencia del sistema se dividió el repositorio de datos en dos conjuntos: entrenamiento y prueba. Para esto, se empleó validación cruzada k-fold, ya que divide la base de datos en subconjuntos mutuamente exclusivos del mismo tamaño, denominados en inglés como *folds;* En el entrenamiento se va iterando un subconjunto como prueba del modelo y los subconjuntos restantes formaran un conjunto de entrenamiento. Al

finalizar todas las iteraciones, se promedian los resultados de precisión y error obtenidos para cada subconjunto de prueba. (Laura-Ochoa, 2019)

Para empezar el entrenamiento de los datos, se aplicó un operador llamado “Smote Sampling” debido a una data no balanceada, como se muestra en la Figura 2, que permite aumentar la muestra minorista creando ejemplos sintéticos. Posteriormente se aplicó validación cruzada k-fold, representada en RapidMiner con el comando “Cross Validation” (Figura 3).

El comando “Cross Validation” requiere de un número de *folds*. Witten, Frank, & Hall (2011) afirman que el número de *folds* recomendado para este comando es 10, ya que permite obtener una mejor estimación del error, actualmente la validación cruzada de 10 subconjuntos se ha convertido en un método estándar para ser utilizada; sin embargo el debate continúa.

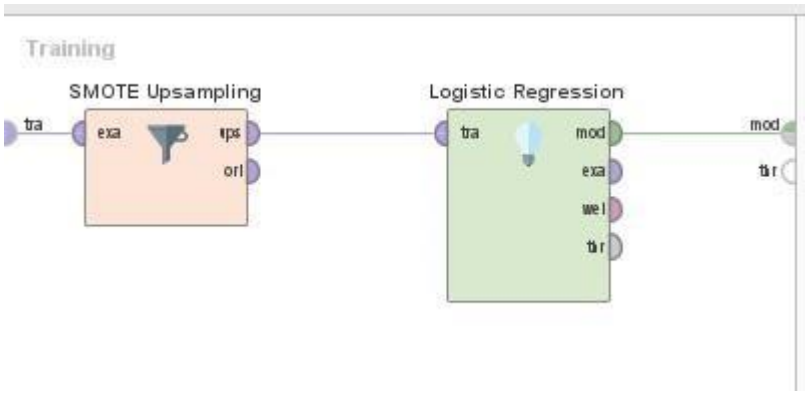

Figura 2 Operador Smote Upsampling para balancear la data.

.

Figura 3 Parámetros del operador "Cross Validation ".

2.5. Modelado RL

En el modelo de regresión logística se estableció un umbral de selección; un criterio aceptable fue clasificar a un individuo como desertor si la probabilidad de que sea desertor es

alta. Existe dos posibilidades de clasificación: éxito o fracaso; Noboa et al (2018) afirman que es común en estos estudios usar un umbral de 0.5, ya que con este valor disminuyen los falsos positivos y el porcentaje de error, por otro lado, un umbral menor aumenta los falsos positivos y el porcentaje de error; mientras que para valores mayores a 0,5 aumenta considerablemente los falsos negativos pero el porcentaje de error se mantiene constante entre valores del 0.5 a 0.8. Con base en lo mencionado anteriormente, se asignó un rango de probabilidad menor o igual a 0.5, para clasificar a la persona como desertora. Para tener control sobre el punto de corte que permite clasificar al estudiante, se aplicó el operador "Create Threshold" que significa crear umbral, como se muestra en la Figura 4. La Figura 5 indica cómo se designó el umbral de clasificación y la Figura 6 detalla el modelo de regresión logarítmica aplicado en este estudio a través de un diagrama de flujo.

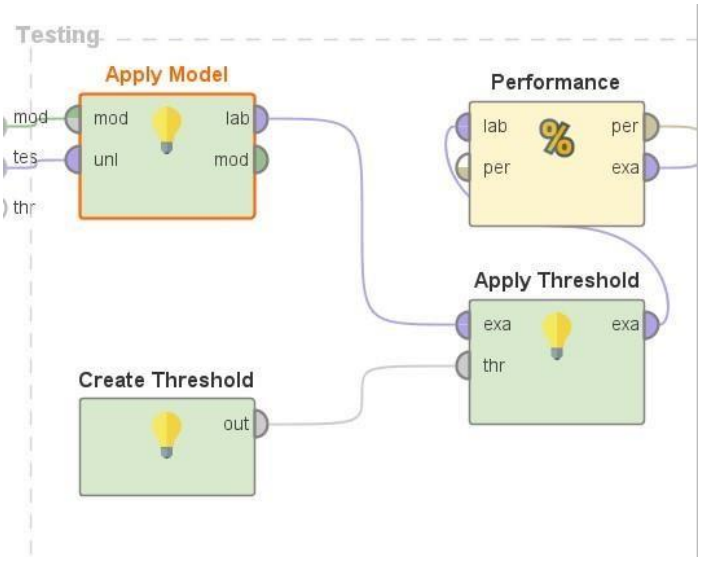

Figura 4 Operador para crear umbra.

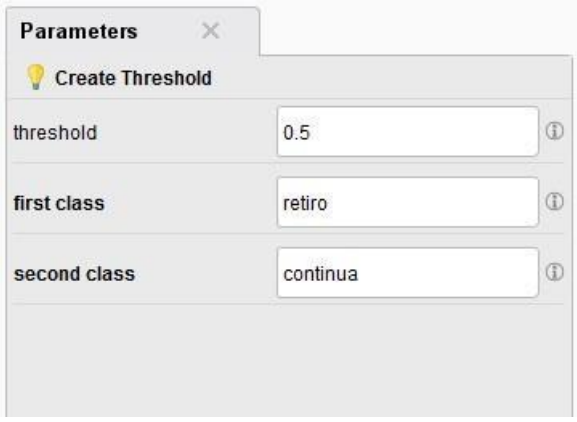

Figura 5 Parámetro de umbral.

El modelo de regresión logística representado por la ecuación 2, donde influyen las variables independientes, estima la probabilidad de éxito (Reyes, Escobar, y Duarte, 2007). Esta ecuación es útil para determinar si *p* es mayor que el valor de corte, en este caso; mayor que 0.5 se considera al estudiante exitoso, caso contrario se le clasifica como fracaso o desertor.

$$p = \frac{e^{b_o + b_1 x_1 + \cdots + b_n x_n}}{1 + e^{b_o + b_1 x_1 + \cdots + b_n x_n}} \qquad (2)$$

Donde:

p: es la probabilidad de éxito. x_n:

Varibles independientes. b_n :

Coeficientes propios de cada variable.

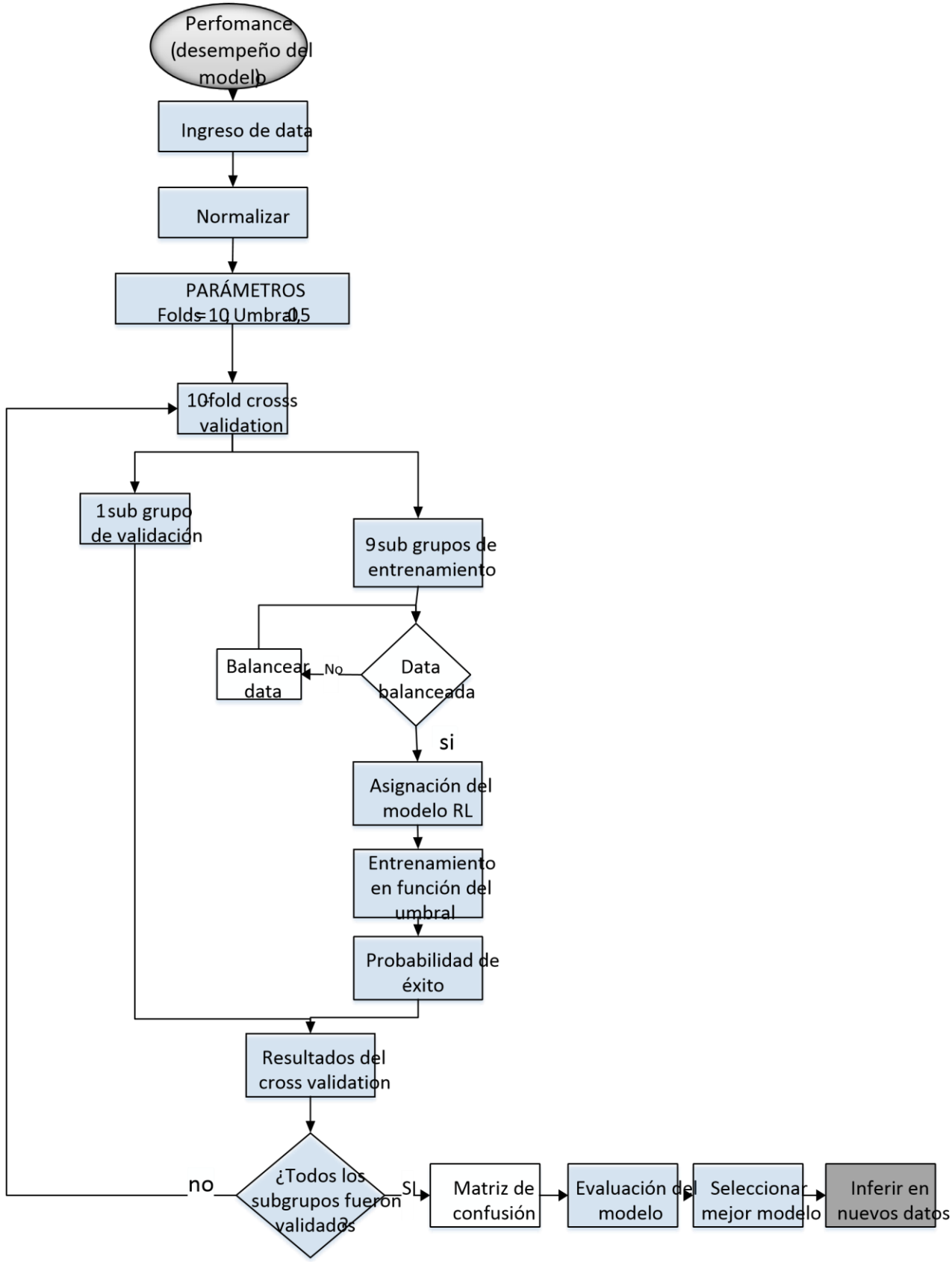

Figura 6 Diagrama de flujo del modelado de regresión logarítmica.

2.6. Modelado Knn

Un parámetro importante en el modelado de knn es el valor de *k,* Lantz (2013) menciona dos maneras de calcular este valor, estas son: calcular la raíz cuadrada del número de ejemplos de entrenamiento, o bien, probar varios valores de *k* y elegir el valor que demuestre mejor rendimiento de clasificación. En RapidMiner se colocó varios valores aleatorios a *k* a través del comando "Optimize Parameters" (Figura 7). Este comando realiza una corrida del modelo con cada *k* tipiada, al final se visualiza una tabla donde indica la exactitud y error obtenido al correr el modelo con cada valor de *k* (ver sección3).

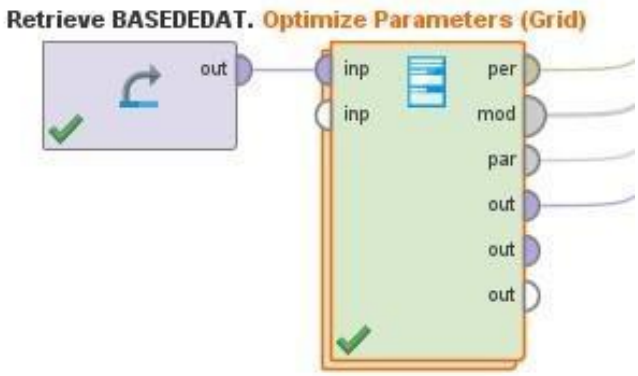

Figura 7 Operador Optimize Parameters.

El algoritmo knn para clasificar una variable necesita únicamente la distancia entre el nuevo dato y los datos existentes en la base de datos. El software empleado cuenta con la capacidad para aplicar la fórmula de distancia euclidiana. Cuando se trabaja con variables categóricas la fórmula es muy compleja, y por motivos de efectividad y optimización de tiempo es mejor realizarla a través de un software.

En la Figura 8 es posible observar un diagrama de flujo donde se detalla el proceso para modelar el algoritmo knn aplicado en este estudio.

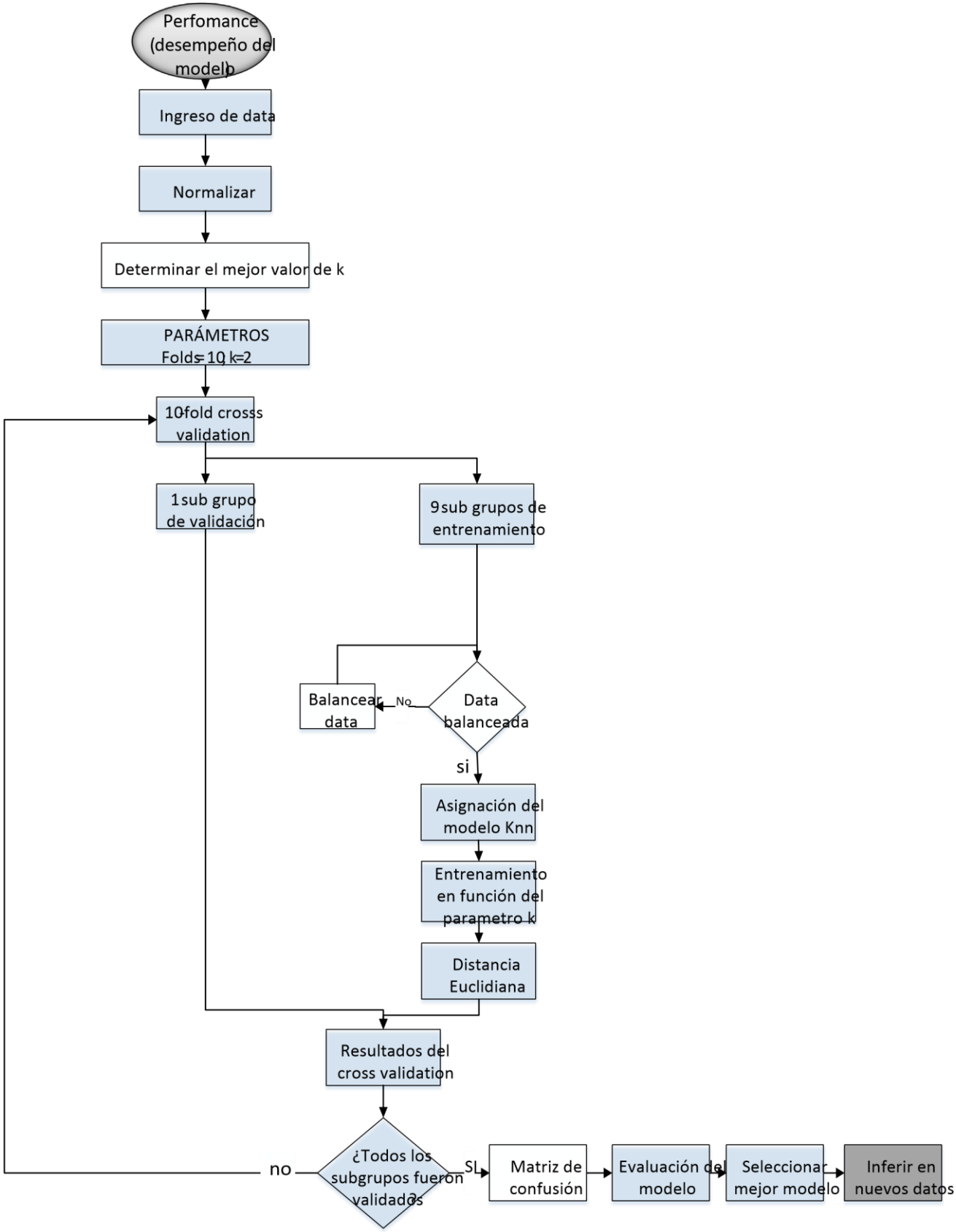

Figura 8 Diagrama de flujo del modelado de knn.

2.7. Evaluación

Se comparó la clasificación entre ambos algoritmos a partir de la matriz de confusión, ya que construye una serie de métricas como: exactitud del modelo, error de clasificación, sensibilidad, especificación, precisión y tiempo. Aunque parezca irrelevante, el tiempo es un factor importante para ser comparado, mientras menos tiempo y más eficiente sea un modelo significa que es más confiable.

La exactitud es otorgada por el modelo a partir de los datos obtenidos en la matriz de la confusión, comúnmente expresada en porcentaje. La exactitud indica la cantidad de predicciones correctas realizadas por el modelo. La ecuación (3) muestra cómo obtener la exactitud del modelo (Lantz, 2013).

$$\text{Porcentaje de exactitud} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \qquad (3)$$

Donde:

TP = Verdaderos positivos.

TN = Verdaderos negativos.

FP = Falsos positivos.

FN = Falsos negativos.

Por otro lado, el porcentaje de error representa la proporción de ejemplos clasificados incorrectamente, para hallar este valor Lantz (2013) propone la ecuación (4):

$$\text{Porcentaje de error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \qquad (4)$$

Donde:

TP = Verdaderos positivos.

TN = Verdaderos negativos.

FP = Falsos positivos.

FN = Falsos negativos.

La sensibilidad de un modelo se le denomina también como tasa positiva verdadera. Calcula la proporción de ejemplos positivos que se clasificaron correctamente. Para hallar este valor se realiza una división de la cantidad de verdaderos positivos (TP) dividido para la sumatoria de la cantidad de falsos negativos (FN) y verdaderos positivos (TP) como se muestra en la ecuación (5), (Lantz, 2013) .

$$\text{Porcentaje de Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100 \qquad (5)$$

La especificidad de un modelo mide la proporción de negativos que se clasificaron correctamente. Contrario a la sensibilidad, esto se calcula dividiendo el número de verdaderos negativos dividido para la sumatoria de verdaderos negativos más falsos positivos, como se muestra en la ecuación (6) (Lantz, 2013).

$$\text{Porcentaje de Especifidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} * 1 \qquad (6)$$

La precisión identifica la proporción de ejemplos positivos que son verdaderamente positivos. Un modelo es muy confiable si predice la clase positiva en casos muy probables de ser positivos. La ecuación (7) indica como calcular la precisión del modelo (Lantz, 2013).

$$\text{Porcentaje de precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}} * 100 \qquad (7)$$

Donde:

TP = Verdaderos positivos.

FP = Falsos positivos

2.8. Inferencia en estudiantes matriculados en el 2019

Una vez definido que el mejor modelo para clasificar fue knn, con los nuevos datos y desconociendo si el estudiante deserta o continúa, se procedió a utilizar dicho modelo para inferir en nuevos estudiantes; demostrando que es posible predecir la deserción estudiantil con la aplicación de algoritmos de clasificación.

En esta fase se necesitaron los comandos “Group Modelo” y “Store”, estos comando permiten agrupar las operaciones y a continuación guardar el modelo para aplicarlo en una nueva base datos como se puede ver en la Figura 9.

Para aplicar el modelo guardado se empleó un comando denominado “Apply Model” (ver Figura 10). Para inferir si un estudiante es desertor o no, se necesita una base de datos con las mismas variables de la base con la que se validó el modelo, es decir la misma cantidad de variables, tipo y que estén nombradas iguales.

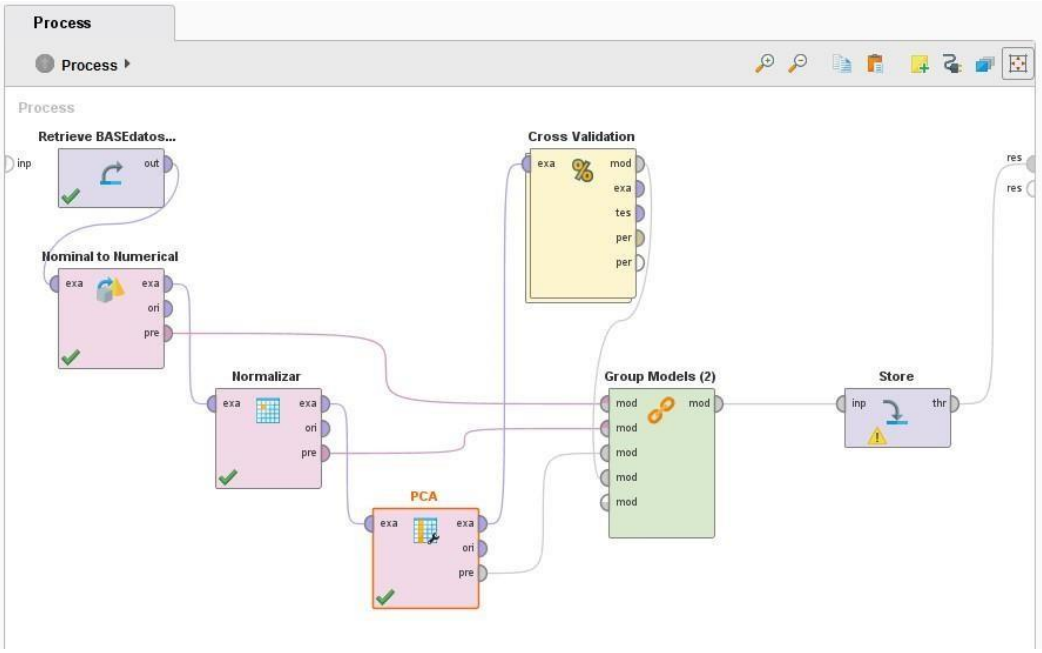

Figura 9 Comando Group Model y Store

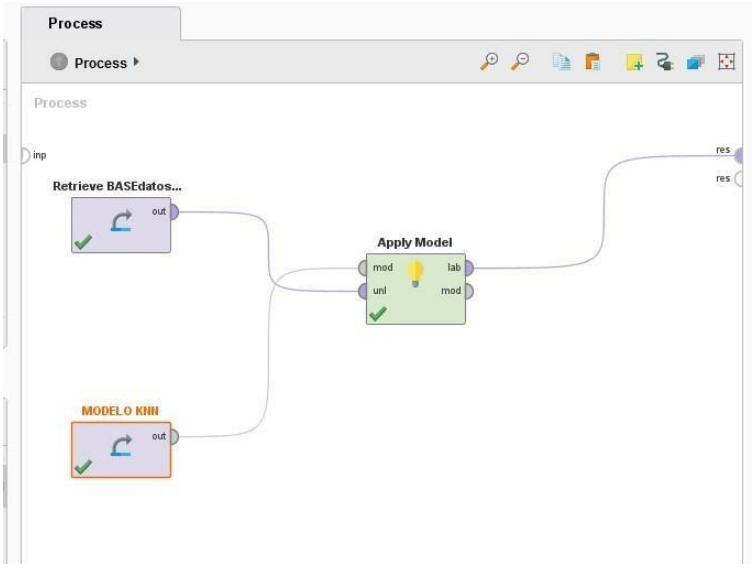

Figura 10 Aplicación del modelo

3. Resultados y discusión

A continuación se muestran porcentajes de la deserción estudiantil de la Facultad de Ciencias Químicas en los periodos comprendidos entre el 2014-2018. La Figura 11 muestra la deserción por años en cada carrera, demostrando que en el 2016 es el año donde existió menor deserción de las carreras de Bioquímica y Farmacia, Ingeniería Ambiental e Ingeniería Industrial, mientras que en el año 2014 la carrera de Ingeniería Química presenta menor deserción. En la Figura 12 se observa que en casi todos los años la mayor deserción se da en personas que vienen de colegios fiscales.

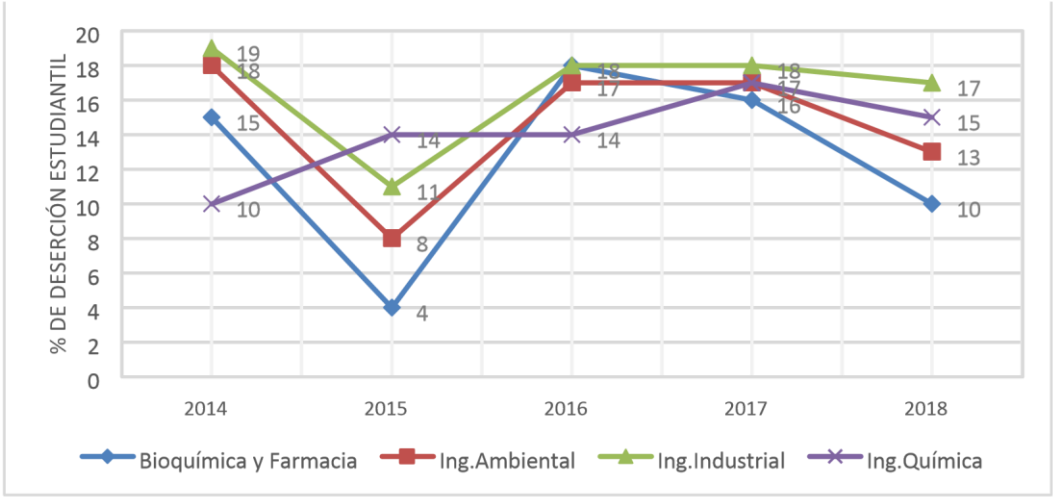

Figura 11 Porcentaje de deserción por carrera desde el año 2014 hasta el 2018.

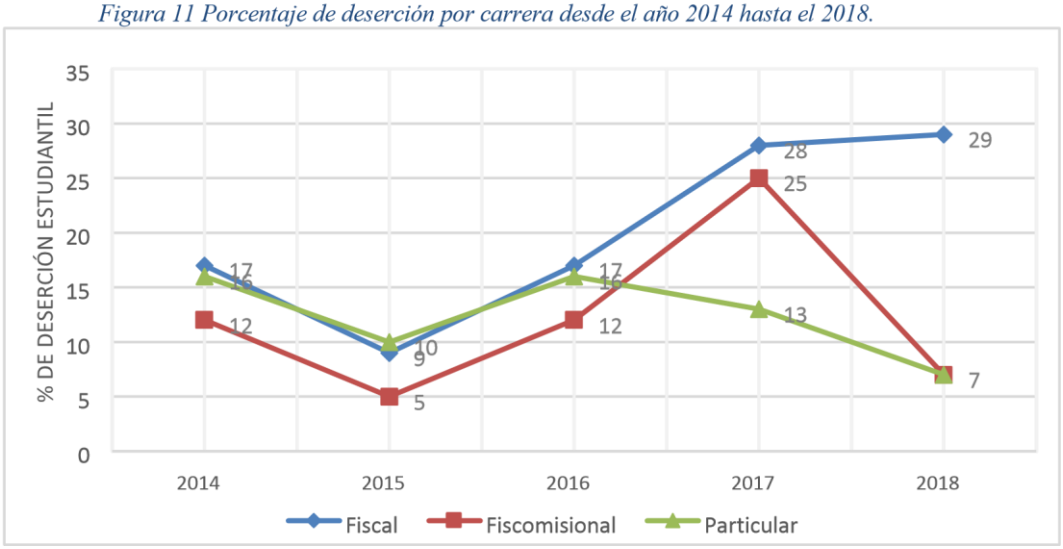

Figura 12 Porcentaje de estudiantes que desertan según el sostenimiento del colegio donde culminaron el bachillerato desde el año 2014 hasta el 2018.

Respecto a la nacionalidad (Tabla 3), la cantidad de estudiantes extranjeros en la Facultad de Ciencias Químicas es escasa y casi todos desertaron, sin embargo, se observa que las personas con nacionalidad peruana no desertaron en el primer año. El lugar de nacimiento es un factor importante en este estudio, debido a que, conforme la evidencia obtenida en esta investigación, muchas personas de otras ciudades migran hacia Cuenca para seguir estudiando en una institución de educación superior. En la Figura 13 a un lado se muestra el porcentaje de desertores con respecto al número de ingresados en esa misma provincia, y al otro lado el número de desertores, en el periodo 2014 -2018. Por ejemplo, de

Azuay ingresan 669 y desertan 96, esto corresponde al 14%., pero de Imbabura ingresan 5 y desertan 2, corresponde al 40%, siendo este el dato más alarmante.

Tabla 3

Cantidad de deserción según la nacionalidad.

Nacionalidad	**Continúa**	**Retiro**
Colombiana	0	1
Cubana	0	1
Española	0	2
Peruana	2	1
Venezolana	1	1

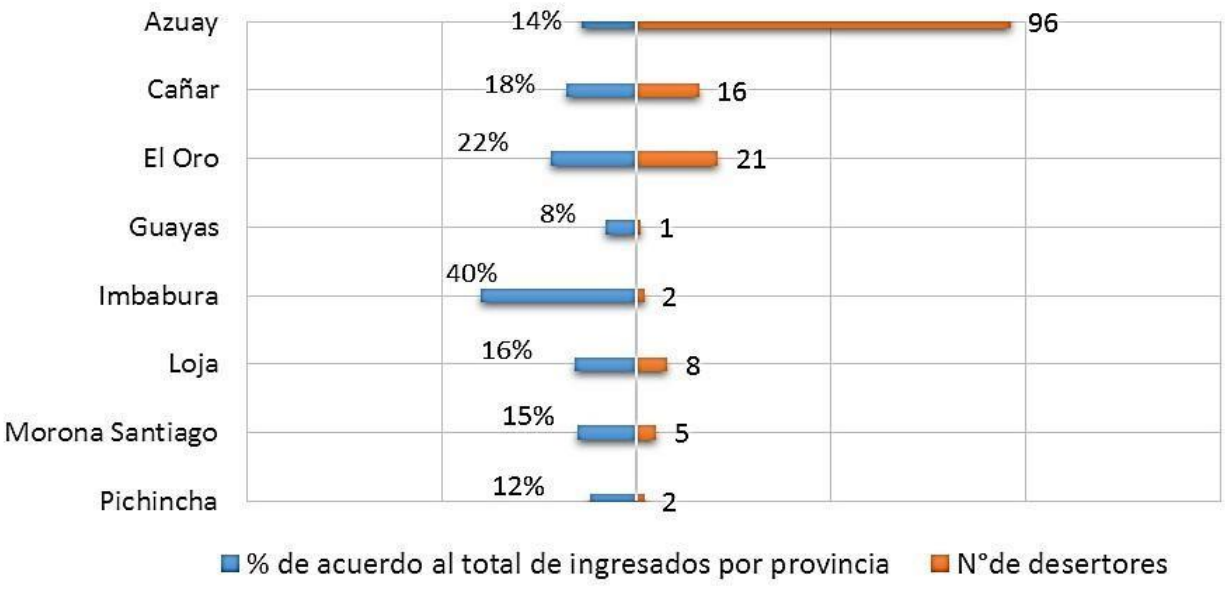

Figura 13. Cantidad de estudiantes nacionales desertores según lugar de nacimiento desde el año 2014 hasta el 2018.

La Tabla 4 demuestra que los estudiantes que terminan el bachillerato a partir de los 20 años de edad en adelante, son más propensos a desertar la universidad.

Tabla 4

Porcentaje de deserción según edad de fin del bachillerato

Edad de fin de bachillerato	**Continúa**	**Deserta**
Menores e igual a 17años	84%	16%
18 años	86%	14%
19 años	78%	22%
20 años	57%	43%
Mayores e igual a 21años	0%	100%

La variable ocupación del jefe familiar fue suprimida en el ACP ya que no influía significativamente en el estudio por tener un bajo valor de comunalidad, es decir no aportan con información en este estudio. Los autovalores de las variables que conforman cada componente principal se puede observar en el Anexo 1. El test KMO aplicado tiene un valor de 0,712 y la Prueba de Esfericidad de Bartlett es igual a cero, lo que permite rechazar la hipótesis nula referente a la existencia de una matriz identidad (Tabla 5). Estos resultados demuestran que es posible realizar el análisis de componentes principales según lo definido en la sección 2.3.

Se obtuvieron 11 componentes con las variables que lo integran (Tabla 6), que representan el 71% de la varianza, como se muestra en la Figura 14. Se agruparon las variables que tienen alta correlación entre ellas para formar un componente principal, sin embargo, existen componentes principales formadas por una sola variable.

Tabla 5

Prueba de KMO y Bartlett

Prueba de KMO y Bartlett		
Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,712
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	8185,206
	Gl	325
	Sig.	,000

Tabla 6

Componentes Principales

Componente	**Variables Agrupadas Principal**
CP1	Total ingreso, total egreso, mensual pago de arriendo, avalúo acumulado de vehículos, colegio.
CP2	Provincia de la vivienda familiar, cantón de la vivienda familiar, lugar de nacimiento.
CP3	Año de graduación, fecha de nacimiento, estudio otra carrera.
CP4	Estudio en otra universidad, Nombre de la otra universidad.
CP5	Número de integrantes en la familia, número estudiantes en la familia.
CP6	Número de hijos menores a 6 años del estudiante, Estado Civil.
CP7	Género, carrera.
CP8	Nacionalidad, título de bachiller.
CP9	Tenencia de vivienda.
CP10	Lugar de residencia y etnia.
CP11	Zona de vivienda.

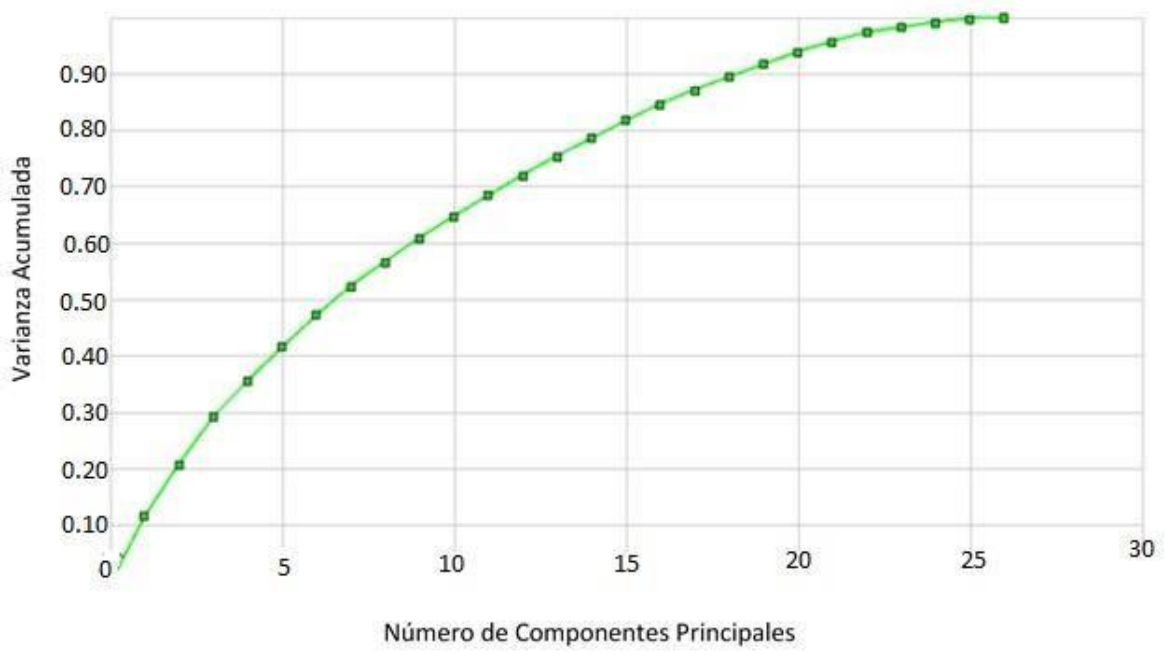

Figura 14 Varianza acumulada según la cantidad de Componentes Principales

En este estudio, al aplicar el algoritmo knn se necesita definir el valor de *k,* para ello se compara con valores al azar (2, 5, 10, 15,20, 30) indicando que el mejor valor para este parámetro es 2, ya que presenta mejor exactitud y menor error de clasificación, como se puede ver en la Tabla 7.

Tabla 7

Valores de "k"

Orden	Knn.k	Exactitud	Error de clasificación
1	2	73.4%	26%
2	5	62.1%	37.9%
3	10	64%	36%
4	15	57.7%	42.3%
5	20	60%	40%
6	30	58.2%	41.8%

En la Tabla 8, se pueden observar los resultados obtenidos de cada modelo, cabe recalcar que previamente se estandarizaron los datos antes de su procesamiento. Así mismo es preciso indicar que el modelo knn nos proporciona mayor sensibilidad y exactitud en los datos, esto

quiere decir que está clasificando e identificando mejor los resultados, además al tener knn una precisión más alta indica que logró mayor número de predicciones correctas.

Tabla 8

Comparativa de los resultados entre los modelos Regresión Logística y Knn

	Regresión Logística	**Knn**
Sensibilidad	53,37%	82,05%
Especificidad	61,78%	25,48%
Exactitud	54,67%	73,30%
Precisión	88,41%	85,74%
Error	45,32%	26,70%
Tiempo de proceso	16 seg	45 seg

Para validar el modelo knn, se usó una nueva data de 138 estudiantes matriculados en el año 2019 de la Facultad de Ciencias Químicas de la Universidad de Cuenca. Este modelo predijo que, de los 138 estudiantes, 118 continúan mientras que 20 desertan en el primer año. Como se muestra en la Tabla 9, el programa coloca el 1 a la clasificación que pertenece el nuevo elemento. Para seleccionar la clase del nuevo elemento utiliza la distancia Euclidiana y los elementos más cercanos, además coloca 0,5 cuando los resultados al momento de clasificar han sido iguales, y los clasifica como si el estudiante continúa sus estudios.

Tabla 9

Sección de los resultados de aplicar el modelo knn para predecir deserción estudiantil.

ID	PREDICCIÓN			Confianza(continua)		Confianza(retiro)	
1	continua	1,0	0,0				
2	continua	1,0	0,0				
3	continua	1,0	0,0				
4	continua	1,0	0,0				
5	continua	1,0	0,0				
6	continua	1,0	0,0				
7	continua	1,0	0,0				
8	continua	1,0	0,0				
9	continua	1,0	0,0	10	continua	1,0	0,0
11	continua	1,0	0,0	12	retiro	0,0	1,0
13	continua	1,0	0,0				
14	continua	0,5	0,5				
15	continua	1,0	0,0	16	retiro	0,0	1,0
17	continua	1,0	0,0	18	retiro	0,0	1,0
19	continua	0,5	0,5				
20	continua	1,0	0,0				
21	continua	1,0	0,0				
22	continua	0,5	0,5	23	retiro	0,0	1,0
24	continua	1,0	0,0				
25	continua	0,5	0,5	26	retiro	0,0	1,0
27	continua	0,5	0,5				
28	continua	1,0	0,0				
29	continua	0,5	0,5				
30	continua	1,0	0,0				
31	continua	0,5	0,5				
32	continua	0,5	0,5				
33	continua	1,0	0,0				
34	continua	1,0	0,0				
35	continua	0,5	0,5				
36	continua	1,0	0,0				
37	continua	1,0	0,0				
38	continua	1,0	0,0				
39	continua	0,5	0,5	40	retiro	0,0	1,0
41	continua	1,0	0,0				

42	1,0	0,0	
43	continua	1,0	0,0
44	continua	1,0	0,0
45	continua	1,0	0,0
46	continua	1,0	0,0
47	continua	1,0	0,0
48	continua	0,5	0,5
49	continua	0,5	0,5
50	retiro	0,0	1,0
51	continua	0,5	0,5
52	continua	0,5	0,5
53	continua	1,0	0,0
54	retiro	0,0	1,0
55	continua	1,0	0,0
56	continua	1,0	0,0
57	continua	1,0	0,0
58	continua	1,0	0,0
59	continua	1,0	0,0
60	continua	0,5	0,5
61	continua	0,5	0,5
62	continua	0,5	0,5
63	continua	1,0	0,0
64	continua	1,0	0,0
65	continua	0,5	0,5
66	continua	1,0	0,0
67	continua	0,5	0,5
68	continua	1,0	0,0
69	retiro	0,0	1,0
70	continua	1,0	0,0
71	retiro	0,0	1,0
72	retiro	0,0	1,0
73	continua	0,5	0,5
74	continua	1,0	0,0
75	continua	1,0	0,0
76	continua	1,0	0,0
77	continua	0,5	0,5
78	continua	1,0	0,0
79	continua	1,0	0,0
80	continua	0,5	0,5
81	continua	1,0	0,0
82	continua	1,0	0,0
83	retiro	0,0	1,0
84	continua	1,0	0,0
85	retiro	0,0	1,0

86	retiro	0,0	1,0
87	continua	1,0	0,0
88	1,0	0,0	
89	retiro	0,0	1,0
90	continua	1,0	0,0
91	continua	1,0	0,0
92	continua	1,0	0,0
93	continua	0,5	0,5
94	continua	1,0	0,0
95	continua	1,0	0,0
96	continua	0,5	0,5
97	continua	1,0	0,0
98	continua	1,0	0,0
99	continua	1,0	0,0
100	continua	0,5	0,5
101	continua	0,5	0,5
102	retiro	0,0	1,0
103	continua	1,0	0,0
104	continua	0,5	0,5
105	continua	0,5	0,5
106	retiro	0,0	1,0
107	continua	1,0	0,0
108	continua	0,5	0,5
109	continua	1,0	0,0
110	continua	0,5	0,5
111	continua	0,5	0,5
112	continua	1,0	0,0
113	continua	1,0	0,0
114	continua	1,0	0,0
115	continua	0,5	0,5
116	continua	1,0	0,0
117	continua	1,0	0,0
118	retiro	0,0	1,0
119	continua	1,0	0,0
120	continua	0,5	0,5
121	continua	1,0	0,0
122	continua	0,5	0,5
123	continua	1,0	0,0
124	continua	1,0	0,0
125	continua	1,0	0,0
126	continua	1,0	0,0
127	continua	0,5	0,5
128	continua	0,5	0,5

129	continua	0,5	0,5
130	continua	1,0	0,0
131	continua	1,0	0,0
132	continua	1,0	0,0
133	1,0	0,0	
134	continua	0,5	0,5
135	continua	0,5	0,5
136	retiro	0,0	1,0
137	continua	0,5	0,5
138	retiro	0,0	1,0

El total de estudiantes desertores del año 2019 inferidos por el modelo knn son cercanos al total real de estudiantes que desertaron en para el año 2020. Según los datos otorgados por la Facultad de Ciencias Químicas de la Universidad de Cuenca desertaron 16 estudiantes, demostrando que el modelo tiene un error del 20% respecto la realidad.

Las variables que más influyen en la deserción académica forman el CP1, estas son: el total ingreso, total egreso, mensual pago de arriendo, avalúo acumulado de vehículos, colegio; mientras que levemente menos importantes son las variables que forman el CP2, tales variables son: provincia de la vivienda familiar, cantón de la vivienda familiar y lugar de nacimiento. El coeficiente principal (CP1) equivale al 12% de varianza de todas las variables, siendo este el mayor valor con respecto a los demás coeficientes principales, y el CP2 equivale al 9,7% de varianza.

Estos resultados guardan relación con lo que sostiene Valero et al., (2010) que la técnica knn es idónea para predecir si un estudiante deserta la carrera, además concuerda con Valero et al.(2010) sobre que entre las variables que influyen mayormente en la deserción se encuentra el ingreso familiar pero se discrepa con Arismendy y Morales (2018) en que aumenta la probabilidad de deserción haber estudiado en un colegio particular, ya que en este influye más en la deserción haber estudiado en un colegio fiscal. Por otro lado Noboa et al.(2018) y Umer et al.(2017) afirman que el mejor modelo para predecir la deserción es regresión logística, en este estudio no se coincide con esos resultados debido a que las variables de cada estudio varían, por ejemplo, Noboa et al.(2018) tiene variables como nivel socioeconómico, número de materias aprobadas y reprobadas; mientras que Umer et al.(2017) aplica variables referentes a cursos online, como son: fecha de inicio y fin del curso, categoría del curso, acceso al fórum del curso, videos del cuso visualizados, módulo, etc Analizando estos resultados podemos ver que

la regresión logística también es buena para predecir la deserción si se tiene variables relacionadas a las notas, cantidad de créditos o número de matrículas, por otro lado, si se tiene variables cualitativas la mejor opción es knn.

4. Conclusiones

A través de la matriz de confusión se evaluaron los modelos (knn y rl) seleccionando al modelo knn como mejor opción. Se concluye que dicho modelo dista un 20% de la realidad ya que según los registros hay menos estudiantes desertores de los que predijo; sin embargo ese error pudiera disminuir si algún estudiante no aprobara la tercera matricula, es decir perdería la carrera, por ende se le consideraría desertor. Al momento de inferir en una nueva base de datos es necesario asegurarse que la nueva base de datos tenga las mismas variables, es decir la misma cantidad de variables, el mismo tipo y nombradas iguales, caso contrario el modelo presenta error. Para obtener resultados más cercanos a la realidad, es importante que al momento de modelar se equilibre la base de datos, en caso de ser necesario.

El software Rapidminer a veces necesita un software complementario, ya que no dispone de algunas funciones como pruebas de KMO, que es un test necesario para realizar ciertos análisis estadísticos, por lo que se usó SPSS. Sin embargo, Rapidminer es un software rápido y eficaz, además muy amigable al momento de modelar,

En cuanto a las variables que influyen en la deserción académica del primer año de la Facultad de Ciencias Químicas, se concluye que: total ingreso, total egreso, mensual pago de arriendo, avalúo acumulado de vehículos, colegio, son los más relevantes para predecir la deserción; seguidas de las variables: provincia de la vivienda familiar, cantón de la vivienda familiar, lugar de nacimiento. Esto demuestra que la educación podría ser un privilegio para personas con cierto poder adquisitivo, o para quienes deben de hacer sacrificios para poder estudiar. En base a esto podríamos colegir que la educación pública no resulta tan accesible para todas las clases sociales ni económicas. Por otro lado, los estudiantes que vienen de otras ciudades o provincias tienen que enfrentar la situación de acoplarse a una nueva ciudad y cumplir con sus estudios; siendo estos escenarios difíciles para ciertos alumnos lo que provoca la deserción de la carrera.

Finalmente, se recomienda realizar una investigación de orden cualitativo, de modo que, mediante un grupo focal o a través de la conducción de entrevistas a profundidad, se pueda saber con mayor detalle las razones de deserción, es decir; qué influyó subjetivamente en cada una de las personas que optaron por no continuar con sus estudios de tercer nivel. A su vez sería muy interesante realizar un estudio comparativo con otros regímenes de educación superior.

5. Referencias Bibliográficas

Arancibia, S. (2005). Análisis Factorial Método de Componentes Principales. *Estadística Aplicada Y Econometría*, *2*, 1–30.

Arismendy, C., & Morales, N. (2018). *Modelo de Regresión Logística como Alternativa para Medir la probabilidad de Deserción Temprana en la Universidad de los Llanos periodo 2015-2018*. Universidad de los Llanos.

Beltran, D., & Poveda, D. (2010, December). RapidMiner. *Universidad Nacional de Colombia*. Retrieved from http://www.fce.unal.edu.co/media/files/UIFCE/Economia/RapidMiner.pdf

educacionecuadorministerio.blogspot.com. (2018). Retrieved June 13, 2020, from https://educacionecuadorministerio.blogspot.com/2018/02/universidades-que-exigenexamen-deingreso-en-ecuador.html

Fernández, S. (2011). *Análisis Componentes Principales*. Universidad Autónoma de Madrid. Retrieved from https://www.estadistica.net/Master-Econometria/Componentes_Principales.pdf

Fernández, T., Solís, M., Hernández, M. T., & Moreira, T. E. (2019). Un análisis multinomial y predictivo de los factores asociados a la deserción universitaria, *23*(1), 1–25. https://doi.org/10.15359/ree.23-1.5

Lantz, B. (2013). *Machine Learning with R Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Retrieved from www.packtpub.com

Laura-Ochoa, L. (2019). Evaluation of Classification Algorithms using Cross Validation. In *Industry, Innovation, And Infrastructure for Sustainable Cities and Communities* (pp. 24–26). https://doi.org/10.18687/LACCEI2019.1.1.471

León, Á., Solano, H., & Tilano, J. (2008). Análisis multivariado aplicando componentes principales al caso de los desplazados. *Ingeniería Y Desarrollo*, (23), 119–142. Retrieved from http://www.redalyc.org/articulo.oa?id=85202310

Montoya, O. (2007). Application of the factorial analysis to the investigation of markets. Case of study. *Scientia et Technica*, *3*(35), 281–286. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=4804281&info=resumen&idioma=ENG

Noboa, C., Ordóñez, M., & Magallanes, J. (2018). *Statistical Learning to Detect Potential Dropouts in Higher Education: A Public University Case Study*.

Quintanilla, C. (2013). *Notas sobre Data Mining*. Managua.

Reyes, J., Escobar, C., & Duarte, J. (2007). Una aplicacion del modelo de regresion logistica en la prediccion del rendimiento estudiantil*, 101–120.

Rodríguez, M. J., & Mora, R. (2001). Estadística informática casos y ejemplos con el SPSS. In *Estadística informática casos y ejemplos con el SPSS* (primera, pp. 134–153). España: Universidad de Alicante. https://doi.org/10.2307/j.ctv893j76.10

Sinchi, E. R., & Gómez Ceballos, G. P. (2018). Acceso y deserción en las universidades. Alternativas de financiamiento. *Alteridad*, *13*(2), 274–287. https://doi.org/10.17163/alt.v14n2.2018.10

Teleamazonas. (2017, July 3). teleamazonas.com. Retrieved June 13, 2020, from http://www.teleamazonas.com/hora25ec/decada-cambios-educacion-superior/

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). Prediction of Students' Dropout in MOOC Environment Hadoop Cluster View project Event Detection View project. https://doi.org/10.18178/ijke.2017.3.2.085

Valero, S., Salvador, A., & García, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Recursos Digitales Para La Educación Y La Cultura*, 33–39. Retrieved from http://ccita2011.itsmotul.edu.mx/documentos/Recursos_digitales.pdf

Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. *Complementary literature None*. Retrieved from http://books.google.com/books?id=bDtLM8CODsQC&pgis=1

Zainal, K., Sulaiman, N., & Jali, M. (2015). *An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka*. *International Journal of Computer Science and Information Security* (Vol. 13).

6. Anexos

Anexo 1

Matriz de componentes rotados, método de rotación: Varimax

Matriz de componente rotado [a]

	Componente										
	1	2	3	4	5	6	7	8	9	10	11
TOTAL_INGRESOS	,912	-,002	,088	,007	,022	-,031	,006	-,014	,034	-,006	,022
TOTAL_EGRESOS	,828	-,012	,041	,000	-,016	-,037	,021	-,060	,061	-,053	-,043
AVALUO_ACUMULADO_VEHICULOS	,681	-,024	,046	,044	,025	,003	,098	-,036	-,267	,011	,085
MENSUAL_PAGO_ARRIENDO	,594	-,001	,121	-,019	,015	-,006	-,104	-,006	,546	-,016	-,026
COLEGIO	,475	-,024	-,106	,017	-,085	,015	-,038	,312	-,033	,104	,233
PROVINCIA_VIVIENDA_FAMILIAR	,001	,921	,039	-,022	,039	-,011	,017	-,022	-,036	,124	-,018
CANTON_VIVIENDA_FAMILIAR	-,017	,891	,049	,032	,054	,003	,016	-,082	-,059	-,014	-,077
LUGAR_NACIMIENTO	-,037	,662	-,036	,030	-,069	-,046	,081	,187	,130	-,216	,096
ANIO_GRADUACION	,113	,028	,898	-,173	,054	-,191	,064	-,003	,054	,059	-,040
FECHA_NACIMIENTO	,111	,024	,895	-,156	,045	-,193	,070	,027	,057	,047	-,047
ESTUDIO_OTRA_CARRERA	,071	-,024	-,514	-,153	,078	-,148	,010	,054	,079	,211	-,356
ESTUDIO_OTRA_UNIVERSIDAD	,023	-,004	-,132	,926	-,030	,004	-,020	,003	-,012	,027	,016
OTRA_UNIVERSIDAD	,027	,037	-,057	,924	,013	-,012	,005	,000	,031	,004	,023
NUM_ESTUDIANTES_FAMILIA	-,052	,014	,061	,009	,862	-,001	-,017	,027	,026	-,004	,024
NUM_INTEGRANTES	,040	,015	-,016	-,026	,849	,007	-,029	-,061	-,092	-,063	-,037
NUM_HIJOS_MEN6_ESTUDIANTE	-,023	-,025	-,076	,055	,108	,792	,033	,070	,104	-,159	-,017

ESTADO_CIVIL	-,024	-,017	-,149	-,062	-,096	,778	,042	-,081	,008	,080	,042
Género	,002	-,069	,048	-,017	-,033	-,007	,828	-,053	,100	-,054	-,010
Carrera	,047	,179	,052	,001	-,014	,083	,751	,080	-,107	,103	,049
NACIONALIDAD	-,014	,065	-,162	-,059	-,078	-,143	,027	,751	,129	-,164	,073
TITULO_BACHILLER	-,019	-,019	,352	,089	,084	,201	,002	,664	-,084	,185	-,174
TENENCIA_VIVIENDA	-,112	,009	,011	,028	-,070	,111	,043	,058	,841	,100	,073
LUGAR_RESIDENCIA	-,014	,125	-,015	-,002	-,099	-,059	-,096	,095	,062	,735	,190
ETNIA	-,003	,225	-,026	-,040	-,020	,014	-,171	,138	-,041	-,588	,152
ZONA_VIVIENDA	,128	-,018	,011	,014	,012	,003	,041	-,006	,071	,062	,863